NCHSR

RESEARCH PROCEEDINGS SERIES

Health Survey Research Methods

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Office of the Assistant Secretary for Health
National Center for Health Services Research

Abstract

This conference report is intended to inform the health research community about recent advances in health survey methods, about continuing concerns of which health survey users should be aware, and about areas requiring further methodological research. The conference concentrated on six major topics: (1) Measures and Correlates of Response Errors; (2) Telephone Survey Methodology; (3) Studies of Survey Measurement Techniques; (4) Use of Records in Health Survey Research; (5) Hiring, Training, and Monitoring Interviewers; (6) Survey Methods for Rare Populations. The conference was supported by conference grants to the Institute for Social Research at The University of Michigan, Ann Arbor, from the National Center for Health Services Research and from the Milbank Memorial Fund, and by services provided by the National Center for Health Statistics.

The editors are Charles F. Cannell and Robert M. Groves, Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan. The views expressed in this publication are those of the authors. No official endorsement is intended or inferred.

Library of Congress Catalog Card No. 84-601003.

National Center
for Health Services
Research

Health Survey Research Methods

Proceedings of the
Fourth Conference on
Health Survey
Research Methods,
Washington, D.C.
May 1982

This conference was jointly sponsored by the National Center for Health Services Research and the National Center for Health Statistics, through NCHSR grant no. HS 04569-01; and through a grant from the Milbank Memorial Fund.

September 1984

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Office of the Assistant Secretary for Health
National Center for Health Services Research

DHHS Publication No. (PHS) 84-3346

Contributors

Lu Ann Aday
Ronald M. Andersen
John P. Anderson
Frank M. Andrews
Morris Axelrod
Martha J. Banks
Terence W. Beed
Charles C. Berry
Gordon Bonham
Norman M. Bradburn
Bengt Brorsson
Fred A. Bryan, Jr.
James W. Bush
Gail Lee Cafferata
Charles F. Cannell
R.A. Carleton
John P. Connelly
Larry Corder
Ronald Czaja
Stephen M. Davidson
K. Downey
Douglas Drummond
Elizabeth Eastman
Jack Elinson
Jacob Feldman
Floyd J. Fowler, Jr.
Joanne Frankel
Robert Fuchsberg
Donald Goldstone
Robert M. Groves
Constance M. Horgan
Daniel Horvitz
Patricia Johnson
Judith A. Kasper
Diane M. Kipp
Phillip R. Kletke
Mary Grace Kovar
Richard A. Kulka
Barbara H. Lacey
Judith T. Lessler
Sara Segal Loevy
Thomas W. Mangione
Alfred C. Marcus
Kent Marquis
Sonja M. McKinlay
Peter V. Miller
Roberta Balstad Miller
Lois A. Monteiro
Jean Morton-Williams
Janet D. Perloff
Gail S. Poe
Stanley Presser
Wornie L. Reed
Dorothy P. Rice
Beth B. Rothschild
Patricia Royston
Maurice Satin
Donald W. Schiff
Laure M. Sharp
Eleanor Singer
Monroe G. Sirken
Susan A. Stephens
Robert J. Stimson
Seymour Sudman
Carol W. Telesky
Owen Thornberry
Diane Tuteur
Carmen Noemi Velez
Lois M. Verbrugge
Daniel C. Walden
Richard B. Warnecke
Donna Watts
Michael F. Weeks
Roy W. Whitmore
Stephen Williams
Lucy B. Wilson
N.N. Woodbury

Planning Committee

Charles F. Cannell, Co-chairman
The University of Michigan

Robert M. Groves, Co-chairman
The University of Michigan

Ronald Andersen
University of Chicago

Joseph L. de la Puente
American Public Health Association

Jack Elinson
Columbia University

Jacob J. Feldman
NCHS-DHHS

Daniel G. Horvitz
Research Triangle Institute

William H. Kitching
NCHS-DHHS

Linda S. McCleary
NCHS-DHHS

Seymour Sudman
University of Illinois

Foreword

The Fourth Conference on Health Survey Research Methods was held in Washington, D.C., in May 1982. Initiated in 1975, with the second in 1977 and the third in 1979, these conferences have been jointly sponsored by the National Center for Health Services Research (NCHSR) and the National Center for Health Statistics (NCHS). For this Fourth Conference, our two organizations were joined by the Milbank Memorial Fund, which provided valuable additional support.

The long-range goal of this series of conferences is to improve the quality of health survey data collected and used by those responsible for shaping health practices, policies, and programs. The more immediate objective is to provide a forum in which methodologists can gather and interact to generate methodological research agenda, identify major needs, and define some major hypotheses to guide research endeavors. There are five specific objectives:

1. Assemble survey researchers from a variety of health and social sciences and statistical disciplines who are involved in methodological research pertinent to the field of health surveys.
2. Stimulate the interests of survey researchers in implementing research programs.
3. Bring unpublished scientific and technical work as well as work in progress to the attention of the professional community of researchers.
4. Suggest and recommend new areas of needed research to improve the quality of health survey data.
5. Share the conference results broadly with the health care professions through publication of the proceedings.

These conferences have been a marked success both in their intended purposes and in publishing conference findings which serve to acquaint health service researchers (whose primary skills are not in survey methods) with the limitations and difficulties inherent in health surveys, as well as to apprise researchers (whose interests and skills are in the area of health survey methodology) of the research needs and priorities identified by conference participants.

Our organizations have been pleased to sponsor these conferences since their beginning. This Fourth Conference was supported through grants to the Survey Research Center of the Institute for Social Research at The University of Michigan from NCHSR and the Milbank Memorial Foundation, as well as by services provided by NCHS.

John E. Marshall
Director
National Center for Health Services Research

Manning Feinleib
Director
National Center for Health Statistics

Acknowledgments

We are pleased to acknowledge the financial and intellectual support for the Fourth Conference on Health Survey Research Methods received from the National Center for Health Services Research (NCHSR), the National Center for Health Statistics (NCHS), and the Milbank Memorial Fund. The first two organizations have supported the conference series since its beginning in 1975; the Milbank Memorial Fund made a generous contribution to this year's conference.

We particularly commend the services provided by James H. Smith, Director, Conference Management Branch, NCHS, and members of his capable staff, Barbara I. Weisel and Mary Gabriel. Their competent and enthusiastic assistance was a major contribution to the success of the conference.

From the Institute for Social Research at The University of Michigan, Sonya Kennedy had the central role in coordinating conference planning and organization and in overseeing the final production of these proceedings. Aimee Ergas skillfully edited the entire manuscript. Marion Wirick prepared the bibliography and index.

The Planning Committee

Contents

Introduction

Charles F. Cannell and Robert M. Groves, The University of Michigan

The Fourth Conference on Health Survey Methods was held in May 1982 in Washington, D.C. Previous conferences were held in 1975, 1977, and 1979; all have been supported through grants and staff participation from the National Center for Health Services Research (NCHSR) and the National Center for Health Statistics (NCHS). When rising conference costs for travel and accommodations threatened to exceed the available Federal grant funds, a generous and timely grant from the Milbank Memorial Fund ensured the continuation of this conference series. The Committee greatly appreciates both the financial and the professional support of these organizations.

The aim of this conference series is to provide methodologists with an opportunity to meet and exchange research ideas, hypotheses, and findings, rather than to disseminate methodological findings to survey practitioners. Methodologists tend to be isolated from one another because they come from many academic disciplines, belong to different professional organizations, and read a variety of journals. These conferences represent the only set of meetings devoted to identifying and discussing the methodological research issues underlying the quality of health survey data.

Because a primary objective of these conferences is to foster informed discussion and an exchange of ideas among methodologists, participation is limited to those who have been invited to present papers or to serve as discussants. These published proceedings are intended to disseminate the information generated in these invitational meetings to all interested survey researchers and practitioners.

Over the years the conferences have undergone some gradual changes from the informal and relatively unstructured sessions of the 1975 conference (with about 50 participants) to the more formally planned and structured sessions of the 1982 conference (with over 80 participants). The Planning Committee for the Fourth Conference established six general topics of current importance to health surveys, and a call for abstracts was announced in the relevant professional journals. After reviewing some 82 submitted abstracts, the Planning Committee selected 23, making up the program of six technical sessions. Each session also included a formal discussion of the papers, with comments prepared by the invited discussants, and an open, informal discussion among all the participants.

Some conference topics are of continuing interest through the years, with a constant need for updating developments and research findings (telephone interviewing, for example); others seem to stimulate little new research (improving response rates, for example). Still other topics are new this session; techniques for sampling rare populations, for example, is a topic of special interest in health research and a developing research literature is now available. This topic has been discussed in each previous conference but this year, because of recent research, it was given an entire session.

Shortly before this conference Gerald Rosenthal had left as Director of the National Center for Health Services Research, and one month following the conference, Dorothy Rice, Director of the National Center for Health Statistics, retired. These two leaders have been strong and enthusiastic supporters of research in health survey methods in general and of these conferences in particular. Both have consistently promoted the importance of furthering methodological research in order to improve survey methods for collecting health data. We, the Planning Committee, and all the future participants will miss their leadership. We wish them well in their new endeavors.

Toward a Research Agenda

This series of conferences was initiated because of the critical need for survey data that are sufficiently accurate and precise to permit sound policy decisions. Ruth Hanft, former Deputy Assistant Secretary for Health Policy, Research and Statistics, Department of Health and Human Services, speaking at the Third Health Survey Research Methods Conference in 1979, emphasized the significance of health surveys with these remarks:

> Health surveys are relatively new in this country, but over the past 20 or 25 years they have helped to shape health policy and programs. For example, when the final push for health insurance for the aged began in the 1960's, survey data on health and use of services by the aged were used to document need and to plan the Medicare program that emerged. Similarly, data on differentials in health status among income groups were important in designing the programs of the "War on Poverty" and other efforts to improve delivery of health services in the U.S.

Surveys are the major source of social and economic data on health, and they involve a sizeable proportion of research funds. Unfortunately, there is much evidence that survey data in general are fraught with errors, both systematic and unsystematic, and there are strong demands for improving the validity of health survey data.

The last decade has seen an increase of interest and research activity in studies of survey errors: to measure their magnitudes and to evaluate techniques for reduction. Survey practitioners have generally become more sensitive to the effects of survey errors and more interested in improving the quality of data. The concept and measurement of total survey error has received increased attention and activity, including efforts to identify, measure, and alter various components of error.

This Fourth Conference on Health Survey Research Methods focused on the development of survey techniques to measure and control survey errors. From the various papers and discussions, we have distilled the following set of statements to serve as suggestions for a research agenda for survey methodologists.

A research agenda for survey methodologists

With the development of estimation procedures for structural equation models containing explicit formulation of measurement errors, survey analysts can measure the effects of some errors due to measurement form. Continual efforts on the part of health researchers to introduce multiple measurements of single phenomena and experimental variation of survey procedures can inform the analyst of different error sources that affect the survey data.

Studies of interviewer effects on the quality of survey data are important to the measurement and reduction of survey error. Measuring interviewer effects should be included in the design of studies, especially those in which new interviewing procedures or new substantive topics are being explored. It is noted that centralized telephone interviewing facilities provide an optimal environment for studies of interviewer effects through tighter control on selection and training interviewers, assignment of cases to interviewers, and the monitoring of interviewer behavior.

Health statistics are often complex combinations of many answers to questionnaire items. In comparing differences across surveys, it is important to consider the design features that affect the statistics—these include respondent rules, reference periods, question form and wording, postsurvey adjustment procedures, and other aspects of the survey. These design features should be documented and referenced in the presentation of results in order to inform comparisons across surveys that appear to measure the same quantities.

As survey methods become more useful to the health profession, there is an increased demand for a wider variety of health measures with an accompanying increase in the need to evaluate these measurement procedures. General population surveys that involve physiological measurements pose problems of gaining respondent cooperation. Cooperation has been demonstrated using a procedure that links the interview to feedback to the the respondent regarding his or her basic health status. Highlighting the benefit of such measurements to potential respondents might improve cooperation in such measurement efforts. In one study physiologic measurements made in a field setting were found to be as reliable and valid as those made at a central site. Further work is needed to replicate this result and to estimate cost savings and limitations of the method. For example, a wide variety of respondent types need to be studied to determine their reactions to such requests and to evaluate types of instrumentation that can be used in household surveys.

The increased use of panels, telephone interviews, and diaries in data collection raises issues that need exploration. Panel studies are particularly sensitive to a decrease in response rates over the life of the panel. If problems of panel attrition can be solved, the method offers potential for cheaper and richer data than single-wave surveys. Similarly, overall cooperation with diary-

keeping in a population sample can be high, but is dependent on formatting the diary in an attractive manner and using survey procedures that make it easy for the respondent to complete the task. This requires careful communication with respondents, with special attention given to groups who tend to have low cooperation rates.

Record check studies are typically performed using either a sample of health records with checks to the persons (reverse record check studies) or a sample of persons with checks on their health records (forward record check studies). Either of the two modes allows the investigator to identify errors of underreporting or errors of overreporting, but not both. The belief that health events tend to be underreported may indeed be a result of the prevalence of reverse record checks. Despite their large expense, full-design record checks need to be mounted in order to quantify the relative sizes of underreports and overreports for health variables.

Subgroups of interest to health researchers often form only a small portion of the society (e.g., persons with specific health experiences). Because they tend to be difficult to locate, these rare populations require the development of new measurement and sampling techniques. Several new developments and refinements of traditional approaches were discussed at the conference. One approach to measure rare subgroups uses administrative records, either to identify eligible persons for the survey or to use data contained on the records for analysis. It was noted that health data are recorded by health service organizations for administrative purposes. They tend to be characterized by limited coverage of the population and by relatively sparse data on those covered. When supplemented by survey data, however, administrative records offer a source of complementary data and the possibility of examination of response errors in both the survey and record measures. Using both the administrative record and the person's self-report about a health experience often gives the researcher a richer understanding of the phenomenon.

Administrative health record systems can also be used as a sampling frame for identifying eligible persons. List frames, when supplemented with other sampling frames, have been found to increase the cost efficiency of health surveys. Another sampling technique in rare population surveys is that of multiplicity sampling. In this technique sample persons provide information on well-defined networks of people attached to them (e.g., families). Eligible persons in these networks are also measured, and through adjustment of statistics for known but unequal probabilities of selection, these multiplicity samples can yield efficient estimates of population parameters. The statistical properties of multiplicity sampling are currently better understood than the practical difficulties of implementing them. More research is needed in the survey methods to accompany multiplicity sample designs.

Even after the rare population is located through these new sampling techniques, there are often serious problems of measurement. Subpopulations of interest to health researchers often have health conditions that make it hard for the individual to supply the survey data because of physical impairment or embarrassment. Special techniques for some populations were discussed, but it was noted that each subpopulation and each topic area seems to present unique difficulties that will tax the inventiveness of the survey methodologist.

The telephone survey is becoming the most prevalent mode of data collection in survey research. As a relatively new mode, it presents many issues and problems needing investigation. Perhaps the greatest need is to develop methods either to improve or to compensate for response rates which generally are lower than in face-to-face surveys. Not all segments of the population are equally likely to have telephones, and thus noncoverage bias is a potential threat to telephone survey validity. More knowledge is needed to map the coverage problems. Nonresponses tend to cluster in some population subgroups, but the characteristics of nonrespondents need to be better identified. The problems of telephone coverage and nonresponse bias are especially important in health surveys since those population segments likely to have lower telephone coverage are also likely to have more illnesses and make less use of health services.

Telephone surveys are less expensive than face-to-face surveys. However, one reason for this is that telephone interviews tend to be shorter (although the permissible length is undetermined), and therefore they can collect less information. Cost-effectiveness studies are needed to ascertain distribution of costs for the two modes. Experimentation is needed to develop procedures to reduce costs for both modes.

Response differences and relative accuracy of reporting in telephone and in face-to-face interviews have been compared in some past studies, but few have been designed to measure response errors without confounding by nonresponse or noncoverage errors. Specialized research is needed to permit general conclusions about the response error properties of the two modes. For studies of response bias, record check studies are recommended. The major issue is not simply the comparison between the two modes, but the level of response validity

of both. For the determination, an accurate record base is necessary.

Telephone interviews from a central location permit close monitoring and control of interviewing. Random assignment of interviewers to respondents is also possible for studies of interviewer variability and their effects. Researchers should take advantage of this more rigorous supervision and should study error attributable to interviewers.

Studies demonstrate that responses can be improved with use of techniques to encourage better respondent performance. More attention needs to be given to improving questionnaire design and developing techniques that will produce more valid responses. Recent developments in computer-assisted telephone interviewing (CATI) introduce new dimensions into telephone surveys. Studies are needed on programming the questionnaire, training the interviewers, and reducing errors and costs.

There has been surprisingly little research on training and supervisory methods in interviewing. However, systematic investigations are now being undertaken and promise to provide useful guidance. Centralized telephone facilities offer the potential for greater control over interviewers' behavior through close monitoring of their work, rapid feedback, and retraining as needed. Face-to-face studies present very different problems. The results on interviewer training reported at the conference deserve more study in order to take advantage of the opportunities they offer.

Conclusions

These comments from the conference represent both continuations of long-existing issues in health surveys and some new evolving ones. Together they present the survey methodologist's view of methodological problems and challenges and represent a series of problems that need to be researched.

SPECIAL SESSIONS

NCHS in Perspective

Chair: Charles F. Cannell, Survey Research Center, University of Michigan

Politics and the Social Sciences—Yesterday, Today, and Tomorrow

Chair: Robert Groves, Survey Research Center, University of Michigan

Health Surveys in Other Countries

Chair: Jacob Feldman, National Center for Health Statistics

Recorder: Jack Elinson, School of Public Health, Columbia University

NCHS in perspective—introductory remarks

Donald Goldstone, Acting Director, National Center for Health Services Research

I have long since been convinced the amount of time that the National Center for Health Services Research (NCHSR) is allotted on any given program is directly proportional to the size of its budget. It was no suprise then when we were asked to make these comments brief.

These are obviously difficult times for social research in general and health services research in particular. Federal agencies that formerly supported studies in these areas have had their budgets reduced as part of the general effort to reduce Federal expenditures. Unfortunately, the budget for NCHSR had declined over several years preceding the present emphasis on controlling the cost of goverment. The new reductions experienced by NCHSR have been all the more consequential. It has never been clear why health services research and, for that matter health statistics, have not been supported with great enthusiasm. Both Centers provide information on programs that are of critical concern to the Administration and to the Congress.

Nor is there any question that this information is being used on a regular basis. Health services research and health statistics programs have had a major impact on Federal and local policies and programs during the past decade. Health services research, for example, provided the basic information that justified the Health Maintenance Organization (HMO) initiative. PSRO was predicated on an early experiment with the Experimental Medical Care Review Organization (EMCRO). In this latter case, an experiment became policy before any evaluation could be completed. This program might have been substantially more effective if we had waited for the research findings. The Rural Health Clinic Act is based on the results of specific NCHSR projects. Medical technology assessment projects supported by NCHSR have provided information that, were it used, could save this country billions now being spent on inappropriate and ineffective diagnostic and therapeutic procedures. New Jersey is now experimenting with a promising new approach to prospective reimbursement that is the result of research funded nearly a decade ago.

These observations have been made in oral and written testimony before budget committees. Yet such evidence has had little influence on the outcome. Research and statistics seem to be viewed as luxury goods—something that we can afford when the general budget is growing and expendable when the budget is being reduced.

The limited growth in the health statistics budget and the reduction in health services research have had a number of untoward consequences. The decline in health services research funds has caused eminent investigators to tranfer their interest elsewhere. New investigators are not entering the field in sufficient numbers. Clearly, the amount of research being done, as well as its quality, has declined. Surveys and studies that might have been undertaken are now not even being discussed. As a result, the information required to inform policy is increasingly out of date. We are drawing down our inventory and it will not be long before the shelves are bare.

The purpose of this conference on health survey research methods is to improve the quality of health survey data. This and earlier conferences have concerned themselves with methodological problems. Sound methodology is a critical prerequisite to the accumulation of reliable data on the cost, content, and use of health services.

To those deeply involved in such substantive problems, budgets may seem a mundane concern. It is important to recognize, however, that in times of reduced circumstances, methodological research is likely to be eliminated before research projects which seem to have more direct, immediate, and identifiable importance in the eyes of those who provide the funds. Methodological research provides the basis for good data and analysis. But this fact tends to be appreciated only by those within the field.

I would hope that the problems of finding support for this area of inquiry and the production of a new generation of investigators prepared to address methodological issues will be treated as no less a matter of concern at this conference than the reports of the progress that has been made in the research itself.

NCHS in perspective

Dorothy P. Rice, Director, National Center for Health Statistics

It is a privilege and an honor to welcome all of you to this Fourth Conference on Health Survey Research Methods. This is the third conference which I have had the honor to address as director of the National Center for Health Statistics, and it is a special occasion for me because it is the last conference I will be addressing as director.

I want to take this opportunity to reminisce a bit about the beginnings of this conference series and to talk about the future of NCHS. I was not involved in the first conference, held on May 1–2, 1975, at Airlie House in Virginia. In his remarks at that conference, Dr. Rosenthal stated several reasons for the importance of methodological research in health surveys to the viability of health services research:

> (1) A significant proportion of analytic work in health services research is based on survey data; (2) The quality of initial requests for research support is diminished by inadequacies in design and inappropriate specification of data pertinent to the research issues; (3) The analyses of data developed by surveys is often deficient because we cannot correct for errors in measurement; (4) The evaluation of demonstration efforts require baseline and follow-up surveys...; (5) There is a need for improved health surveys in terms of the time frame of the research in order to avoid delays in the completion of the studies. Investment in the overall design of the survey could result in significant savings and improved quality of the data being obtained.

Those reasons are no less important in 1982 than they were in 1975. The successful conduct and outcome of the past three conferences and the large turnout at this one are testimony to the continuing interest in and need for improvements in health survey research methodology.

The first conference held in 1975 was semi-structured and focused on four main topics: research instruments, interviewing, problems of validity, and total survey design. The important new areas that were discussed included telephone interviewing and network or multiple respondent surveys. The following conferences were more structured than the first and covered a multitude of important topics, including issues of respondent burden, standardization of survey items or measures, ethics of social research in health, panel surveys, provider surveys, and implications of survey research for health policy programs. By the third conference in 1979, we had lost prematurely two founders of the conference series—Elijah White and Leo Reeder—as a result of terrible accidents. We still keenly feel and mourn those losses.

This fourth conference in the series will cover the following timely and significant topics: telephone survey methodology, use of records in health survey research, survey methods for rare populations, measures and correlates of response errors, health surveys in other countries, studies of survey measurement techniques, and the hiring, training, and monitoring of interviewers. These topics and earlier ones are essential for the continued growth and development of the National Center for Health Statistics.

One illustration of the impact of these conferences on the work of the Center is the research we have conducted on the telephone interview. NCHS interest in the telephone interview as a mode of data collection began with the first conference. The following year, 1976, I established a committee to assess and report on the potential applicability of the telephone interview and the telephone survey methodology to the data collection needs of the Center. A number of projects grew out of the NCHS work groups, including the development of an in-house Telephone Health Interview System at the Center, a major contract with the Survey Research Center of the University of Michigan for methodological research on the telephone interview, and a national random-digit-dialed telephone survey of personal health practices. Finally, we are now looking at the use of the telephone as a component of our redesign plans for the Center's population-based surveys. The papers this afternoon will be devoted to telephone survey methodology, illustrating the progress and evolution over the years of this important subject at these conferences.

During my six years as director of the Center, I have seen much progress and many accomplishments in survey research as well as in the conduct of old and new surveys. We have come a long way in our health data collection efforts. Few people now dispute the need for data as a basic requisite for the development and implementation of policy and programs in every area of our national life, whether it be health, welfare, defense, agriculture, transportation, or any other. The question now is: "What is the appropriate role of the federal government in statistics?" and the complementary question is: "What is the appropriate role of the private sector?" Dr. Edward Brandt, the Assistant Secretary for Health, recently formed a task force, which I cochair along with Dr. Martin Cummings, Director of the National Library of Medicine. We are investigating the role of the private sector in the collection, analysis, and distribution of health data and information. We are asking each of the Public Health Service agencies the extent to which there is duplication of effort between the public and private sectors and the extent of reliance by these agencies on the data produced by the private health organizations.

What is the current climate with respect to federal statistics? Questions are being raised about the need for data. All the federal statistical agencies are facing extreme budget pressures that will have a severe impact on the federal statistical system. Several recent articles in *The New York Times* and *The Washington Post* summarized the budget pressures and problems succinctly and accurately. The October 18, 1981, issue of *The New York Times* contained an Op-Ed article entitled "Data on Cuts Imperiled by Cuts in Data," which stated the following:

> Yet even as the White House devours all this data, the agencies that provide them—under extreme budget pressure—are busy eliminating or scaling down the censuses and surveys used to gather them. Figures from the 1980 census are oozing out of the Census Bureau like cold molasses, and the Bureau is likely to curtail its many economic and agricultural censuses, industrial reports, and educational and fertility surveys. The National Center for Health Statistics has dropped some of its health reporting and is contemplating more cuts. The story is much the same at the Bureau of Labor Statistics and other agencies.
>
> All of this is occurring at a time when many argue that good reliable statistics are more needed than ever, given the Reagan drive to reshape federalism.

On October 19, 1981, *The Washington Post* published an editorial entitled "Truth in Numbers," which stated the following:

> Numbers can be powerful weapons, especially against those who don't understand them. And no administration has made more effective use of numbers to promote its social policies than the current one. Budget director David Stockman, the architect of many of those policies, is an avid consumer of federal statistics. You may question the quality of some of his data or disagree violently with his conclusions, but you cannot deny his numerical ability. His sources are the federal statistical agencies that produce data on income, employment and other measures of economic and social well-being.
>
> In recent years, pressure on these agencies has increased for more and better data. As the economy moves through a period of rapid structural change, it is increasingly important to look behind the old familiar statistics to see what is really going on.
>
> Because the administration's policies can be expected to have different consequences for different areas and different types of people, measuring their progress will require more detailed data than is now available. These new demands will come at the same time that statistical agences will be taking substantial cuts in budget and personnel. Some important surveys have already been cancelled, and the Bureau of Labor Statistics has announced delays in future releases on employment, earnings, prices, and other indicators.
>
> Agencies will face a difficult choice between the quality and the quantity of their products—and the heavy pressure is likely to be on the side of quantity. There is a saying among seasoned players in the policy game that "some numbers beat no numbers every time." That's quite true. But there is an important corollary, which is that bad numbers are much more dangerous than no numbers. As the nation enters a period of major change, the greatest care must be taken to ensure not only the availability but the quality of the numbers that guide and measure government policy.

The April 17, 1982, issue of *The National Journal* stated in an article entitled "Bad Numbers, Bad Government," the following:

> In modern society, one of government's jobs is to keep us informed about ourselves. Like defense, this task naturally (and appropriately) falls to government, because it is in everyone's interest to provide it. The Reagan Administration compromises this role when it skimps on statistics and charges for everything (including basic budget documents) in sight.
>
> All this is as dry as dust, but what ultimately is at issue is the competence of government. Future [*Statistical Abstracts of the United States]* will be no less thick, but with current trends, the numbers inside may be less believable. Good information may not create good government, but bad information risks bad government.

A final blow to the federal statistical system is the dissolution last week of the Statistical Policy Branch in the Office of Management and Budget. The Statistical Policy Branch served as a coordinating mechanism for all of the federal statistical agencies, set statistical standards, promulgated uniform classification systems, and was the focal point for international statistical activities. Support for this office has eroded over the years. Last week's action to dissolve the remnants of what should be a high priority activity clearly illustrates the low priority given to the coordination of federal statistics.

The National Center for Health Statistics is also facing organizational problems that may have a serious impact on us. The National Center for Health Statistics and the National Center for Health Services Research are being reorganized into a new agency that will combine the Health Resources Administration and the Health Services Administration. I feel strongly that the credibility of the Center's statistics may be damaged if it is moved from its present position in the Office of the Assistant Secretary for Health, which has jurisdiction over all components in the Public Health Service. If NCHS were to be moved to the new agency, its priorities and resources might be subject to new and more narrow interests. We have attained a great deal of prestige in our current organizational location; it is important that we do not lose ground.

As director of one of the federal statistical agencies, I believe that the *best* way to provide objective high quality information on the demographic, economic, social, and health characteristics of our population and trends in those characteristics is through agencies specifically established for that purpose. These agencies have "no axe to grind," can usually guarantee confidentiality to respondents, and hence are able to produce unbiased, quality information acceptable to a wide array of users both within and outside of government. However, even in the best of economic times it is difficult to obtain adequate budgets to support the necessary data collection and analysis activities. Recognizing the philosophy that the federal government should only be in the business of doing things that cannot be done adequately by states and/or the private sector, it may be necessary to

reassess the core programs of the federal statistical system.

Regardless of what changes must be made in the core programs we must ensure that an information base continues to be available that will provide baseline data, that will be useful for monitoring trends, and that will have the ability to quickly detect any changes or aberrations in the economic, social, or health characteristics of the nation. I believe that the appropriate federal role in statistics is to produce national-level data useful for those purposes as well as providing norms to which subnational data can be compared. The data must be of high quality, produced in a timely manner, and relevant to the issues of the day. Federal statistical agencies must assume responsibility for activities that cannot reasonably or feasibly be assumed by individual states, local governments, and the private sector. The federal role must include the development and promulgation of standards and procedures for assuring the validity, reliability, comparability, and quality of statistical products and the provision of technical assistance in these areas. Federal statistical agencies also have to anticipate future needs for information and design today's systems to meet those needs.

In conclusion, in considering the future prospects for improved health statistics to meet the nation's needs, we must recognize that resources will not grow parallel to demands for data and services. The demands for health data are greater than our ability to produce them. Budgetary pressures are requiring reassessment of current data collection and dissemination procedures. Statistical agencies must make choices among data collection, research, and analysis, and among needed data sets.

As we move closer to our objective of a national and systematic approach to meeting the information needs of policy development and program evaluation, we also need to coordinate our data collection activities both within the federal establishment and between government and the private sector. Although considerable progress has been made in coordination, we must continue to avoid unnecessary and costly duplication, to encourage comparability of information collected by different systems, and to use the ongoing data collection programs to provide specific information for many organizations. More effort is needed to provide essential data, yet reduce the burden on individual and institutional respondents.

What emerges is a challenge. The hallmark of the National Center for Health Statistics has been a responsiveness to changing needs resulting from advances in survey and statistical methodology, as well as from changes in requirements for data. As we look ahead, it is clear that the Center must continue to evolve if it is to accommodate the data requirements of the future as it has those of the past.

As I plan for my retirement after 30 years as a federal career civil servant, I look back with great pleasure and pride on the many wonderful professional and personal friends with whom I have been privileged to be associated. As Director of the National Center for Health Statistics, I have been most fortunate to work with a dedicated group of highly professional, competent, talented, knowledgeable, and capable individuals who are committed to the production of high quality, reliable, timely, and useful health statistics and health information. I have been truly fortunate to work with such dedicated and professional individuals, both in the Center and in the many public and private organizations with whom we have had contact. I know that the Center will continue to grow and flourish in the future as it has in the past.

My best wishes for a successful conference!

Politics and the social sciences—yesterday, today and tomorrow

Roberta Balstad Miller

ROBERT GROVES:

Social scientists have banded together in an attempt to influence the political process that determines levels of support for research. We are obviously new at the game but as the short track-record shows, we are learning fast how to pump information into the system regarding the worth of social science research for the general society.

One of the groups at the forefront of this effort is the Consortium of Social Science Associations (COSSA). This is an organization supported by ten separate disciplinary associations, and the group's offices act as a clearinghouse for getting social scientists involved by writing to members of Congress, testifying in front of committees, and so on. At the forefront of COSSA's activities is Roberta Balstad Miller, our speaker for tonight. Dr. Miller received her Ph.D. in history from the University of Minnesota. She has published widely on issues of regional development and the measurement of scientific productivity. The popular press has in several articles noted how quickly and effectively social scientists have banded together to save funding in basic and applied areas. Much of the credit of this success goes to Roberta Balstad Miller, and we are privileged to have her speak tonight.

ROBERTA BALSTAD MILLER:

Thank you very much. I think what I will do is to step back and talk very broadly about the general topic of politics and the social sciences rather than talking specifically about this year's political situation. Partly, this is because the question of politics and the social sciences is too seductive—it is very easy to get caught up in the political battles of the day, without ever stepping back to see just how these battles relate to longer-term issues. I think this stepping back can be useful, particularly at the present time, because the political status or situation of the social sciences has been in such turmoil over the past year. Yet, the Reagan administration certainly has not been the first group to note the political vulnerability of the social sciences, and it has not been the first to try to take advantage of it.

As I begin, I should perhaps warn you of my own bias on the subject of politics and the social sciences. I think that it is a dangerous mistake to see the Reagan budget cuts as an isolated or unusual political act or as a series of events that took place this year because of the current administration in Washington. Quite the opposite. I think we in the social science community will face great political and professional problems if we fail to see the budget cuts in long-term perspective. The political vulnerability of the social sciences is not new, and the dislike of social science by people like David Stockman, William Proxmire, and the late John Ashbrook is but a recent manifestation of something that has been going on for some time.

There are, I believe, several ways in which politics and the social sciences have been interwoven in the past. One of these is in political theory; a second is through the person of particular social scientists; and a third is through the role of social scientists in the government. A fourth is in the role of research in public policy, but a discussion of a subject that broad is beyond the scope of this paper. I would, however, like to look at each of the first three a bit more carefully.

Social theory and political theory have frequently been influenced by social science ideas and methods. One of the best known examples of this impact of the social sciences is in the works of Karl Marx, which attempt to apply the scientific method to historical and social analysis. A second example is Darwinism, which is essentially an adaptation for the natural world of ideas on the nature of cultural change from the social sciences. This adaptation was later, as natural science, used to confirm the social science ideas that it had originally drawn upon, ironically, as Social Darwinism. This adaptation of social science ideas had a pervasive political influence in the nineteenth century that was generally opposed to Marxism and socialism.

A second link between politics and the social sciences has been through individuals—scholars whose political involvement seasoned and frequently informed their scholarship. The direction of influence here is the opposite: from politics to the social sciences. A very good example of this kind of interaction between politics and the social sciences is in the life and work of Benedetto Croce, the leader of the Italian Liberal Party for many years and the major historian of modern Italy.

But there is also a third link between politics and social sciences, one that was dominant in the United States and that has colored public perceptions of the social sciences for many years. In some ways this is a cross between the other two types of relationships where social scientists are independent scholars and scientists and, at the same time, advisors to and participants in the government.

It is this third link that I wish to discuss tonight, the role of social scientists as government advisors. For the roots of the current political vulnerability of the social sciences and, perhaps, its vulnerability in the scientific

community as well, lie in the long-term practice of using social scientists as government advisors. I should make clear here that in looking at the role of social scientists in government I am not criticizing the social sciences for taking part in government. I think it very important for national policy to be informed and guided by the best and the most current social science research available to it. I do see, however, that there has been confusion in the past between social science research and the government policy that is derived from it, between the problems and institutions that have been studied and those who have studied them. It is this confusion—confusion on the part of politicians, confusion on the part of both natural and social scientists, and even, at times, confusion on the part of the public—that has created political problems for the social sciences in the past. It is, in fact, because we are beginning to deal with this confusion today that I am rather optimistic about future relationships between the social sciences and politics.

However, our situation today is still shaped to some extent by our condition in the past. For this reason I want to step back tonight and look at the changes that have taken place over the past 50 years to give us some perspective on where we are at the present.

In the early twentieth century, social science, that is, the systematic study of social conditions, was closely related to local reform movements. C. R. Henderson, a Chicago sociologist, went so far as to say that "to assist us in the difficult task of adjustment to new situations God has providentially wrought out for us the social sciences and placed them at our disposal."

In the 1920s, however, social scientists turned away from social reform as an impetus to research. Instead, they tried to make the social sciences more "scientific" by looking to the methods of the natural scientists to explain and describe social phenomena. In addition, they broadened their scope to encompass the study of national rather than local social conditions. This changed approach coincided with an interest in national social planning in the federal government. President Herbert Hoover, more than any of his predecessors, believed that social policy should be based on the accumulation and analysis of social data, a belief which led in 1929 to his establishment of the President's Research Committee on Social Trends. This Committee employed leading social scientists in a massive three-year study project to survey with statistics the nation's social resources. While the study was still underway, Hoover regularly asked the social scientists working on his committee for information, assessments, and data on which to base his policies. For Hoover, who was in some ways the quintessential Republican President of the century, both the technical capability and the analytic product of the social scientists were essential to modern governance.

This tradition of looking to the social sciences continued beyond the Hoover administration and flourished in a more active guise during the New Deal of the 1930s. Social scientists were among the most influential members of the circle that surrounded President Franklin D. Roosevelt. Some of these social scientists moved from advisory positions in the Hoover Administration directly into the Roosevelt Administration. During this period the role of the social scientists came to involve more direct political (as opposed to technical and educational) participation in the government. Yet, among the applications of social science research was again an emphasis on scientifically rigorous national planning.

Natural scientists viewed the New Deal impulse toward planning and the employment of science in the national interest with some skepticism but also with interest, particularly when they saw that it might involve research funding for their own projects. The vehicle proposed by certain scientists for such support was the Science Advisory Board (SAB) which was established in 1933 and headed by the physicist Karl Compton. The story of the SAB has been told with great frequency and abundant detail elsewhere and I do not intend to go into it here. I do wish to point out, however, that the failure of the SAB to accomplish its aims or even to survive beyond its initial two-and-one-half-year period was perceived at the time to be due to the jealousy and resentment of social scientists in the administration because of the absence of social scientists on the SAB.

The failure of the SAB consequently became a source of antagonism on the part of natural scientists for social scientists. The strength of this antagonism is a measure of the perceived power of the social sciences at that time. More importantly, its residues undoubtedly colored relations between the natural scientists and social scientists in the 1940s, particularly since some of the leading antagonists to social science research in the postwar period (such as Vannevar Bush and Isaiah Bowman) were intimately involved in the SAB.

The blame borne by the social scientists for their role in the SAB was somewhat misplaced. Recent scholars have suggested that the failure of the SAB was due to the fact that the natural scientists involved in the SAB had very different political and economic ideas than the administration and that it was very unlikely from the start that Roosevelt would ever give them control over their own research funds.

Regardless of the real cause for the failure of the SAB, the result of its failure was that natural scientists began to distrust the social scientists for their perceived influence in the failure of the SAB, for their power in the Administration, and for their tendency to force individuals into what Robert A. Milliken (a member of the SAB) called "the soft bosom of the State." This distrust was not without consequence in the later evolution of discussions on the National Science Foundation in the mid 1940s.

During World War II, the U.S. government engaged in research on a scale far broader than ever before. Most of this research took place under the aegis of the Office of Scientific Research & Development (OSRD) estab-

lished in 1941 and headed by Vannevar Bush. However, quite separate from the research carried out in OSRD, there was an additional expansion of governmental support for social science research. The Army Research Branch, the War Relocation Authority, the Office of Strategic Services, and the Office of War Information supported major programs in social science research. In part because of the administrative diffusion of this research, social science research never attained the wartime visibility of the research in the physical sciences undertaken at OSRD. More importantly, perhaps, it was overlooked by those who were responsible for drawing up U.S. science policy in the postwar period. In November, 1944, as World War II neared its end, FDR sent Vannevar Bush a letter asking him to recommend what the federal government could do to aid research and improve the training of young scientists and engineers in the postwar period. Seven months later Bush replied in a report entitled, *Science—The Endless Frontier,* which he submitted to then-President Harry Truman. In this report, Bush recommended the establishment of a permanent Science Advisory Board and a new agency which he called the National Research Foundation, later to be called the National Science Foundation (NSF), as a vehicle to provide government funds for scientific research. His recommendations for the new foundation included no provision for the support of research in the social sciences. He stated:

> It is clear from President Roosevelt's letter that in speaking of science he had in mind the natural sciences, including biology and medicine, and I have so interpreted his questions. Progress in other fields, such as the social sciences and the humanities, is likewise important; but the program for science presented in my report warrants immediate attention.

One reason why Bush was so certain that Roosevelt's letter referred simply to the natural sciences might be that Bush himself originally drafted the Roosevelt letter. Certainly the text of the letter itself provides the reader with less certainty than Bush displayed. The only way the letter can be read to suggest that the social sciences are to be excluded from consideration is to arbitrarily define the term "science" in the letter to exclude the social sciences. This is what Bush appears to have done.

Despite Bush's reluctance to include the social sciences in his recommendations for the new foundation, the question of their inclusion was explicitly raised two months later by the new president. In September 1945, Truman sent a message to Congress urging that the legislation for a science foundation include provision for the promotion and support of research in the basic sciences and the social sciences. In response to the President's request, the compromise Senate bill for the National Science Foundation included a provision for the social sciences in the foundation.

In the first day's hearings on the bill, Isaiah Bowman, an alumnus of the SAB controversy during the New Deal, who had originally proposed the establishment of the SAB in 1933 and been appointed a member of the Board, turned the tables in the hearing and began to question the Senators about the presence of the phrase "including the social sciences." He was told by Senators Magnuson and Kilgore that it was included because of Truman's interest in the social sciences. Kilgore himself welcomed and strongly supported the president's interest in this issue. When Bowman was later asked to testify in hearings in the House of Representatives, he was asked by Congressman Clarence Brown whether there was, as he perceived, some antipathy toward the social sciences. Bowman replied, "The remarks are a summary of the views of the scientists who have testified before the Senate Science Committee."

Yet, despite the opposition of Bowman and other prominent natural scientists to the inclusion of the social sciences in the National Science Foundation, most of the scientific community did not share his feelings. In congressional testimony on an earlier version of the bill, both James B. Conant and J. Robert Oppenheimer had suggested that legislation for the Foundation be amended to include the social sciences. Similarly, the executive secretary of the American Association for the Advancement of Science (AAAS) testified at Senate hearings that an AAAS poll showed "a substantial number of physical scientists believe that the social sciences should have an integral place in the program." Moreover, a survey of more than 4,000 scientists conducted somewhat later by *Fortune* magazine showed that 81% of all academic scientists, 83% of all scientists employed by the government, and 76% of scientists employed in industry believed that the federal government should support social science research.

Although there was broad scientific and governmental support for the social sciences, spokesmen for science who opposed or only lukewarmly supported the inclusion of the social sciences in the foundation ultimately prevailed. They opposed the social sciences for a number of reasons. They were unfamiliar with social science research, skeptical about its scientific base, and distrustful of both social scientists and the political implications of their research.

In this latter area, the fears of some conservative physical scientists reinforced the fears of conservative politicians in such potentially explosive areas as race relations. For example, social science research undertaken by the Army Research Branch had looked at the question of racial discrimination in the military during the war. Moreover, Gunnar Myrdal's monumental study of racism in American society, *An American Dilemma,* was published in 1944, just before these congressional hearings began.

As a result, for many Americans there was an association between social science research and attempts to improve the status of blacks in America, between things and people studied and those studying them. This identification became overt in the Senate debate on the National Science Foundation bill when Senator Hart, the

author of the amendment striking the social sciences from the bill for the National Science Foundation, said in defense of his amendment, "No agreement has been reached with reference to what social science really means . . . it may include all the racial questions." The racial issue had been raised even earlier by one of Vannevar Bush's closest aides, who told the dean of a Texas university that the social sciences would endanger funding for the entire foundation because of Congressional opposition to, for example, "studies designed to alleviate the conditions of Negroes in the South."

It was perhaps this undercurrent connecting the social sciences with the promotion of racial equality that led Harry Truman in later years to blame the defeat of his attempt to include the social sciences in the foundation to the opposition to his Committee on Civil Rights. Truman's memory was faulty. The Committee on Civil Rights was established six months after the passage of the Senate amendment which struck the social sciences from the foundation; yet the association Truman made between civil rights and the exclusion of the social sciences from the National Science Foundation was made by many others at the time. In fact, the first Congressional objection to a "silly" research title was a study entitled "Integration of an Oak Forest Community." Members of Congress thought this was a social science study of Oak Park, Illinois; in actuality it was a study of biological changes in oak forests.

Opposition to Truman's attempt to include the social sciences in the National Science Foundation came from both certain prestigious and influential natural scientists and socially conservative members of Congress. But equally important in the passage of the amendment striking the social sciences from the bill was the social science community itself. Timidity, uncertainty, and inactivity characterized the actions of many social scientists during this period. The absence of a strong, well-articulated position in the social science community on behalf of government support for its research was not due to disinterest. On the contrary, postwar federal support for research had been discussed by social science leaders since 1943. There was some concern, however, about the dangers government support posed to scientific inquiry—fears that social scientists might uncritically take on the methodological coloration of their new colleagues in the foundation. There was also some disagreement among social science leaders whether government support was necessary for future advances in research. Most significantly, many social scientists shared the natural scientists' doubts about the scientific base of social science research and about its political vulnerability.

With the publication of the Bush report in July 1945, social scientists were faced with difficult choices. If they supported the report, despite its failure to deal with the social sciences, they would have seen their own research support decline relative to that of the natural sciences. If, on the other hand, they advocated the inclusion of the social sciences in the new foundation, they risked alienating potential allies in the natural sciences and, perhaps, undermining Congressional support for the foundation itself. To take no position was impossible given the important role that social science had played in the wartime research effort. Truman's decision in September 1945 to press for social science involvement in the foundation provided a way out of this dilemma.

But at this juncture, when social scientists had an advocate in the White House and an opportunity to take a strong stand in support of the President's decision, they failed to do so. In testimony before the Senate committee holding hearings on the foundation, the economist Wesley Mitchell and other social scientists acknowledged "that the social sciences...[had not] reached a stage comparable to that of some of the other scientific disciplines." They argued that to reach that stage was reason enough for the government to support research for the social sciences. This was hardly the posture to effectively counter Bush's argument that the social sciences did not require immediate attention. Nor did it allay the fears of Congressional conservatives who felt that untrammelled social science research might undo the existing social order.

In the end the social sciences were excluded from the new foundation. Final defeat of the attempt to include them was due to three factors. The first was the opposition of key scientists to NSF support for social science research. Second, conservative fears that social science research would emphasize such potential political problems as racial inequality undermined Congressional support for social science research. Third, despite strong support from such Senators as Harley Kilgore, the social science community did not make a strong bid for inclusion with the natural sciences in the new foundation. Any of these factors alone could probably have been surmounted; taken together they were impossible to overcome.

The National Science Foundation, as finally established in 1950, provided for some support for social science research if desired but such support was not to be mandatory. It was not until 1958 that the NSF actually established an Office of Social Sciences and not until 1968 that the Organic Act of the foundation was amended to include the social sciences within the legislated jurisdiction of the NSF.

In some sense, the problems facing the social science community over the past year are similar to the problems facing it nearly 40 years ago. But in certain key respects, they are very different and the response, certainly, has been very different. Had the Reagan Administration been less virulent in its original attack on the social science community and had the administration been less successful in its initial months in office, we might have seen a replay of events of the 1940s. As it was, the Reagan Administration's budget cuts were very severe, very deep, and perceived (quite rightly) as a threat to the integrity of social science research itself. As a result, we had support

from the natural scientists that we did not have in the 1940s. The impact of this support cannot be underestimated.

A second difference is that in the 1980s, unlike the 1940s, we were able to win strong support from a broad spectrum of members of Congress. The very first vote on a National Science Foundation issue, and on any basic research funding issues, was at the end of July 1981. The vote was on an amendment to decrease funding at the foundation, to cut out some funding that the Appropriations Committee restored to the foundation's appropriation in order to help the social sciences. The administration's measure was defeated by a 112-vote margin of a combination of Democrats and Republicans. This issue was considered by Congress and the scientific community to be largely a social and behavioral science issue. Here, members of Congress as well as the scientific community came to the defense of the social sciences.

But a final difference—and I think the most critical difference—is that in the 1981 budget fights there was a very strong, militant, and even angry social science community. It was a community that was willing to undertake active budget lobbying and other kinds of political activities to defend its professional interests, a community that made clear that there was a difference between the professional interests of social science researchers and the political uses to which their research might be put. This budget-lobbying effort was based on providing information about research and educating the public about research. So at the same time that we were trying to get political support, we were educating people about the importance of social science.

We have not been successful in all areas. The funding situation at the National Institute of Mental Health, which is still not good, is due largely to the fact that key individuals in OMB perceive social science research to be social advocacy research. They have a hard time separating the two, and as a result less social science research is being funded today.

It is ironic that the Reagan Administration is responsible for more unanimity of action and purpose in the social science community than we have ever experienced before. We find right now that researchers from many different disciplines are working together; we find that scholars are beginning to discuss priorities for social science research, asking what kinds of things are worth perserving, what kinds of activities we should seek federal support for, and what kinds of support would benefit the community the most. This is evidence of a new unity and a new maturity in the social science community. Over the past year, social scientists engaged in political discussions maintained a steady emphasis on the substance of social science research. I think this emphasis is very critical. Budget cuts were opposed because of their implications for specific kinds of research. Restorations of research budgets were requested because of the importance of the research that could be funded with the additional support. This emphasis on substance and the emphasis on the contributions that social science research makes to the nation have led to a greater understanding in Washington, most certainly in Congress, of social science research and the role that it plays in government and the economy. Clearly, the social sciences are not immune from further political attack, but as a result of the debates and the discussions resulting from budget cuts in the first year of the Reagan Administration, it is far more likely that future budget incursions against social science research will be cast in terms of the substance and usefulness of the research rather than in terms of political and ideological considerations, as they were in 1981.

Under these conditions, social science research will stand or fall on its own merits. We can ask no more; we should expect no less.

An overview of survey research activity and the context of health surveys in Australia

Terence W. Beed, Director, Sample Survey Centre, University of Sydney

Robert J. Stimson, Director, Centre for Applied Social & Survey Research, The Flinders University of South Australia

Introduction

An insight into the current position of health survey research in Australia can be gained by placing it in the context of survey activity generally, noting particularly the status and operational constraints of academic, government, and commercial survey practice. Certainly, survey research activity is flourishing in Australia, but the research community is coming under increasingly strong pressure as a consequence of funding shortages. The prospect for upgrading the generally low level of investment in fundamental methodological research appears gloomy.

The academic sector plays only a minor role in fielding surveys and conducting research into survey methods. Survey research centres were first established in Australia in the mid 1970s in only a few universities and they operate on a very small scale. The University of Sydney has the only fully underwritten survey centre, the Sample Survey Centre (SSC) with a tenured staff of three. At The Flinders University of South Australia and at Griffith University, Queensland, small scale centres are run by collaboratives of academics who hold tenured and full-time teaching appointments in other departments. These are the Centre for Applied Social and Survey Research (CASSR) and the Institute for Applied Social Research (IASR), respectively. Their survival depends on securing research grants and contracts from bodies outside the university. A fourth centre at the Australian National University (ANU) has recently undergone a transition from survey research to survey data archiving activity and is known as the ANU Social Science Data Archives.

None of these organizations offers continuous national or regional surveys, their work being of a more ad hoc nature and responsive to client demand across a wide range of disciplines. In this setting, the approach towards methodological research is opportunistic, depending on what the client budget can afford and on the availability of personnel after other client-directed priorities have been met. Typically, each of these centres is committed to running several surveys at once and resources must of necessity be thinly spread.

The principal organization for high quality data collection continues to be the Australian Bureau of Statistics (ABS), which has developed an increasing commitment to nationwide social surveys in recent years. For the survey research community, however, access to de-identified unit record data collected by the Bureau has been precluded by a remarkably conservative Act of Parliament dating back to 1905. This has affected not only scientific data analysis, all of which must be carried out at "arms length" by Bureau officers, but in addition, few if any outsiders have ever been invited to participate in the Bureau's survey design or experimental work. The ABS does however consult with the statutory Australian Statistics Advisory Council, on which academics and others are represented. This body also assists in ordering work priorities.

The private sector has been the scene of explosive growth in survey activity, but most of the collections are confidential and not accessible to other researchers. In many cases, they have poorly documented designs. The public opinion polls and syndicated media surveys in Australia are somewhat of an exception, with a steadily increasing degree of data access, but their documentation still leaves much to be desired.

Survey research activity: the general setting

The academic sector. As in other countries, academic surveys depend on various external sources for funds to cover the costs of data collection, preparation, and analysis. In Australia there are several major avenues of funding: a national academic research granting authority, the Australian Research Grants Committee (ARGC), and a similarly constituted body, the National Health and Medical Research Council (NH and MRC). These bodies fund most of the exclusively academic surveys in Australia. Under austerity measures, the federal government recently dismantled a third source of academic survey support, the Educational Research and Development Committee, which had in the past funded surveys in the field of educational research.

The funds awarded by ARGC and NH and MRC for academic surveys have been modest by international standards. The record award made by ARGC for a survey program in the last ten years was A$70,000, and grants at even half this level are exceptional. It must be remembered, however, that these grants are made to cover direct costs only (interviewing, data preparation, etc.) and do not cover the salaries of principal investigators or the use of computers. As Australian universities are funded by federal and state allocations, government argues that it should not pay a researcher's salary or buy a computer twice over. This makes for difficulties, especially where an investigator might desire time release from teaching commitments.

It is fair to say that the ARGC and NH and MRC have

given only scant attention to the rigor of survey designs in the projects it has funded despite its frequent recourse to referees in the approval process. To their credit, they have recently advised applicants that they should be prepared to document their surveys and deposit data in reputable archives for general access by the research community. We are unaware of any ARGC grant ever being made to support methodological research for surveys alone, either in an experimental or an evaluative mode. Indeed, no category exists for survey research methodology in the ARGC application protocols: proposals would have to be associated with a specific discipline.

Research contracts written with federal, state, and local government departments, Royal Commissions, and statutory committees of Inquiry have proved to be an increasingly fruitful area of funding for academics, especially for high budget surveys. Perhaps the most promising situation for the academic is the contract which allows publication of results, public archiving of data, and opportunities for methodological innovation. This is not always feasible, especially where short deadlines are imposed.

With the general scarcity of funding sources for academic surveys, the goal of an ongoing, high quality national survey vehicle seems far off despite burgeoning demand. Recently, proposals were put to the federal government's Research Centres of Excellence Committee by the Sample Survey Centre to underwrite the development of such a facility. Of the nine centres funded earlier this year, only one had a remote connection with the social sciences.

A further handicap to progress in survey methodology research is the generally low profile of survey education in the Australian universities and colleges. Multidisciplinary service courses are indeed a rarity, and there are no national survey methods training programmes open to graduate students. In the past five years, Australia has sent about 20 scholars and graduate students to the ICPSR Summer Training Program at the Institute for Social Research at The University of Michigan, but there is a serious need for us to send more. The potential for local training programmes is considerable, and the centres at Sydney and Flinders have made strong progress in the area of jointly convened advanced training workshops, the first of which were offered to about 100 participants in Sydney and Melbourne last summer. They were open to government and commercial survey researchers as well as academics, and they found a ready clientele. A further major workshop was conducted on the subject of crime victim surveys in Sydney last March, and a large national workshop on sampling methods is in an advanced stage of planning for early 1983.

Another move which will stimulate interest in methodological research is the growing interest and commitment to survey data archiving in the universities. The Australian Consortium for Social and Political Research, Inc. (ACSPRI), was formed four years ago and is cooperative of 19 universities and colleges dedicated to acquiring local and overseas datasets. It is the Australian corporate member of ICPSR and strong ties have now been set up between the two organizations. The Social Science Data Archives at the ANU was set up in response to lobbying by ACSPRI, for which it is now the secretariat. It is performing a very useful "clearing house" function in monitoring academic surveys, an outgrowth of a program established by its predecessor, the ANU Survey Research Centre (Mugford, 1979). This activity is complemented by opinion poll indexing and monitoring at the SSC in Sydney (Beed et al., 1978) and roundups of surveys publicised in the Australian mass media, reported quarterly in the *Newsletter of the Sample Survey Centre*. A strong survey information system is beginning to emerge as a result of these initiatives. It has attracted the attention of considerable numbers of academic and nonacademic survey researchers, and now encompasses several thousand entries.

Despite the constraints imposed by the difficult funding situation and the low level of investment in support facilities, Australian academics have nevertheless proved themselves capable of high quality work. Several important large-scale survey datasets are now on deposit in the leading international academic data archives (Aitkin et al., 1967; Beed, 1981), and we can expect many more to be documented to these standards in the future.

The government sector. At federal government level, the Australian Bureau of Statistics conducts a regular monthly workforce survey which is the basis for estimating levels of unemployment. On a few occasions each year, special supplementary surveys are conducted on topics such as health, household expenditures, crime, and travel. These use national .05% three-stage cluster probability sample designs covering about 30,000 dwellings. The Bureau has its own vote of funds to underwrite these surveys. It also offers to design samples and questionnaires, and at times actually collects and analyzes data for various government departments (both state and federal) as one-off surveys on a wide range of topics. However, the resources of the ABS are severely strained and its activities in this regard have been greatly curtailed in recent years. Demand greatly exceeds the supply of ABS survey facilities.

Until recently, it was very rare for the ABS to collect attitudinal data. Furthermore, relatively little experimentation in methodological issues and basic survey research is undertaken by the Bureau. Very little formal documentation of experiments exists. There are no parallels to the U.S. Census Bureau contracts for methodological research.

Other federal government agencies are relatively active in conducting surveys usually related to specific industries, specialized topics, or specific populations.

They involve using national, state, or regional sampling frames. Such agencies include the Bureau of Agricultural Economics, the Bureau of Industry Economics, the Bureau of Transport Economics, and the Bureau of Labor Market Research. Various survey methods are used, including the personal interview, diaries, self-completion mail questionnaires, and telephone interviews. Most of these surveys are contracted out to the private sector, and it is very rare for methodological research exercises to be built into these undertakings.

Relatively few government departments at the state level conduct regular surveys or maintain a survey research facility. In consequence, much is commissioned to the private sector. In general these surveys are one-off. They use a variety of sampling procedures and survey data collection methods. Quality with respect to control for sampling error and other biases is quite variable. Currently, there is considerable activity on the part of state transport authorities in the running of travel surveys, surveys on topics such as equal employment opportunity in the Public Service, socio-legal surveys by Justice Departments and Law Reform Commissions, State Health Commission surveys on a variety of health topics, and tourism surveys. State social service and community welfare departments often run surveys dealing with community needs and disadvantaged or handicapped subgroups of the population.

Local government is the lowest level in the three-tier hierarchy of government in Australia. There are many of these bodies representing highly variable numbers of people. During the last decade they have assumed expanded functions, particularly in the human services area. There has been a proliferation of so-called community needs studies which invariably contain a survey component; in most cases this work is contracted out. The quality of survey design and especially the questionnaires is generally extremely poor, and error and bias are largely ignored.

In recent years there has been increased activity in government-sponsored special purpose surveys of a "highly political" nature. These usually have strong policy objectives and are typically associated with a Royal Commission, such as The South Australia Royal Commission into the Non-medical Use of Drugs, official studies and inquiries such as the Agent Orange (Viet Nam ex-servicemen) inquiry, and the National Inquiry into Education and Training. Some of these surveys are concerned with trivial topics, such as the search for a new national anthem for Australia. Again, it has been usual for surveys dealing with these topics to be commissioned to the commercial area or academia, but at times the ABS has assisted.

Commercial surveys. In 1981, the market research industry was estimated to be billing about $55 million in survey fees. From this viewpoint, and in terms of interviews completed, this is the most important area of survey research in Australia. Ironically, this is the area about which least is known on methodological matters. Most organizations do not publish anything more than the briefest of reports about sample designs and field methods, usually in the context of a client report which is destined to remain confidential for years to come. Public scrutiny is therefore fraught with problems of accessibility.

While the industry has bodies to represent the interests of its members (employers and employees), efforts to attend to fundamental research problems have been tokenistic only. A current major concern is with the problem of apparently declining response rates in personal interview surveys. There have been meetings of the Market Research Society and the Association of Market Research Organizations at which the problem has been discussed, but little has emerged by way of a serious plan to research the problem, outside of an essay-writing competition on the subject. There remains a lack of any systematic approach toward recording non-response, so that reliable industry-wide comparisons of field performance might be made. There seems to be much potential for industry, government, and academics to conduct joint research into this problem which confronts essentially every practitioner in the field. At this stage only tentative feelers have been put out for such an approach, and it may take some time to achieve it.

Unlike the academic and government sectors, commercial researchers in Australia are plunging into the area of telephone interviewing. This move raises serious methodological issues, not the least of which is a less than complete level of telephone ownership in the population, currently about 75%. Some limited fundamental research in this area (Jones, 1981) suggests an abundance of problems in adequately representing the population in any telephone sample currently being drawn in Australia. In addition, the telephone system is operated by a single governmental authority. Its attitude toward the use of the system for survey research is far from crystallised as it foresees considerable technical problems in the use of techniques such as random digit dialing. Its own research into the characteristics of owners and non-owners (Cutler and Sharp, 1981) again underlines the dangers inherent in trying to represent the population in telephone samples. Privacy authorities in the state governments are also concerned about widespread use of the telephone for survey research purposes. Yet against this background, some commercial survey researchers in Australia have set up shop, offering an all-telephone survey research facility. That they are apparently viable reflects on the level of discrimination exercised by some sectors of the community of research buyers. The survey centres at Sydney and Flinders have thus far been unsuccessful in floating a proposition to look into the serious methodological issues which will have to be solved before practitioners can

proceed with confidence in the area of telephone surveys.

A brighter report can be made regarding progress in the acquisition of satisfactory documentation and machine-readable datasets from the Australian opinion polls. Through the pressure that ACSPRI has been able to bring to bear on the pollsters, we have now accepted the first batch of three years of machine-readable poll data from one of these organizations, and a second organization is on the verge of releasing a seven-year series of data drawn from over 120,000 interviews. This will have a considerable impact on the academic research and is a long awaited development. Nevertheless, very little is known about the efficacy of the pollsters' methods. Again, there is a lack of systematic data on survey performance and design qualities. At least one pollster has steadfastly refused to allow any detailed scrutiny of methods, despite strong pressure from academics and indeed from the politicians who are often the target of releases of poll results in the media.

Brief overview of Australian health surveys

The 1970s saw a proliferation of research in the health and medical areas in Australia that incorporated survey methods. There has not emerged, however, a progression of data collection that has generated time series data on the nature of incidence of illness and disease and use of health services by any government agency, and there is nothing akin to the U.S. Department of Health national surveys of households.

It is not possible to adequately review the full range of survey activity in the health and medical area in Australia in this paper. What we have set out to do is to highlight the broad areas of activity at the government and academic research levels, placing emphasis on population (rather than patient) surveys.

Federal government level

The Australian Bureau of Statistics 1977–78 Australian health survey. This is the most comprehensive general population health survey conducted in Australia for the purpose of providing basic data on the health of the population and the use of and need for various related services and facilities. It was the first health interview survey of a national scale conducted in the country and ran over a 12-month period.

Objectives of the survey were to provide a wide range of data using the health-care model which proposed that people's use or nonuse of health services would be influenced by their perceived state of health, their social and economic background, and their attitudes to and the availability of health-care services. Data collected related to episodes of recent illness; days of reduced activity due to illness or injury; incidence of accidents; compilation of a general well-being index; incidence of chronic conditions; doctor consultations and use of other health delivery personnel; episodes in hospital in the last 12 months; use of medication; child vaccination status; health insurance status; and usual demographic data.

The survey was national, designed to provide variable estimates for 21 regions within the states plus territories. Private dwellings were the sampling units, and all persons in the sample household were interviewed (confined to the Australian domestic population). The total sample size was 15,000 dwellings, selected on the basis of a stratified multistage area sample using the usual method of selecting blocks within sample CDs within LGAs, with probability of selection proportional to the number of dwellings within each block (these containing 20 to 60 dwellings). Systematic random selection of dwellings occurred within blocks.

Questionnaires were administered by personal interview. There were separate household and personal interview schedules. Usual ABS Population Survey and/or Household Expenditure Survey trained interviewers were used. Each interviewer had a 40 to 50 dwelling workload for each two weeks in the field. All persons aged 15 years and over were interviewed, and for those aged 14 years or less, proxy interviews were held with the parent or guardian. Each personal interview took at most 25 minutes. Aids to response included prompt cards, maps, and calendars. Rigorous follow-up procedures were used.

Response rates for the survey were 86% of all eligible households (plus 8% additional partially completed questionnaires). A total number of 40,650 persons responded to the survey.

Data analysis and publication of results gave rigorous attention to details of sampling and nonresponse errors or interviewer errors, plus aids to interpretation of results. Results have been published in a series of releases, with special reports on accidents chronic conditions, Sabin and triple-antigen vaccination, episodes in hospitals, recent illness, doctor consultations, days of reduced activity due to illness or injury, consultations with health professionals, an information paper, and outline of concepts, and methodology procedures used.

The ABS has conducted periodically other health surveys on the Australian population using its national sampling frame covering specific aspects of health, such as sight, hearing, and dental care. These more limited aspect health surveys are usually conducted in September.

Other federal initiatives. The federal Department of Health has been responsible for funding a wide range of research into a multitude of health-related matters. During the period of the Whitlam Labor Government, 1972–1975, the Hospitals & Health Services Commission was established, and it initiated a comprehensive Community Health Program aimed at taking initiatives in the fields of community health centres development, in areas of health services deficiencies, and in a variety of other programs aimed at specific purpose populations and areas. Since 1976, under the Fraser Liberal-Na-

tional Country Party Government, these initiatives have virtually ceased.

Through the Department of Health during the 1970s many evaluation studies were commissioned to monitor these initiatives under the Community Health Program. In general they were area specific and project specific; often they contained a survey component based on either users of a specific facility or the total population of a catchment area of a specific facility. They were typified by great variability in rigour of sampling design and other methodological issues and rarely were longitudinal in nature. Both private consultants and academic researchers were used to conduct these studies.

It is fair to state that despite the explosion of activity in this area financed by the Department of Health, virtually no research was supported specifically on methodological issues, experimentation in survey design, or management in the health area, unlike the program sponsored by the U.S. Department of Health. Budgets for projects rarely exceeded $50,000.

State government level. Most state governments have established health commissions to administer hospitals and other health services which the states have constitutional responsibility to provide. They have also tended to develop rather broad-based research capabilities, and there has been an emerging interest in the use of survey methods.

For example, the New South Wales (NSW) Health Commission took a major initiative in 1975 when it commissioned the ABS to conduct a Health Care Survey in the Central Coast and Illawarra areas of NSW. This was conducted during the period October to December 1975. It was based on a 3,000-household probability sample of households in which all residents participated. The model underlying the survey focused on illness and use of health services; this assumed a relationship between them influenced by predisposing factors, enabling factors, and the nature of the health-care system as intervening between perceived morbidity and use of services. The survey sought to collect data that would develop measures to monitor health needs and well being of the population, to provide information to assist the planning of regional and local health services, and to investigate the correlates of illness and the determinants of health services use. As the survey was conducted by the ABS, excellent methods of survey design and management were used.

This example of the use of survey research at the state health-authority level was relatively rare. More typical were the proliferation of specific population, area, and illness or disease studies, using both private consultants and academic researchers, plus health commission research personnel. Many studies focused on evaluating community health needs of specific areas, evaluating community health program initiatives in new forms of service delivery, and investigation of specific health problems, such as psychiatric services or need for women's shelters. Survey designs were quite variable in quality and type. In general samples were small, often nonrandom, and generally concern over errors and biases was not an area of high priority.

Academic research

General situation. Academic research on health using survey research methods is not in an advanced state of development in Australia. Few social scientists with survey methodology expertise are employed in medical schools, and it is uncommon for departments of epidemiology and community medicine to have nonmedical people teaching or doing research. Within the social sciences themselves only a handful of economists, geographers, sociologists, and psychologists work in health-related areas, and few have conducted extensive work using survey methods.

At an academic level, there was a stimulus to work for both medical and social science researchers with the formation in the early 1970s of the Australia & New Zealand Society for Epidemiology & Research in Community Health (ANZSERCH). Because of research grant arrangements in Australia, individual project grants to conduct research on health that involve a survey component are relatively small; they would range between $3,000 and $50,000.

There are few organized research groups that conduct multidisciplinary research into health matters that typically involve research. Probably the greatest amount of activity in survey work was incorporated in research projects on specific patient groups, and usually nonrandom sampling designs were used. Consumer-based studies on health typically are of patients using a specific facility, where some attempt at random sampling can be made, or of populations in designated catchments of a specific facility, where area probability sampling is most appropriate but not always employed.

Apart from isolated instances of research groups such as the Centre for Applied Social & Survey Research at Flinders University, there is a lack of methodological research in health survey research.

Selected summary. At an academic research level, very few surveys that are general health surveys have been conducted at a large-scale level. An exception was Krupinski & Stoller's 1971 survey in Melbourne.

What follows is a selected summary of research, mainly of a spatial nature in that the concerns were with health-care behaviour of populations in specific areas. They have a consumer-oriented behavioural base in which survey methods have been used.

An approach to the study of accessibility factors. Payne et al. (1977) conducted a study on two health centres in the Brisbane area, Queensland, at Inala and at Ipswich. A further control sample was also taken. They used a behavioural-functional model of health-care use to look at accessibility factors:

$$U = f(H, D, S, A, P, R) + E$$

where the notation is

U = utilization of health services;
H = perceived health status of the individual, for example, reported symptoms;
D = demographic factors, for example, age, sex;
S = socioeconomic factors, for example, occupation, income, education;
A = accessibility factors;
P = price factors, for example, transport costs, out-of-pocket costs of health care, opportunity cost of time;
R = predisposing factors, for example, availability of regular source of care, attitudinal factors;
E = residual error term.

Emphasis was placed on accessibility (specifically measure of time-distance to usual source of health care), waiting time, use of public transport, problems encountered in traveling to usual source of health care, and perceived accessibility of care. The study referred to certain difficulties inherent in this type of approach, particularly the problem of respondent recall for periods exceeding six months and the problems of collecting detailed information on the cost of obtaining medical care, especially as the study bridged the introduction of the initial Medibank scheme of universal health insurance. The study provided detailed data on the health-care behaviour of people in two areas, which is invaluable for evaluating the operation of innovations in the delivery system. It also furnished empirical data on the distance-decay function with respect to GP services, showing that in an outer metropolitan setting, about 35% travelled less than one mile, 30% travelled 1 to 5 miles, only 5% travelled 6 to 10 miles, and 30% travelled more than 10 miles (these being from rural areas).

The type of approach to the analysis of accessibility to health services necessarily restricts the researcher to looking at the *usual* service used. Various scaling procedures are employed to gain measures on attitudinal questions about the perceived accessibility of health care, whereas the other variables are readily quantifiable. Payne et al. (1977) refer to the problems of multicolinearity in a multivariate model of utilization. The approach is innovative in the Australian context, but the collection of this type of data from household interviewers is an expensive method of conducting research, as are most behavioural approaches.

Suburban health-care behaviour. A major study of the health-care behaviour of households and individuals in outer surburban areas has been conducted in Adelaide and reported by Stimson and Cleland (1975) and Cleland et al., (1977b). Figure 1 illustrates the aspects of health-care behaviour that were covered in the 1974 survey of 650 households (2310 individuals). Three areas were studied, two of which were to receive community health centres and the third of which was a control area, in what was planned as a longitudinal quasi-experimental design study to evaluate the impact of an innovation of the delivery system.

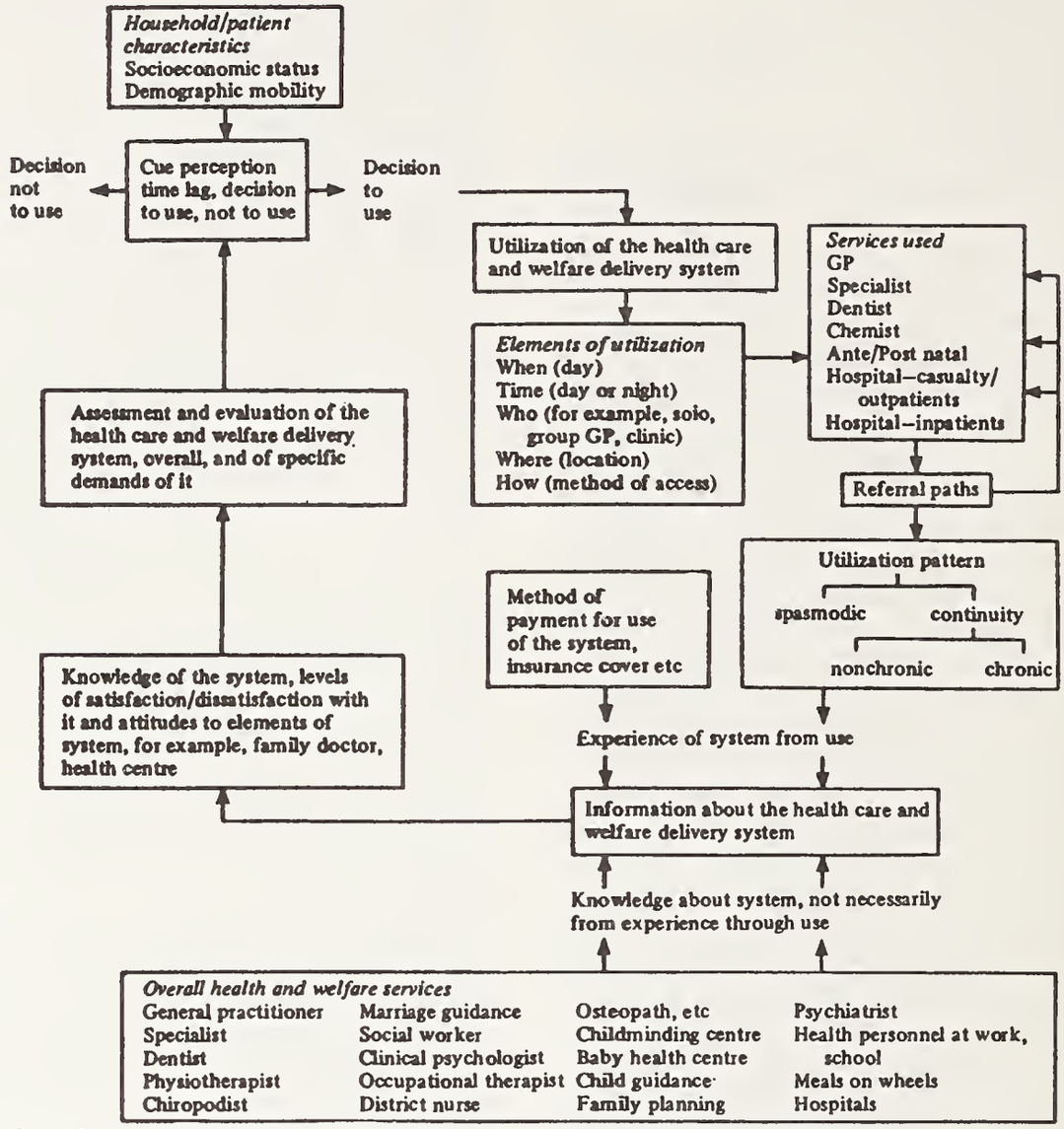
The results of this study showed that the majority of people had consulted a GP within the previous month, and 40% had been to a medical specialist during the past year. Paramedical services rarely had been used. Although there were high levels of satisfaction in general with health-care services, there were widespread complaints about queueing and temporal access problems, especially for GP services. Most people gained access to health-care services by private car, but the distance travelled to consult a GP varied widely between areas, depending on the spatial and temporal availability of local GP services, and it is difficult to make definitive statements about the distance-decay function of these trips For services such as dental treatment, trips were markedly longer. The general location of a range of health-care services used by households indicates a decreasing local service (that is, proximate) use as the hierarchical level of the service increases. About one-third of the trips to the GP were multipurpose, being conducted in association with activities such as shopping and collecting children from schools. Consultation with GPs also led to referral to higher-order services in over 20% of the cases, and 70% resulted in the prescription of medication, thus emphasizing the degree to which further travel and facility use is generated by GP visits.

Studies of this type are important in understanding the nature of consumer behaviour in the health-care services system, and from them greater insights may be developed into the nature of health-care service utility functions and how they vary between subgroups of the population.

Catchment studies and consumer behaviour. Studies of the catchment areas of the higher-order health-care service facilities are relatively easy to conduct by using hospital records that give the home addresses of patients. However, a behavioural research framework is required if one is to understand why certain use patterns develop for facilities such as hospital outpatients departments, which are largely used for primary-care purposes.

Such a study has been conducted for the Adelaide Childrens' Hospital casualty-outpatients department (Cleland et al., 1973). It was demonstrated that southern European migrants tended to have higher proportionate rates of use compared to Australian and other migrant groups, such as the British. There were no significant differences in the use rates between socioeconomic status groups. However, it was found that about one-third of the trips were made because of the actual or expected (perceived) unavailability of a local GP, and a similar proportion were referred there by a GP. About 30% saw the department as having the role of providing after-hours service and as offering a source of second opinion to that of a GP. The actual catchment area showed marked biases to the lower socioeconomic-status areas and the outer areas of the northwestern, northern and southern sectors of Adelaide, where rapid growth was occurring and where there was shortage of GPs or a lack of temporal availability of their services.

Figure 1
Summary of elements involved in a consumer-based approach to the evaluation of health care and welfare services and their use



Source: Cleland et al., 1977b, p. 80.

Another study of factors influencing the image of hospitals was conducted in the Coffs Harbour district of NSW by Walmsley (1978). He showed how there was a marked negative correlation between outpatient per capita usage and distance, and a slightly less negative relationship between inpatient usage and distance. In addition it was found that the distance a patient lives from the hospital has no significant effect on the duration of hospitalization per admission. A number of planning implications are evident from this study. The most important is that the distance-decay phenomena means that spatial injustices occur that prevent some people from using hospitals for primary-care purposes when local GPs are not available because of low population densities. Thus some are considerably disadvantaged in terms of time, money, and transport where hospitalization is required. Walmsley also questions the appropriateness of the hierarchy of base, district, and community hospitals for rural areas. On the basis of empirical evidence, either the hierarchy needs to be reviewed or the thresholds adjusted to more realistic proportions. A district hospital located in a town of 15,000 would need a catchment of 80 km on the NSW tableland and 40 km on the coast. The study also suggests that a mobile health-care facility could be appropriate in rural areas, especially if the service was provided by nurse practitioners.

The behavioural impact of a disaster cutting access to services. A unique study has been conducted by McGlashan (1978) assessing the effects on access to health care following a major disaster in Hobart, Tasmania. The city was divided in 1975 when a ship collided with the harbour bridge, causing it to collapse, thus isolating the eastern suburbs, which contained about one-third (41,000) of the population of the city, from the western shore where the vast majority of medical services were located. A sample of 1 in 20 households (N = 500) was surveyed on the eastern shore eight months after the disaster to assess the effects it had had on health-care behaviour and to evaluate the response of the govern-

ment and private practitioners to the servicing of the cut-off section of the population. A follow-up survey was conducted a year later.

McGlashan (1978, p. 1) refers to one of the major problems highlighted by the disaster, namely the lack of relevant information on which to base action after the bridge collapse, particularly data on use patterns of medical facilities by the eastern shore population and its needs for various types of paramedical services. The study showed that the immediate emergency response by local volunteers and by federal medical-services personnel was excellent. The state government also responded quickly with the provision of clinics at new locations. However, the GPs did not emerge from the events with credit (McGlashan, 1978, p. 25); they opposed government provision of services while at the same time claiming excessive workloads. At the time, most of the GP surgeries (serviced by sixteen doctors) located on the eastern shore were in the high socioeconomic status areas, and, as expected, basically they serviced local populations. However, after the disaster, 11% of eastern shore households found it necessary to change from a western shore GP to an eastern shore GP, even though 14% continued to attend a western shore GP. A total of 60% of households had at some stage since the disaster sought medical care from someone other than a GP. It is worth noting that the eastern shore had a young and fertile demographic structure unlike that of Hobart as a whole, and that there had been long-standing complaints of inadequate provision of GPs and hospital services on the eastern shore. Use rates were high, with 64% of households having had one or more of its numbers consult a GP within the preceding month (as of the August 1975 survey). Households in the more remote and lower-status suburbs often tended to prefer to use the free round-the-clock public hospital outpatient treatment at the Royal Hobart Hospital on the western shore. After the bridge collapse one would expect that households from these areas would tend to use the government clinic. McGlashan (1978, p. 11) relates how the Department of Health Services, acting on little quantitative information, quickly established four salaried GP positions to help alleviate the GP shortage on the eastern shore. The overall effects of the disaster were to increase the workload of eastern shore GPs by redirecting demand there from the western shore facilities, which led to increased time difficulties of access, a decline in home visits, and considerable conflict over methods of charging patients at a time when the original Medibank had just been introduced. It is worth noting that although the provision of a private hospital on the eastern shore has been much discussed, the hospital remains unbuilt.

This study is particularly important in demonstrating the way in which consumer behaviour is constrained by the locational decisions and functional organization of supply service personnel, both public and private. The bridge collapse in Hobart created major access problems for the eastern shore populations. The private sector was unable to respond adequately to meet *in situ* the demand for primary care. This necessitated the provision of a government-salaried medical-officer facility, the location of which, it would seem, was appropriate in serving a low-status government-housing area remote from access to other public and private medical facilites (particularly following the disaster), despite the relatively high level (24%) of dissatisfaction with the service it offered. Thus, behavioural responses of the consumer are to a large extent determined by the decisions of the actors on the supply side of the health-care system.

Health-care behaviour in isolated areas. Research has concentrated on studying health-care behaviour and service provision in the metropolitan or larger industrial cities of Australia. The problems of country residents in gaining access to health-care services have been sadly neglected. The magnitude of the problem was neatly summarized by the then Prime Minister, E.G. Whitlam, in his "1975 Rural Policy in Australia": "The provision of health services in rural areas is often complicated by the great distances involved, the sparseness of the population and the absence of towns of sufficient size to provide the infrastructure to support a modern hospital" (from Brownlea and Ward, 1976, p. 174). Although the Flying Doctor Service is famous worldwide for servicing those living in the outback areas, those who live in areas between the outback and the eastern and southern coastal crescent of closer settlement and major urban development face considerable access problems and have been neglected by health-care services planners.

The magnitude of the problem has been documented in two recent studies, one by Cleland et al. (1977a) on the access costs of households in isolated areas in using city-based, high-order health-care services in South Australia, and the other by Brownlea and Ward (1976) on health-care access problems in relatively isolated communities in northern Queensland and the Darling Downs.

Cleland et al. (1977a) focused their study on the Port Lincoln area of the Eyre Peninsula, a town of 10,200 that services a huge area containing an additional 9,000 people. Over a period of one year, they collected detailed cost data from households in which someone had been referred to Adelaide (the state capital) for specialist medical or hospital services by a GP in the one large group practice of the town. These accounted for about 35% of all such referrals to high-order health-care services, the remainder being to visiting specialists or the local hospital. Access to the city is difficult, involving an all-day trip by private car, an overnight trip by bus, or a fifty-minute flight over water. Invariably the person referred had to be accompanied by another member of the family. Often the consultation led to either a further consultation, a referral to other city-based health services, or both. This necessitated numerous trips. It was

shown that on the average the access costs (travel plus accommodation and living expenses) involved were $132 per trip at 1975 prices, or a total of $260 when medical costs were also included (note: 85% of medical costs could be recouped). It was estimated that for this type of isolated area, 3 trips per 100 population to city-based, high-order health-care services will be generated each year, representing an outlay for access costs of $301 per 100 population.

Brownlea and Ward (1976) took a somewhat different approach in their study of access problems to health services in the Darling Downs and Brigalow Development Scheme Areas in Queensland. They set out to document some of the overt features of health-services provision and to explore some covert disparities that manifest themselves in fatalistic attitudes to illness and disadvantage. Participatory planning of rural health services was seen as a not unrealistic expectation in their research (Brownlea and Ward, 1976, pp. 174-175). Their study highlights the difficulties of working in rural areas because there "exists a *grey* area of rural living as far as reception of services in general is concerned. This *grey* area is that geographic area where the distance to drive to services is such that given proper conditions people can get to health services (approximately in the range of one to three hours driving time) and as such are not included in services geared specifically to remote areas, like the Flying Doctor." (Brownlea and Ward, 1976, p. 175.)

In these *grey* areas, there is a lack of two-way radio links, and the grazier or rural worker relies on his car as the only communication link with health services. Air travel is difficult because of cost and the lack of landing strips, and telephone links are not always available during the day or at night. The predominant cause of people feeling isolated in these areas was sudden illness or accident which may be aggravated by the conditions of the roads and the distance to medical attention.

Brownlea and Ward used a variety of data collection methods. Data on child illness and health-services use was obtained from hospital records. Mail questionnaires for families were sent through local school principals. They held community seminars followed by a mail questionnaire. Structured interviews were held with health-care providers and members of the Family Medicine Programme. All data were analyzed in conjunction with a range of 1971 census data. The approach is innovative and highlights the research design problems confronted by researchers working in areas where existing records are deficient.

The study showed that distance itself was not the main problem; the roads and weather were. Other difficulties included the problem of gaining access to distant services, many of which are only open for the periods 10 a.m. to 12 p.m. and 2 p.m. to 4 p.m. each weekday, and time lags and cost in providing rural telephone links and the restricted hours of local exchange operators. There is a lack of training to handle emergency situations such as heart attack, internal injuries, head injuries, and snakebite. Mobile itinerant services, especially prenatal clinics, are needed. Spatial techniques, such as the travelling salesman solution, could be useful in planning these services. There are possibilities for providing government-employed nurse practitioners and other paramedics, some of whom would be grazier's wives, as substitutes for doctors who are notoriously loath to settle in remote areas. Another problem is that often following a long distance of travel to the nearest doctor, referral is then made to higher-order services in the large regional towns or cities, which involves even more costly and distant travel.

Survey research on health topics in Britain

Jean Morton–Williams, Social and Community Planning Research, London

Survey research as a specialist technology has suffered from low prestige in the health field in Britain. Even today there is a considerable amount of survey work undertaken by universities and medical schools that is conducted by researchers with little or no expertise in questionnaire design, sampling, or survey analysis. This arises, I believe, from a preoccupation with the esoteric nature of many health topics and a conviction that "anyone with a bit of common sense and a textbook can do a survey"—or, at least, the belief that anyone with a degree in sociology can do a survey.

The Department of Health and Social Security (DHSS), by far the largest sponsor of health research in Britain, has an explicit policy of putting most of its social research projects through universities on the grounds that academic personnel will be concerned with relevant problems even when not undertaking DHSS research, whereas a specialist survey organization would be dealing with quite other topics. Some branches of DHSS have only recently begun to take on interest in ensuring that the survey aspect of a project is competently carried out by professionals and all too often they think still that "survey work" starts with fieldwork and not with questionnaire design. The result of this lack of interest in survey technology is that there has been very little methodological research on the enormous problems of collecting accurate information on many important and heavily researched topics.

Nonetheless, health-related research forms quite a large proportion of the work of the two major social-research survey organizations, the Social Survey Division of the Office of Population Censuses and Surveys (OPCS) and Social and Community Planning Research (SCPR), an independent research institute. In addition, there is the small Institute of Social Studies in Medical Care, directed by Anne Cartwright, which specializes entirely in survey research on health and allied topics (e.g., contraception) and is supported by a rolling grant from DHSS, which also funds most of its research.

There is also, of course, a lot of health-related survey research undertaken by market research companies connected with the marketing and advertising of pharmaceutical products to both the public and the medical profession. The major market research companies also are sometimes invited to tender for surveys on health topics sponsored by government departments and other public bodies.

A feature of British survey research that is markedly different from the U.S. situation is that no British university has an ongoing survey facility of any size. Although some university researchers, particularly medical school researchers, still hire their own interviewers and set up an organization to process a survey, most fieldwork of any scale is now carried out by the professional survey organizations. But these organizations are frequently relegated to the role of service agency with a minimum of involvement in the survey design and none in the interpretation of the findings.

The main funders of survey research on health topics are: the DHSS, the Medical Research Council, the Social Science Research Council, the Health Education Council, and private foundations. I'll describe these briefly but first it is necessary to fit the Social Survey Division into the picture. The Social Survey Division of OPCS is at present provided with a budget by Central Government. Government departments and allied bodies (such as the Health Education Council) negotiate for a share of this budget to cover some of their research needs (this research being, of course, conducted by Social Survey Division), so they draw funds from Central Government via SSD. Social Survey Division is asked also sometimes to fund (or channel the funds for a department) and supervise survey work subcontracted to other survey organizations, thus performing the role of upholding standards. This role of Social Survey Division and its budgeting system are currently under review by the present government which tends to favour the private sector and is threatening to put the Division in the position of having to tender for research in competition with other organizations.

Other funders of health research are the medical schools, some of which have a certain amount of financial independence through donations and bequests, etc., and private industry ranging from the drug and contraceptive manufacturers to the cigarette companies looking for ways of diversifying their products. There are also one or two other minor government departments whose interests impinge on the health field, such as the Department of Prices and Consumer Protection which is collecting data on accidents in the home caused by products people buy, and the Health and Safety Executive, which is concerned with problems of environmental pollution from industry and with industry accidents and illnesses. We have recently carried out a survey for them on attitudes to hazards and risks related to these problems.

DHSS

The Department of Health and Social Security, as its name implies, covers a very wide area including community and hospital health care, social services, and the

provision of welfare benefits. In 1972 Lord Rothschild prepared a report for the government on the organization of government-funded research in which he recommended that it be tied more closely to policy objectives. This led to the establishment of internal systems that became particularly formalized in the large government departments that deal with very large research funds (notably DHSS and the Department of the Environment). Social research in general, and survey research in particular, is only a very small proportion of their research.

Each of the main subdivisions of DHSS has its policy staff, serviced by a research manager, and also a Research Liaison Group consisting of representatives of policy staff directly involved, representatives of other subdivisions whose interests overlap, the research manager, and sometimes also senior academics of standing in that particular area. The Research Liaison Group's function is to review all the research needs of the division (not just the social science needs) and to decide priorities and policies. Stemming from the Rothschild recommendation and the function of the Research Liaison Groups, there is a growing tendency for research to be fairly tightly specified within the Department and to be put out to competitive tender. Academic researchers complain that the 'customer-contractor' principle and the emphasis on relevance to policy has had the effect of largely confining social research in the health field to the *current* interests of *specific* Research Liaison Groups. Research that looks at longer-term issues, that covers the interests of more than one Research Liaison Group, or that straddles areas of more than one government department, has difficulty in obtaining funding.

The objectives and scope of most DHSS survey research is thus defined within the Department. Sometimes it is only broadly defined and put out to one or more universities for tender; sometimes it is defined more precisely and carried out by the Social Survey Division of OPCS or put out to tender with a number of survey agencies. However, there is limited opportunity for DHSS-sponsored research to be initiated outside the Department. Some senior academic researchers who act as advisors are in a position to present ideas for research to the Department, and the DHSS also runs a Small Grants Scheme (up to £40,000) through which proposals for research can be presented.

Medical Research Council

This is a body set up and funded by Central Government in the same way as the University Grants Council, the council consisting of eminent academic researchers in the medical field. The vast majority of the research it funds is biomedical, and social science research is awarded only a very small proportion of its resources. But, perhaps because most medical research requires the setting up of long-term research units, it has sponsored some of the most interesting social research in the health field. Its main social research units are: the National Survey of Health and Development (Bristol University), which has followed a cohort since birth for 36 years; the Institute of Medical Sociology (Aberdeen University), which has conducted a program of research on abortion, fertility, and family life; and the MRC Unit at the Institute of Psychiatry (London University), which has carried out projects on the effects of environmental stress on mental health. The MRC also provides funds for ad hoc projects but usually only to researchers in accredited medical research establishments.

Social Science Research Council (SSRC)

The SSRC is financed by Central Government and is the main body to which academic researchers can apply for grants to carry out research that they have initiated. Because of the large amount of research funded by the DHSS and the existence of the Medical Research Council, the SSRC is reluctant to give grants for research in the health field unless the research clearly deals with long-term theoretical or fundamental issues (rather than those of immediate interest to policymakers). It will fund "program" research (five years or more) and also smaller, more specific projects. Because of its alignment with academic disciplines, the SSRC to some extent presents a counterbalancing influence to the DHSS, as it enables research to be funded that cuts across the lines set by the Research Liaison Groups.

A research board also has been set up within SSRC to consider whether there are particular research issues of fundamental importance that are being neglected. The board is in a position to define these issues and to invite one or more university departments to put up proposals. The initiative for this kind of research has sometimes come from government departments (including DHSS) who have collaborated with SSRC in developing the brief and sometimes also in funding the research.

The Health Education Council

This is a 'Quango' (quasi-national government organization) financed by Central Government and working closely in collaboration with the DHSS. Its main function is, as its name implies, to improve the health of the public through publicity, education programs in the schools, doctors' surgeries, etc. It has some funds for research, used mainly in developing publicity content or in evaluating the effects of their programs. Social Survey Division has also conducted major surveys for the Council (in conjunction with DHSS) on such topics as smoking, drinking, dental care, and attitudes toward the deaf.

Private foundations

Although they provide only a small proportion of the total spent on social research on health topics, private foundations have played an important part in financing

some of the most imaginative and cross-disciplinary research. Most of them, however, will not fund salaries and overheads and are thus accessible only to academics in tenured posts.

Conclusion

Although the amount of money spent on survey research on health topics has grown enormously in the last 10 to 15 years, many of those engaged in carrying out the research do not feel that the situation is satisfactory. Academics feel that the government, although it has become the main source of funds, takes little responsibility for the welfare of academic researchers on whom it draws at will. Independent survey organizations feel that they are treated as fieldwork agencies and that scant respect is paid to their expertise in survey and questionnaire design. The situation is exacerbated by the current pressure to cut government expenditures.

Health survey research: Some experiences in Latin America

Carmen Noemi Velez, School of Public Health, Columbia University

Jack Elinson, School of Public Health, Columbia University

Introduction

We report on our experience with health survey research in Latin America, specifically on research projects in Puerto Rico and in the Dominican Republic.

During the years 1973 to 1975, Dr. Velez was the project director and Dr. Elinson was consultant on an Island-wide study of drug use among high-school students in Puerto Rico (Robles, 1974). This project was in itself modeled on a nation-wide study of teenage drug use (TADS) in the continental United States done by Columbia University researchers between 1971 and 1974 (Elinson et al., 1977), of which Dr. Elinson was the principal investigator. During the fall of 1974, the Columbia team decided to conduct a special substudy of the TADS project in two New York City high schools having a large number of Puerto Ricans enrolled, for the purpose of comparing it with data being collected simultaneously in various high schools on the Island. At the time, Dr. Velez was invited to observe the data-collection procedures in the two New York schools. This gave her the opportunity to observe the differences and similarities in surveying the same population (i.e., Puerto Rican high-school students) in two different settings (New York and Puerto Rico).

The different ways in which the socio-cultural and socio-political environments determined the shape the surveys took was evident at almost every step of the research.

The Puerto Rican study

Sample selection. The first difference between the two surveys, emerging from the differences in the socio-political environment of both societies, was observed in the sample selection procedure.

In Puerto Rico, it was decided to study a representative sample of all public and private schools on the Island. This decision was facilitated by the fact that the public-education system on the Island has a centralized organization directed by an education department. To get access and entry to the schools selected at random, permission was needed from the main office; with this permission the entry to the different schools was assured.

In the mainland-U.S. TADS study, the schools were not selected at random; rather the investigators decided to concentrate the research efforts on the two coasts, assuming that the prevalence of drug use was higher there. For purposes of comparison, other regions (Midwest & Southeast) were also selected in addition to various other schools serving socioeconomically or ethnically different communities. However, school accessibility and willingness to participate in the study turned out to be important factors in the selection of schools. Unlike Puerto Rico, the authority within the continental U.S. school system tends to be located at local community level. For example, two of the largest West-Coast city schools systems refused to participate in the study after being selected to participate, and substitutes had to be found. Moreover, in some other cities selected for the study, the researchers were not allowed to select the schools themselves but had to use schools selected for them by school officials. The final selection of schools included in the study was therefore largely based on the schools' cooperation and willingness to participate, in addition to other criteria.

Questionnaire construction. Another difference in the surveys conducted in each country was in the questionnaire construction phase. In both studies in the continental U.S. (i.e., TADS and the substudy in New York City), the researchers decided to leave out of the questionnaire certain issues that, although theoretically important, could have raised concern among members of the P.T.A. Specifically, questions about the students perception of the relationship between their parents (i.e., frequencies of fights and arguments, threats of separation and divorce, and even parental smoking behavior) were not asked out of anticipated objections by the parents.

In Puerto Rico, on the other hand, it was decided without hesitation to include various questions on such issues, since the researchers were convinced of the theoretical importance of the students' perceptions of the relationship between their parents and since they anticipated no special concern or opposition from the parents. It is interesting to observe such differences in reactions on the part of the researchers in relation to the same set of questions. Evidently, in the social context of Puerto Rico, the P.T.A. is not perceived to be as concerned or as powerful as it is in the United States. As mentioned, the schools in New York City and in all of the U.S. are much more community controlled than they are in Puerto Rico. In addition, these differences might also reflect differences in the researchers' ideas about what are "sensitive" or "private" areas of people's lives. It should be mentioned however that permission for the students' participation in the study was requested from parents in both Puerto Rico and the U.S.

Methods of data collection. Differences were also observed in the actual procedures for administration of the questionnaires in the schools in both settings. In Puerto Rico, with the permission of the Department of Education, it was possible to arrange to administer the questionnaires to the whole school during the same time period. In doing this, problems of bias because of previous knowledge about the questionnaires were avoided, and the number of nonrespondents due to cutting classes was kept at a minimum. This procedure also facilitated the arranging of additional time in a quiet setting for slow readers to finish the questionnaire. In Puerto Rico the item schedule for the school day on which the data collection was done was rearranged to accommodate the needs of the survey. In the U.S. study, it was the survey that had to accommodate the school schedule. It was not possible to arrange to survey the whole school during the same time period. The questionnaires were usually to be administered in the English and Social Sciences clases which met throughout the day in the schools. This meant that the interviewers had to stay the whole day in the school, that special arrangements for slow readers were not possible, and that the students who had completed the questionnaires had the opportunity to discuss the questionnaire and their answers with those in later classes.

The impact of these differences on the arrangements for data collection were especially noticeable in the completion rates of the questionnaires (Velez, 1981) of the Puerto Rican students surveyed in Puerto Rico and those surveyed in New York. While the missing values for questions reached almost 30% in some items toward the end of the questionnaire in the New York City study, it was never more than 3% for the students surveyed in Puerto Rico. In the New York City sample, the percentage of missing values increased consistently with the position of the question in the questionnaire, suggesting that many students did not have enough time to answer all of the questions.

The interviewers in New York City were also found to have a much harder time than the interviewers in Puerto Rico. Since the population of high-school students in Puerto Rico is homogeneous, most of the student's questions emerging during the data-collection procedure were anticipated by the researchers, and the interviewers were prepared and trained to handle them in a more standard way during the training session. The homogeneity of the school population in Puerto Rico made the respondents very predictable. The situation in the schools surveyed in New York City was very different; the school population was very heterogeneous, including various and diverse ethnic groups such as Dominicans, other Latin-Americans, British West Indians, Haitians, Black Americans, White Americans, and others. This heterogeneity provoked a wider range of questions to the interviewers which were not anticipated in the training session. In addition, not only did many of the non-Puerto Rican Hispanics complain that they did not understand certain words in the questionnaire, but, in terms of the wording of the questionnaire, there also were differences between the various ethnic groups, ranging from the street names of some of the drugs to the "proper" way to inquire about the sex of the person (femenino/masculino vs. varon/hembra vs. hombre/mujer).

The Dominican Republic study. Another perspective on how the socio-cultural setting shapes the health survey methods in different countries comes from our experience as consultants to a project on socioeconomic aspects of malaria in the Dominican Republic. This project was funded by the World Health Organization (WHO) through the Special Programme for Research and Training in Tropical Diseases. The proposal was submitted to WHO by the National Services for the Eradication and Control of Malaria in the Dominican Republic (SNEM). Although the fight against malaria was initiated in the Dominican Republic in 1941 as part of the Public Health Department, the SNEM was created in 1964. The SNEM has a total of 346 workers, more than 80% of whom are assigned to field activities and parasitoscopic tests.

As far as we know, this project is the first of its kind to be done by the SNEM, since its main activities so far have been the provision of services for the eradication and control of malaria in the country. The development of the research proposal was facilitated by the Regional Office of the Pan-American Health Organization (PAHO) in the Dominican Republic, which helped put together an ad hoc team of researchers to develop the proposal. Since there were no such research teams within the country, various consultants were brought in by PAHO. At various times the team consisted of nationals and non-nationals, malariologists, sociologists and agronomists working together.

The overall objective of the project is to study prospectively the relationship between the modes of production, migration patterns, macro and micro socio-ecological conditions, the activities of the SNEM Program, and the incidence of malaria in the Dominican Republic. The way the SNEM program is organized throughout the country greatly facilitates the sampling and data collection process of the study. For operative purposes the SNEM has divided the country into North and South zones; each zone is in turn divided into 171 operative areas. The study will exclude the two most urban cities of the country (Santo Domingo, the capital, and Santiago), since they have no problem of malaria transmission. The remaining 157 operative areas will be stratified according to the rates of malaria incidence into high, medium, and low areas. Two operative areas will be selected from each malaria stratum. In each of the operative areas a random 10% sample of the households will be surveyed for an estimated total of 2,500 households. These households will be interviewed with a detailed

schedule at time 1 and then will be followed up monthly with shorter schedules.

The sampling process within each operative area will be easy and efficient because all areas in the country are mapped out by the local SNEM programs and are kept up to date continuously by SNEM field workers. The main responsibility of the field workers is to visit each household regularly to perform the usual SNEM program activities, such as searching for new malaria cases, spraying the houses with insecticides, taking blood samples, and administering treatment. As part of their regular activities, they update the area maps, keeping well informed about new construction and other changes in the housing composition of the area. As a matter of fact, the SNEM organization is so efficient that the Census Bureau "hires" the SNEM personnel and uses the SNEM area maps for their population and agricultural and animal husbandry censuses. This organization of the SNEM throughout the country will make it feasible to reach remote areas in the country to which access would otherwise be very difficult or which would be missed altogether.

The data-collection process will also be facilitated by the SNEM organization. The SNEM field workers will be used as interviewers in this study. They are males between 25 and 35 years old, with 8 to 12 years of formal education, and have worked as field workers in the area for at least 10 years. They are well known and accepted in their communities, so problems of entry to the households can be ruled out. Since these field workers regulary visit the households for their search and malaria control activities, the initial interview as well as the follow-ups will be easily integrated into their routine work.

The organization of the health care system in the rural areas of the Dominican Republic includes not only the network of SNEM personnel but the "promotores de salud" as well. These are health promoters whose work is similar to the work of "barefoot doctors." This network of people have contact with most or all of the rural population on the Island on a regular basis, so it is easier to assure a better coverage of the rural population in health-related surveys than of the metropolitan areas, especially the marginal areas in the metropolitan sectors of the country.

Population census in the Dominican Republic. Finally, we would like to share with you some of the features of the last population census done in the Dominican Republic. First, the census was preceded by an apparently effective mass campaign, in which among other things the population was asked to stay at home during the period the census was to take place. According to the personnel participating in the census, the compliance with the request was very high and the streets in the different towns were deserted while the census was taking place. Second, all of the public employees were asked to participate in the census, which required attending the training session. In most cases, their tasks in the census enterprise paralleled their positions as public employees, that is, chiefs of divisions and departments would be in charge of supervising the census taking in large areas, while clerical personnel would do the actual census interviewing. These arrangements, according to various sources, turned out to be very successful.

Another interesting characteristic of the Dominican census was the way ethnicity (Dominican or Haitian) was ascertained in the census. The interviewers were requested not to ask the question but to establish ethnicity (basically, if Haitian) based on the physical appearance of persons and their pronunciation of the Spanish language. When we inquired about this strange nonverbal way of ascertaining ethnicity, we were told by some of the census personnel that the personal offense and insult that the question would imply to a Dominican was so great that they preferred to obtain the information through observation rather than lose all rapport with and cooperation from the respondents.

SESSION 1:
Measures and correlates of response errors

Chair: Ronald Andersen, Center for Health Administration Studies, University of Chicago

Recorder: Larry Corder, Health Care Financing Administration

The construct validity and error components of survey measures: Estimates from a structural modeling approach*

Frank M. Andrews, Institute for Social Research, University of Michigan

Introduction

There is growing recognition that measurement errors in any kind of data—including survey research data—can have profound effcts on statistical relationships. Some kinds of measurement errors make simple relationships appear stronger than they really are; others make simple relationships appear too weak. The effects of measurement errors on multivariate relationships can be great and also complex. Under certain combinations of error, an observed relationship provides no information whatsoever about the true linkages among the concepts being assessed: The observed relationship can be "wrong" in both direction and magnitude. However, if one has information about the validity and error composition of the measures being analyzed, more informed judgments can be made about the underlying relationships that are of primary interest.

Insightful survey researchers have always been interested in the quality of their data, and new information about data quality has been a major contributor to the deveopment of survey technology. Much attention has been devoted to *sampling errors,* and there now exist good ways to estimate their magnitudes and much knowledge about how to reduce them (Cochran, 1977; Kish, 1965; Sudman, 1976a). One kind of *measurement error,* bias (a consistent tendency for a measure to be higher or lower than it "should be" for the particular respondents to a survey), has also received considerable attention. (Sudman and Bradburn, 1974, provide an extensive review.) However, while bias can produce serious distortions in percentages, means, and other measures of central tendency, and hence is a threat that must always be considered, a bias that is constant for all respondents does not affect linear relationships at either the bivariate or multivariate level. It is other kinds of measurement errors that intrude on relationships—*random* and *correlated measurement errors.* (Key terms are defined below.) Despite their impact on relationships and near-universal presence in survey data, until recently these kinds of errors have received relatively little attention from survey researchers. These are the kinds of measurement errors investigated in this study.

This study has four major goals, none of which has been pursued previously in a large-scale and systematic way: (1) test the feasibility of incorporating a particular kind of methodological supplement in regular on-going national and organizational surveys and of using structural model estimation techniques to generate estimates of measurement quality; (2) provide descriptive information about estimated construct validity, method effects, and residual error for a broad range of survey measures as implemented by the standard data collection procedures of a respected survey organization; (3) account for the reasons why some survey measures have higher (or lower) measurement quality than others; (4) provide a means for predicting the construct validity and error components of other survey measures not actually examined in this study. Achievement of these goals would, of course, lead to a more general outcome of considerable importance to survey researchers and other users of survey data: more knowledge about how to produce higher quality data.

As things worked out, the approach used here proved highly feasible; quality estimates were generated for more than a hundred survey measures (from five national surveys and one organizational survey—all conducted by the University of Michigan's Survey Research Center), and subsequent analysis showed that over two-thirds of the variance in the quality estimates could be accounted for by considering various characteristics of the survey design and of the respondents.

The next section of this report discusses the basic aspects of data quality that are central to this study—construct validity, correlated error, and random error—and describes the role of method effects in generating correlated error. Part 3 presents information about the 106 measures whose quality was estimated—topical content, response scale used, and the survey in which each measure was included. Part 4 describes the structural modeling methods that were used to estimate the quality components of each measure and some of the analyses that gave us confidence in these estimates. The major results of the investigation are presented in part 5—first, descriptive information about the quality of the mea-

---

* Gerald A. Cole and Mary Grace Moore made numerous and substantial contributions to the work reported here; they were valued and much appreciated members of the project team. I am grateful to David Bowers, Angus Campbell, Charles Cannell, Philip Converse, Richard Curtin, Daniel Denison, and Robert Groves for allowing us to include methodological supplements in some of their surveys. Robert Caplan kindly permitted Henry Law and our project group to reanalyze data relating to social desirability from a study by himself and others. Many of my colleagues at the Institute for Social Research have provided useful comments and advice. Skillful typing was done by Jo Wilsmann.

Earlier versions of this paper were presented at the 1980 Annual Meeting of the American Psychological Association, and the 1982 Annual Meeting of the American Association for Public Opinion Research.

This research was supported by grant #SOC78-07676 from the National Science Foundation.

sures and, second, extensive multivariate analyses of factors that relate to measurement quality. Part 6 discusses some of the implications and uses of the study and part 7 summarizes the major points.

Basic notions about validity and measurement error

The literature on validity and measurement error is rich and varied, but also imprecise and inconsistent. It will be helpful to discuss some of the key notions about measurement quality that are used in this study.[1]

The importance of assessing measurement quality. For each of the measures included in any particular anlaysis, one would like, ideally, to be able to apportion the total variance into three components: valid variance, correlated error variance, and random error variance. From this, one could know the extent to which the true bivariate relationships (i.e., the relationships among the concepts being investigated) were being attenuated—because of random measurement error—and/or inflated—because of correlated measurement error. In addition, one could sort out the complex effects that random and correlated measurement errors have on multivariate statistics such as regression coefficients, multiple and partial correlation coefficients, and path coefficients.

The important impact that measurement errors have on statistics of relationship has received some attention in recent years (e.g., in sociology by Bohrnstedt and Carter, 1971; in psychology by Linn and Werts, 1973; in political science by Asher, 1974; in statistics by Cochran, 1970), but it still goes unrecognized by many data analysts.

A pair of examples, taken from the data of this study, will illustrate how misleading even a simple bivariate relationship between observed measures can be when allowance is not made for the effects of measurement errors. In Survey #2, the observed product-moment correlation between items having to do with perceptions about changes in business conditions over the past year and in the coming year averaged .41.[2] After allowing for measurement error, however, the true relationship between respondents' perceptions was estimated to be .70. Thus in terms of overlapped variance, the observed relationship was only about *one-third* of what it should have been (17% versus 49%). In this case, random errors led to a gross deflation of the relationship. While this is a common occurrence, it does not always occur, as illustrated in the second example. In Survey #5 a pair of ladder-scale measures showed, for respondents rated as relatively uninterested in the survey, a relationship of .44 between evaluations of own health and of work that had to be done around the house; but after allowing for measurement error, the true relationship for these respondents was estimated to be .30. Here, correlated error overwhelmed random error, and the observed percentage of overlapped variance was more than *double* what it should have been (19% versus 9%).

If measurement errors acted only to deflate relationships, one could at least infer something about the underlying true relationships, but—as just illustrated—this is not always the case. It is obvious that conventional corrections for attenuation due to unreliability cannot solve the problem. (This is because reliability calcuations do not distinguish between validity and correlated error.) Thus even at the simple bivariate level, observed relationships may provide little if any information about what investigators really want to know about—the underlying relationships—unless they are accompanied by estimates of measurement error. At the multivariate level, the information provided by observed relationships becomes even weaker, and measurement error estimates are even more important.

Definitions. What precisely is meant by the terms "validity," "correlated error," and "random error"?

By "validity" we refer explicitly to *construct validity*—the extent to which an observed measure reflects the underlying theoretical construct that the investigator has intended to measure (Cronbach and Meehl, 1955; American Psychological Association, 1974).[3] As noted by Zeller and Carmines, construct validity is different from several other types of validity—content, concurrent, predictive—and involves different notions from those of reliability, but "is the most appropriate and generally applicable type of validity used to assess measures in the social sciences" (1980, p. 83).

The difficulty in estimating construct validity arises from its explicit linkage to an unmeasured theoretical construct, and many discussions of construct validity stress the importance of a theoretical model in the construct validation process. Recently developed structural modeling techniques allow theoretical models to enter the analysis in much more explicit and powerful ways than could be achieved before and, as will be described in part 4, such models play a fundamental role in the present study.

"Random measurement error" refers to deviations (from the true or valid scores) on one measure that are statistically unrelated to deviations in any other measure being analyzed concurrently. Conversely, "correlated measurement error" refers to deviations from true scores on one measure that *do* relate to deviations in another measure being concurrently analyzed.[4]

Several points implied by the above definitions need attention. Our use of the term "correlated error" emphasizes an important general aspect of these errors. Other phrases that usually refer to the same phenomenon include "systematic error" and "halo effects." However, our use of "correlated error" is different from the way the term has sometimes been employed in investigations of "interviewer effects" (Bailar, 1976; Fellegi, 1974; Hansen, Hurwitz, and Bershad, 1961; Krotki, 1978; Krotki and MacLeod, 1979). In this latter usage interest is on the extent all respondents interviewed by a single

person systematically score too high or too low on a single measure. Interviewer effects are represented in our analysis, but their classification as random or correlated errors depends on how errors in one measure relate to errors in other measures, not on the circumstances of data collection.

Note, also, that whether an error gets classified as correlated or random explicitly depends on what other measures happen to be in a given analysis. This is reasonable, given that the purpose is to examine how measurement errors affect *relationships,* which involve concurrent consideration of two or more variables.

Method effects and correlated errors. A major reason that correlated errors appear in survey research is because analysts examine multiple measures derived by the same method. When the method by which a measure was obtained affects scores on that measure (a form of measurement error that is very likely to be present to some degree), and when measures reflecting the same method effects are analyzed together, these similar method effects produce correlated errors. Thus it is important to know how big is the effect that the measurement method has had on the scores. This is one of the types of errors that is estimated in this study.

A brief example will illustrate the nature of a method effect and how it can generate correlated error if two items using the same method are analyzed together.

Imagine a survey item that taps respondents' evaluations of their own health. The evaluation is obtained by asking respondents to pick one of several answer categories ranging from "very good" to "very bad." We can expect the answers will vary. This is partly because people differ in the way they perceive their own health, which generates valid variance. In addition, the answers may vary because people interpret the answer categories differently: Some respondents may interpret "very good" to mean something more positive than do other respondents. Hence, two people with the same underlying evaluation of their health might give different answers. This is measurement error attributable to the method or methods variance.

Now, if a second survey item using the same response scale was included in an analysis with this item on health, and if each respondent was consistent in the way he or she interpreted the meaning of these categories (as could be expected), the measurement errors attributable to the method would be the same in both items. Respondents who tended to be "too high" on the first item—because of the way they interpreted the answer categories—would also tend to be "too high" on the second item—because it used the same categories. Of course, this overlap in method effects generated covariation between the items, and this covariation gets added to any covariation that may exist between the concepts tapped by the items. The covariation attributable to common method effects—which is correlated error—strengthens the observed correlation if a positive relation exists among the concepts or weakens the observed correlation if a negative relation exists among them.

Although this example of a method effect producing correlated error focuses on attitude assessments, method effects can occur for any type of survey item if there can be variation in the interpretation of or reaction to the introduction, the question, and/or the response scale. Nearly all survey items are subject to some such variation in interpretation.

Of course, method effects need not be the only source of correlated errors. Correlated errors will appear whenever respondents differ; these differences affect the way respondents answer two or more items, and these differences are not linked to the concept(s) the item was intended to tap. However, two lines of thought lead us to the judgment that method effects are probably the major source of correlated error in survey data. First, in our own empirical investigations—described in part 4 of this paper—we have not found substantial and systematic correlations among the measurement errors that could be attributed to anything other than method effects. Second, other survey methodology studies have not produced compelling evidence of the actual presence of substantial correlated errors arising from other sources—at least up to now.[5] Nevertheless, sources of correlated error in survey data is a topic that merits further investigation.

Ideal versus obtained measurement quality components. It is important to understand the relationship between the measurement quality components that one would ideally like to have and the quality components actually obtained in this study. For each measure included in an analysis one would like firm figures for valid variance, random error variance, and correlated error variance (with information about the patterns of error correlation for various subsets of the measures). What this study actually obtained are *estimates* of valid variance, method variance, and residual variance for each meaure. As noted previously, the estimates of method variance can be used to infer at least minimum figures for correlated error when two or more measures using the same method are in a single analysis. And, the estimates of residual error indicate the maximum possible amount of random error. (Moreover, since no substantial and systematic correlations have been found among the residual errors, we believe this upper bound is itself a reasonable estimate of total random error.)

Sources of Data

Surveys and methodological supplements. The data used in this study come from six different surveys. Five of these surveys contacted probability samples of American adults who lived in households; the sixth was a survey of members of a large Canadian corporation. Table 1 lists these surveys, the population each represents, the number of respondents, the method of data collection, and the substantive content of the measures used in the present investigation. As may be seen there,

Table 1
Data sources

No.	Survey date and method	Population represented	Number of respondents	Number of measures	Concepts assessed in multimethod-multitrait design
1.	August 1978 personal interview	A	3767	19	Quality of life: assessments of housing, standard of living, self, family life, community, health, life-as-a-whole
2.	January 1979 telephone interview	A	884	24	Reports about past and anticipated changes in business conditions, personal finances, health, keeping up with the news
3.	March 1979 telephone interview	A	560	9	Behavioral reports on eating too much, drinking beer, watching television
4.	July 1979 telephone interview	A	1173	18	Reports about past and anticipated changes in business conditions, personal finances, health, keeping up with the news
5.	September 1979 telephone interview	A	946	12	Quality of life: assessments of financial security, housework, health, life-as-a-whole
6.	Fall 1979 group-administered questionnaire	B	376	24	Ratings of organization of work, firm's interest in workers' welfare, improvement of working conditions, group members' knowledge about jobs, quality of groups' response, group decision making, behavioral reports of eating too much, drinking beer, watching television
	SUMS		7704	106	

Key to populations represented: A = American adults living in households; B = Employees of Canadian business firm

the total number of respondents was 7,704 (the median number per survey was about 900; range: 376 to 3,767). Table 1 also indicates that four of these surveys used telephone interviews to collect the data, one used face-to-face interviews, and one used a group-administered questionnaire. All these surveys were conducted as part of the regular on-going research activity of the University of Michigan's Survey Research Center.

The basic strategy for the present investigation was to select several important concepts from the regular content of each survey, and then to add a few (6–20) additional items tapping these concepts in such a way that a multimethod-multitrait data design would be achieved (Campbell and Fiske, 1959). In other words, each of the selected concepts ("traits") was to be assessed by several different methods—i.e., by several distinctively different response scales. Careful attention was devoted to ensuring that the several items intended to tap a single concept all assessed exactly the *same* concept despite their using different response scales. The supplementary items were incorporated into the questionnaire in such a way that they constituted an integral part of the interview.[6]

Measures for which quality estimates were obtained. To provide a broad base for our methodological findings, the survey measures used in this study—the "primary measures"—are a large and intentionally heterogeneous set. In all, there are 106 primary measures that tap 26 different concepts. As may be seen in Table 1, these concepts include assessments of life quality (housing, standard of living, family life, health, etc.), attitudes about economic matters (changes in business conditions, personal finances), behavioral reports (drinking beer, over-eating, watching television "to get away from it all"), and ratings of the respondent's employing organization and work group (interest in workers' welfare, improvement of working conditions, quality of decisionmaking, etc.). Many of the topical content areas were addressed in two different surveys to reduce the chance that measurement conditions in any one survey would be confounded with characteristics of the concept itself.

In a further effort to broaden the base for our findings (and, at the same time, to make the measurement model estimable), a large number of different but commonly-used response scales were included. In each survey, at least three different response scales were used. Across all six surveys, there were a total of 14 different scales. Included were simple yes/no formats (a 2-point scale); 3-point better/same/worse formats; and 4-, 5-, 7-, 9-, 11-point and graphical scales of agreement, satisfaction, goodness, etc.; and actual frequency reports of certain kinds of behaviors. Table 2 provides the details.

Table 2
Measurement methods (response scale formats)

No.	Short name	Nature of scale	Survey[a]
1.	Yes/no	Two categories labeled "yes," "no"	3,4
2.	Better/worse-3	Three categories labeled "better," "same," "worse," plus—for some but not all measures—a "don't know" category	2,4
3.	Better/worse-unfold	A two-stage sequence in which respondent indicates a general response (e.g., "better," "in between," or "worse") and then answers a second question to refine the position (e.g., "a lot better" or "somewhat better") or to indicate he/she hadn't "thought much about it"	2,4
4.	Goodness	Five categories labeled "very good," "fairly good," "neither good nor bad," "not very good," "not at all good"	1
5.	Satisfaction	Seven categories ranging from "completely satisfied" through "neutral" to "completely dissatisfied" with unlabeled intermediate categories	1,5
6.	Delighted/terrible-7	Seven categories labeled "delighted," "pleased," "mostly satisfied," "mixed," "mostly dissatisfied," "unhappy," "terrible," plus an off-scale "no feelings at all" category	1
7.	Delighted/terrible-unfold	A two-stage sequence in which respondent first indicates a general response ("good," "bad," or "mixed") and then answers a second question to refine the position (e.g., "delighted," "pleased," "just mostly satisfied")	5
8.	Ladder	Description of a ladder with 10 steps ranging from "worst feelings" to "best feelings"; respondent indicates his/her position	5
9.	Graphical assessment	Line ranging from 100 (labeled "perfect") to 0 (labeled "terrible") with each decile marked	1
10.	Agree/disagree	Five categories labeled "agree strongly," "agree moderately," "in the middle," "disagree moderately," "disagree strongly"; alternatively: "agree a great deal," "agree somewhat," "mixed feelings—not sure," "disagree somewhat," "disagree a great deal"	2,6
11.	Extent	Five categories labeled "to a very little extent," "to a little extent," "to some extent," "to a great extent," "to a very great extent"	6
12.	Frequency-4	Four categories labeled "almost every day," "every few days," "once or twice," "not at all"	3
13.	Frequency-9	Nine categories labeled "never," "hardly ever," "some of the time," "somewhat less than half," "about half the time," "somewhat more than half," "most of the time," "nearly all of the time," "all of the time" plus an off-scale "don't know" category	6
14.	Frequency-days	Respondent reports actual number of days per month	3

[a]The numbers in this column refer to the surveys listed in Table 1.

To minimize confounding the effects of response scale and survey, or response scale and concept, many of the response scales were used in more than a single survey, and all were used with several (3–8) different concepts.

Example items and terminology. Let us indicate how a couple of actual survey items fit within the preceding discussion and clarify the way several key terms are used in this report.

Figure 1 presents a pair of items as they appeared in the interview schedule for the January 1979 telephone interview. As indicated in Table 1, this is Survey #2, which sought information from a representative sample of American adults living in households and obtained data from 884 respondents. The items illustrated in Table 1 ask about two concepts, past and anticipated changes in health, and use one response scale, a 3-point better/same/worse response scale (which is Scale #2 in Table 2).

Figure 1. Example Items

B1. Now here are a couple of questions about your health:

B1a. Would you say that your health is *better* or *worse* than it was a *year ago?*

1. BETTER NOW	5. WORSE NOW	(IF VOLUNTEERED) 3. ABOUT THE SAME

B1b. Do you think that *a year from now* your health will be *better, worse,* or just about the *same* as now?

1. WILL BE BETTER	5. WILL BE WORSE	3. WILL BE ABOUT THE SAME

In considering characteristics of items, it is useful to distinguish three distinct parts, and these can easily be seen in the example items in Figure 1. There may be an *introduction,* which is followed by one or more *questions,* and each question has a *response scale,* i.e., a set of answer categories. Together, the introduction (if any), the question, and the response scale constitute the *survey item.* The survey item, when answered by a set of respondents, provides a *survey measure,* i.e., data for analysis.

(The complete set of primary measures from one survey—#3—is presented in the Appendix.)

Other measures. In addition to the measures whose quality was to be assessed (the "primary measures"), this investigation makes use of a substantial number of other measures to account for variation in the quality assessments. These include reports by respondents about some of their own personal characteristics, ratings by respondents about the survey in which they participated, ratings by interviewers about the respondents, judgments by study staff about characteristics of the primary measures, and objective information about the respondents, the primary measures, and the design of the survey in which each primary measure was included. In all, 46 such other measures were examined. These are described in part 5 of this report.

Estimating the quality components of survey measures

General strategy. The measurement quality estimates were derived from a structural model of the measurement process. This model is based on a set of causal assumptions which are grounded in classic measurement theory and which involve the basic notions about data quality discussed in part 2 of this report.

In accord with classic measurement theory (and with what seems intuitively reasonable), a respondent's recorded answer to a survey item is assumed to reflect three types of influences: (1) the way that particular respondent feels about the concept the survey researcher intended the item to tap (e.g., the respondent's perception about changes in his or her health); (2) the way that respondent reacts to the method used for obtaining the data (particularly, in our case, the response scale); and (3) everything else that might affect a recorded answer (e.g., lapses of memory by the respondent, misunderstanding by the reviewer, etc.). These simple ideas can be represented by a causal model of the measurement process. Given appropriate data that include sufficient "cross-checks," estimates of measurement quality can be obtained from such models.

A schematic version of the measurement model used in this study appears in Figure 2. This hypothetical example, which is too small for actual use, shows two survey measures that tap Concept A, one obtained using Response Scale X and one using Response Scale Y, and two others that tap Concept B, one using Scale X and one using Scale Y. These four measures are indicated in Figure 2 by rectangles. The sources of variance ("latent" or "underlying" variables) that influence the measures are shown as ovals in the exhibit—Concepts A and B, Response Scales X and Y, and four sources of residual error. The one-way arrows in the exhibit indicate direct causal influences; two-headed arrows indicate relationships between concepts where no assumption about causality is made.

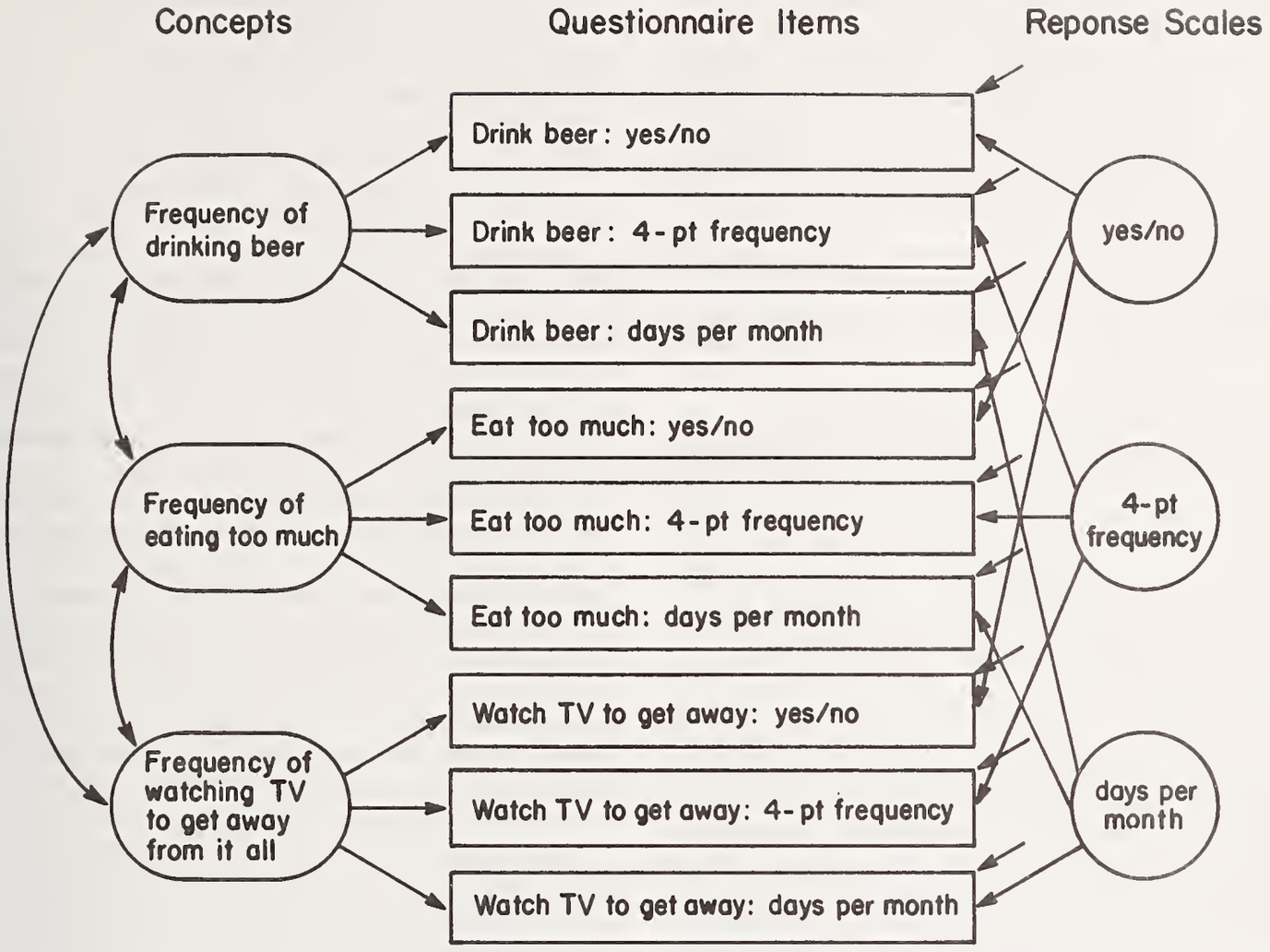
The model is a general one, and it may help to consider a specific example. This model might be applied where one had obtained information on perceived changes in financial well-being (Concept A) and health (Concept B) using a 3-point better/same/worse scale (Response Scale X) and a 5-point extent-of-improvement scale (Resonse Scale Y).

If one has confidence in such a model (a topic addressed below), some of the parameters can be interpreted as measurement quality assessments because they indicate the extent to which variation in a given measure reflects (1) differences among respondents with regard to the underlying concept, an estimate of construct validity (parameters a–d in Figure 2), (2) differences among respondents in the way they interpret and use the response scale, a major source of correlated error (parameters e–h), and (3) the extent the measure reflects other influences, primarily random measurement error (parameters j–m). Other parameters can be interpreted as estimates of the true relationships among the concepts (parameter i).

The process of estimating the structural model involves finding a unique set of strengths for the causal linkages (parameters a–h), imputed correlations among latent concepts (parameter i), and variances for the residual error sources (parameters j–m) that will produce a set of *predicted* relationships among the observed measures that come as close as possible to the *observed* relationships among these measures. In this study these parameter values were obtained by a maximum-likelihood estimation technique incorporated in the LISREL computer program (Joreskog, 1978; Joreskog and Sorbom, 1978).

The schematic model shown in Figure 2 is for illustration only and is too small for actual use. To obtain unique estimates for the parameters, more measures, concepts, and methods were included in the models actually used for this study. In addition, an equality constraint was imposed on each set of method effect parameters,[7] and each method effect was constrained to be independent of all other latent variables. These constraints ensured that the model's estimates of method effects reflected the assumptions that each method effect was a statistically unique phenomenon and that it had an equal impact on all measures based on that method. These constraints also helped to identify the other parameters in the model. (Thus by definition the method effect equalled the portion of variance in each measure that was associated with response scale and

Figure 2
Schematic form of structural model



independent of the question topics.)

Generating the measurement quality estimates that are analyzed in part 5 of this report required running 125 measurement models. Each of the six surveys involved a series of models of a given form, and within each series runs were made for different (and sometimes overlapping) groups of respondents. The nature of the model applied to each survey can be determined from the information available in Tables 1 and 2. For example, Table 1 shows that Survey #3, the March 1979 telephone interview, included nine measures whose quality was assessed and that these involved three concepts (having to do with eating, drinking, and watching TV); Table 2 shows that three response scales were used in this survey (yes/no, frequency-4, and frequency-days); hence, this is a simple 3 × 3 design (three concepts each assessed by three methods), which accounts for the nine measures.[8] The model actually used for Survey #3 is presented in the Appendix.

Relevant precedents. Although this study represents the first broad-scale attempt to use structural modeling technology to generate measurement quality estimates for data from regular on-going surveys, this investigation is grounded on prior work and exists within a larger research framework.

The basic idea of using a multimethod-multitrait data design for assessing validity and method effects was proposed by Campbell and Fiske (1959). How best to analyze such data was much debated prior to the development of powerful structural modeling techniques (e.g., Jackson, 1969; Conger, 1971; Levin, 1974; Golding and Seidman, 1974). However, as the potential of structural modeling for handling multimethod-multitrait data became recognized, many investigators advocated its use (e.g., Alwin, 1974; Boruch et al., 1970; Schmitt et al., 1977). Other more general discussions of using measurement models for social data include those by Wheaton et al. (1977) and Alwin and Jackson (1979). The general usefulness of structural models for illuminating the nature and quality of social data now seems well established, and measurement models receive extensive attention in a number of recent texts on

social measurement and analysis (e.g., Zeller and Carmines, 1980; Bohrnstedt and Borgatta, 1981; Sullivan and Feldman, 1979; Namboodiri et al., 1975).

Although theoretical and methodolgical discussions about the use of measurement models are no longer rare, only recently have investigations that actually use the new modeling technology to estimate data quality begun to appear. Our own work on the measurement quality of subjective social indicators was one of the first (Andrews and Withey, 1974, 1976; Andrews and Crandall, 1976). Other implementations include those by Mason et al. (1976), Kluegel et al. (1977), Robins and West (1977), Bielby and Hauser (1977), Bielby et al. (1977a, 1977b), Andrews (1979), Mare and Mason (1980), and Corcoran (1980).

Reasons for confidence in the estimates. Beyond the fact that other investigators are using and advocating the same general approach as has been used in this study, what empirical evidence is there that suggests one should treat the results from our modeling analyses as providing reasonable estimates of data quality? The issue is, obviously, an important one, and it has occupied a central role in the analyses of this investigation.

Appropriateness of the data for the measurement model. The measurement model assumes the data can be adequately explained within the confines of an additive linear system. Furthermore, given that product-moment correlations were input to the LISREL computer program, they are assumed to appropriately represent the actual relationships in the data. To see whether these assumptions were met, the data from each survey were scanned for instances of marked skews, nonlinearities, and nonadditivities. In most cases we concluded that the data in their original form were appropriate for further analysis; in a few instances a transformation was effected to reduce skew and/or curvilinearity before the correlations were computed.

Adequacy of the measurement model. Models that work well (1) are theoretically reasonable, (2) manage to closely account for the observed relationships in the data, and (3) do so with parameters that are themselves of reasonable magnitude. How well did the models used in this study rate on these criteria?

1. The theoretical relevance of the model has already been discussed. The model represents a direct implementation of classic measurement theory, and the primary measures (those represented in the multimethod-multitrait matrices) were designed specifically for use with this model.

2. Models of the form displayed in Figure 2 proved to fit the data well, i.e., they did a good job in accounting for the actual relationships among the measures. In over a hundred runs involving the application of this model to different sets of variables and/or different sets of respondents, this model consistently produced predicted relationships that tended to be close to the observed relationships. The mean absolute deviation between predicted and observed correlations was always less than .10, and in most runs it was less than .05.[9]

3. Although a very wide range of parameter values was obtained, in nearly all cases these values were reasonable in the light of theoretical expectation. Specifically, validities, methods effects, and residual errors should all be within 0 to 1. Although there was no constraint on the estimation of parameters to prevent the occurrence of unreasonable values, they very rarely occurred. Out of 2,115 validity estimates that were generated, none was less than 0 and only 19 (0.9%) exceeded 1.00 (the highest was 1.13). Of 2,115 method effect coefficients, none was less than 0 and none exceeded 1.00. Of 2,115 residual error coefficients, 32 (1.5%) were less than 0 and none was greater than 1. In short, out of more than 6,000 measurement quality estimates, 99.2% had reasonable values, and the few that fell outside the reasonable range were not very far outside.[10]

Thus, on the criteria of theoretical reasonableness, adequacy of fit, and reasonableness of estimates, the model seems good.

Exploring modifications to the model. Of course this is not to say that this model always produced the best possible fit. In many cases, minor modifications could be made in the model that would modestly improve the fit. However, despite considerable searching, we have been unable to find any way to change the model that would produce consistent and substantial improvements in fit. Two extensive explorations of this matter merit attention here.

1. Sorbom (1975) has proposed a procedure in which one successively frees certain previously fixed or constrained parameters so as to improve the fit. Using data from the January survey, we pursued this approach for 11 iterations and observed that while the fit did improve modestly, the measurement quality parameters which are of primary interest for this study—i.e., the estimates of validity, methods effects, and residual errors—changed very little. Furthermore, 10 out of the 11 modifications involved allowing covariation among selected pairs of the residual errors, but none of these covariations was substantial. This provides one important piece of evidence for our belief that the residual errors provide reasonably good estimates of random error.

2. The second exploration took advantage of the fact that all of the respondents to the July survey had also participated six months earlier in the January panel—i.e., this pair of surveys constitutes a two-wave panel. The data were collected in this way specifically to allow investigation of whether there was stable, consistent variance present in the residual components that was unique to a given measure.[11] Thus this provides another check on the meaning of the residual errors. It allows us to ask whether the residual variance really looks like random error, or whether there is some part of it which, while independent from that of any other measure assessed in any one survey, might be a reliable aspect of that particular item. Results of this panel analysis were clear: There

was little, if any, reliable unique variance for any of the 12 measures that had been repeated. (The estimates of the reliable unique variance components ranged from 0% to 11% with a mean of 2%). This is further important evidence that it seems appropriate to use the residual parameters in our measurement models as estimates of random error.

All of these considerations, then, led us to believe that the schematic model portrayed in Figure 2, suitably modified to reflect the particular concepts and measurement methods included in a particular survey, provides an appropriate and generally applicable means for generating estimates of construct validity, method effects, and random error.

Note on correlations versus covariances. Readers who are sophisticated with regard to structural modeling may wonder how the use of correlations as input to the model interacted with the equality constraint imposed on each set of method effect parameters (the *only* reason that it would make any difference whether input was correlations or covariances). A check showed that all measures obtained using any one method tended to have similar variances, and hence it made no difference whether correlations or covariances were used as input to the models.

Note on absence of some linkages involving response scale factors. As indicated in Figure 2, a response scale factor is defined to be a source of variance that is shared by measures using the same response scale but that is statistically independent from other response scale factors and from all of the substantive concepts. (Note the absence of linkages between each response scale factor and all other factors.) Constraining these linkages to zero helps to ensure that the response scale factors operate as they should if we wish to interpret them as method effects, but reflects a set of assumptions that might not actually hold in the data. The fact that the models estimated for this study fit as well as they did, however, suggests that it was not unreasonable to assume that these response scale factors were statistically independent from one another and from the substantive concepts being measured. This result is consistent with those from a previous study (Andrews and Crandall, 1976) where relationships among response scale factors were left unconstrained and empirical estimates showed them to be very weak.

Analysis of the measurement quality estimates

Analysis strategy. The preceding stages of this research resulted in the generation of a large number of measurement quality estimates. Specifically, there were 2,115 sets of estimates, where each set consists of a validity estimate, a method effect estimate, and a residual error estimate for a particular survey item as it was answered by a particular group of respondents on a particular occasion. The next step in the research was to analyze these quality estimates to determine under what conditions they tended to be higher and under what conditions lower.

This was done in three steps. First, simple descriptive statistics showing the level and variation of these measurement quality estimates were examined. Second, the effects of survey design characteristics were examined. A series of multivariate analyses explored how aspects of survey design—characteristics of response scales, item wordings, the topic investigated, and questionnaire design—related to each of the three measurement quality estimates. Third, after the effects of survey design characteristics had been removed, the effect of respondent characteristics—age, education, sex, and many others—were examined.

This strategy provides a direct way of addressing some of the key questions in survey research—how to achieve more valid, less errorful measures. The approach, however, is unusual. Few previous studies have had a sufficient number of measurement quality estimates, together with information about the survey and respondent characteristics associated with each estimate, to make a direct multivariate analysis of those estimates feasible.

It is important to note that this analysis strategy involves a shift in databases. When the measurement quality estimates were being generated, the "cases" were individual respondents—as is conventional for most survey analysis. Here, a different database is being analyzed. Now the "cases" are survey-items-as-administered-to-a-particular-set-of-respondents. For example, one of the "cases" in this dataset is a survey item about satisfaction with own health that was answered on a 7-point scale of satisfaction in August 1978 by a representative sample of Americans 71 to 90 years old.[12] This is just one among more than 2,000 such cases.

Data quality—level and variation. Table 3 presents basic descriptive information about the measurement quality levels, and the variation in those levels, observed in this research. Presented there are the univariate statistics for all 2,115 estimates of each of the 3 measurement quality components that will be "dependent variables" in this analysis.[13] These results are of considerable interest in their own right.

Table 3
Measurement quality estimates—average level and variation

	Validity	*Meth. eff.*	*Res. error*
Mean	.81	.16	.53
Median	.81	.16	.55
Standard deviation	.10	.11	.16
Number of estimates	2115	2115	2115

Note: The estimates of validity and method effects are based on lambda parameters from LISREL; the estimates of residual error are based on the *square roots* of the theta parameters from LISREL.

The first row presents the means for the estimates of validity, method effect, and residual error. Note that the average validity is .81, the average method effect is .16,

and the average residual error effect is .53. These figures, when squared, indicate that the "typical" survey measure examined in this research consisted of 66% valid variance, 3% method variance, and 28% residual variance. (These sum to 97%, which is reasonably close to the expected 100%. The discrepancy presumably reflects minor imperfections in the way the measurement model used to generate the quality estimates fit some of the data.) Although no claim is made that the set of survey items examined here is representative of all items used in current surveys, this set of items is broader and more heterogeneous than any other we know whose quality has been estimated, and hence these estimates for the quality of our "typical" item probably provide the best available information about these aspects of measurement quality for single-item survey measures tapping attitudes and behavior.

The second row in Table 3 presents the median quality estimates. Since these figures are very close to the means presented in the first row, one can infer that the quality estimates are approximately symmetrically distributed on either side of their respective means. Furthermore, examination of skew and kurtosis coefficients for these quality estimates confirms that their distributions are approximately normal.[14]

The third row in Table 3 presents the standard deviations of the quality estimates. These range from .10 to .16. For the validity estimates, which have a standard deviation of .10, one can infer that about two-thirds of all validity estimates fell in the range .71 to .91. Hence, about two-thirds of the survey measures examined here contained between 50% and 83% valid variance; roughly one-sixth contained more than 83% valid variance and about one-sixth had less than 50% valid variance. Comparable calculations for method effects suggest that two-thirds of the measures had between 0% and 7% method variance. Similarly, two-thirds of the measures had between 14% and 48% residual variance.

Survey design characteristics and data quality. The next step of the analysis was to perform a series of multivariate analyses to attempt to statistically explain the observed variation in the quality estimates. Given the nature of the available predictor variables and the lack of previous exploration in this area, it was important that the multivariate analyses be able to incorporate nominal-scale predictors, non-linear relationships, and interactive (i.e., non-additive) effects. Accordingly, multiple classification analysis using pattern variables (Andrews et al., 1973) and SEARCH (formerly known as AID—Sonquist et al., 1971; Survey Research Center Computer Support Group, 1981) were selected as the primary multivariate analysis methods.

Predictive power of survey characteristics. After considerable exploration using various combinations of the available independent variables, a final multivariate anlaysis was selected that used 13 predictors. (One of these predictors was a pattern variable based on two more basic

Table 4
Summary results from multiple classification analysis using thirteen aspects of survey design to predict validity, method effects, and residual error

Predictors	Quality component being predicted						
	Validity		Meth. eff.		Res. error		
	adj. eta²	MCA beta²	adj. eta²	MCA beta²	adj. eta²	MCA beta²	Mean beta²
Characteristics of response scale							
Number of answer categories	.19	.56	.20	.68	.25	.74	.66
Explicit "Don't know" option	.06	.31	.09	.45	.02	.30	.35
Category labeling	.00	.27	.04	.28	.00	.15	.23
Explicit midpoint	.01	.01	.03	.06	.01	.00	.02
Characteristics of the item							
Absolute versus comparative	.00	.28	.15	.15	.02	.33	.25
Length of introduction & question	.12	.13	.27	.35	.12	.10	.19
Questionnaire design, data collection							
Battery length	.10	.17	.09	.19	.13	.44	.27
Position of item in questionnaire	.24	.13	.22	.18	.24	.16	.16
Data collection procedure	.02	.03	.18	.24	.01	.02	.10
Characteristics of the topic							
Sensitivity to social desirability	.09	.07	.04	.00	.11	.08	.05
Content specificity	.06	.06	.04	.00	.07	.04	.03
Experience versus prediction	.04	.01	.04	.00	.05	.01	.01
Content salience	.08	.01	.09	.00	.08	.00	.00
Joint explanatory power of 13 predictors (R^2 adj.)	.66		.72		.67		
N (number of quality estimates)	2115		2115		2115		

Notes: Eta², the squared correlation ratio, shows the proportion of variance in a dependent variable explained by one predictor variable considered alone. The coefficients shown here include an adjustment for shrinkage likely to occur upon replication.
MCA beta² is a measure of the strength of relationship between a dependent variable and a predictor while holding constant the effects of all other predictors included in this analysis.
R^2, the squared multiple correlation coefficient, shows the proportion of variance in a dependent variable explained by all predictors jointly. An adjustment for likely shrinkage upon replication has been incorporated.

Table 5
Effects of survey design characteristics on data quality
(Bivariate coefficients show deviations from the mean associated with membership in the designated category; multivariate coefficients are similar except effects of other predictors have been held constant by multiple classification analysis. Also see notes at end of table.)

	Number of estimates	Validity		Meth. eff.		Res. error	
		Biv.	Mult.	Biv.	Mult.	Biv.	Mult.
Characteristics of the response scale (4 variables)							
Number of answer categories							
2	120	−.00	−.06	.04	.11	.02	.04
3	364	−.09	−.13	−.07	−.05	.14	.22
4–5	542	.01	.04	.00	.01	−.00	−.06
7	650	.02	.00	−.03	−.02	−.01	.04
9–19	208	.04	.01	.09	.21	−.09	−.07
20+ or actual frequency	231	.05	.14	.07	−.13	−.14	−.28
Explicit "Don't know" option							
No	1516	−.02	−.04	.02	.04	.01	.06
Yes	599	.04	.09	−.05	−.11	−.03	−.14
Category labeling							
All categories labeled	1502	−.00	−.04	−.01	.04	.00	.03
Some categories unlabeled	613	.00	.08	.03	−.09	−.01	−.10
Explicit midpoint							
No	532	.02	.02	−.03	.05	−.03	.01
Yes	1583	−.01	−.01	.01	−.02	.01	−.01
Characteristics of the item (2 variables)							
Absolute versus comparative							
Absolute	1275	.00	−.04	.03	.01	−.02	.07
Comparative	840	−.01	.07	−.05	−.02	.03	−.11
Length of introduction and question							
Short intro., short question	56	−.02	−.05	−.04	−.01	−.09	−.02
Short intro., medium question	231	−.03	−.05	−.03	.03	.06	.08
Short intro., long question	259	−.06	−.01	−.04	−.03	.09	.05
Medium intro., short question	249	−.03	.03	−.01	−.08	.05	−.04
Medium intro., medium question	365	−.02	.06	−.00	−.08	−.01	−.07
Medium intro., long question	44	.12	.06	−.02	−.09	−.16	−.05
Long intro., short question	351	.00	−.00	.11	.05	−.04	−.00
Long intro., medium question	219	.03	.00	.02	.07	−.05	−.00
Long intro., long question	341	.03	−.03	−.06	.07	−.19	.06
Questionnaire design (3 variables)							
Battery length							
1 item (i.e., not in battery)	537	−.05	.03	−.03	−.04	.08	−.10
2–4 items	519	.03	.05	.03	.02	−.08	−.11
5–9 items	736	.02	−.03	−.03	−.03	−.02	.09
10 or more items	323	−.00	−.07	.04	.10	.02	.14
Position of item in questionnaire							
1–5 (i.e., among first five items)	148	−.09	−.01	−.06	.03	.15	.03
6–25	280	−.09	−.02	−.03	−.02	.13	.05
26–35	336	.01	.04	.00	−.01	−.01	−.07
36–39	276	.04	.02	−.07	−.03	−.04	−.03
40–100	493	.05	.04	.00	−.05	−.10	−.06
101–200	309	.02	−.04	−.00	.07	−.01	.08
201–348	273	−.02	−.06	.11	.06	.01	.08
Data collection procedure							
Telephone interview	1332	.01	−.01	−.03	−.02	−.02	.01
Face-to-face interview	399	−.02	.00	.09	.10	.03	−.01
Group administered questionnaire	384	.01	.04	−.01	−.05	.02	−.04

(continues on next page)

Table 5 continued

	Number of estimates	Validity		Meth. eff.		Res. error	
		Biv.	Mult.	Biv.	Mult.	Biv.	Mult.
Characteristics of the topic (4 variables)							
Sensitivity to social desirability							
Low or medium	1587	−.02	−.02	.01	.00	.03	.03
High	528	.05	.05	−.04	−.00	−.09	−.08
Content specificity							
Low	700	−.02	−.03	−.01	.01	.04	.04
Medium	401	−.02	−.01	.04	−.00	.04	.02
High	1014	.03	.02	−.01	−.01	−.04	−.03
Experience vs. prediction							
Actual experience	1753	.01	.00	.01	−.00	−.02	−.01
Predictions	362	−.04	−.02	−.05	−.00	.08	.04
Content salience							
Low	445	−.05	.01	−.02	−.01	.08	−.01
Medium	1056	.03	.01	−.02	−.00	−.03	.00
High	614	−.01	−.01	.05	.01	.00	.01

Note: The estimates of validity and method effects are based on lambda parameters from LISREL; the estimates of residual error are based on the *square roots* of the theta parameters from LISREL.

Statistical Significance: By conventional tests of significance, a difference between most pairs of multivariate coefficients of .02 or more is significant at the p = .05 level. Standard errors for the multivariate coefficients range from about .002 when N is 800 to about .01 when N is 50; see text.

variables.) These 13 predictors are listed in Table 4, together with an indication of the explanatory power of each, both singly and in combination with all others, and with an indication of the total explanatory power achieved by the entire set. Table 5 presents more detailed results from this anlaysis and shows the effect on each of the three measurement quality estimates of each category of each predictor variable, both before and after holding constant the other predictors. After describing the nature of each predictor variable and the results of the multivariate anlaysis, brief mention will be made of several preliminary analyses whose results led to the design of this main analysis.

One of the first things to note in Table 4 is that the characteristics of survey design represented in this analysis account for a large part of the variation in the estimates of validity, method effects, and residual error. As shown by the adjusted R^2s, the survey design characteristics account for 66% to 72% of the variance in the dependent variables. This is an important finding, for it shows that much of the variation in measurement quality can be explained, and hence measurement quality is subject to prediction in other surveys and perhaps to improvement.

As indicated in the table, the characteristics of survey design have been grouped into four conceptually distinct sets. The first included four aspects of the response scale used with the survey item: the number of answer alternatives, whether a "Don't know" alternative was presented to the respondent, whether each answer category carried its own lable, and whether the response scale included a midpoint.

Next comes a set of variables that tap several characteristics of the survey item itself. Included here is whether the item called for a *comparative* judgment (e.g., "Would you say that your health is better or worse now than it was a year ago?") or an *absolute* judgment (e.g., "How do you feel about your health and physical condition?"—answered on a seven-point scale of satisfaction). Also assessed is the length of the question and the length of any general introduction to the question. (This is a nine-category pattern variable that includes all combinations of short, medium, and long questions and introductions.)

The third set of survey characteristics has to do with the design of the questionnaire or the interview schedule and the data collection mode. Measures here include whether the item was part of a battery (i.e., in a set of items having a common introduction and using similar response scales) and if so, the length of that battery, the position of the item in the questionnaire (i.e., how far from the beginning the item was located), and whether the data collection was by face-to-face interview, telephone interview, or group-administered questionnaire.

The fourth set of survey characteristics includes four variables tapping various aspects of the topical content of the survey item. One is the extent the topic was judged likely to be subject to social desirability biases. (For example, reports about "eating too much" or "keeping up with the news" were expected to be more subject to social desirability bias than were perceptions about national business conditions.) Next comes the specificity of the thing being asked about (where questions about satisfaction with own housing, state of own health, and frequency of drinking beer were judged to be more specific—i.e., more concrete, less abstract—than questions

about life-as-a-whole or national business conditions).[15] A third measure distinguishes between (1) survey items that asked about things the respondent was currently experiencing or had previously experienced and (2) items that asked for predictions about the future. The last measure in this set taps the salience of the topic to the respondent—i.e., its importance or immediacy. (High salience topics included own health, own standard of living, own family life; low salience topics included business conditions and satisfaction with community.)[16]

The beta^{2}s in Table 4 indicate the relative importance of the various survey characteristics in accounting for validity, method effects, and residual error. (The eta^{2}s, which reflect the simple bivariate relationships are also interesting, but since these sometimes include spurious effects arising from the particular combinations of items and surveys assembled for this study, the multivariate results—reflected in the beta^{2}s—are the more useful.) The most important aspects of survey design, as indicated by average beta^{2}s of .25 or more, are the number of answer categories in the response scale, whether these answer alternatives included a "Don't know" category, battery length, and whether the item uses an absolute or comparative perspective. Making appropriate choices with respect to these design characteristics can, apparently, have an important impact on the measurement quality in a survey. Survey characteristics with a more moderate and/or less general effect (average beta^{2}s in the range .15 to .24) include whether answer categories are all labeled, the length of the question and its introduction, and the position of an item in the questionnaire. Choices with respect to these matters can also have important effects on measurement quality. Equally interesting are the design characteristics that did *not* prove to have substantial effects on data quality. These include all four aspects of the topic being asked about, whether the answer scale includes an explicit midpoint, and the data collection procedure. The effects of some of these design matters have been the subject of considerable debate among survey researchers, and it is of real interest to find that—at least in this study—they have relatively little impact on measurement quality.

Specific effects of survey design. In Table 5 one can see the way each of these characteristics of survey design relates to validity, method effects, and residual error. Presented there are the effects of each category of each predictor variable on each component of measurement quality, both before and after controlling for the effects of the other predictors. The coefficients in the "Biv." columns are results from the simple bivariate analyses (the relationships that are summarized by the eta^{2}s in Table 4), and the coefficients in the "Mult." columns come from the multiple classification analyses (and are summarized by the beta^{2}s in Table 4). As noted earlier, the multivariate analysis results probably are the more useful.

The multivariate coefficients in Table 5 show the amount by which the quality component would go up or down from the mean (presented in Table 3) if a measure had the characteristic indicated and there were no effects from any of the other predictor variables. For example, the −.06 effect on validity of using an item with a two-point answer scale means that, holding everything else constant, validities of such items can be expected to be six "points" lower than that of the average item—i.e., .75 (=.81 − .06). Although the coefficients presented in Table 5 may appear small to the uninitiated reader, in many cases they show sharp and important effects.[17]

1. Number of answer categories. The number of answer categories is shown in Table 4 to have the biggest effects on data quality, and the multivariate coefficients in Table 5 show that, in general, as the number of answer categories goes up, data quality goes up—i.e., validity tends to increase and residual error tends to decrease. (The trend for method effects is less clear, though at its extremes it follows the general trend.) The validity and residual error results show an interesting and possibly important curvilinearity at the low end of the scale: Both two-point and three-point scales give less good measurement quality than scales with four or more categories, but two-point scales are not as bad as three-point scales.

These results favoring use of greater numbers of scale points are not entirely expected. They are consistent with a wide-ranging and uncoordinated literature showing that use of more categories (at least up to five to seven) produces a more accurate reflection of the underlying variation. See, for example, Bollen and Barb (1981), Cochran (1968), Conner (1972), Cox (1980), Lissitz and Greene (1975), Martin (1978), Pearson (1913), and Ramsey (1973). This previous body of literature, however, would not have predicted the marked superiority of our "20+" category, and it is possible that this aspect of our results is an artifact.

2. Explicit "Don't know" option. According to Table 4, the second most important survey characteristic is whether the answer categories include an explicit "Don't know" option. The effect of this design matter is clear and consistent: Inclusion of an explicit "Don't know" category is associated with better data—higher validity, lower method effects, and lower residual error. The reasonable idea that one should let respondents "opt out" if they lack the requisite information or opinions receives strong endorsement.

3. Battery length. The third most important predictor denotes whether the item is included in a battery with other items, and if so, how long that battery is. The primary effect here is with respect to residual error, but the results for validity and for method effects, though weaker, tend to follow the same trend. The results show that not being in a battery or being in only a short battery is associated with higher quality data than being in a medium length battery (five to nine items), which is not as bad as being in a longer battery.

Including items in batteries where all share a common introduction and/or identical answer scale offers obvious advantages with regard to efficiency and speed, but these results suggest that such gains come at the cost of reduced measurement quality if the battery consists of more than just a few items. It is possible that respondents and/or interviewers recognized the "production line" character of this survey strategy and that it promotes carelessness in the way questions are asked and answered.

One might have guessed that what mattered would not be the total length of the battery in which an item is included but rather how far into a battery the item is located. Both characteristics of items were examined, and, to our surprise, battery length showed stronger relationships to each of the measurement quality components than did position in battery. By logical necessity, the two variables were highly correlated, and it was not feasible to retain both in the multivariate analysis.

4. Absolute versus comparative. The fourth most important predictor taps whether the item uses an absolute or comparative perspective. The results clearly favor the comparative approach: As shown in Table 5, this is associated with higher validity and lower residual error (and, to a minor extent, lower method effects as well). It many be that the provision of some "anchor points," as is required in the comparative approach, helps respondents give more precise answers.

5. Length of introduction and of question. Two item-designation characteristics, the length of the introduction to a question and the length of the question itself, were combined into the single nine-category pattern variable that was used in the multiple classification analysis because preliminary exploration had shown these characteristics to have both curvilinear main effects and a first-order interaction.[18] These are reflected in the multivariate coefficients presented in Table 5. Note that validity tends to be highest and both types of error lowest when questions are preceded by a *medium*-length introduction (defined as an introduction of 16 to 64 words). Furthermore, given a medium-length introduction, medium or long questions (16–24 or 25+ words, respectively) yield higher validity and less of both kinds of error than shorter questions. With respect to validity, given a short introduction, it is better to follow it with a long question; on the other hand, given a long introduction, it is better to follow it with a short or medium-length question. The overall pattern of these results suggests that short introductions followed by short questions are not good (perhaps the respondent does not have an opportunity to get a clear understanding of what is being asked and/or does not have time to develop a precise answer) and neither are long introductions followed by long questions (respondents may lose track of what is being asked and/or get bored while waiting to answer).

6. Position of item in questionnaire. Table 5 shows a consistent, moderate-strength tendency for data quality to be lower when items are at the beginning of a questionnaire (within the first 25 items) or far into a long questionnaire (beyond the 100th item); in either of these situations validity tends to be lower than average and both types of errors tend to be above average. Better data quality comes from items that fall in the 26th to 100th positions.

It is not hard to imagine how this effect might come about. Items that come early in a questionnaire may be presented before the respondent is "warmed up" to the task and, in an interview, before rapport between interviewer and respondent has been developed. On the other hand, after the 100th item, respondents and/or interviewers may begin to suffer from fatigue or become careless.

7. Category labeling. The moderate-strength relationships associated with category labeling were a surprise and are not yet fully understood. The contrast is between items whose answer categories were fully labeled—i.e., an explicit meaning was indicated for every possible answer—and items where some of the answer categories were left unlabeled—as in a format where only the end points are labeled and some intermediate points take their meaning from their relative position on the page. Contrary to expectation, the results of the multivariate analysis suggest that data quality is below average when all categories are labeled.[19]

8. Predictors showing weak links to quality. All of the remaining predictors have average beta^{2}s of .10 or less, and with just a single exception none of the individual beta^{2}s associated with these predictors exceeds that level.[20] The fact that some of these relationships between survey design and data quality are weak is of great interest. It is helpful to know that the oft-debated issue of whether to allow respondents an "easy out" by including an explicit mid-point ("neutral," "pro-con," etc.) had only slight effects on data quality. It is also interesting to observe that it made little difference whether an item asked about things the respondent had already experienced or asked for predictions about the future. (The small effects that do appear here are in the expected direction—i.e., favoring things the respondent has experienced.) And the finding of only very small effects on validity and residual error attributable to whether data were collected by telephone interviews, face-to-face interviews, or group-administered questionnaires will also be encouraging to many survey researchers. That none of the characteristics of the substantive topic being asked about in a survey item had an important effect on data quality was in some respects a disappointment, for we had clear expectations about how these variables might relate. However, the fact that the relationships all proved very weak might be seen as a desirable outcome by many, for it suggests that—other things equal—data of at least average quality can be obtained about a wide range of topics.

Supporting analyses. The multiple classification analysis (MCA) reported in Tables 4 and 5 was designed on

the basis of results from a series of preliminary explorations. Key questions addressed in these preliminary analyses had to do with (1) an appropriate and feasible set of predictors for the MCA, (2) whether important interactions were present that, without special treatment through the pattern variable approach, would be unrepresented in the additive MCA model, and (3) whether the predictors were defined in a sufficiently general way that they could be used in future studies of measurement quality. In addition to the survey characteristic variables that have already been discussed, 11 others were also examined.[21]

Two analytic approaches were helpful in designing the final multivariate analyses. Careful attention was given to all the two-way (and in some cases higher order) tabulations involving the predictors. Although the survey items assembled for this investigation constituted a large and heterogeneous set, there were—as expected—certain combinations of survey design characteristics that were perfectly or near-perfectly confounded with others. Of course these had to be identified and handled in some way (by eliminating one of the potential predictors from the analysis or by combining appropriate categories).

In addition, a series of SEARCH[22] runs was made to explore the predictive power of each predictor in combination with others and to identify statistical interactions. These were helpful in selecting the more useful predictors when some had to be omitted and in assuring that no major interactions were being missed. In fact, given that the proportions of variance explained in the final SEARCH runs were extremely close to the R^2s achieved by the multiple classification analyses presented in Table 4, we can be confident that the use of the simple additive MCA model in the way we have done is appropriate for these data.[23]

Respondent characteristics and data quality. The idea that characteristics of respondents may relate to data quality is an appealing one and can be readily investigated in a sophisticated way with these data. This analysis is possible because estimates of data quality were obtained for many different subgroups of respondents. For example, data quality estimates were obtained separately for young respondents, middle-aged respondents, and elderly respondents, and hence it is possible to see how validity, method effects, and residual error—each averaged across many survey items—varies with age.[24]

Analysis strategy. Because the respondent subgroups for which data quality estimates could be obtained varied from survey to survey, and because survey design characteristics (e.g., topics investigated, answer formats used, data collection procedure, etc.) also varied from survey to survey—and because survey design characteristics have a major impact on data quality, as discussed in the preceding section—it is necessary to remove the survey design effects before looking at the effects of respondent characteristics. This was accomplished through a process of residualization.

As with any analysis of "residual scores," the dependent variables in the analyses to be described are the deviations of the actual estimates of an item's validity, method effects, or residual error from *what would be predicted* to be that item's measurement quality given the design of the survey in which that item occurred. Another way to think about these analyses is to note that the measurement quality estimates begin with the variances shown in Table 3 (the square of the standard deviations), then a certain portion of this variance was explained by the multiple classification analyses (as indicated by the R^2s in Table 4), and the goal now is to explain some of the remaining (unexplained) variance using characteristics of respondents.

Table 6 presents the results. In interpreting the information shown there the nature of the dataset must be clearly understood: As noted previously, a "case" is a survey-item-as-answered-by-a-particular-group-of-respondents. Thus, for example, the "N" of 106 for the All-respondents-together group indicates that measurement quality estimates for 106 survey items were computed for the total set of people responding to one or another of the surveys represented here. (It does *not* mean that there are 106 respondents; as noted in part 3, more than 7,600 respondents participated in these surveys.) In some instances the number of items for which quality estimates were computed is rather small, and of course effects based on small numbers of cases will be less stable than others.[25]

Table 6 was derived by performing three simple (but large scale) bivariate analyses: A 53-category respondent-group variable was related to the (residualized) estimates of validity, method effects, and residual error. The resulting coefficients, shown in Table 6, indicate the effect on data quality when items were answered by members of the designated group. For example, note the −.04 effect on validity for people with only 0–11 years of education and the +.03 coefficient for people with at least a Bachelor's Degree. This means validity was 4 "points" lower than it otherwise would have been when items were answered by respondents with less than a high school education, and that it was 3 "points" higher than otherwise when respondents had completed college. This suggests a clear positive relationship between education and validity, a finding that many survey researchers would expect.

Predictive power of respondent characteristics. Before looking at Table 6 in detail, it is instructive to consider the explanatory power of all the respondent characteristics taken together. As shown at the end of the exhibit, the eta^{2}s, adjusted for likely shrinkage upon replication, range from .05 to .16. Specifically, these respondent characteristics can explain about 12% of the remaining unexplained variance in the validity estimates, 16% for the method effects, and 5% for residual error. Clearly, respondent characteristics are not a major predictor of

Table 6
Effects of respondent characteristics on data quality
(Coefficients show deviation from the mean associated with membership in the designated category after effects of 13 survey design characteristics have been removed. See text and notes at end of table.)

	Number of estimates	Validity	Meth. eff.	Res. error
Group				
All respondents together	106	.01	.00	.00
Education				
0–11 years	82	−.04	.01	.06
High school (or HS plus tech.)	63	.00	.01	−.01
Some college	63	.03	−.04	−.02
Bachelors degree or more	87	.03	.00	−.05
Grade school to some college	24	−.00	.01	.00
Some college or more	19	.04	−.03	−.03
Age				
18–34 (or 18–30)	106	.02	−.02	−.02
35–54 (or 31–56)	87	.00	−.01	−.00
55–90	82	−.04	.03	.05
65–70	19	−.06	.05	.05
71–90	19	−.08	.11	.03
Race				
White	82	.01	−.01	−.01
Black	82	−.04	.01	.04
Sex				
Female	106	−.00	.00	.00
Male	106	.01	.00	−.01
Where respondent grew up				
Rural	24	−.01	−.01	.01
Suburban	24	−.01	−.04	.02
Urban	24	.02	.01	−.04
Seniority in Firm X				
0–4 years	24	.00	.01	−.01
5 or more years	24	−.00	−.01	.00
Interviewer's ratings				
R's interest high	82	.02	−.01	−.03
R's interest low	82	−.03	.02	.02
R's intelligence high	19	.03	−.02	−.03
R's intelligence low	19	−.00	.01	.01
R's sincerity high	19	.02	−.02	−.01
R's sincerity low	19	−.01	.03	.01
R's suspiciousness high	19	.01	−.01	−.02
R's suspiciousness low	19	.01	−.04	−.00
R's reluctance high	9	.01	−.08	.02
R's reluctance low	9	−.00	.03	−.01
Respondent's own ratings				
Interview seems long	9	−.03	.04	.03
Interview seems short	9	.01	−.02	−.01
Interest in survey high	9	−.02	−.06	.03
Interest in survey low	9	.01	.04	−.03
Importance of survey topics high	24	−.00	.02	.00
Importance of survey topics low	24	−.00	−.02	−.00
Survey's expected impact high	24	−.00	.01	−.00
Survey's expected impact low	24	−.00	−.01	.01
Assistance by interviewer				
None or once	12	.02	−.01	−.02
Twice or more	12	−.05	.08	.05
Clarifications requested				
None or one	9	.01	.03	−.02
Two or more	9	−.03	−.12	.06
Questions repeated				
None	9	−.01	.04	−.00
A few to many	9	.00	−.01	−.01
Was R interviewed by SRC before?				
No	54	.01	.00	−.01
Yes, within 6 months	54	.01	−.00	−.00
Number of attempts to reach R				
One	73	.01	−.01	−.01
Five or more	73	.01	−.01	−.01
Special interviewing techniques				
None, standard methods	9	.01	−.03	.01
Spec. instructions, commitment, etc.	9	−.01	.04	−.01
R's concern for social desirability				
High	51	−.01	.01	.02
Low	51	.02	.01	−.03
Explanatory power of 53 groups above				
eta² adj.		.12	.16	.05

Statistical Significance: By conventional tests of significance, a difference between these means is significant at the p = .05 level if: the difference is at least .02 and N's are at least 50, or the difference is at least .03 and N's are at least 25, or the difference is at least .05 and the N's are about 10. Standard errors for the coefficients are about .007 when N is 100, .010 when N is 50, and .020 when N is 10. See text.

variation in the quality of measurement in these data. However, this conclusion reflects the particular characteristics we were able to examine and the number of survey items available for each subgroup. There are some respondent characteristics that show potentially important links to measurement quality.

Specific effects of respondent characteristics. 1. Education, age, race. Education, age, and race each show intriguing relationships with data quality. Table 6 shows that validity was higher for more educated respondents, for younger respondents, and for whites. Residual error showed exactly the opposite trends. (Method effects had a clear trend only with respect to age, where it showed sharp increases with increasing age.)

Of course these demographic characteristics are known to be substantially correlated among United States adults, and an immediate question is whether these trends would hold up under various controls. A series of subsidiary analyses (not shown) indicated that most of the race effects in Table 6 disappeared when education was controlled. Apparently most of the "race effect" is attributable to the fact that blacks tend to have less education than whites. However, the age and education effects persisted as strongly as ever when the other variable was controlled. This suggests that age and education each has its own independent impact on data quality.[26]

2. Sex. Table 6 indicates virtually no relationships between the sex of the respondent and the indicators of data quality—a result that seems reasonable.

3. Interviewer ratings of the respondent. At the conclusion of an interview in some surveys, interviewers recorded their impressions about the respondent—e.g., interest in the survey topics, general intelligence, sincerity with which questions were answered, suspiciousness, reluctance to participate, etc. How do these kinds of characteristics relate to data quality? As shown in Table 6, most of the effects are in the expected direction—i.e., high interest, intelligence, and/or sincerity were associated with higher validity and lower errors—but none of the effects is very large. (In 15 relationships examined in this set, the only marked exception to what one might expect involves respondent reluctance and correlated error: Reluctant respondents tended to produce answers with *fewer* method effects. It is not clear why, but the number of cases is very small and the anomaly does not seem important.)

4. Respondent's own ratings. Ratings by respondents themselves about their interest in the survey, the importance of the survey topic, etc., show mainly weak and conflicting relationships to data quality and hence did not prove very useful.

5. Indications of respondent difficulties. In some surveys, records were kept regarding whether the respondent had difficulty in coping with the interview. These included instances of assistance provided by the interviewer, requests for clarification, and/or repetition of questions. The number of survey items for which this kind of information is available is very small, but most of the trends evident in Table 6 go in the expected direction—i.e., respondents who had greater difficulty with the interview tended to give answers with lower validity and higher residual error. (Trends with respect to method effects are conflicting.)

6. Prior participation in surveys. Does prior participation in another survey conducted by the Survey Research Center relate to data quality? Table 6 clearly suggests there is no relationship.

7. Difficulty in reaching respondent. A persistent concern of survey organizations is whether it is "worth it" to try to contact hard-to-reach respondents. Not to do so raises the possibility of nonresponse bias affecting the data, to do so risks the inclusion of data that some observers have suspected might be of lower quality. The coefficients in Table 6 are clear on this point: Absolutely no quality differences are observed between respondents contacted on the first attempt and those contacted only after five or more attempts.

8. Respondent's concern for social desirability. In a few of the surveys we included five items selected from the Crowne-Marlowe (1964) scale of social desirability in order to obtain an indication of respondents' concern for presenting themselves in socially desirable ways.[27] Although this scale proved to have low internal homogeneity, it did produce results in the expected direction: Table 6 shows that respondents who scored relatively high on this concern had a modest tendency to give data that was below average in validity and above average in residual error.

Further explorations. A legitimate concern is whether the effects on data quality of the various survey design characteristics explored in the previous section are the same for all types of respondents. In formal statistical terms, are there interactions involving respondent characteristics, survey design characteristics, and data quality? Or can respondent effects simply be added on to the survey design effects to get good predictions of data quality? With 22 sets of contrasting respondent characteristics (see Table 6) and 13 sets of survey design characteristics (Table 4) and 3 data quality assessments, there were over 800 first-order interactions that might potentially occur. Eight of these that promised to be most interesting and for which data were available in sufficient depth were checked in detail. These included combinations of (a) age, education, or respondent concern for social desirability with (b) number of answer categories, data collection procedure, length of introduction and question, or item sensitivity to social desirability, as related to (c) mean levels of validity. For example, we checked to see whether the general finding that validity improved as the number of answer categories increased was as applicable for people who had not completed high school as for those with a college education. (It was.) The general result of this exploration was that no major interactions were found. While of course we cannot be sure that interactions do not exist where we have not checked for them, we have increased confidence that most of the survey design effects described in the preceding section will be generally applicable to a wide range of different types of respondents.

Implications of the study

There are at least four ways in which the outcomes of this investigation may prove useful. These have to do with (1) implementation of the technique in future surveys, (2) empirically-based recommendations about survey design, (3) prediction of measurement quality for survey items not included in this study, and (4) using measurement quality estimates to enhance the meaningfulness

of observed relationships. Some comments and advice about each of these topics seem warranted.

Implementing the technique

One of the most important outcomes of this research is the discovery that it was indeed feasible to generate measurement quality estimates using the new measurement modeling techniques in regular on-going surveys. As noted in part 2 of this report, basic ideas about measurement modeling and some limited applications preceded this study, but never before has there been an attempt to implement this technology in a broad way in an operational setting.

The key components of the approach, that we believe could readily be implemented in other surveys to generate measurement quality estimates, include the following.

Identifying key concepts. First, one must identify a few (perhaps two to eight) concepts that are of sufficient importance in the survey to justify some modest investment in assessing the quality of their measures. Given a choice, one would prefer concepts that are distinctly different from one another (i.e., having only low or moderate statistical interrelationships) and concepts that lend themselves to measurement by similar methods. In our work, this aspect of the implementation usually proved easy.

Developing a multimethod-multitrait design. Next, one must develop a multimethod-multitrait data design for the selected concepts. The designs used by us ranged in size from 3 × 3 (three methods, three concepts) up to 4 × 7 and 3 × 8, and our experience suggests that designs much larger or smaller than these will probably not be attractive. A 2 × 2 design cannot provide unique estimates (technically, it is "unidentified"), and we believe designs much larger or smaller than those we used will prove unacceptably burdensome to respondents and/or interviewers. Except in the smallest designs, it is not necessary that every concept be assessed by every method. However, every concept must be assessed by at least two methods, and every method must be used with at least two concepts, and there must be sufficient "interweaving" of methods and concepts that there is no subdesign included that is as small as 2 × 2.

The range of possibilities for different "methods" is potentially very great. Ideally, the methods should be as distinct as possible. In the present study, we attempted to achieve that by using distinctly different response scales, and this seemed to work well. In previous implementations (Andrews and Withey, 1974, 1976; Andrews 1979) we have also used information from entirely separate sources—e.g., friends and neighbors of the respondent with whom the respondent gave us permission to consult or external evaluators of research teams who could inform us of the teams' performances. Other alternatives, not used by us, include such external sources as hospital archives, voter registrations, psychological tests. Obtaining information by such strategies may not be feasible in some surveys, but a wide range of alternative "methods" should at least be considered.

It is crucial that all alternative methods for assessing a concept do in fact tap the *same* concept. This fundamental rule would be violated if, for example, one method assessed the frequency of doing something and another dealt with the importance of doing it; if the concept involves frequency, then alternative methods for assessing frequency must be found.

We found that assembling an appropriate multimethod-multitrait design, while not hard, often required considerable careful thought.

Collecting the data. Having assembled items for an appropriate multimethod-multitrait design, these items need to be built into a questionnaire or interview schedule in a way that does not obstruct the smooth flow of the data collection. As noted in part 3 of this report, sometimes we found it helpful to briefly acknowledge to respondents that a set of questions was exploring the same topics as a previous set and to note that these questions did it in a different way.

Although it was our practice to administer the methodological supplement to all respondents, there may arise situations in which it is reasonable to administer the extra items to just a subset of the respondents (e.g., a representative subsample of the total sample). There is, however, a minimum number of respondents required for any particular measurement model. (The more parameters included in the model, the more respondents needed.) For most of the models estimated in this study, experience suggests one would like to have at least 100 respondents and preferably 200 to 300. Of course, if one wanted to generate measurement estimates for particular groups of respondents (e.g., elderly people), the total number of people answering the methodological supplement would have to be large enough to include the above minimum numbers of respondents in the desired subgroup.

Obtaining measurement quality estimates. After the data have been collected and prepared for computer analysis, one should check for instances of marked skews in the distributions and for curvilinearities in the relationships. If present, these might be reduced or eliminated by appropriate transformations. Next, a correlation (or variance-covariance) matrix is computed. (If there is more than a trivial amount of missing data, we suggest deleting cases only for the relevant variables rather than for the entire matrix.) Then this matrix is inputted into a computer program that will obtain parameter estimates for a structural model appropriately adapted from the one used here. The LISREL or COFAMM computer programs are the ones we used.[29] As noted in part 4, the estimates we have reported for construct validity, method effects, and residual error are what LISREL put out as concept lambdas, method lambdas, and theta parameters (of which we took the square root), respectively. Of course, before interpreting these parameters as esti-

mates of measurement quality, one should ensure that the solution meets the various criteria for model fit discussed in part 5 of this report.

Comment on cost. The marginal cost of obtaining measurement quality estimates for key concepts assessed in a survey can be quite low. Two components of cost need to be considered. One is the cost of collecting the extra data. Since the number of additional items is not very large (rarely more than 10–20 for designs of the size recommended above), this is not a major cost. Furthermore, in addition to buying the means for estimating measurement quality, this cost often can be partly justified on the basis that it buys *better* measures of the survey's key concepts—because the multiple measures of a concept can be combined into a composite scale which is likely to have higher reliability (and validity) than a measure based on just a single item.

The other component of cost arises from the staff time and computing charges for the subsequent anlaysis of those data. For a professional with the requisite skills (or with access to skilled consultants), neither of these need be a major task.

Recommendations about survey design. A second general implication of this investigation is the promise it holds for generating knowledge about how to design surveys that will yield higher quality measures. The multivariate analysis described in part 5 of this report itself includes many suggestions for ways to enhance measurement qualtiy, and it could be a prototype for similar analyses performed on other measurement quality estimates that may become available for other survey measures, for other national or cultural settings, and/or for specialized groups of respondents. The potential of generating quality estimates for a particular survey item as it is responded to by a particular set of respondents, and then linking that information to various characteristics of the design of the survey in which that measure was implemented, opens exciting possibilities for making survey design less of an art and more of a science.

Our discussion of the multivariate results in part 5 of this report has already noted the specific characteristics of survey design which, in this study, related to higher levels of measurement quality, and there is no need to repeat that discussion here. What we would emphasize, however, is the general point that the trends shown there—particularly if confirmed and extended in future methodological studies—provide specific answers for many of the very practical questions that survey designers decide on every day, often on the basis of less good information.

Predicting the measurement quality of other items. One of the potential uses for the detailed set of coefficients presented in Table 5 of this report is to generate predictions of measurement quality for survey items that were not actually included in this study. In many instances, as noted above, we believe it would be feasible for a survey team to develop its own multimethod-multitrait methodological supplement and generate its own estimates of measurement quality. However, in addition or perhaps instead, measurement quality estimates could be obtained by extrapolating the results from the present investigation. The accuracy of such estimates is, of course, open to question. However, almost certainly it would be better to use them than to totally disregard measurement error. Assuming perfect construct validity and zero random and correlated errors is practically always wrong! The predictions derived by extrapolating from results in this study should be more accurate as the items whose measurement qualities are being predicted are more similar in type of content and format to the items examined here, as the surveyed population is more similar to a general American or Canadian adult population, and as the surveying organization uses methods and procedures more similar to those of the University of Michigan's Survey Research Center.

To actually make a prediction of the measurement quality of a survey item, one would use the information in Tables 3 and 5, beginning with the means shown in Table 3. These would be adjusted upward or downward according to the particular combination of survey design characteristics that pertain to the item for which one is making the prediction. The appropriate adjustment is determined by adding together relevant coefficients selected from Table 5. Of the 13 survey design characteristics presented in Table 5, all except the final 4 should be easy to determine for any item. The final 4 have to do with characteristics of the topic, and if one did not want to include these in the prediction, one would not go very far astray if they were simply neglected. (This is true because none of these 4 correlate very strongly with other predictors, and none has a strong effect on any of the measurement quality estimates.)

Measurement quality and observed relationships. This report began by observing that measurement errors influence observed relationships. It is appropriate to conclude with a brief discussion of how one can use information about measurement error to make inferences about the true relationships. First we consider the matter with respect to a simple bivariate relationship between measures based on single items, then consider measures based on combinations of items, then note how multivariate relationships can be handled.

Bivariate relationship between single-item measures. The basic assumptions of measurement modeling which can be represented in the algebra of path analysis (Zeller and Carmines, 1980), predict that a simple observed relationship will be equal to the true relationship between the concepts being tapped times the product of the validities of the measures, plus the proportion of correlated error in the measures. Algebraically,

$$r_{AB} = r'_{AB} V_A V_B + E^c_{AB}, \qquad (1)$$

where: r_{AB} is the observed product-moment correlation between measures A and B,

r'_{AB} is the true correlation between the concepts tapped by measures A and B,
V_A is the construct validity of measure A,
V_B is the construct validity of measure B, and
E^c_{AB} is the correlated error shared by measures A and B.

This formula can be transformed to provide predictions of the true relationship based on information about the observed relationship and measurement quality:

$$r'_{AB} = (r_{AB} - E^c_{AB})/V_A V_B. \quad (2)$$

A couple of examples will illustrate how the kinds of measurement quality estimates obtained in this study can be combined with information about an observed relationship to predict the true relationship between the underlying concepts. (1) Assume a correlation of .40 is observed between 2 measures that have estimated validities of .6 and .7, respectively, and that use different methods (hence we assume correlated error is 0). In this case, Formula 2 predicts a true relationship of .95, which is obviously much higher than the observed .40 relationship, which reflects the effects of random measurement error. (2) Assume a similar relationship, .40, is observed between 2 measures that have validities estimated at .93 and .95, respectively, and that use the same response scale and include method effects estimated at .36 (on the basis of which we assume correlated error is .13—which is .36 squared). In this case, Formula 2 predicts a true relationship of .31, which is somewhat lower than the observed relationship.

Multi-item scales. The estimates of validity and error components obtained in this study are for measures based on *single* survey items. Many survey analyses, however, use scales derived by combining several items. An important rationale for using such scales is that they usually have higher construct validity than single-item measures. (Depending on how a scale is constructed, it may also reflect higher method effects.) If one has an observed relationship involving one or more multi-item scales and one wishes to predict the underlying true relationship, the measurement quality of the scale(s) must be determined before Formula 2 can be used.

In the simple and common situation where a set of items, all of which are assumed to tap the same underlying construct and have about equal validities and method effects, are added together to form a scale, a standard psychometric formula can be used to predict the validity of the scale. Guilford's (1954) Formula 14.37 can be adapted for this purpose as follows:

$$V_S = V_I/[(1 - V_I^2 - M_I^2)/N + V_I^2 + M_I^2]^{1/2}, \quad (3)$$

where: V_S is the estimated construct validity of the scale,
V_I is the estimated construct validity of a single item,
M_I is the estimated method effects in a single item, and
N is the number of items in the scale.

This same formula can be adapted to provide a prediction of the method effects reflected in a scale:

$$M_S = M_I/[(1 - V_I^2 - M_I^2)/N + V_I^2 + M_I^2]^{1/2}, \quad (4)$$

where: M_S is the estimated method effects in the scale, and all other terms are as above.

Once one has obtained estimates of the construct validity and method effects for a scale, the residual error effect can be obtained by the following formula: $(1 - V_S^2 - M_S^2)^{1/2}$. This is based on the definition that residual error is what is left after validity and method effects have been taken into account.

The first several items combined to form a scale will produce the biggest enhancement in validity, but gains taper off with further items. For example, if 2 items meeting the assumptions above and with validities of .7 and methods effects of .1 were combined to form a scale, Formulas 3 and 4 would predict that the scale would have validity of .81, method effects of .12, and hence residual error of .57. With 3 items being combined, these values would be .86, .12, and .50, respectively.

Predicting the validity and method effects for a scale is more complicated if the scale is not constructed by a simple addition of the items, if the items have different validities, and/or if the items reflect different method effects. There is some discussion of such situations in the literature (e.g., Green and Carmines, 1979), but the problem is complex and not fully solved. The new technology of structural modeling with latent variables, on which we comment below, offers a particularly useful way of handling some of these situations.

Multivariate relationships. Just as measurement errors affect bivariate relationships, so also do they have impacts on multivariate statistics—and here the effects are often harder to sort out. One approach for getting true multivariate relationships, uncontaminated by the effects of measurement error, is to use the procedures presented above for obtaining predictions of the true bivariate relationships, and then use these relationships as input to the calculation of the multivariate statistics. Another approach is to use structural modeling with latent variables, which is briefly discussed next.

Comment on structural modeling with latent variables. Within the last decade there has developed a powerful new approach for estimating the true relationships, either bivariate or multivariate, among underlying concepts (i.e., among latent variables). This is the structural modeling technology implemented in computer programs such as LISREL (Bentler, 1980; Joreskog, 1978; Joreskog and Sorbom, 1978). If one had suitable data, one could use survey measures in such models and obtain useful estimates of underlying relationships without first making the corrections for measurement error

just described. This requires, in effect, doing two things at once: Estimating a set of measurement quality parameters and performing an analysis on the latent variables.

However, sometimes one may not have data that permit simultaneous solution of both problems. In such circumstances, prior information about the validity and error components of the measures—information that might have been obtained using the approaches described in this report—can be incorporated into the structured equation model and will let one proceed to an insightful analysis.

Summary

There is growing recognition that measurement errors can have profound effects on statistical relationships. Some kinds of measurement errors make a simple *bi*variate relationship appear stronger than it really is, others make it appear too weak; the effects of measurement errors on a *multi*variate relationship are both complex and substantial. To make better inferences about the "true" relationships among the concepts being studied, survey researchers need a more complete understanding of the error components of their measures, and this requires new and better methods for assessing data quality.

The research reported here addresses these goals by applying a new technology—structural measurement modeling of specially collected multimethod-multitrait survey data—to generate quality estimates for a heterogeneous set of survey measures. For each measure, three quality estimates are examined: (a) percentage of valid variance (based on estimates of construct validity), (b) percentage of method effects variance (a major source of correlated error), and (c) percentage of residual variance, mainly random measurement error. Over 2,000 sets of such estimates are generated for 106 survey measures assessed on a wide range of demographically defined respondent groups. The data come from methodological supplements included in five representative national surveys of American adults and one survey in a Canadian corporation—7,704 respondents in all.

It is important to note that these quality estimates are different from the estimates of measurement bias, certain kinds of correlated error attributable to interviewers or coders, nonresponse bias, and, of course, sampling error, that have been the focus of previous methodological research. The new quality estimates provide an important complement to these other indicators of data quality.

This study makes two kinds of empirical contributions: the basic descriptive summaries derived from these quality estimates are of considerable interest in their own right, for they provide the broadest assembly of such information available to date. According to these results, a "typical" survey measure—when administered by a professional survey organization to a general population sample—consisted of 50%–83% valid variance, 0%–7% method effects variance, and 14%–48% residual variance. (The ranges reported here include about two-thirds of all the quality estimates examined and reflect the interval extending one standard deviation on each side of the mean.) Note that these quality estimates refer to single-item survey measures, and that many survey researchers enhance measurement quality by combining single items to form scales or indices; given quality estimates for a specific set of items, the amount of such enhancement can be calculated. This paper presents formulas for making this calculation and also presents other formulas for "correcting" observed relationships for the effects of measurement error.

Of at least equal interest are the results from the multivariate analysis of the quality estimates themselves. Over two-thirds of the variance in each of the 3 quality estimates could be explained by taking account of 13 design characteristics of the survey in which the measure was implemented. These design characteristics focus on such practical matters as the number of categories in the response scale, whether a question was phrased in an absolute or comparative way, how far along in the interview the measure was located, whether the item was embedded in a battery with other similar items, the length of the introduction to the item and of the question, etc. The results from this anlaysis provide information on the design conditions that were associated with better (or worse) measurement quality and provide empirically based suggestions for ways to improve measurement quality in future surveys. These analyses also provide a statistical model that could be used to estimate the measurement quality of a wide range of *other* measures not included in this particular study. The practical use of this model is discussed.

In addition to these two types of empirical contributions, this study provides an example of the implementation, in regular on-going surveys, of a new technology that can be applied in a straightforward and economical manner in other surveys to generate estimates of measurement quality. Some advice on how to do this, based on the experience gained in this study, is included.

Appendix

This appendix details how the structural modeling of multimethod-multitrait data was actually implemented in one of the surveys. Because Survey #3 (see Table 1) had the smallest number of primary measures, it is the simplest to present and is the one illustrated here. Similar procedures were used for all surveys.

Table 1 shows that the multimethod-multitrait design in Survey #3 involved nine measures that tapped three concepts—frequency of drinking beer, of eating too much, and of watching TV to get away from the ordinary cares and problems of the day. Table 2 shows that the three response scales used for the multimethod-multi-

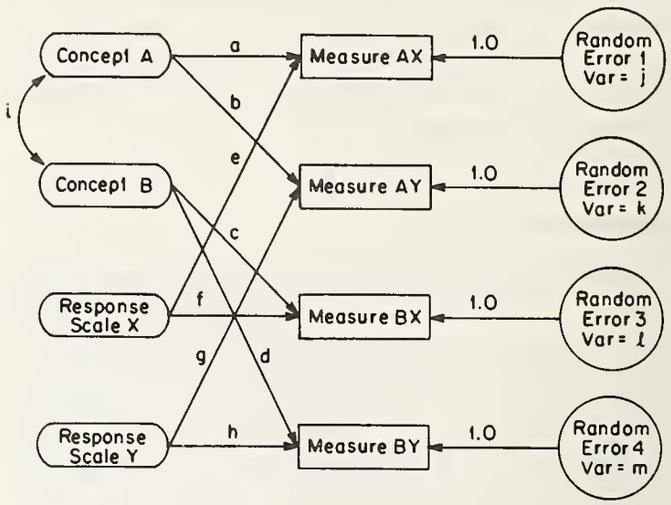
Table 7
The nine primary measures from Survey #3

RA9.	Now turning to things you eat and drink. Some people feel they eat too much. During the past month, how often do you feel you ate too much? Almost every day, every few days, once or twice, or not at all?
RA10.	During the past month were there more than ten days when you drank some beer? (Yes/No)
RB7.	During the past month were there more than four days when you watched TV just to get away from the ordinary cares and problems of the day? (Yes/No)
RB8.	As you know we're trying to get the most accurate information we can so I'd like to ask you about a few things we have already talked about. These may sound like the same questions, but they give you different answers to choose from. Please tell me how often each has been true for you over the past *month*.
RB8A.	During the past month *how often did you drink beer?* Almost every day, every few days, once or twice, or not at all?
RB8B.	During the past month were there more than two days when you *ate too much?* (Yes/No)
RB8C.	How often during the past month did you watch TV just to *get away from the ordinary cares and problems of the day?* Almost every day, every few days, once or twice, or not at all?
RB10.	Here are the last questions about things we asked earlier. On about how many days during the past month did you drink at least one glass of beer?
B10A.	On about how many days during the past month did you eat too much?
B10B.	On about how many days during the past month did you watch TV just to get away from the ordinary cares and problems of the day?

trait design in Survey #3 consisted of a yes/no format, a four-point frequency scale, and reports of the actual number of days. The actual items that implement the design are presented in this appendix in Table 7. These nine items were only a small part (less than 20%) of the total interview and were interspersed with other material.

Figure 3 in this appendix shows the structural model used to generate measurement quality parameters for the primary measures in Survey #3. Functionally, this model is identical to the simpler schematic model presented in Figure 2. (To make Figure 3 more readable, the response scale factors, which appeared on the left in Figure 2, have been moved to the right; and the random error factors, which appeared on the right in Figure 2, have been indicated by short diagonal arrows entering each measure, as is conventinal in path diagrams.) As noted in part 4 of this report, the strength of all linkages emanating from the same response scale factor were constrained to be equal, and each response scale factor was constrained to be statistically independent from all other factors.

Figure 3
Structural model used for Survey #3



The model shown in Figure 3 was run 28 times—once for all respondents together and additional times for various (sometimes overlapping) subgroups of respondents—e.g., males, females, young adults, older adults, etc. In each run, the strengths of the linkages connecting content factors to measures were used as estimates of construct validity, the strengths of the linkage connecting response scale factors to measures were used as estimates of method effects, and the square roots of the variances of the residual factors were used as estimates of the random error effects.

Footnotes

[1] The conceptualization of measurement quality used here is similar to that discussed by Heise and Bohrnstedt (1970) and by Zeller and Carmines (1980).

[2] This is an average of nine different correlations (ranging from .24 to .55) that assessed this relationship.

[3] As is common (but not universal) practice, in this study a measure's validity is expressed as the correlation between the measure and the underlying construct; hence the square of this figure will indicate the proportion of valid variance in the measure. (For example, a measure having a validity of .8 would consist of .64—i.e., 64%—valid variance.)

[4] The deviations that constitute random or correlated errors each average to zero across the respondents being examined. If deviations do not average to zero, it is possible to apportion these deviations into two components: A constant, which is the "bias" discussed previously—an important source of error in its own right, but which does not affect relationships and is not addressed in this investigation—and the remaining part, in which deviations do average to zero across respondents.

[5] This is not to deny, however, that discussions of social desirability effects, yea-saying effects, and the like have suggested that these *might* be sources of correlated error.

[6] We expected respondents would recognize topics they had discussed previously, so explanations such as the following were sometimes included: "This next section asks about some of the same things we have already talked about, but the questions are different. We are doing

research to find the best way to ask these questions. Just tell me which answer seems to fit your situation."

[7] Thus, in Figure 2, parameters e and f would be constrained to equal each other, and so also would parameters g and h; e and f, however, might differ from g and h.

[8] This same approach can be applied straightforwardly to Surveys 2, 5, and 6, which involve complete 3 × 8, 3 × 4, and 3 × 8 method-by-trait designs, respectively. Surveys 1 and 4 are complicated by the fact that the multimethod-multitrait data design was not complete—i.e., some concepts were not assessed by all methods.

[9] This examination of the congruence between observed and predicted relationships, as well as our general approach to the topic of model fit, is in accord with perspectives advocated by Bentler and Bonett (1980) in a recent article on assessing how well models fit data. Their article was not published until after all computing for this study had been completed, but had we been able to compute the fit indices they propose for each of the 125 models we ran, there is little doubt that the indices would show that our models account for most of what is going on in the data.

[10] One might ask why *any* unreasonable estimates were generated. Our guess is that they result from minor inconsistencies in the input correlations arising from some respondents being omitted from the calcuation of one correlation (because their data were missing on one or both of those particular variables) while different respondents were omitted from other correlations. Although the use of pairwise missing data deletion from correlation matrices has this drawback, it generally leads to less serious problems than what can occur with casewise deletion, in which a great many respondents can be lost from the entire matrix.

[11] Factor analysts sometimes refer to this as "unique true variance."

[12] Creation of this new data set required transcription of a large number of LISREL-produced measurement quality estimates and assembly of many descriptive characteristics for each survey item. These data were then punched, verified, and built into a computer data file. The reliability of this coding process was checked for a sample of the cases. The coding accuracy rate was found to be 99.4%, a level that was judged to be highly satisfactory.

[13] The measurement quality estimates are taken directly from the structural modeling analyses described in part 4. The validity and method effect estimates are the LISREL-produced lambda parameters, and the residual error estimate is the *square root* of the LISREL-produced theta parameter. The square root transformation of the theta parameter was used so all three measurement quality estimates presented in the tables that follow would be on the same scale—i.e., the *square* of each of them indicates the percentage of variance of the indicated kind in the survey measure. A preliminary check when data from only four surveys were available showed that transforming the LISREL-produced theta parameters by taking the square root would result in a more normally distributed variable for subsequent analysis (lower skew and lower kurtosis). Subsequently, however, the measurement quality estimates from the two other surveys led us to question whether the square root transformation would always produce a more normally distributed variable.

[14] The skew coefficients for the estimates of validity, method effects, and residual error are, respectively: −.48, .21, and −.77. The corresponding kurtosis coefficients are: .42, −.46, and 1.14.

[15] Judgments about content speci-ficity were made on a three-point scale by three project staff members working independently after an initial discussion of the nature of this judgment. Their initial agrement rate was 69%. This agreement rate is much better than the pure chance rate of 33%, but still is only moderately satisfactory. Although the notion of specificity seemed clear in the abstract, its actual application to the topics included in these surveys proved difficult in some instances. Where disagreements occurred, a final classification was agreed upon after discussing the reasons for disagreement.

[16] Salience classifications were derived in the same manner as the specificity classifications described in the previous footnote. The initial agree ment rate for salience was 74%.

[17] In the example just given, the original mean validity figure (.81), which implies that a measure consists of 66% valid variance ($=.81^2$), is 16% better than the 56% valid variance expected from a measure using a 2-point scale ($.75^2 = .56$). Using the formula $[v(1 - R^2)/N_i]^{1/2}$ to estimate the approximate standard errors of the multiple classification analysis coefficients (where v = the variance of the dependent variable, R = the multiple correlation coefficient, and N_i = the number of cases in the category) (Hill, 1979), one obtains standard errors ranging from about .002 for categories with large numbers of cases (800) to about .01 for categories with small numbers of cases (50). It follows that if one makes standard assumptions for computing statistical significance, a difference between almost any pair of coefficients in Table 5 that is .02 or larger is "significant" at or beyond the p = .05 level.

[18] Andrews et al. (1973) describe how interactions can be handled in a multiple classification analysis through the use of pattern variables.

[19] Two bits of evidence lead us to doubt this result. First, as can be seen in Table 5, the multivariate results here are markedly different from the bivariate results. This is unusual in these data (but not unprecedented—it also occurred for the absolute versus comparative predictor). In addition, however, category labeling has a rather strong relationship to another predictor, number of answer categories. (Cramer's V = .79, the strongest relationship between any pair of predictors included in the analysis.) It is possible that the surprisingly good quality associated with answer scales having 20+ categories (many of which consisted of reports of actual frequencies-per-month of doing various things and hence which count as "fully labeled") has interacted with the category labeling predictor so as to produce an *overestimate* of the quality of data from 20+ answer categories and a corresponding *underestimate* of the data from fully labeled categories. Unfortunately, the data are not sufficient to clarify this matter. While the analysis could have been rerun omitting the category labeling predictor, to do so would have been to hide a seeming anomaly that instead merits further investigation.

[20] The exception is an indication that face-to-face interviews result in a slight enhancement of method effects, but this is *not* accompanied by any reduction in validity for measures obtained in face-to-face interviews.

[21] These were: questionnaire length, relative position of item in questionnaire (e.g., half-way through), position of item in battery, relative position of item in battery, immediacy of the topic (whether the respondent would be expected to have direct personal experience), temporal stability of the topic (extent of fluctuation over time of the phenomenon addressed by the item), type of answer scale (verbal, numeric, pictorial/graphical), clarity of item wording, skewness of responses to the answer scale, topic addressed by the item, and actual answer scale used by the item.

[22] SEARCH is an analytic routine included in the OSIRIS IV software system. It is an updated version of the Automatic Interaction Detector (AID) program described by Morgan and Sonquist (1963) and Sonquist et al. (1974).

[23] The SEARCH runs mentioned here included all of the predictor variables listed in Table 4 plus two others (clarity of item and scale type) and explained the following proportions of variance: validity, .69; method effects, .71; residual error, .67.

[24] A separate paper (Andrews and Herzog, 1981) provides a detailed analysis of the relationship between survey data quality and respondent age.

[25] Using the same procedure for estimating standard errors as described in a previous footnote, one obtains standard errors of about .007 for categories with about 100 cases, about .010 for categories with 50 cases, and about .020 for categories with 10 cases. A note in the table itself describes how large differences between selected pairs of coefficients need to be in order to reach statistical significance.

[26] To accomplish these controls it was necessary to go back to the original survey data and compute some additional measurement quality estimates for respondents having certain *combinations* of demographic characteristics (e.g., whites who had not completed high school). These analyses were performed only for the two surveys with

the largest number of respondents.

[27] The items were selected on the basis of exploratory and confirmatory factor analyses applied to data from 2,000 American men that had been collected and analyzed for other purposes (Caplan et al., 1975, 1980). The 5 selected items all described characteristics of the repondent (e.g., "I am always courteous, even to people who are disagreeable") and were answered either "true" or "false."

[28] LISREL and COFAMM are distributed by National Educational Resources, P.O. Box 1025, Chicago, Illinois, USA, and are available at many major computing installations.

Effects of interviewer characteristics and interviewer variability on interview responses

Bengt Brorsson, Department of Social Medicine, Uppsala, Sweden

Introduction

Nathanson (1978) states in a comprehensive review of sex differences in health interviews that women, in general, report higher sickness rates and greater use of health services than do men. She also gives examples of factors in the data collection process that can contribute to the observed sex differences. One of these factors is assumed to be the sex of the interviewer. The author cites data indicating that female interviewers succeed in providing a more nearly complete report than male interviewers, even when the respondents are male.

The first part of this paper describes a study of effects on interview responses of interviewers' sex, age, length of time employed as an interviewer, and number of interviews carried out. Because a substantial effect of interviewers' sex on interview responses was found in this study, the question arose whether this was really the major source of variability among interviewers. Because of the great interest and importance of this question, a second study was carried out (Brorsson, 1980). The objective was to study variability in results between individual interviewers. The second part of this paper describes an interviewer variance study designed to measure the contributions of interviewers to the variability of health statistics.

Significant between-interviewer variance in the reporting of health data has been observed in a number of studies. Data from other studies also indicate that interviewer effects may operate differently for different statistics.

Material

The data were collected by the Survey of Living Conditions (SLC), a continuous interview survey carried out by the National Central Bureau of Statistics (NCBS) since the autumn of 1974. During 1975 the survey covered the five welfare components—health, employment, housing, education, and finance. The Survey of Living Conditions is nationwide and covers the Swedish population aged 16–74. The sample size is about 10,000 individuals per annum based on simple random sampling of individuals. In cases where the sample individual lives with another individual aged 16–74 (mainly married people), both are interviewed. The sampling probability is therefore doubled for co-habitating adults. The country is divided into interviewer districts generally having about 20,000 inhabitants aged 16–74. The NCBS tries to have one interviewer per district, carrying out all types of interviews within his/her district.

The data for the present studies were collected during 1975. The material consisted of 264 interviewers who together carried out 10,026 inerviews, an average of 38 interviews. Some 1,500 interviews conducted by telephone were excluded. Among the 264 interviewers, 32 were men and 232 women. Results obtained for 54 central variables in the health portion of the questionnaire were used for comparisons.

The interviewers' mean age was about 50 years. Each male interviewer carried about 35 interviews and female interviewers about 38 interviews. Generally they had been employed by the NCBS for quite some time; a quarter of them had worked for the organization for at least 10 years, another quarter for two years or less. There were no differences between male and female interviewers in age or interviewing experience.

Each SLC interview takes about 60 minutes to complete. The health portion of the Questionnaire comes second, after questions about housing. The 54 variables used for this study fall into the following areas:

Longstanding illnesses and their consequences—6 variables
Prevalence of the most common longstanding illnesses—6 variables
Functional capacity in the areas of vision, hearing and mobility—7 variables
Incidence of acute illnesses and accidents and psychic well-being—7 variables
Use of medicines—11 variables
Use of medical care services—10 variables
Other questions related to health—5 variables
State of dental health and use of dental care services—2 variables

As must be clear from the above, the actual studies rest on a secondary analysis of data which originally were gathered for the purpose of describing health status and use of health care services in the Swedish population. For the current investigation the respondents ideally should have been randomly distributed among interviewers. Through such random distribution, the underlying true value is assumed to be the same for different interviewers. With such a procedure, statistically significant differences between the interviewers' results could be interpreted as indicative of an interviewer effect. However, it is physically impossible to distribute respondents randomly among interviewers in large nationwide investigations. An investigation could, however, have been conducted within a defined area

such as a large town. The number of interviewers taking part must then as a rule be small. Likewise, there is a risk that local conditions may influence the results, which limits the ability to make generalizations.

For the present studies it was necessary to neutralize the effects of differences in age and sex composition among respondents who were interviewed by different interviewers. This was achieved by indirect standardization. A standardized difference was then defined as the difference betweeen observed minus expected relative frequency. Standardized differences were computed for every interviewer for all 54 variables. These standardized differences were then used throughout the analyses.

Effects of interviewers' sex, age, length of employment, and number of completed interviews on interview responses

The first study was limited to comparisons of the results obtained by groups of interviewers differing in sex, age, length of time employed as an interviewer, and number of interviews completed for the SCL investigation in 1975. No other data were available concerning interviewers employed by the NCBS. For the variables age, length of time as an interviewer, and number of completed interviews, the comparisons were done in such a way that the results obtained by the lower and upper quartiles were contrasted; for example the 25 percent oldest and 25 percent youngest were compared.

Method. The statistic used for the test of differences was (approximately t-distributed if no true difference exists):

$$t = \frac{\bar{a} - \bar{b}}{[(S_a^2/n_1) + (S_b^2/n_2)]^{1/2}}$$

Comparisons of results with respect to age, length of time employed as an interviewer, and number of completed interviews were done by contrasting the upper and the lower quartiles.

In order further to scrutinize differences in results between male and female interviewers, each of the 32 male interviewers was matched with a female interviewer. This matching was carried out independently by the NCBS interviewer staff. The criteria were that the female interviewer should have been working in an adjacent, structurally comparable district and if possible have the same interviewing experience. An interviewer index was created to investigate further the differences between male and female interviewers. The interviewer index summarized the results obtained by individual interviewers. First, however, those variables which were highly intercorrelated ($r \geq 0.70$) were omitted. After this, 45 variables remained. To get comparable results from different variables, a further standardization was carried out. For each interviewer, each variable was multiplied by a weight which consisted of the inverse value of the mean error for each variable as it had been observed among all interviewers.

Results. The differences observed between male and female interviewers were statistically significant ($p<0.05$) for 18 of the 54 variables. For all of these variables male interviewers recorded smaller proportions and amounts of illness conditions and use of health services than female interviewers. The size of the differences is illustrated in Table 1, which shows results obtained by male and female interviewers.

In Table 2, results from the comparison between male and female interviewers are summarized.

The results from the matched cases are almost idential with the results presented in Table 2.

As can also be seen, statistically significant differences seemed to occur irrespective of the subject of the question, with one exception. When the interviewer index was applied, a statistically significant difference was found ($p<0.01$). This significance level is surprisingly low, given the above results. The explanation probably lies in the fact that relatively large correlations exist between many of the variables included in the index.

Five statistically significant differences were found when the results obtained by older interviewers were

Table 1
Longstanding illnesses (LSI) and their consequences by sex of interviewers and respondents (percent)

	Respondents reporting:			
	LSI	*consultation with e doctor within 3 months beceuse of LSI*	*severe suffering beceuse of LSI*	*serious reduction in working cepecity beceuse of LSI*
Male interviewers (n = 32)	35.3	17.0	12.0	9.5
Resp. male	38.2	16.4	11.9	9.9
Resp. female	32.2	17.7	12.0	9.1
Female interviewers (n = 232)	40.0	20.5	14.4	10.3
Resp. male	38.9	18.9	13.2	10.9
Resp. female	41.2	22.2	15.5	9.7
Value of t	−1.46	−1.97	−2.50	−0.57

Table 2
Cumulative distribution of 54 t-values from the comparison of results obtained by male and female interviewers

	t-values				
	<−3.29 (p<0.001)	*<−2.58 (p<0.01)*	*<−1.96 (p<0.05)*	*<0.0*	*<1.96*
LSI and their consequences		1	4	6	6
Prevalence of the most common LSI		1	4	5	6
Functional capacities				1	7
Incidence of acute illnesses and accidents			2	5	7
Use of medicines		4	5	10	11
Use of medical care services	1	2	2	8	10
Other questions related to health	1	1	1	4	5
State of dental health and use of dental services				1	2
All	2	9	18	40	54

compared with those obtained by younger interviewers. Four of these indicate that older interviewers obtained smaller proportions and amounts of illness conditions than younger interviewers. Five statistically significant differences were observed when results were analyzed by length of time employed as an interviewer. For all these variables, the interviewers employed for the shortest time reported larger proportions than the interviewers with the longest service. Comparisons by number of completed interviews produced six statistically significant differences. Four of these indicate that those interviewers who had completed fewer interviews provided a larger proportion of reports of ill health and use of health care services than those who had completed a large number of interviews. The number of statistically significant differences found when results were compared among interviewers grouped according to their age, time employed as an interviewer, and number of completed interviews was so small that these differences may well have occurred as a result of chance alone.

Discussion

Nathanson states in a comprehensive review of sex differences in health interviews that women, in general, report higher sickness rates and a greater use of health care services than do men. The present report confirms that female respondents report a worse health status and a higher use of health services than do men, irrespective of the interviewers' sex. This tendency is, nevertheless, more pronounced when the interviewer is female. The differences in results obtained by male compared with female interviewers follow the same pattern and are in many cases even greater than the differences between male and female respondents.

The conclusions from the first study were that the sex of the interviewer affected responses to a large number of the variables in the study, and that these differences could not be accounted for by referring to the content of the questions. This conclusion also seems probable against the background of prevailing sex patterns. Questions associated with health can be assumed to be more engaging and meaningful to female than to male interviewers.

What are the practical consequences of this lower reporting obtained among male interviewers? Let us take an example from the present LSI investigation. Male interviewers had completed about 1,000 interviews and female interviewers about 9,000. The results obtained by the male interviewers were more often than not some 20% lower than the corresponding results obtained by the female interviewers. Then, if in one question 40% of the female interviewers' respondents reported a certain state of health, male interviewers could be expected to obtain 32% among their respondents. If the interviewers' sex is not taken into account, a jointly weighted estimate of the actual state of health of 39.2% will result, i.e. 0.8% lower than what would have been reported if all of the interviewers had been women. This difference appears at first to be of a trivial size. However, according to the results published by the NCBS for the material that is discussed here, the upper limit for the size of the 95% confidence interval for 10,000 respondents is 1.2%. So there is a danger that if the sex distribution of the interviewers changes from one year to another and other measures are taken to improve results, the joint effect may be so large as to risk the occurrence of a statistically significant difference wrongly being interpreted as reflecting a change in health.

Interviewer variability study

As already noted, the above analysis is rough and limited because only a few and, in this context, perhaps less important variables have been studied. Against this background the question arose whether these variables really constituted the major source of variations between interviewers. Because of the great interest and importance of this question, a second study was carried out. The objective was to study variability in results between individual interviewers. Thus, an analysis of variance was employed which in this case was a one-way hierarchical nested classification with three levels: regions, interviewers, and respondents.

Method. The fact that respondents were not randomly distributed by interviewers is here—more than in the study reported earlier—a weakness. The variance that can be estimated between interviewers consists of a true interviewer effect and effects of local conditions that may exist between different interviewer districts. It is therefore important—as far as possible—to sort out the

proper interviewer effect from characteristics that may be inherent in different interviewer districts. To be able to do this a regional division was used. An existing standard based on the degree of urbanization of Swedish communes was used for this purpose. The standard contains six classes. Applying this standard does not, however, solve all problems. Further variations may still exist between interviewer districts which cannot be controlled for.

As in the study just reported the results obtained by different interviewers were standardized for differences in the age and sex distributions among their respondents. The analysis was carried out by means of a one-way hierarchical analysis of variance. The choice of a hierarchical classification rests on the necessity to consider that sampling had been done stepwise on three levels. In this study regions and interviewers are given. The model is thus one with fixed effects where the samples of regions and of interviewers are fixed and only the sample of respondents is at random. Further, the number of interviewers varies between regions and the number of respondents varies between interviewers. Snedecor and Cochran (1967) have suggested a way to handle such data with analysis of variance. When computing components of variance it was, however, assumed that both respondents and interviewers could be considered as small random samples from large populations, from which follows that an analysis of variance of model II was applied.

The estimated values of the MSS, which are common to results from all variables, are in Table 3.

Table 3
Estimated values of the MSS

Source of variation	*Degrees of freedom*	*Expected values of MSS*
Regions	5	$\sigma_0^2 + 45.2\sigma_1^2 + 1{,}582\sigma_2^2$
Interviewers	258	$\sigma_0^2 + 37.8\sigma_1^2$
Respondents	9,762	σ_0^2

F-tests are approximate because the assumptions concerning normal distribution and equal dispersion for all individuals are not complied with. Also note that the constant for σ_1^2 is larger for regions than for interviewers. The F-test is, however, approximate for ratios of coefficients which are near unity (Tietjen, 1974).

Results. Table 4 shows sums of mean-squares and components of variance for those 14 variables for which regional results were easily available.

The results show a statistically significant strong variability ($p<0.001$) between interviewers for 13 of the 14 variables included in the study. The results from the F-test of regions against interviewers show that 12 of the 14 F-tests are *not* statistically significant.

One possible explanation of these results is that some interviewers work during the daytime while others carry out most of their interviewing during late evenings and nights. Those interviewers working during the daytime could thus be expected to do more interviews with respondents who are ill. In order to test the plausibility of this suggestion, a new analysis of variance was undertaken in which the regional division was replaced by a binary classification of employment, i.e., under the extreme assumption that individual interviewers had had either only employed or only unemployed respondents. A statistically significant strong variability ($p<0.001$) between interviewers was found for all 14 variables when this assumption was put to a test. A comparison of the estimated components of variance showed that they were of almost indentical size in the two analyses.

Table 4
Analysis-of-variance according to the scheme for hierarchical classification. Sums of mean squares, estimated components of variance and results of F-tests for interviewers against respondents and of regions against interviewers

	$\overline{RSS}$	$\overline{ISS}$	$\overline{ESS}^a = S_0^2$	S_1^2 *(Interviewers)*	S_2^2 *(Regions)*
Longstanding illness	1.330	0.592	0.323	0.0071^{xxx}	
Reduced working capacity	0.238	0.203	0.135	0.0018^{xxx}	
Reduced mobility	0.178	0.250	0.163	0.0023^{xxx}	
Incidence of acute illness	0.076	0.156	0.112	0.0012^{xxx}	
Emergency room visits	0.124	0.213	0.109	0.0028^{xxx}	
Ordered visits to doctors	0.896	0.485	0.281	0.0054^{xxx}	
All doctors' consultations	2.132	0.620	0.376	0.0065^{xxx}	0.0009^{x}
Have a family doctor	2.020	0.951	0.361	0.0156^{xxx}	
Hospitalized	0.082	0.084	0.061	0.0006^{xxx}	
Use of anodyne	0.108	0.316	0.119	0.0052^{xxx}	
Use of tranquillizers	0.240	0.124	0.081	0.0011^{xxx}	
All medicines used	0.612	0.649	0.351	0.0079^{xxx}	
Visits to dentists	2.158	0.386	0.277	0.0029$^{N.S.}$	0.0011^{xxx}
Difficulties chewing	0.118	0.300	0.109	0.0051^{xxx}	

[a]The number of degrees of freedom have been reduced with a factor of 1.69 to compensate for the fact that the sampling probability for cohabitants is doubled (see Introduction).

Discussion. The results obtained from the study of variability among individual interviewers are surprising. As stated earlier, these studies rest on a secondary analysis of data that were orginally collected for the purpose of describing health status and use of health care services in the Swedish population. Ideally, respondents should have been randomly distributed among interviewers. Therefore, it may be premature to conclude that characteristics of interviewers contribute disproportionately to the variability in results from health interview surveys. It may very well be that one can only generalize from these results to a limited extent, and that to arrive at a definitive and convincing answer, a new study must be undertaken in which respondents are randomly distributed among interviewers. If they are confirmed, these results have wide implications for the usefulness of data on

health status and health services use derived from the SLC investigations for monitoring health in the community.

Of further concern in the present survey is the actual contribution of the interviewers to the total sampling variance of an observed mean. For arithmetic means, the variance can be expressed as follows (Horvitz, 1952) (disregarding sources of variability on higher levels and assuming that all interviewers have the same number of respondents):

$$\sigma^2_{mean} = \frac{\sigma^2_{respondents}}{N} + \frac{\sigma^2_{interviewers}}{K}$$

$$= \frac{1}{N}\left[\sigma^2_{respondents} + \sigma^2_{interviewers} \times \frac{N}{K}\right]$$

where N = number of respondents
K = number of interviewers

Table 5 illustrates the size of the total variance when considering the variance observed among respondents and among interviewers simultaneously and allowing N/K to vary from 1 to 100 interviews. The variable "longstanding illness" is used in this example. Estimates of $\sigma^2_{respondents} + \sigma^2_{interviewers}$ are obtained from $S_0^2 + S_1^2$ in Table 4, which for the variable "longstanding illness" are 0.323 + 0.0071.

The average interviewer carried out 38 interviews during 1975. As can be seen from Table 5, the sum of the variances between respondents and between interviewers when N/K = 38 is 0.59. The interviewer contribution is in this case 45%. This clearly demonstrates that the more interviews an interviewer performs the more pronounced the so-called correlated enumerator variance will become. The effect is therefore maximized in regional comparisons, where the findings of individual interviewers are all classified in a specific region. The effects ascertained are so great that the differences in health status and use of medical care, which appear in comparisons of, e.g., the county councils, might well be artifacts caused by the interviewers' nonidentical ways of working. For similar reasons the differences in working methods make it very difficult indeed to trace changes over time. However, the SLC surveys are used in many contexts where the interviewer effect is of minor importance. This pertains, for example, to the distributions by occupational category, educational level, and income. In these subgroups any interviewer will have performed only a limited number of interviews.

We conclude that the SLC surveys cannot at present be used extensively for some of the purposes for which they were started, for example, describing regional differences or monitoring changes in health status and health services use. However, in this as in so many other situations, it should be remembered that the "best must not be the enemy of good." The indicated shortcomings ought to become the objects of detailed analyses in order to find methods for solving the problems. This process has, in fact, already begun.

Table 5
Sum of variances between respondents and interviewers when the number of respondents per interviewer rises

Number of respondents per interviewer N/K	*Variance for mean Longstanding illness*
1	0.33/N
10	0.39/N
20	0.47/N
38	0.59/N
50	0.68/N
100	1.03/N

Discussion: The construct validity and error components of survey measures *and* Effects of interviewer characteristics and interviewer variability on interview responses

Eleanor Singer, Center for the Social Sciences, Columbia University

If what distinguishes social science from more impressionistic observations of human affairs is the quest for unifying principles to bring order out of chaos, then today marks one of those special, happy, and all too rare occasions when significant progress is made toward that goal. For what Frank Andrews has described in his paper is an elegant way of synthesizing a great deal of the research that has been done on methods effects and of providing an organizing framework within which much future work can be located.

Throughout, Andrews describes his work as a "technological" advance, and I suppose that is correct: the method used here for estimating construct validity was proposed some time ago and the statistical model for partitioning variance was, too. What was needed was the conjunction of powerful new methods of analyzing the data derived from a multi-trait multi-method matrix with the resources provided by the National Science Foundation for collecting the supplemental data. But I think we should give some credit to Andrews for seeing the splendid possibilities.

The topics of the two papers given so far in this session fall into the general area of "nonsampling error," which has replaced sampling error at the forefront of methodological concerns. Validity, reliability, and bias are key concepts of this new frontier. But, as is not uncommon, other concepts abound, as well: response variability, nonresponse bias, response error, to name just a few; and one of the needs for progress in this area, I think, is the development of a common set of concepts and terms in which discussion can take place.

In the past, researchers have relied on information obtained by means other than surveys to provide validating information—a criterion against which the magnitude of nonsampling error, and conversely the validity of the information obtained by surveys, could be assessed. Andersen, Kasper, and Frankel, for example, use this approach in their landmark investigation of *Total Survey Error.* The charm of Andrews's approach is that *it can be applied to data for which no external criterion of validity exists.* It is, thus, ideally suited for assessing error in the measurement of subjective phenomena, although it can also be used to assess errors in reporting about the external world. In fact, one of the things I was curious about as I read Andrews's paper was whether there was any difference in methods effects on the two types of items. But the relation of Andrews's model of "correlated and residual error" to concepts of "total survey error" remains to be clarified and is beyond the scope of this discussion (and, I might add, of this discussant!).

What Andrews has done is, first, to estimate values for the construct validity, correlated error, and random error components of more than 100 survey measures, by applying structural modeling techniques to data derived from a multi-method, multi-trait design; and, second, to investigate the effect of a large number of survey design elements on each of these components. I will not attempt to comment on the techniques involved in doing all this. Instead, I will talk about some of the implications of the findings and some possible ways of extending them.

As I read the paper, apart from the sheer excitement of the undertaking as a whole, I was struck by two things: (1) that the total amount of variance accounted for by methods effects, as these are conceptualized and measured in this study, is relatively small, and (2) that the proportion of variance accounted for by elements of survey design is very large.

I will start by talking about the implications of the second finding and then go back to speculate about the first. I'll conclude with some comments about the second paper, on interviewer effects.

Andrews's analysis of the role of survey design in correlated error begins with the estimates of valid variance, correlated error variance, and random error variance which are derived from the structural model, and then attempts to *predict* these estimated values from elements of survey design—for example, whether a D.K. answer category is included, whether respondents were interviewed by phone or in person, whether respondents were reached on the first attempt or required several callbacks, whether the introductions to the questions were long or short, whether the item was included in a battery of similar items, and so on. As I read them off, I'm sure you all recognize them as having been subjected to extensive and often inconclusive research on methods effects. The beauty of Andrews's analysis is that it can tell us, quite precisely, *which* of these elements in fact make a difference, so far as correlated error is concerned, how *much* difference they made in this investigation, and furthermore, *which values* of the variables are "better," so far as improving construct validity and reducing error variance are concerned.

It is comforting to know, for example, that such elements of survey design as mode of administration, number of callbacks required to reach a respondent, and various aspects of the content of the item have relatively

small effects on estimates of correlated error. It is perhaps even more important to know that long batteries of items, the failure to provide an explicit D.K. category, or the failure to provide a substantial number of answer categories (although one suspects that in this last case open-ended responses may be responsible for the predicted increase in validity) all increase methods effects, increase random error, and reduce validity.

It is for me especially interesting to learn that a comparative phrasing of questions—e.g., Is your health better (or worse) than it was a year ago, or better (or worse) compared to other people your age—yields more valid information than a question which provides no reference point for the respondent (e.g., Would you say your health, in general, is excellent, good, fair, or poor). This indicates that people need the anchor provided by a comparative frame of reference in order to be able to answer questions of this sort with *precision* and *reliability*. I would like to remind you that the early theory and research on social comparison processes by Festinger and his students, and on level of aspiration by Kurt Lewin and his students, predicted exactly these effects. Although I have not done so, I wonder whether thefindings reported here might not be fruitfully linked with various social psychological theories to provide a more systematic grounding for the methods effects observed.

It is, I think, fair to say that none of the specific findings about the effects of design elements contradict findings from prior research. In that sense, they are not new. But they are grounded on a wider sampling of survey contexts and survey items, and they yield more precise estimates that can be compared across items and across survey designs. Nevertheless, a few caveats are in order.

One of these, of course, is the standard caution that the findings derive from a limited sample of items and of survey design elements, but this caveat does not worry me unduly. Compared with most prior research on methods effects, which is based on a few items in a single survey, Andrews's analysis is indeed rich and diverse. Nor should we be more than normally skeptical about those hefty R^2's, which have been swelled somewhat by being fitted to a particular set of data (although also preshrunk). What I am, rather, concerned about is the temptation to *overgeneralize* the findings.

1. Andrews states, for example, that mode of administration leads to minimal *correlated error* effects. But this does not mean that mode of administration has no effects on *total* survey error. Response rates, for example, tend to be somewhat lower on the phone than in person, and some people have no phone. How does this model deal with these types of errors?

2. More generally, bias—a source of measurement error which does affect estimates of central tendency even though relationships among variables remain unaffected—was not estimated in this investigation at all. But researchers are often interested both in measures of central tendency and in relationships among variables, so that we need ways of evaluating a method with respect to both types of error. Can we minimize both simultaneously or must we trade one against another?

This brings me to the second observation—namely that while survey design variables account for a large fraction of the estimated variance of the quality estimates in this analysis, the amount of variance associated with methods effects, as these are conceptualized in this study, is very small. If it were small only in an absolute sense, that would be a cause for rejoicing. But it is also small relative to what is estimated in the model as "random error," which means that there is room for a good deal of improvement, so far as validity is concerned.

Now the question is, Why are measured methods effects as small as they are in this analysis relative to the effect of random measurement error? One possible explanation, it seems to me, is the similarity of the methods entering into the matrix. The fact that questions about number of doctor visits and amount spent for health care are asked with a long or a short introduction constitutes a variation in method, to be sure, but a far less drastic variation than questioning vs. check auditing, for example. If it were possible to get at the information we ordinarily ask about in interviews by such disparate methods, what would happen to the relative size of the validity, correlated error, and random error components? My guess is that validity would remain relatively unchanged, but the share of error attributable to methods would increase and that attributable to residual error would decrease.

And then what? What, really, do we mean by methods effects? Do we mean that amount of error attributable to the way in which we ask a set of questions, or do we mean the amount of error attributable to the fact that we obtain certain information by means of questioning rather than in some other way?

In some sense, it seems to me that what one would really like to reduce in social research is residual error—variance that cannot be accounted for. Andrews has broken out one source of error from the residual error component—namely correlated errors associated with different ways of asking questions. Brorsson has looked, instead, at another source of error—namely that associated with different interviewers asking the questions, although in principle some interviewer effects could also be examined by means of the same structural modeling techniques used by Andrews.

Brorsson reports on two investigations. The first evaluates the effect of certain interviewer characteristics on responses to questions about health. Of these, only interviewer sex consistently affected responses, and although the effect is small, it is large enough in this particular sample potentially to jeopardize conclusions about changes in the health status of the population. The other interviewer characteristics investigated—age, experience, and number of interviews completed—signifi-

cantly affected only a small number of the responses to the questions about health, and Brorsson concludes that these effects could have occurred by chance. His second study, like other investigations of interviewer effects (e.g., Sudman et al., 1977, and Tucker, forthcoming), indicates that the contribution of variation among interviewers to the variance of a given item is relatively small, but that it cumulates to unacceptably high levels as the number of interviews completed by any one interviewer increases, and especially when the comparisons one is interested in—in Brorsson's case, regional variations—are confounded with interviewer variations.

The limitation of Brorsson's study, which he himself acknowledges, is that interviewers were not assigned to respondents at random, necessitating after-the-fact controls. This dilemma has traditionally plagued research on interviewer effects. Either one restricts the investigation to one city and interpenetrates interviewer assignments, in which case the number of interviewers is generally very small and one can't be sure that the findings are generalizable to other settings; or one studies interviewer effects in a naturally occurring setting, in which case one risks confounding interviewer effects with geographic and associated variations.

The proliferation of telephone interviewing, and especially of computer-assisted telephone interviewing, means that truly experimental investigations of interviewer effects are increasingly within the reach of survey researchers, though even here true randomization of interviewing assigments is difficult because not all interviewers can work at all hours. (For a discussion of the practical effects of this limitation, see Singer and Frankel, 1982, and Tucker, 1983.)

In my own recent study of the effect of survey introductions of response (Singer et al., 1983), interviewers, though few in number, were randomly assigned to respondents. My results agree with those of Brorsson in finding little, if any, consistent or significant effect of interviewer age or experience on response. Sex of interviewer was held constant in my study; only female interviewers were used. Education, on the other hand, was significantly related to item non-response, with those interviewers having more education obtaining the lowest item nonresponse rates. Furthermore, age, experience, and number of interviews completed all were significantly related to achieved response rates on the survey. The youngest interviewers had the lowest response rates, those with larger interviewing assignments had lower response rates, and the relationship of response rate to experience was curvilinear, being highest among those with a year's experience and lower both among those with less and among those with more experience. In addition, interviewers' expectations of the ease of obtaining an interview were strongly predictive of the actual response rates they obtained, varying from 60% to 78% among those with the most pessimistic and the most optimistic expectations.

All of these findings suggest that interviewer effects, relatively neglected of late, deserve more systematic attention than they have yet received. Even more important, the time is ripe for a new consideration of the interrelations among these various sources of survey error. Such an undertaking will require (1) agreement in a uniform terminology, and (2) agreement on a model incorporating these various components of total survey error, which in turn will permit their estimation.

Methodological issues in the measurement of health policy outcomes*

Phillip R. Kletke, Department of Health Systems Research and Development, American Academy of Pediatrics

Stephen M. Davidson, Center for Health Services and Policy Research, Northwestern University

Janet D. Perloff, Department of Health Systems Research and Development, American Academy of Pediatrics

Donald W. Schiff, American Academy of Pediatrics

John P. Connelly, Department of Health Systems Research and Development, American Academy of Pediatrics

The reliability of respondent answers is a continual worry for survey researchers. Threats to reliability include lapses of memory and inaccurate estimates, not to mention willful distortion. In this paper we present two measures of the same phenomenon, physician participation in Medicaid, and discuss their reliability and relative utility in light of our findings. Previous studies have relied on physicians' self-reported estimates of the extent of their Medicaid participation (Held et al., 1978; Sloan et al., 1978) and have tacitly assumed that the doctors' estimates accurately measured the true extent of their participation. In the research reported here we examine this assumption by comparing two different measures of participation for the same physicians. Obviously this issue has important implications, since the results of previous studies must be questioned if it is found that doctors' self-reported estimates do not measure their Medicaid participation reliably. Moreover, this issue is of more than academic interest since public policy recommendations increasingly are based on the findings of social science research.

Survey and data collection methods

The data presented in this analysis are from the Survey of Pediatrician Participation in Medicaid conducted by the National Opinion Research Center in 1979 and 1980 under the direction of the American Academy of Pediatrics.[1] Original data were obtained from personal interviews with physicians as well as from encounter forms completed for samples of their patient visits. Thus, we have alternative sources of information on Medicaid participation.

A three-stage sampling plan was used to collect these data. In the first stage a sample of thirteen states was drawn using a method designed to maximize variation in state Medicaid policies. In the second stage the Physician Masterfile of the American Medical Association was used to draw a random sample of nonfederal, office-based pediatricians in each of the study states. A total of 1,457 physicians were included in the original sample, but in telephone-administered screening interviews, only 879 physicians were found to be eligible for the survey.[2] Of the eligible pediatricians, 814 participated in the personal interview, yielding a response rate of 93%. In the third stage, samples of patient visits were selected (the methods of selection will be discussed in detail below), and the doctors were asked to complete a one-page patient record on each of the approximately 35 patient visits. The patient record was a 16-item form which asked for information on various aspects of the patient visit, including the expected source of payment. This part of the sampling methodology was adapted from the National Ambulatory Medical Care Survey (U.S. Department of Health, Education, and Welfare, 1974). A total of 710 pediatricians completed patient records for a response rate of 81%.

Alternative indices of Medicaid participation

Using these methods, we obtained two measures of the physicians' participation in Medicaid. First, in the interview the physicians were asked to estimate the percentage of their patients whose care was paid for by Medicaid. We refer to the responses to this question as the physicians' self-reported estimates of Medicaid participation. Second, the doctors were asked to indicate on each patient record form the expected source of payment for the patient visit. By aggregating the patient records for each physician, we were able to calculate the proportion of patients in his sample for whom Medicaid was expected to pay, which we have called the physician's behavioral estimate of Medicaid participation.[4] Data for both self-reported and behavioral measures were obtained for each of 660 pediatricians.

Neither of these indices is a perfect measure of a physician's Medicaid participation. The self-reported estimate was subject to error when the physician did not have accurate knowledge about the source of payment for his patients. Various factors may have caused the doctor to have a false impression of the extent of Medicaid participation. For instance, one group of patients may have stood out in the physician's mind relative to others and caused him to overestimate their true pres-

* This work was prepared under Grant #18-P-97159/5 from the Health Care Financing Administration, Department of Health and Human Services, to the American Academy of Pediatrics.

ence in his practice. Thus, if a physician's Medicaid patients had more complex clinical problems or were more difficult to communicate with, it may have seemed to him that he had more Medicaid patients than he actually did. For similar reasons, doctors may have been inclined to overestimate their Medicaid participation if they found it especially difficult to complete Medicaid claims or if they found Medicaid to take a relatively longer period to make payments.

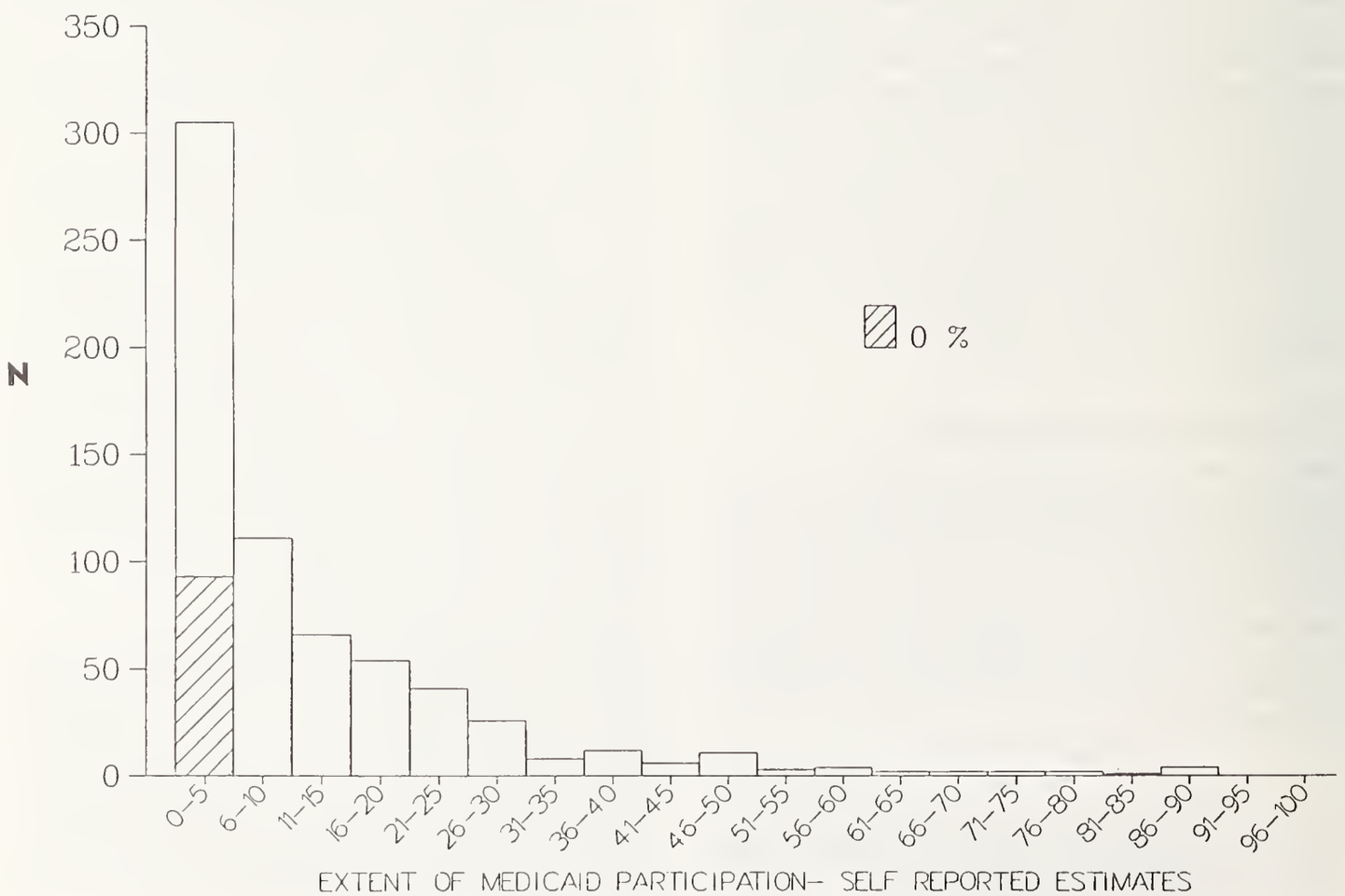
The behavioral index, which is based on aggregated patient record data, is subject to sampling error. Because of chance variation in the selection of visits for which the doctor completed patient records, the sample of patients may not have been representative of the doctor's practice. As a result, the estimate of Medicaid participation based on the aggregated patient record data may not equal the doctor's true rate of participation. However, since the index is an unbiased estimator, the more patient records a doctor completed, the more accurate this measure of Medicaid participation will be. (This assumes that the selection of patients for whom patient records were completed was random, an assumption to be analyzed in more detail below.) The amount of sampling error for an individual doctor might be large if he filled out only a few patient records, but it would be relatively small if he completed many of them. Moreover, when we take an average of the index for all 660 doctors, the separate sampling errors tend to cancel each other out. While the estimates for individual physicians may be substantially higher or lower than the actual extent of participation, the average value of the index for all 660 doctors should be very close to the average of their true extent of Medicaid participation (assuming that there was no systematic bias in the way the doctors' samples of patients were chosen.)

If doctors did in fact have accurate knowledge of their participation in Medicaid, the self-reported and the behavioral indices of Medicaid participation would be approximately equal. In other words, the accuracy of the doctors' perceptions about their Medicaid participation can be determined by comparing these two indices.

Descriptive data for the two indices

Examination of the two indices of Medicaid participation shows that, in fact, doctors tended to overstate their participation in Medicaid. The average value of their self-reported estimates (SR) is 13.0%, whereas the average value of the behavioral estimates (B) is 7.7% or only 60% of the self-reports for the 660 doctors for whom

Figure 1
Histogram for doctors' self-reported estimates for their extent of Medicaid participation (SR)

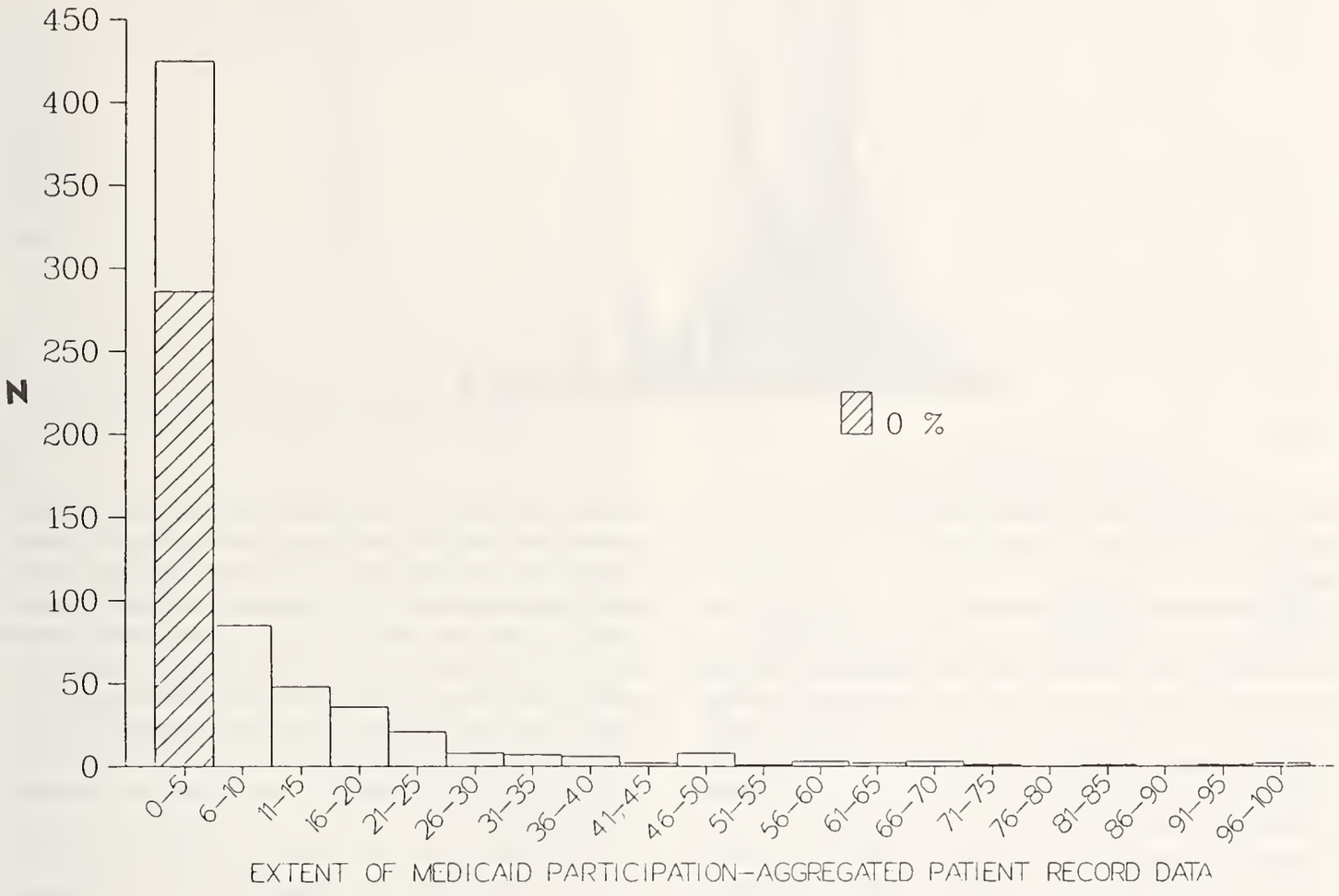


there are valid data for both indices.

Figure 1 is a histogram showing the number of pediatricians by their self-reported estimates of participation; and Figure 2 is a histogram for the behavioral estimates. Both have very skewed distributions, but the index based on the patient record data (B) is skewed to a greater degree because more doctors are concentrated in the 0%–5% category.

reported estimates exceeded the behavioral index. Figure 3, a histogram for DIFF, shows a distribution ranging from -40 to +70 with the greatest concentration in the values above zero. DIFF was positive for 78% of the sample and negative for only 17%, showing that the vast majority of physicians overstated their Medicaid participation. The average value of DIFF was 6.2% and the median value was 5%, indicating that half of the physi-

Figure 2
Histogram for doctors' extent of participation as measured by aggregated patient record data (B)



According to their self-reported estimates, 93 of the pediatricians were nonparticipants and claimed to devote 0% of their practice to Medicaid patients. In this analysis, we assume that these doctors consciously decided not to participate in the Medicaid program and, as a result, could estimate their participation in Medicaid with complete accuracy. In contrast, participants had to estimate the extent of their participation.[4] Consequently, we have eliminated the nonparticipants from the sample for the rest of the analysis. When they are excluded, SR has an average value of 15.1% and B has an average of 8.9%.

The variable DIFF was computed by subtracting the behavioral index (B) from the self-reported index (SR) in order to measure the amount by which the doctors' self-

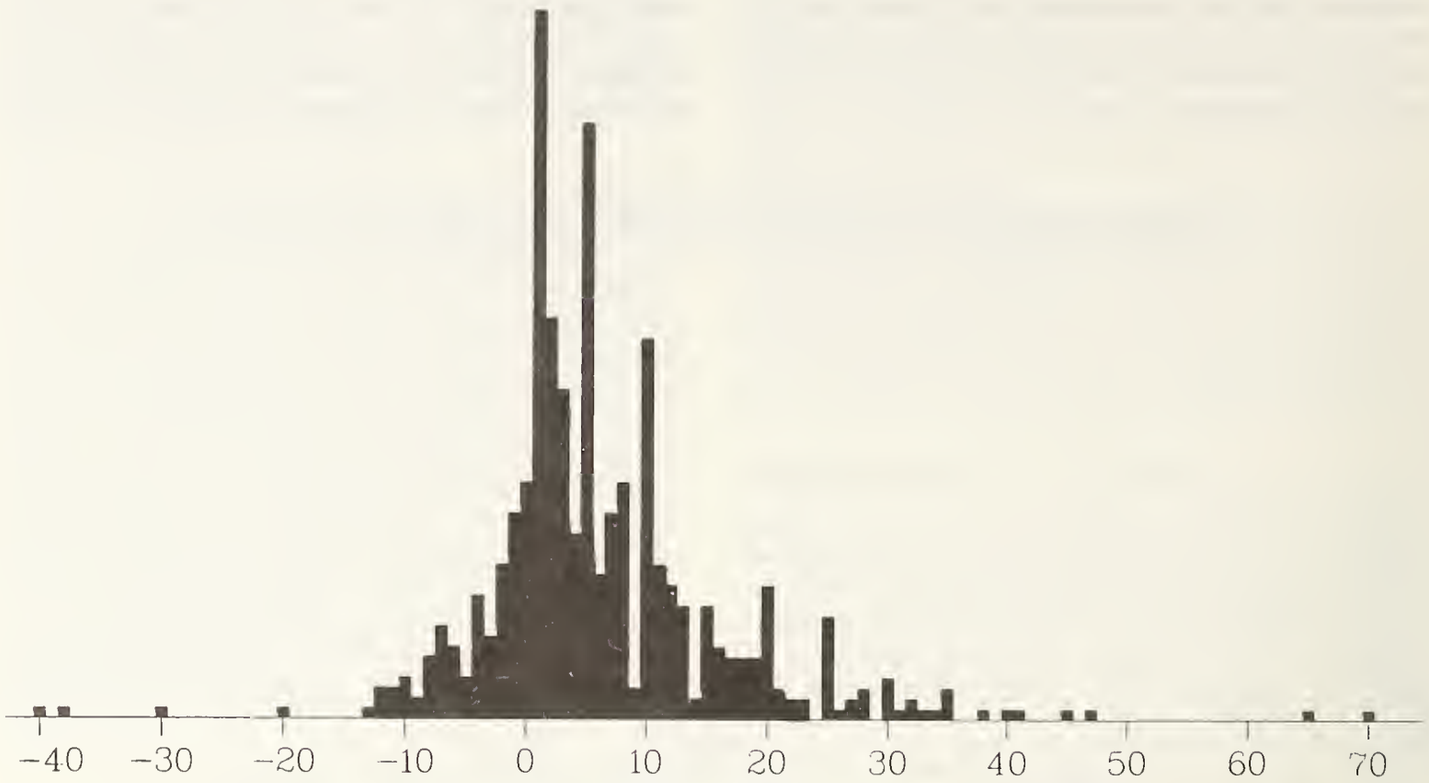
cians overstated their Medicaid participation by at least 5 percentage points. Thirty percent of the physicians overstated their participation by 10 or more percentage points.

Methodological issues

Before analyzing further the discrepancy between the values of these two indices, two methodological issues will be discussed. The first concerns a conceptual difference between indices SR and B and the second, the possibility of sampling bias in the patient record data.

A conceptual difference between the self-reported and behavioral indices. Index SR is the doctor's self-re-

Figure 3
Frequency distribution for DIFF



ported estimate of the percent of his *patients* who are paid for by Medicaid, whereas index B is based on a sample of the doctor's *patient visits*. One could argue that even if doctors had perfect knowledge of their participation in Medicaid, index SR would equal index B only in the unlikely event that the doctor's Medicaid patients made visits with precisely the same probability as his non-Medicaid patients. An analysis of secondary data sources suggest, however, that the Medicaid children in our sample in fact may have made physician office visits at a higher rate than non-Medicaid patients. In the Appendix we present data from the Health Interview Survey which indicate that among children who make office visits, those from low-income families tend to make more visits than those from high-income families. Thus, index B is likely to *overstate* the percent of the doctor's patients who are actually paid for by Medicaid. Consequently, the difference between the doctor's self-reported estimate and his true rate of participation may actually be *understated* by DIFF.

Possible bias in the patient record data. In order for the aggregated patient record data to be an unbiased estimator of a doctor's true extent of Medicaid participation, it is necessary to assume that the doctor's sample of patients was selected randomly with no systematic bias. The patient record data were gathered in the following way: After the personal interview in which he gave his self-reported estimate of Medicaid participation, each doctor who agreed to participate in the second phase of the study was asked to indicate the number of patients he expected to see during the week following the interview. Based on his response, the physician was then asked to complete a patient record form for every fifth, every third, every second, or every single patient visit of the survey week.[5] The sampling fractions were assigned to physicians so that a target number of 35 patient records would be obtained from each.

The question of whether this method produced a systematic bias in the sample can be divided into two parts: (1) Were the patients chosen representative of all the patients the doctor saw *that week?* (2) Were the patients that week representative of the patients the doctor saw *that year?*

It is unlikely that the method by which patients were chosen would create any bias in the representativeness of patient record data for the survey week. One might argue on technical grounds that choosing every nth patient does not constitute a random sample. However, this method of selecting a sample (known as sequential sampling) is a source of bias only in unlikely situations—such as a receptionist organizing a doctor's schedule so that every other patient seen was a Medicaid patient. Furthermore, this method of sampling patient visits has been used by the National Ambulatory Medical Care Survey, which has successfully collected data on ambulatory medical care for many years.

The second issue—whether or not the patients se-

lected were representative of the physician's patients for the entire year—is a more probable source of bias. The patient records completed by the doctors were for visits between July and December 1978. The percent distribution of the patients by month shows that the majority of the patients were concentrated in August (Table 1). Is it possible that the concentration of patient visits in this month distorts our picture of the doctor's year-round practice? And, if so, could this nonrepresentativeness have caused the discrepancy between the self-reported and behavioral estimates of Medicaid participation?

Table 1
Distribution of patient records by month

Month	Number	Percent
July	2,313	8.1
August	16,004	56.2
September	4,609	16.2
October	1,162	4.1
November	4,154	14.6
December	244	0.9
Total	28,486	100.0

One way to answer these questions is to examine for each month the percent of patient records for which Medicaid was the payer. These data are provided in Table 2, which shows that Medicaid participation does appear to vary by month. The percent of patient records paid for by Medicaid was higher in the summer months of July and August and lower in the fall.[6] The data do *not* suggest, however, that the patient records understate Medicaid participation. In fact, Medicaid participation was somewhat above average in August, the month given heaviest weight in the analysis. Since the available data do not support the notion that seasonality of the patient record data caused Medicaid participation to be understated, there appear to be no compelling reasons to believe that sampling bias caused the large discrepancy between the self-reported and behavioral estimates of Medicaid participation.

Comparing the two indices

In this section, we compare the self-reported index of Medicaid participation and the behavioral index to answer two questions: (1) How closely are the two related? (2) What factors are associated with physician overstatement of Medicaid participation? Linear regression analysis is used to answer both of these questions.

The relationship between self-reported and behavioral estimates. Although the values of the self-reported index (SR) are higher than the behavioral index based on the patient record data (B), these two indices have a strong positive relationship—a Pearson correlation of +.77. The relationship between the two indices can be further described through the use of regression analysis. The first column of Table 3 shows a regression equation in which SR is regressed on B. Note that the one independent variable explains almost 59% of the variance. The second column of Table 3 displays the regression equation when 10 outlyers are removed from the sample.[7] The regression constant of 6.66 is the expected value of SR for participating doctors who filled out no patient records for Medicaid patients—that is, doctors for whom B equals 0. For these doctors, the difference between the self-reported index and the aggregated patient record data was, on average, 6.7%. Also note that since the regression coefficient is .90, the expected value of SR increases by only 0.9 for every unit increase in B. As a result, the discrepancy between the expected values of SR and B decreases for doctors who were heavy Medicaid participants.

Table 2
Percent of patient records for which Medicaid is the source of payment by month

Month	% of patient records paid by Medicaid	n
July	9.0	2,313
August	8.5	16,004
September	5.8	4,609
October	7.0	1,162
November/December	6.3	4,398
Total	7.7	28,486

Table 3
Regression analysis of the doctor's self-reported estimate of Medicaid participation (SR) on the aggregated patient record index (B)

Independent variable	Entire sample: regression coefficient	Sample with outlyers removed: regression coefficient
Participation index based on the aggregated patient record data (SR)	0.856[b]	0.897[b]
Constant	7.479	6.660
R^2	0.587	0.655
N	567	557

[b]significant at $p < .01$

Factors associated with the overestimation of Medicaid participation. Earlier we defined the variable DIFF as the value of SR minus the value of B. The variable DIFF measures the amount by which doctors overstated their participation in Medicaid. This section analyzes the determinants of DIFF.

Several factors affect the accuracy with which doctors perceive their participation in Medicaid. These factors can be divided into three categories: (1) the degree of the doctor's involvement in the billing of patients; (2) problems with Medicaid participation, which cause the doctor's experiences with Medicaid patients to stand out in his mind; and (3) the doctor's attitudes and preconceived notions.

The involvement of doctors in the billing of patients. Doctors who are most active in the billing of patients should have the most accurate knowledge about the source of payment for most of their patients. We expected the following three variables from the personal interview to be associated with the doctors' involvement in billing procedures.

1. *Percent of time spent in administrative activities.* Doctors who spend a large amount of their time in administrative activites should have more accurate knowledge about which patients are being paid for by Medicaid. Thus, we expected that the percent of time the doctor reported spending in administrative activites would be negatively associated with DIFF.

2. *Type of practice.* Group practices hire more nonphysician personnel than solo practices or two-person partnerships and have greater division of labor. In such practices employees are often given the task of billing patients. The doctors are less involved in the billing process and consequently have less accurate information about the sources of payment for their patients. Thus, we expected that the value of DIFF would be greater for doctors in group practices. This variable is dichotomous, having a value of one if the doctor was in a group and zero if he was not.

3. *The number of full-time-equivalent employees per doctor.* In practices with a high ratio of nonphysician personnel to doctors, the physicians are less likely to be directly involved in the billing of patients. Consequently we expected the number of full-time-equivalent employees per doctor to have a positive effect on DIFF.

Problems with Medicaid participation. Many doctors believe that Medicaid provides inadequate reimbursement for the services rendered and that the bureaucratic red tape associated with participation is especially burdensome. These problems may cause doctors to exaggerate their experiences with Medicaid patients and, hence, unwittingly overestimate their Medicaid participation. We used four variables to analyze the effects of these problems of DIFF.

1. *The percent Medicaid payment is of the current usual fee for a routine follow-up office visit.* Doctors were asked to report both their current usual fee and the expected payment from Medicaid for a variety of services. These two values were used to calculate the percent that the Medicaid payment was of the doctor's current usual fee for a follow-up office visit. We expected this variable would have a negative effect on DIFF.

2. *The percent of Medicaid claims returned to the doctor's office for additional work.* This variable, which reflects the difficulties the physician had in filling out Medicaid claims, was expected to have a positive effect on DIFF.

3. *The number of weeks for payment from Medicaid.* The questionnaire asked the doctors to report the average number of weeks between the date that Medicaid was billed and date payment was received. This variable was expected to have a positive effect on DIFF.

4. *The number of minutes required to fill out a Medicaid claim.* The questionnaire asked the doctor to report the number of minutes required to fill out a Medicaid claim. We reasoned that the longer it took the physician to complete the form, the more prominent his Medicaid experience would be in his consciousness. Therefore we expected this variable to have a positive effect on DIFF.

Attitudes and preconceived notions. A doctor's attitudes and preconceived notions about Medicaid may influence his perception of his Medicaid experiences, and thus bias his self-reported estimate of participation. Several possible sources of this sort of bias are discussed below.

1. *Opinions about the responsibility of government to provide medical care to the poor.* The doctor was asked to agree or disagree with the following statement: "It is the government's responsibility to ensure that medical care is available to those who cannot afford it." Doctors who do not believe that government has the responsibility to provide medical care to the poor may feel that their experience with Medicaid is especially distasteful, causing them to exaggerate their participation in Medicaid. This variable is dichotomous, having a value of one if the doctor agreed with the statement and a value of zero if he did not. We expected this variable to have a negative effect on DIFF.

2. *Per capita income in the doctor's zip code area.* Doctors who do not have accurate knowledge of their participation in Medicaid may be influenced in part by characteristics of the patients they treat. For example, doctors who practice in low-income neighborhoods may assume that most of the patients they treat are eligible for Medicaid and, therefore, have a tendency to overestimate the extent of their Medicaid participation. Consequently, we expect that the per capita income in the doctor's zip code area would have a negative effect on DIFF.

3. *Percent of patients who are Black or Hispanic.* Because Blacks and Hispanics have lower average incomes than the white population, they are disproportionately represented among Medicaid recipients. Doctors who have little knowledge about the sources of payment for their patients may be under the false impression that a high proportion of their Black and Hispanic patients are Medicaid recipients. The questionnaire asked the doctor to estimate the percentages of his patients who are Black and Hispanic. We expected the sum of these two percentages would have a positive relationship with DIFF.

4. *The length of time practicing in the community.* The accuracy of a doctor's knowledge about sources of payment should be directly related to how well he knows his patients. Since we assumed that the doctor's knowledge of his patients would be associated with the number of years he had practiced in the community, we expected this variable to have a negative relationship with DIFF.

The results of the regression analysis. Table 4 shows the results of the regression analysis in which the dependent variable DIFF was regressed on the eleven independent variables listed above. In general, the 11 variables did a

poor job of explaining the variance of DIFF. Only 4.6% of the variance was explained, and only two of the independent variables, the per capita income in the doctor's zip code area and the percent of his patients who were Black or Hispanic, were statistically significant. It thus appears that the degree to which doctors overestimate their participation in Medicaid is determined to a limited degree by the characteristics of the patients they treat.

Table 4
Regression analyses of DIFF—the difference between the doctor's self-reported estimate of Medicaid participation and the aggregated patient record index

Independent variables	*Beta*
Percent of time in administrative activities	−0.004
Type of practice	0.035
Number of full-time employees per doctor	0.065
Percent Medicaid payment is of current usual fee for follow-up office visit	0.041
Percent of Medicaid claims returned for additional work	0.019
Elapsed weeks between billing and payment of claims	−0.051
Minutes spent completing a Medicaid claim	−0.050
Agree/disagree that government should provide medical care to the poor	0.026
1970 per capita income in doctor's zip code area	−0.111[a]
Percent of patients who are Black or Hispanic	0.100[a]
Number of years doctor has practiced in the community	0.003
R^2	0.046
N	485

[a]significant at $p < .05$

Implications for the evaluation of past research

It has been established that doctors' self-reported estimtes greatly exaggerate the amount of their participation in Medicaid. We now consider the implications of this finding on the evaluation of previous studies of the determinants of Medicaid participation which have analyzed self-reported estimates. The conclusions of this section are based on two regression equations—one analyzing the self-reported estimate of Medicaid participation (SR) and the other, the behavioral index (B). If the results of the two analyses differ, it would be reasonable to conclude that the use of self-reported estimates as the dependent variable in a multi-variate analysis produces misleading results and that the findings of past research must be reconsidered. On the other hand, if the results of the two regressions are similar, they would provide additional support for previous findings.

The determinants of Medicaid participation. The regression analysis presented here is based on the 13-state study of the determinants of Medicaid participation described earlier. The regression equations in this analysis contain 14 independent variables, which can be grouped into 3 categories: (1) personal and practice characteristics; (2) service area characteristics; and (3) policy variables. The operational definitions of these variables and their expected effects on Medicaid participation are presented below.[8]

Personal and practice characteristics. The characteristics of a physician and his practice may influence Medicaid partcipation in several ways. First, personal characteristics such as age, place of graduation, and board-certification status may affect demand for his services. Demand, in turn, affects Medicaid participation, because the greater the non-Medicaid demand, the less likely the physician will be to choose to accept patients from the less lucrative Medicaid market. Medicaid participation may also be affected by the physician's opinions and predispositions, including his attitudes toward poor people and his view of the government's proper role in social welfare. Finally, a physician's participation in Medicaid may be affected by the operating expenses of his practice. As practice costs rise, the extent of Medicaid participation may decline if he excludes the least lucrative patients from his practice in an effort to increase his share of other patients. The salaries paid to nonphysician personnel represent an important component of operating costs. The following five variables represent personal and practice characteristics which were expected to affect the extent of Medicaid participation.

1. *Age.* This variable, defined as the age of the physician on December 31, 1978, was expected to have a negative effect on Medicaid participation.

2. *Place of medical education.* This dichotomous variable, which equals one if the doctor is a foreign medical graduate and zero if he is a U.S. medical graduate, was expected to have a positive effect on Medicaid participation.

3. *Board certification status.* This is also a dichotomous variable, which equals one if the doctor is board certified and zero if he is not. It was expected to have a negative effect on Medicaid participation.

4. *Opinions about the role of government in ensuring that the poor have medical care.* This is a dichotomous variable that equals one if the doctor agreed and zero if he did not agree with the following statement: "It is the government's responsibility to ensure that medical care is available to those who cannot afford it." It was expected to have a positive effect on Medicaid participation.

5. *Nonphysician personnel costs.* This variable is a hospital wage index developed by the Health Care Financing Administration (U.S., Federal Register, 1980). The index, based on 1978 Bureau of Labor Statistics data for the hospital industry, yields an average monthly wage figure for hospital employees in each Standard Metropolitan Statistical Area and for all nonmetropolitan counties in each state. It was expected to have a negative impact on participation.

Service area characteristics. The characteristics of the area in which a physician practices may affect the demand for his services in the nonMedicaid market, which according to theory is inversely related to the extent of his Medicaid participation. The demand from the non-Medicaid market has a positive relationship with per

capita income and a negative relationship with the proportion of the population who are Medicaid eligible and the availability of alternative sources of health care, including the presence of other physicians in the community. Finally, the Medicaid demand for physician services in metropolitan areas is expected to be greater than that in rural areas and consequently the nonMedicaid demand is expected to be less. The following four variables represent characteristics of the physician's service area which were expected to affect the extent of his Medicaid participation.

1. *Zip code area per capita income, 1970.* This variable, based on data from the Fifth Count of the 1970 Census of Population, equals the 1970 per capita income of the population residing in the zip code area in which the physician practices.

2. *Estimate of proportion of zip code area population on Medicaid, 1978.* The 1970 census data on the proportion of the zip code area population below poverty were used to develop an estimate of the proportion of the zip code area residents receiving Medicaid benefits in 1978. For each physician the estimate is calculated as follows:

$$\frac{\%\text{ of Zip Code Population Below Poverty, 1970}}{\%\text{ of State Population Below Poverty, 1970}} \times \%\text{ of State Population Receiving Medicaid Benefits, 1978}$$

Data pertaining to the poverty population in 1970 were obtained from the 1978 Statistical Abstract of the United States.

3. *Active physicians per 100,000 county population, 1976.* The data to compute this variable came from the U.S. Bureau of Health Manpower Area Resource File. This variable was expected to have a positive effect on the extent of Medicaid participation.

4. *Size/type of community.* This dichotomous variable, which equals one if the doctor practiced in a Standard Statistical Metropolitan Area and zero if he did not, was expected to have a positive effect on the extent of Medicaid participation.

Policy variables. The policy variables measure aspects of state Medicaid policy which may foster or hinder pediatrician participation in Medicaid by affecting two key factors in his participation decision, economic incentives and professional autonomy. The extent of Medicaid participation is directly related to positive economic incentives, such as high Medicaid reimbursements, and inversely related to the administrative costs associated with collecting Medicaid reimbursements (e.g., the proportion of claims returned to the doctor for additional work, the number of weeks it took to be paid, and the number of minutes needed to complete a claim.) In a comparable manner, the extent of Medicaid participation is expected to have a direct relationship with positive professional incentives and a negative relationship with professional costs. That is, Medicaid participation is expected to be higher in states which offer a broader range of optional services, do not place arbitrary limits on the amounts of services that can be provided, and offer Medicaid benefits to the medically needy. The following five variables represent aspects of state Medicaid policy which were expected to affect the extent of Medicaid participation.

1. *State Medicaid reimbursement for a follow-up office visit.* This variable, which is an average of the amounts pediatricians in a state reported receiving from Medicaid for a follow-up office visit, was expected to have a positive effect on Medicaid participation. An Area Price Deflator was used to control for geographic variations in prices.

2. *Percentage of time a Medicaid claim is returned to the physician's office for additional work.* The questionnaire asked physicians to report what percentage of their Medicaid claims were returned for additional work. Their responses were expected to be correlated negatively with the extent of Medicaid participation. An instrumental variable was used in the regression analysis to minimize the possibility of a spurious relationship due to reverse causality.

3. *Elapsed weeks between billing and payment of Medicaid claims.* The questionnaire asked the physicians to report the number of weeks required to receive reimbursement from Medicaid. This variable was expected to have a negative effect on Medicaid participation.

4. *Minutes spent completing a Medicaid claim.* The physicians were asked how many minutes were required to complete a Medicaid claim. Their responses were expected to have a negative relationship with Medicaid participation. An instrumental variable was used in the regression analysis to minimize the possibility of a spurious relationship due to reverse causality.

5. *Revised Medicaid Program Index (RMPI), 1978.* The RMPI is a composite measure of six different aspects of state Medicaid policy which affect the professional autonomy of physicians: the number of optional services covered by Medicaid; limitations on the provision of basic procedures; the income eligibility level for AFDC; eligibility of the medically needy; reimbursement procedures; and the presence of a state Medicaid Management Information System. This index, which is a revision of an earlier measure developed by Davidson (1978), was expected to have a positive effect on Medicaid participation.

The results of the regression analyses. The results of the two regression analyses are shown in Table 5. These regression analyses are based on the sample of 525 doctors for whom we had valid data on SR, B, and all of the independent variables of the regression equations. In many respects the two regression equations in Table 5 are quite similar. The signs of the coefficients are identical for all variables in both equations. Further, the regression coefficients are nearly equal for many of the independent variables. With only one exception, the set of variables significant at the .05 level or better is identical for both regression equations. The lone exception,

the number of active physicians per 100,000 population, had a significant effect for the self-reported estimates, but not for the behavioral estimates.

There are several other minor discrepancies between the two equations. The regression constant is -11.3 for the self-reported estimate and -16.9 for the behavioral index. The difference in these values apparently reflects the difference between the mean values of the two dependent variables. Another important difference is that the regression coefficient for Medicaid reimbursement in the analysis of B is only about 60% of what it is in the analysis of SR. In other words, the effect of Medicaid fees on the extent of participation is less in the analysis of the behavioral estimates than it is for the self-reported estimates.

Table 5
Regression analyses of the self-reported estimate of Medicaid participation (SR) and the aggregated patient record index (B)

Independent variables	*Self-reported estimate: regression coefficient*	*Aggregated patient record data: regression coefficient*
Age	−0.0235	−0.0180
Place of medical education	12.379[b]	12.699[b]
Board certification status	−1.581	−1.549
Agree/disagree that government should provide medical care to the poor	2.699	1.851
Nonphysician personnel costs	0.0130[a]	0.0144[b]
1970 per capita income in doctor's zip code area	−0.00261	−0.00194
Active physicians per 100,000 county population, 1976	−0.01190[a]	−0.00235
Estimate of proportion of zip code area population on Medicaid, 1978	1.294[b]	1.150[b]
Size/type of community	2.171	0.874
State Medicaid reimbursement for a followup office visit	1.050[b]	0.584[a]
Percent of Medicaid claims returned for additional work	−1.201	−1.211
Elapsed weeks between billing and payment of claims	−1.031	−0.330
Minutes spend completing a Medicaid claim	0.140	0.870
Revised Medicaid Program Index, 1978	0.785[b]	0.667[b]
Constant	−11.304	−16.872
R^2	0.262	0.244
N	525	525
F(14,510)	12.950	11.788

[a]significant at $p < .05$
[b]significant at $p < .01$

The striking similarities of these two regression equations suggest that in fact self-reported estimates can be used reliably in research on the *determinants* of the extent of Medicaid participation. In this respect, the analysis supports the findings of past research. However, as pointed out earlier in this paper, self-reported estimates greatly overstate the amount of actual physician participation in Medicaid, possibly by a factor as large as 60%. Thus, when the purpose of a study is to measure the amount of participation or to predict the precise degree to which participation will be increased by a particular policy change, self-reported estimates are unreliable. For example, past studies have reported elasticities for the independent variables in their analyses, which were computed using the mean of the physicians' self-reported extent of Medicaid participation. Since that mean is inaccurate, it follows that the values of the elasticities must also be inaccurate. This result does not mean that physicians are unresponsive to Medicaid fees, but only that without more accurate estimates of participation it is not possible to predict how much participation would increase as a result of a given increase in Medicaid fees.

Conclusion

In this paper we have compared two measures of an important policy outcome, the extent of physician participation in Medicaid. The results show that while physicians tend to overestimate their participation, their self-reported estimates are strongly correlated with their observed rates of participation. Thus, either measure can be used to identify the determinants of particiption. On the other hand, only the behavioral measure should be used when the purpose is to estimate the precise effects of a particular policy change.

Social scientists are increasingly called upon to undertake research on health policy. Our research suggests that these invesigators must be sensitive to the issues raised by the discrepancies between self-reported and behavioral measures. As our findings indicate, self-reported estimates may suffer from systematic bias and their accuracy should be validated whenever possible. Such efforts to validate data are particularly important when research findings have implications that go beyond academic theorizing to the practical arena of public policy decisions. Here, the stakes associated with error in the measurement of health policy outcomes are obviously much greater.

Moreover, much could be learned from further research comparing self-reported with other indices, not only for outcome variables such as Medicaid participation, but also for independent variables. The study reported here, for example, affords the opportunity to conduct analyses comparing two different measures of variables such as fee levels. The study reported here yielded data not only of physician-reported Medicaid fees, but also on the amounts actually recorded in the physician's financial records. An analysis comparing these two very different sources of information about physician fees will allow some very interesting comparisons. We will be able to examine the accuracy of the physicians' perceptions regarding the reimbursement aspect of Medicaid policy, to investigate the effects of these perceptions on their Medicaid participation, and

to explore further the implications of different measurement strategies for the findings of health policy research.

Appendix: The use of office-based services by Medicaid and nonMedicaid children

The central finding of this paper is that the average value of the behavioral measure of Medicaid participation is only about 60% of the value of the self-reported measure. As already stated, this discrepancy may be due to the conceptual differences between the two measures. The self-reported measure is defined as the doctor's estimate of the percentage of his *patients* who were paid for by Medicaid. In contrast, the behavioral measure is defined as the proportion of the doctor's sample of *office visits* for which he expected Medicaid to pay. If Medicaid patients, on the average, made fewer office visits than nonMedicaid patients, the discrepancy between the behavioral and self-reported indices of Medicaid participation might be a statistical artifact. That is, the discrepancy might be due to how the indices were defined and not to physicians overstating their participation.

In this appendix, we argue that Medicaid children do, in fact, make as many office visits as nonMedicaid children and, hence, the discrepancy between the behavioral and self-reported measures is not a statistical artifact.[9] This is an empirical issue which can only be resolved by a closer look at relevant data. Unfortunately, since the 13-state Medicaid participation study had a cross-sectional design, it does not provide information on the rate of office visits. Consequently, we must rely on a secondary source of data which is comparable to the data from the 13-state study. We therefore need a secondary source of data with the following set of criteria:

First, we need data on the frequency with which *office visits* are made, not physician contacts in general. This is an important consideration since several sources of use data (e.g., the Health Interview Survey) encompass in their definition of "patient visits" almost any sort of consultation between doctor and patient, including telephone calls, house calls, and visits to hospital emergency rooms and clinics.

Second, the data should be limited to the patients who made *at least one office visit* because the unit of analysis for the data from the 13-state study is the office visit. People who did not make at least one office visit, therefore, could not have been included in that data set.

Third, the data should be limited to *children*. This restriction is important because use data for the total population often show different patterns than data for just the child population.

Fourth, the data should preferably be available for both the Medicaid and nonMedicaid populations. Unfortunately, the necessary data on the use of physician offices by Medicaid and nonMedicaid patients are not available. Thus, we must use income-category data instead and assume that Medicaid patients behave in a manner similar to other low-income patients. This represents a shortcoming in the suitablity of the available data for our purposes.

Fifth, the data should be limited to patients seen by pediatricians. Unfortunately, since there are no suitable data on use rates of pediatric patients broken down by income, we must draw inferences from the patients seeing physicians of all specialties. One should keep in mind that pediatric patients may differ from this more general population in terms of their age and income distributions and, consequently, in their use patterns.

Given these criteria, the most useful source of available data is the third volume of *Better Health for Our Children, The Report of the Select Panel for the Promotion of Child Health* (Kovar, 1981). This volume provides a statistical profile of child health in the U.S., drawing data from a wide variety of sources, including the Health Interview Survey. The data in this volume are not cross-classified exactly as needed, but they nevertheless provide a base from which we can make inferences.

In order to draw inferences from the HIS data published in *Better Health for Our Children*, it was necessary to make several computations. The computations are presented in the worksheet below (Table 6). Column (1) of the worksheet displays "the number of physician contacts per year for children under 18 years of age". These data represent the 1975–1976 annual average. The data in column (1) represent all physician contacts, including emergency room visits and telephone calls as well as office visits. We can adjust the data in column (1) to represent the frequency of *office visits* alone by multiplying the data by "the percent of physician contacts which are office visits." Unfortunately, data on "the percent of physician contacts which are office visits" are not available for the various income categories. However, we were able to find a fairly close proxy, "the proportion of children under 18 whose last physician contact was an office visit." These data are displayed in column (2).

This proxy is acceptable for two reasons. First, the two variables are closely related, and there do not appear to be any compelling reasons to believe that the proxy might deviate systematically from the preferred variable.[10] Second, the data available for the two variables are in close agreement. The first column of Table 7 displays HIS data on "the place-of-visit of physician contacts for children under 15"; and the second column of Table 7 displays data on "the place-of-visit of the last physician contacts made by children under 18," i.e., the proxy variable. The two percentage distributions are in close agreement, supporting the notion that one variable can be used as a proxy for the other.

Returning to Table 6, column (3) displays estimates for "the number of *office visits* per year for children under 18." The values in column (3) were computed by multiplying the values in column (1) by those in column (2). The data in column (3) represent the average frequency of office visits for *all* children under 18 years of

Table 6
Worksheet to compute the number of office visits per year for children with at least one office visit by income category

	(1) Number of physician contacts per year for children under 18, 1975-76 annual average[a]	(2) Proportion of children under 18 whose last physician contact was an office visit, 1975-76 annual average[b]	(3) Estimated number of office visits per year for children under 18, 1975-76 annual average (3) = (1) × (2)	(4) Proportion of children under 18 who made at least one office visit in the preceding year, 1974[c]	(5) Estimated number of office visits per year for children under 18 who made at least one office visit in the preceding year (5) = (3) ÷ (4)
Income Category					
Under $5,000	4.3	.558	2.40	.495	4.85
$5,000–$9,999	3.7	.574	2.12	.547	3.88
$10,000–$14,999	4.1	.651	2.67	.648	4.12
$15,000 or more	4.4	.671	2.95	.707	4.18
Total	4.1	.634	2.60	.620	4.19

[a]from Kovar, *Better Health for Our Children*, Vol. III, Table 74, p. 237.
[b]Ibid, Table 75, p. 239.
[c]Ibid, Table 79, p. 245.

age, including those children who made no office visits. As stated above, the children with no office visits should not be included in this analysis because they are not included in the data from the 13-state study.[11] We therefore adjust the data in column (3) to represent only those children who made at least one office visit by dividing column (3) by the "proportion of children who made at least one office visit in the previous year." Column (4) displays this proportion for each of the income categories.[12]

Column (5) displays estimates for the "number of office visits per year for children under 18 who made at least one office visit in the preceding year." The values in column (5) were computed by dividing column (3) by column (4). The data in column (5) show that the frequency of office visits among children with at least one office visit is actually higher in the lowest income category than any of the higher income categories.

Table 7
A comparison of the percent distribution of the preferred and proxy variables by place-of-visit (all income categories combined)

Place of visit	*All* physician contacts made by children under 15, 1975[a] (in percent)	*Last* physician contacts made by children under 18, 1974[b] (in percent)
Physician's office	61.5	63.4
Hospital clinic or emergency rooms	14.3	14.2
Telephone	18.6	16.9
Home	0.5	0.6
Other	5.1	4.9
Total	100.0	100.0

[a]from *Vital and Health Statistics*, Series 10, Number 128, "Physician Visits: volume and interval since last visit, 1975," Table 17, p. 30.
[b]from Kovar, *Better Health for Our Children*, Vol. III, Table 75, p. 239.

The available data show that low-income children who made at least one office visit in the previous year made at least as many office visits as high-income children (and quite possibly more). It follows that there is no reason to believe that Medicaid patients were less likely than non-Medicaid patients to be included in the doctor's sample of office visits, and in fact they may have been somewhat more likely to have been included. Thus, the discrepancy between the self-reported and behavioral measures of particiation does not appear to be the result of a statistical artifact.

Footnotes

[1] For a more complete description of the survey methodology, see Davidson et al. (forthcoming).

[2] To be eligible, the respondent had to be an office-based pediatrician in practice at least 20 hours per week; he had to have practiced in the same community for all of the preceding year, and he could not be in a group practice of 10 or more physicians.

[3] Before calculating the behavioral index, we eliminated from the sample of doctors those who filled out fewer than 15 patient records because we believed that percentages based on fewer than 15 patient records would be unstable.

[4] This difference between participants and nonparticipants is borne out by the patient record data. None of the declared nonparticipants completed patient records for Medicaid patients. Thus, the two indices are in perfect agreement at 0% for nonparticipants, while there is considerable disagreement between them for the participants.

[5] If the doctor said that he would not see any patients in the week following the interview, he was asked to complete the patient records in the next week in which he would see patients.

[6] Because of the small number of patient records in December, the data for this month are combined with those for November.

[7] An analysis of the residuals of this regression showed that 10 doctors were statistical outlyers, defining statistical outlyers as those doctors for whom the residual for the regression equation is greater than ±3.0 standard deviations. These are the doctors who did a particularly bad job of estimating the extent of their Medicaid participation. Removing these outlyers from the analysis strengthened the relationship between the two indices, increasing the Pearson correlation from +.77 to +.81.

[8] Since it is not pertinent to the primary issue of this paper, we will not present a detailed description of the theoretical perspective on which

this study is based. A complete discussion can be found in Davidson et al. (forthcoming).

[9] We wish to express our appreciation to Mary Grace Kovar for raising these questions about the analysis and encouraging us to pursue them further. We believe our conclusions are stronger as a result.

[10] The use of the proxy requires the assumption that "the percent of children whose last physician contact was an office visit" equals "the percent whose first contact (middle contact, etc.) was an office visit." There may be reasons why this is not true, but in the absence of empirical evidence, it appears to be a reasonable assumption.

[11] The inclusion of these children in the data of column (3) puts a negative bias on the frequency of office visits. This is especially the case for the lower income categories in which a large proportion of the children made no office visits.

[12] Note that the data in column (4) are for 1974, whereas the rest of the data in Table 6 are for 1975-1976. This lack of comparability could cause some bias, especially since we are dealing with income categories which are subject to change due to inflation. However, the difference between these two time periods is fairly small, and consequently we expect the amount of bias to be minor.

A comparison of estimates of out-of-pocket expenditures for health services obtained from the National Health Interview Survey Family Medical Expense Supplement and the National Medical Care Expenditure Survey*

Gail S. Poe, Division of Health Interview Statistics, National Center for Health Statistics

Daniel C. Walden, Division of Intramural Research, National Center for Health Services Research

In 1977 and 1978, two national health surveys were conducted of the civilian, noninstitutionalized population of the United States. These were the National Health Interview Survey (NHIS) and the National Medical Care Expenditure Survey (NMCES). Both the reference period (1977) and many of the concepts measured in both surveys were similar or identical, e.g., health status and use of health services. Methods of data collection, however, differed. NMCES was a panel survey and NHIS a cross-sectional survey. The focus of this paper is a comparison of the two surveys with regard to data collection and the resulting estimates of out-of-pocket health expenditures, and the methodological and policy implications of this comparison.

Accurate and efficient measurement of out-of-pocket health expenses, including health insurance premiums, is of considerable policy importance because out-of-pocket expenditures are the most widely used measurement of the individual's burden of health care costs. In the past two decades the cost of health services in the United States has increased at a rate that substantially exceeds the general rate of inflation. This trend has raised many issues important to policymakers such as methods of cost containment, the role of the public sector in the financing of medical care, and the distribution of the financial burden of health care among the population. Currently pending legislation (House of Representatives Bill 850 introduced by Representative Gebhardt and former Representative Stockman) requiring the annual collection of national estimates of out-of-pocket expenditures indicates that policymakers recognize the important contribution of timely and reliable health-care information to decision making.

It is unlikely that NMCES-type surveys will be funded on an annual basis because of their cost. If, therefore, estimates from a less expensive survey method such as the NHIS-type self-administered form are found to be of adequate reliability, such a survey may well suffice.

To meet anticipated information needs, evaluation and assessment of alternative collection, processing, and analysis procedures are critical. The major reason for including questions on out-of-pocket expenses in the 1978 NHIS was methodological, so that comparisons could be made of the resulting data with those from NMCES. The estimates for out-of-pocket expenditures are expected to be the same for the two surveys because of their similar target populations, reference period, and basic definitions. NHIS survey planners also took advantage of the opportunity to undertake a questionnaire design study of the effects of using different forms on the survey estimates and response. It is hoped that the comparison of NMCES and NHIS methods and overall results and a discussion of findings from the NHIS questionnaire design study will improve measurement techniques for future data collection efforts of out-of-pocket health expenditures.

Previous research

A few research efforts conducted over the past 20 years have provided information on the level of accuracy of household-reported health expenses. In 1960, NCHS sponsored a study conducted by the National Opinion Research Corporation (NORC) to test two alternative data collection strategies (U.S. National Center for Health Statistics, 1963). A random half of 442 NHIS households were given a short set of direct, in-person questions administered as part of the NHIS interview. The other half of the households were given a short set of questions on a self-administered form left with the household. The criterion measurement was a lengthy set of questions on utilization and expenditures administered by NORC three to four weeks following the initial data collection. In both the test procedures and the criterion measurement, respondents were asked expenditure information for a 12-month reference period. The reported level of total medical expenditures was similar for the test procedures and the criterion measurement.

Another methodological study conducted in 1970 by the Center for Health Administration Studies and the National Opinion Research Center of the University of Chicago (CHAS-NORC) made an important contribution to the understanding of response error in house-

*The views in this paper are those of the authors and no official endorsements by the National Center for Health Statistics or by the National Center for Health Services Research are intended or should be inferred. We gratefully acknowledge the helpful comments of Robert Wright, Clinton Burnham, Roger Hitchner, Judith Kasper, and Louis Rossiter, the questionnaire design work of Joyce Stevens, the editorial assistance of Renate Wilson, the secretarial support of Diane Cord and Evelyn Stanton, and the programming support of Sue Hsiung and Amy Bernstein.

hold-reported health expenditures data. Early in 1971, a nationwide probability sample of 11,619 persons were interviewed on their utilization and expenses for the calendar year 1970. Verification data were collected from family physicians, clinics, hospitals, insurers and employees about the families' medical care and health insurance for the survey year. In the analyses, family-reported expenditures for individual types of utilization were compared to the verification data. Andersen et al., (1976) reported that the 1970 NCHS estimates of out-of-pocket expenses were slightly higher than their "best estimates" for total, hospital, and physician expenses.

In a NCHS pilot study of health expenditures conducted by Johns Hopkins University and Westat in 1975–1976 in Maryland with a panel of 691 households, there was some evidence of underreporting of out-of-pocket expenses. Overall, the household report of expenses was 92.1 percent of the "best estimate." (Health Services Research and Development Center, 1977)

The National Medical Care Expenditure Survey (NMCES)

Data collection. Data for the National Medical Care Expenditure Survey (NMCES) were obtained in three separate, complementary stages which surveyed (1) about 14,000 randomly selected households in the civilian, noninstitutionalized population, each household being interviewed 6 times over an 18-month period during 1977 and 1978; (2) physicians and health care facilities providing care to household members during 1977; and (3) employers and insurance companies responsible for their insurance coverage.

Funding for the NMCES was provided by National Center for Health Services Research (NCHSR), which cosponsored the survey with NCHS. Data for the survey were collected by Research Triangle Institute of North Carolina and its subcontractors, National Opinion Research Center of the University of Chicago and ABT Associates, Inc., of Cambridge, Massachusetts. Data processing support was provided by Social and Scientific Systems, Inc., of Washington, D.C.

The survey sample was designed to produce statistically unbiased national estimates that are representative of the civilian, noninstitutionalized population of the United States. The survey reference period was January 1 to December 31, 1977. To this end, the study used the national multistage area samples of the Research Triangle Institute and the National Opinion Research Center. Sampling specifications required the selection of about 14,000 households. Data were obtained for about 91% of eligible households in the first interview and 82% by the fifth interview. Approximately 11% of all survey participants provided data for only some of the time in which they were eligible to respond. Information for these respondents was adjusted to account for this partial nonresponse (Cohen, 1981). For a detailed description of the survey sample and of sampling, estimation, and adjustment methods, including weighting for nonresponse and poststratification, see Cohen and Kalsbeek (1981).

The first round of household interviews began in January 1977, and the following interviews were conducted approximately every three months thereafter. For the interview instruments, see *NMCES Instruments and Procedures 1* (Bonham and Corder, 1981). All interviews in the first, second, and fifth rounds were held in person. About 80% of the third and fourth round interviews and about 50% of the sixth round interviews were conducted by telephone. In the first five rounds information was elicited on the use of and expenditures for health services. Data supplied by the respondent were amended in each round through the household summary update process, which allowed the respondent to correct or add to the information provided in previous interviews. Interviewers then updated a computer-generated summary of both expenditures and sources of payment for health care previously reported by the respondent. In the fifth interview, respondents reviewed with the interviewer each reported event of health care shown in the household summary.

The household questionnaire elicited use, expenditure, and source of payment data for eight types of health care use and/or expenditure: inpatient hospital services, "other hospital" services, inpatient physician services, prescription drugs, dental care, medical equipment and supplies, purchase or repair of eyeglasses and contact lenses, and ambulatory care provided by physicians or other practitioners. For each of these types of health care, a common set of questions was administered about total charges and sources of payment (see Figure 1).

Imputation of missing data. To account for data on charges and sources of payment that remained missing or incomplete at the end of all six rounds of interviewing, several types of imputation procedures were employed.

For hospital services and for care provided by physicians to either ambulatory or hospitalized patients, data from the NMCES Medical Provider Survey (MPS) were used. First, the visit and/or hospitalization data for household members included in the MPS sample were examined. For incomplete records from the household survey that matched MPS records, the missing component in the household record was supplemented by data from the MPS record. However, where events with complete total-charge data were contained in both the household survey and MPS record, no attempt was made to develop a "best-estimate." Thus, the MPS was used as the source of data only for missing household data of this type. Where data remained incomplete in the household survey records, the remaining missing total-charge and/or source-of-payment for physician services or hospitalizations were imputed using a method developed at

Figure 1
National Medical Care Expenditures Survey questions relating to expenditures and financing of hospital stays

The following questions are about the charge for this hospital stay--not about any separate bill from the doctor or surgeon.

8. How much was the total hospital charge for this stay including any amount that may be paid by health insurance or other sources?

$_____ (10)

No charge 00 (9)

Don't know. . . . . .94 (A)

A. Do you expect to receive a bill for this stay?

Yes 01 (14)

No. . . . . .02 (B)

B. Why don't you expect to receive a bill?

Free from hospital. . . . . .01 (A)

Included with mother's bill (new baby only) 02 (14)

Other source will pay 03 (10)

Already paid. 04 (10)

9. Why was there no charge for this stay?

Free from hospital. 01 (14)

Included with mother's bill (new baby only) 02 (14)

Other source will pay 03 (13)

10. How much of the (CHARGE) did you or your family already pay?

Partial _____ %

All 01 (12)

None. 00

11. How much (more) of this charge will you or your family pay?

Partial _____ %

All 01

None. 00

D.K.. 94

IF FAMILY PAID/WILL PAY ANY AMOUNT (Q's. 10 AND/OR 11), ASK:

12. Do you expect any source to reimburse you or pay you back?

Yes . . . . . 01 (A)

No. 02 (13)

A. Who will reimburse or pay you back? ENTER UNDER SOURCE. Anyone else?

B. How much will (EACH SOURCE) reimburse or pay you back?

SOURCE	AMOUNT
	%
	%
	%

IF ALL IN Q's. 10 AND/OR 11, SKIP TO Q. 14.

13. Who else paid or will pay any part of the charge? ENTER UNDER SOURCE. Anyone else?

A. How much will or did (EACH SOURCE) pay?

SOURCE	AMOUNT
	%
	%
	%

No other source 01 (14)

the U.S. Bureau of the Census for their *Current Population Reports* series. This procedure, often called hot-deck imputation, randomly imputes data from individuals with complete information to individuals with missing data but otherwise similar characteristics.

Hot-deck procedures were also used to impute most of the missing data on charges and sources of payment for the other types of health care and/or expenditures for which verification data were not collected. A different strategy was used for imputation of Medicaid payments for dental care and prescription drugs because the Medicaid programs often pay the provider less than the fee-for-service charge. Here, Medicaid fee schedules for 1977 were used to impute missing charges and Medicaid payments according to U.S. Census division, therapeutic categories, and type of dental service.

Where the sum of percents reported for several sources of payment exceeded 100 per visit or event, the amounts paid out of pocket by the family were reduced to make the sum of percents paid by all sources equal to 100%. If the sum of percents reported paid by all sources was less than 100% per visit or event, the remainder was assigned to the "other" source of payment category.

The National Health Interview Survey

The National Health Interview Survey is a household interview survey of a national-probability sample that has been continuously conducted snce 1957. It provides national data on the incidence of acute illness and accidental injuries, the prevalence of chronic conditions and impairments, the extent of individual disability, the use of health care services, and related topics. Interviews are conducted each week of the year by the Bureau of the Census interviewers. The survey covers the noninstitutionalized, civilian population of the United States living at the time of the interview. For technical and logistical reasons, patients in long-term care facilities, persons on active duty with the Armed Forces, United States nationals living in foreign countries, and persons who have died during the calendar year preceding the interview are excluded from both the sample and the survey estimates.

The sampling plan follows a multi-stage probability design which permits continuous sampling of households. The first stage consists of a sample of 376 primary sampling units (PSU's) drawn from approximately 1,900 geographically defined PSU's covering the 50 states and

the District of Columbia. A PSU consists of a county, a small group of contiguous counties, or a Standard Metropolitan Statistical Area. Within PSU's, smaller units called segments are defined such that each segment is expected to contain four households. The sampling plan is designed to yield national estimates, although separate estimates can be obained for the four geographic regions of the U.S.

The households selected for interview each week are a probability sample representative of the target population. Each calendar year, data are collected from approximately 40,000 households containing a total of about 110,000 persons. The annual response rate of NHIS is usually at least 96% of the eligible households in the sample. The 4% nonresponse is divided equally between refusals and households where no eligible respondent could be found at home after repeated calls. (For a detailed description of the sample design, see U.S. National Center for Health Statistics, 1958.) The questionnaire consists of two parts: (1) a core set of health, socioeconomic, and demographic items and (2) one or more sets of supplementary health items. The supplements change, usually on a yearly basis, in response to current interest in health topics.

The Family Medical Expense Supplement (FMES). The NHIS has included supplements for the collection of out-of-pocket health expenditures in 1963, 1966, 1971, 1975, 1976, and 1978. A short self-administered questionnaire is left with the household at the end of the personal interview during the first calendar quarter. (There were a total of 10,272 households in 1978.) The household respondent is requested to fill in the form and mail it within five days in a preaddressed, postage-paid envelope provided to the household. The form contains questions about direct out-of-pocket health expenditures for the previous calendar year for each person in the household. Each family unit or unrelated individual is given a separate questionnaire. Information is requested on spending for hospital, physician, dental, and optical services, for prescribed medicines, and for other health expenses for each person. Also obtained is the amount the family paid for health insurance premiums, either directly or as deducted from paychecks during the past calendar year. Respondents are encouraged to use any records such as bills, receipts, or check stubs in answering the questions, but may give their best estimates if they cannot supply exact amounts.

If the FMES is not received within a week to 10 days, an identical questionnaire is mailed to the household. The cover letter reminds the respondent of his or her participation in the survey and stresses the importance of completing the questionnaire and mailing it. After two more weeks, attempts are made by telephone to reach all households that have not returned a form and the questionnaire information is obtained over the telephone. No follow-up work is done in person, due to the high cost involved.

A supplement fails the edit criteria if either the health insurance item is incomplete or three or more other form entries are left blank. Households with such failed-edit questionnaires are called by telephone to obtain the data. The data are coded, key punched, and edited by computer for missing information and inconsistencies, and then merged with the household and person information obtained from the main NHIS questionnaire.

Questionnaire design experiment. The FMES form used in fiscal years 1963 and 1966 and calendar years 1971, 1975 and 1976 consisted of a cover letter, instructions on the back of the letter, one page for dollar amount entries for each of up to nine family members, and a back page for entries for amounts paid for health insurance, payments made for nonhousehold, nonfamily members, and names of persons who participated in filling the form. Figure 2 shows the first two pages, one person page, and the last page of this version.

NHIS questionnaire designers were concerned that the form was too long and contained more instructions than needed, causing it to appear overly complex. On the assumption that a shorter, more attractive, and less verbose form would increase response rates and decrease edit failure rates without loss in response quality, a short form was designed which employed a matrix format for the individual expense items. The result was a more attractive, simple, and less formidable looking document, while the information collected remained almost identical. Figure 3 contains a copy of this short form. The dimensions of the short form are 8½ by 11 inches whereas the long form is 8½ by 14 inches.

The 1978 first quarter sample was divided; a random half received the long form and the other half, the short form. Within sample segments, however, all households received the same form. Identical interviewing followup and editing procedures were used for both forms.

Estimation procedures. In the NHIS, a complex multistage probability sample, the data are weighted by the reciprocal of the probabilities of selection, adjusted for nonresponse, and ratio adjusted to Census estimates for age, sex, color, and residence classes. The effect of this ratio-estimating process is to make the sample more closely representative of the civilian, noninstitutionalized population thereby reducing sampling variance. (For a detailed description of the estimation procedure, see U.S. National Center for Health Statistics, 1970.)

Although there is additional nonresponse for the FMES (in 1978, for approximately 12% of the household responses to the core NHIS, an FMES response was not received), no further allocation or imputation techniques for missing data are employed. Nor are such techniques employed for item nonresponse in which the respondent either did not know the answer to one or more questions or failed to complete these items on the questionnaire. Thus, all estimates are based only on

Figure 2. FMES long form (actual size 8-1/2" x 14")

FORM HIS-2B(a)

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
HYATTSVILLE, MARYLAND 20782

NATIONAL CENTER FOR
HEALTH STATISTICS

Dear Friend:

Your household has just taken part in a health interview conducted by the Bureau of the Census for the U.S. Public Health Service. We greatly appreciate your cooperation in providing us with this information.

Another area of great concern today is the cost of health care in our country. We, therefore, ask you to provide us with information about the amount of money you, your family, and other relatives living with you spent for medical care during the past 12 months, that is, from January 1, 1977 to December 31, 1977, by answering the few questions on this form. Please use any records such as bills, receipts, or check stubs, that would help you in answering the questions. If you cannot supply the exact amounts from your records, give the best estimate you can.

We would appreciate your completing the attached questionnaire within FIVE DAYS, and returning it in the enclosed preaddressed envelope which requires no postage. If a delay cannot be avoided and you cannot answer and return your form during this time, please fill in the information and return it as soon as possible. Since this study is based on a scientific sample of the total population, it is important that each household return a completed questionnaire.

Please be assured that the Bureau of the Census and U.S. Public Health Service hold as confidential all the information you provide. Thus, the results of this voluntary survey will be issued only in the form of statistical totals from which no individual can be identified.

Thank you for your cooperation.

Sincerely yours,

Robert R. Fuchsberg

ROBERT R. FUCHSBERG
Director
Division of Health Interview Statistics

ASSURANCE OF CONFIDENTIALITY: Information contained on this form which would permit identification of any individual or establishment has been collected with a guarantee that it will be held in strict confidence, will be used only for purposes stated for this study, and will not be disclosed or released to others without the consent of the individual or the establishment in accordance with section 308(d) of the Public Health Service Act (42 USC 242m).

FOR INTERVIEWER USE ONLY:

a. PSU	b. Segment	c. Serial	d. Col. of head	e. Interviewer's name	code	f. Follow-up

HRA-74-3a

O.M.B. No. 68-R1600
Approval Expires March 31, 1979

GENERAL INSTRUCTIONS

1. Fill a separate page for the family member whose name is entered at the top. Answer all questions on the page even though the person may not have had any medical or dental expenses during the past 12 months. If the person did not have any expense of a certain kind during that period, mark the "No bills paid" box. The amounts you give should only include what THIS FAMILY paid, NOT any payments made by health insurance or some other person or agency. IF EXACT AMOUNTS ARE NOT KNOWN, PLEASE ENTER YOUR BEST ESTIMATE.

2. Do NOT include any amounts paid (or to be paid) by:

   - Health insurance
   - Workmen's compensation
   - Non-profit organizations such as the "Polio Foundation"
   - Charitable or Welfare Organizations
   - Military Services
   - Veterans Administration
   - Federal, State, City, or County Governments

3. If there are any babies in the household who were born during the past 12 months, the hospital and doctor bills relating to the baby's birth should be reported on the page for the mother. All other medical expenditures relating to the baby's health should be reported on the page for the baby.

4. **PLEASE COMPLETE THE BACK PAGE BEFORE MAILING.**

Figure 2 continued

Please answer the following questions for _____ Person No. ☐☐

DENTAL BILLS PAID

1. How much did **THIS FAMILY** spend on dentol bills for this person during the post 12 months, thot is, from January 31, 1977 to December 31, 1977?

DOLLARS $ | CENTS

or

☐ No dentol bills poid for this person

INCLUDE omounts spent for:

Cleonings
Fillings
Stroightening
X-roys
Dentol surgery
Extroctions
Bridgework
Dentol loborotory fees
Other services from o dentist or hygienist

DOCTORS' BILLS PAID

2. How much did **THIS FAMILY** spend on doctor bills for this person during the post 12 months?

DOLLARS $ | CENTS

or

☐ No doctor bills poid for this person

INCLUDE omounts spent for:

Routine doctor visits
Treotments
Check-ups
Doctor fees while o potient in o hospitol
Operotions
Deliveries
Pregnoncy core
Loborotory fees
Shots
Other services by o medicol doctor

HOSPITAL BILLS PAID

3. How much did **THIS FAMILY** spend on hospitol bills for this person during the post 12 months?

DOLLARS $ | CENTS

or

☐ No hospitol bills poid for this person

INCLUDE omounts spent for:

Room ond boord
Operoting ond delivery rooms
Anesthesio
Tests
X-roys
Speciol treotments
Any other hospitol services

PAYMENTS MADE FOR PRESCRIPTION MEDICINE

4. About how much did **THIS FAMILY** spend on medicine for this person during the post 12 months thot wos purchosed on a **DOCTOR'S OR DENTIST'S PRESCRIPTION?**

DOLLARS $ | CENTS

or

☐ No prescribed medicines bought for this person

INCLUDE omounts spent for:
Medicines ONLY if they were prescribed by o doctor or dentist

PAYMENTS MADE FOR EYEGLASSES, CONTACT LENSES OR OPTOMETRIST'S BILLS

5. During the post 12 months, how much did **THIS FAMILY** spend on eyeglosses, contoct lenses, or optometrists' fees for this person?

DOLLARS $ | CENTS

or

☐ No omount poid for these items

PAYMENTS MADE FOR "OTHER" MEDICAL BILLS

6o. How much did **THIS FAMILY** spend on other medicol expenses for this person during the post 12 months?

DOLLARS $ | CENTS

or

☐ No omount poid for these items

DO NOT INCLUDE ony expenses which you hove olreody recorded. DO NOT INCLUDE omounts spent for medicines of ony kind.

INCLUDE omounts spent for such expenses os:

Chiroproctors' or Podiotrists' fees
Heoring oid
Speciol broces, trusses, wheelchoir or ortificiol limbs
Physicol or Speech Theropy
Speciol nursing core
Nursing Home or Convolescent Home core

6b. Whot type of medicol expenses did this person hove?

_____
Type of Medical Expense

REFERRED TO RECORDS

7. Check one of the following boxes:

1 Referred to records for **ALL** dollor omounts entered on this poge. ☐ All

2 Referred to records for **SOME** but not oll dollor omounts entered on this poge. ☐ Some

3 Did **NOT** refer to **ANY** records. ☐ None

Figure 2 continued

HEALTH INSURANCE

1. During the past 12 months, that is, from January 1, 1977 to December 31, 1977 how much did THIS FAMILY spend on health insurance premiums for plans that pay for any part of a hospital bill or doctor's bill?

DOLLARS	CENTS
$	

or

☐ This family did not pay any insurance premiums

INCLUDE:

Amount deducted from paycheck for health insurance premiums
Amount deducted from Social Security check for Medicare
Amount paid directly to health insurance plans or to Social Security for Medicare

DO NOT INCLUDE:

Health insurance plans that pay only in the case of accidents
Employer or union contributions

PAYMENTS MADE FOR PERSONS NOT LISTED ON THIS QUESTIONNAIRE

2. During the past 12 months, that is, from January 1, 1977 to December 31, 1977 did THIS FAMILY pay any medical expenses for anyone whose name does NOT appear on this questionnaire?

This might include expenses for children now away at school or parents, other relatives or friends now in nursing homes or elsewhere, or who are deceased.

These expenses may include bills from doctors, dentists, optometrists, hospitals, nursing homes, health insurance premiums, cost of prescription medicine, eyeglasses, and so forth.

☐ No

☐ Yes

(Check one box)

TYPE OF MEDICAL EXPENSE	Amount This Family Paid
_____	DOLLARS $ / CENTS
_____	DOLLARS $ / CENTS
_____	DOLLARS $ / CENTS

3. Please print below the name of the person or persons who completed this form

Name _____

Name _____

NOTE: Before returning this questionnaire, please check to see that you have filled in an answer for EACH question for EACH person listed on the questionnaire, even though the person did not have any medical or dental expenses during the past 12 months, that is, from January 1, 1977 to December 31, 1977.

Figure 3. FMES short form

FORM HIS-1B(a) FAMILY MEDICAL EXPENSE

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
HYATTSVILLE, MARYLAND 20782

NATIONAL CENTER FOR
HEALTH STATISTICS

Dear Friend:

Your household has just taken part in a health interview conducted by the Bureau of the Census for the U.S. Public Health Service. We greatly appreciate your cooperation in providing us with this information.

As you know there is a great concern in our country about providing health care to persons who need it. There is also an urgent need to know how much money persons are spending for medical care. Only you can provide accurate information about the amount you pay for medical expenses. We, therefore, are asking you to tell us the amount of money you and your family have spent for medical care during 1977 by answering the few questions on this form. If you cannot give the exact amounts from your records, give the best estimate you can.

The survey is authorized by title 42, United States Code, section 242K. The information collected in this voluntary survey is confidential and will be used only to prepare statistical summaries. No information that will identify an individual or a family will be released.

Because this is a sample survey, your answers represent not only your household, but also hundreds of other households like yours. For this reason, your participation is extremely important to ensure the completeness and accuracy of the final results. Each unanswered question reduces the accuracy of the information collected.

Please answer all the questions as soon as possible, preferably within FIVE DAYS, and return the questionnaire in the enclosed postage-paid envelope.

Thank you for your cooperation.

Sincerely yours,

Robert R. Fuchsberg

Robert R. Fuchsberg
Director
Division of Health Interview Statistics

FOR OFFICE USE ONLY:

a. PSU	b. Segment	c. Serial	d. Col. of head	e. Interviewer's name	code	f. Follow-up

HRA-74-2a

O.M.B. No. 68-R1600
Approval Expires March 31, 1979

Figure 3 .continued (actual size 11″ x 17″)

HEALTH CARE EXPENSES PAID FOR PERSONS IN THIS FAMILY

- For each person listed please enter the amount you or this family paid for that person's medical care. Subtract any amount you got back from health insurance.
- Count only the amount you paid between **January 1, 1977** and **December 31, 1977**. Subtract any amount you got back from health insurance during this period.
- Please check your bills, receipts or checkstubs.
- If you do not have bills, receipts or checkstubs, please enter your **best estimate**.
- If the person did not have any expenses, mark the "none" box with an "X."

	1	2	3	4	5	6
	first name / last name	first name / last name	first name / last name	first name / last name	first name / last name	first name / last name
1. AMOUNT PAID FOR **DOCTOR EXPENSES** Include all expenses related to doctor office visits **and** the amounts paid for doctors and surgeons while this person was a patient in the hospital.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none
2. AMOUNT PAID FOR **HOSPITAL EXPENSES** Include all hospital charges **except** doctor and surgeon fees while this person was a patient in the hospital.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none
3. AMOUNT PAID FOR **DENTAL EXPENSES** Include all expenses related to dental office visits for this person.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none
4. AMOUNT PAID FOR **OPTICAL EXPENSES** Include all expenses for having this person's eyes examined for glasses **plus** the cost of eyeglasses or contact lenses.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none
5. AMOUNT PAID FOR **PRESCRIPTION MEDICINES** Include all expenses for medicine obtained with a doctor's or dentist's prescription for this person.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none
6. AMOUNT PAID FOR **OTHER MEDICAL EXPENSES** Include any other medical expenses which are not included above. Do this for each person.	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none	$ or: ☐ none

PLEASE GO TO NEXT PAGE

Figure 3 continued

7. HOW MUCH DID THIS FAMILY SPEND ON **HEALTH INSURANCE** in 1977 for persons listed on page 2?

Include:

- paycheck deductions for health insurance
- Social Security check deductions for Medicare
- amounts paid directly to health insurance plans or to Social Security for Medicare

Do not include:

- accident insurance
- amounts your employer or union pays for your health insurance

$ ____

or: ☐ none

8. DURING 1977 DID THIS FAMILY PAY ANY MEDICAL EXPENSES FOR **PERSONS NOT LISTED ON PAGE 2?**

Some examples:

- persons living here now, but not listed on page 2
- anyone who lived here in 1977, but does not live here now
- children now away at school or elsewhere
- parents, other relatives, or friends in nursing homes or elsewhere
- parents, other relatives, or friends now deceased

☐ YES → Please go to question 9.

☐ NO → Please go to question 10.

9. PLEASE ENTER THE **TOTAL AMOUNT** THIS FAMILY PAID DURING 1977 FOR ALL TYPES OF MEDICAL EXPENSES FOR **PERSONS NOT LISTED ON PAGE 2.**

(Remember to subtract amounts you get back from health insurance.)

$ ____

or: ☐ none

10. MARK **ONE** BOX WITH AN "X."

☐ Checked records for all dollar amounts entered on this form.

☐ Checked records for some amounts.

☐ Did not check records.

11. ENTER **YOUR NAME** AND THE NAMES OF ALL PERSONS WHO HELPED YOU ANSWER THE QUESTIONS ON THIS FORM.

If any of these persons are not listed on page 2, do not write their names. Instead, write "friend," "sister," "doctor," and so forth.

_____ your name

_____

_____

_____

Thank you very much!

those records for an expense item (or items) which contain dollar or zero dollar amounts. This procedure in effect imputes to persons with unknown amounts or to persons not responding the same distribution of health expenses as that for persons with known amounts. Each individual type of expense is based on known dollar or zero amounts. However, the sum total for all types of expense, including health insurance, is based on known amounts for all seven types of expense measured; that is, each person must report known data for all seven items of expenditure before the reported data are included in the total personal expenditures.

Because of the data collection procedures used in the NHIS and FMES, it is not appropriate to relate out-of-pocket expenditures to use levels, episodes of illness, or to total expenditures that include public and third-party payments. Whereas the reference period for some types of use and illnesses is the previous calendar year, expense information is collected on payments made during the previous calendar year for expenses incurred during that year or earlier.

Comparison of surveys estimates

Methodological and definitional differences. While NMCES and FMES used the same definitions of type of care and/or services underlying dental and prescription expenses, there were differences with respect to hospital, physician, optical, and other expenses between the two data sources. These differences in defining hospital expenses arise because it is unclear in FMES in which category respondents would report hospital expenses other than inpatient expenses. For instance, if billing for outpatient services was done by the hospital, as opposed to a physician or other health professional, the family may have reported this as a hospital expense. In NMCES, only inpatient expenses, excepting physician fees, were included under hospital expenses. For optical expenses, the NMCES included only the expenses for glasses and contact lenses while the FMES included examination expenses as well. Examination expenses in NMCES were included with physician expenses if they were provided by ophthalmologists. In FMES, the "other" medical expense category was the residual item in the series of questions about expenditures for health care and was directed at expenditures not included in other categories, whereas in NMCES "other" expenses included expenditures for ambulatory care not provided by physicians, "other hospital" expenditures such as those for ambulance services, and expenses for medical equipment and supplies such as wheelchairs, corrective shoes, or hearing aids.

The two surveys also differed with respect to the payment of health care expenses by the family. The focus of the FMES questionnaires was on actual monies paid out of pocket in 1977; these payments could have been for care received before 1977. In the instructions for the FMES long form, respondents were asked to report out-of-pocket payments made in 1977 for services received in 1977 and earlier, less amounts received or expected to be paid by third-party payers for 1977 services. The corresponding instructions for the FMES short form asked respondents to report out-of-pocket payments made in 1977 less insurance reimbursements in 1977, but did not request insurance payments received in 1978 for 1977 services. Furthermore, item missing data in FMES were not included in the estimates. Had they been given a zero value, the FMES estimates of average out-of-pocket expenses might have been less than those shown in the tables.

In contrast, the NMCES data collection was on care received only in 1977 and on charges and sources of payment for that care. NMCES focused on allocating the responsibility for each charge among all sources of payment, making sure the amount attributed to the family reflected payments by third parties. NMCES counts as out-of-pocket expenses all amounts for which the family was liable for care given in 1977. Whether a family actually paid in 1977 an amount for which it was liable is not a distinction made in NMCES data. Also, the editing procedures employed with the NMCES data reduced the amount paid by the family if the amounts paid by all payers combined exceeded 100%.

No unqualified mention of a difference between estimates is made unless the difference cited is statistically significant at the .05 level based on a t-test of significance. Relative standard errors for estimates for the NHIS-FMES may be found in U.S. National Center for Health Statistics (1979). The relative standard error for an estimate from either the short form or the long form is approximately 1.36 times the relative standard error for both forms combined. Relative standard errors for estimates from NMCES may be found in National Center for Health Services Research (1981-83).

Because the differences in estimates between the FMES long and short forms were not large enough to invalidate a comparison of NMCES and FMES estimates, the following comparisons are first made between the NMCES estimates and the FMES for both forms combined. Individual out-of-pocket health insurance expenses for NMCES are not presented in this paper. (See Walden, Horgan, and Cafferata, Session 4 in this volume.)

General findings. There were differences between the NMCES and FMES estimates of the percents of persons with out-of-pocket expenses (Table 1) and average expenses for persons with expense (Table 2). These differences in survey estimates were not unexpected given major differences in design and definitions of particular expense components. While the direction and magnitude of nonsampling errors in the estimates produced by each survey are for the most part unknown, additional reasons that must be considered by way of explanation are failure to take into account all third-party payments by FMES respondents, telescoping of pay-

Table 1

Percent of persons with out-of-pocket expense, by type of expense according to NMCES and NHIS—FMES by form type, sex, race, age, and family income: United States, 1977[a]

Survey form type and sex, race, age, and family income	All types of expense		Health expenses							
	Including insurance premium	Excluding insurance premium	Hospitel	Doctor	Dental	Prescription medicine	Optical	Health insurence premium	Other	Populetion (in thousends)
	Percent of Persons with Out-of-Pocket Expense									
All persons										
NMCES		75.3	4.7	56.9	34.8	50.2	9.1		15.3	
FMES—both forms	86.4	77.6	11.3	58.0	41.9	53.1	23.7	57.9	8.2	213,195
Long form	86.8	77.2	11.7	57.6	41.9	52.6	23.2	59.0	6.1	
Short form	86.0	78.0	10.9	58.3	41.9	53.6	24.2	56.9	10.2	
Under 17 years										
NMCES		70.0	2.0	53.1	31.5	42.4	4.0		8.4	
FMES—both forms	80.2	70.6	8.7	54.9	38.4	45.5	11.1	50.0	4.1	59,348
Long form	80.9	69.6	9.1	52.9	37.9	44.1	10.8	51.5	1.9	
Short form	79.7	71.6	8.3	56.9	38.8	47.0	11.5	48.4	6.4	
Under 6 years										
NMCES		72.7	3.1	62.8	9.6	54.7	0.2		8.1	
FMES—both forms	78.5	70.9	12.4	62.9	14.8	55.3	2.1	45.6	5.0	18,453
Long form	78.7	69.8	12.4	61.7	14.8	54.8	2.1	46.4	2.5	
Short form	78.1	71.9	12.4	63.9	14.8	55.6	2.2	44.6	7.3	
6–16 years										
NMCES		68.9	1.5	48.9	41.1	37.0	5.6		8.5	
FMES—both forms	81.0	70.4	7.1	51.4	48.8	41.2	15.1	51.9	3.8	40,896
Long form	81.8	69.5	7.7	49.1	47.9	39.4	14.6	53.8	1.7	
Short form	80.3	71.4	7.5	53.7	49.6	43.1	15.7	50.1	5.9	
17–44 years										
NMCES		74.4	4.9	54.4	37.4	47.0	9.0		14.3	
FMES—both forms	85.8	77.7	11.7	56.5	45.7	50.6	22.6	52.8	6.8	87,866
Long form	86.3	77.6	11.7	56.2	46.1	50.5	22.5	53.5	4.7	
Short form	85.4	77.8	11.6	56.8	45.4	50.7	22.7	52.1	8.8	
45–64 years										
NMCES		80.1	5.7	62.0	38.0	57.7	14.9		20.4	
FMES—both forms	91.4	83.4	12.3	61.7	44.8	60.1	36.6	64.4	11.7	43,382
Long form	91.0	83.2	13.0	62.7	45.2	59.6	35.5	65.5	9.7	
Short form	91.6	83.6	11.6	60.8	44.3	60.4	37.5	63.3	13.6	
65 years and over										
NMCES		83.3	9.6	67.3	27.4	68.5	11.6		27.5	
FMES—both forms	95.0	83.9	14.3	63.8	31.2	67.8	33.7	84.2	16.5	22,598
Long form	96.4	83.8	14.9	64.4	30.8	67.5	32.5	84.6	14.2	
Short form	94.3	83.7	13.7	63.3	31.6	67.8	35.1	83.6	18.8	
Sex										
Male										
NMCES		70.7	3.5	50.5	32.4	43.3	7.7		13.1	
FMES—both forms	85.1	74.6	10.1	52.9	40.6	47.7	21.2	57.8	7.5	102,870
Long form	85.5	74.5	10.5	52.5	41.1	47.5	20.7	58.8	5.6	
Short form	84.6	74.7	9.7	53.4	40.1	47.9	21.7	56.9	9.3	
Female										
NMCES		79.5	5.9	62.9	37.1	56.6	10.4		17.3	
FMES—both forms	87.7	80.4	12.4	62.7	43.1	58.1	26.1	58.0	8.8	110,324
Long form	88.0	79.8	12.8	62.4	42.7	57.4	25.6	59.2	6.5	
Short form	87.4	81.0	12.0	63.0	43.4	59.0	26.5	56.8	11.1	
Race										
White										
NMCES		78.2	4.8	59.4	37.6	52.5	9.6		16.5	
FMES—both forms	88.6	80.6	11.7	60.4	44.6	55.0	24.8	59.5	8.7	184,611
Long form	89.0	80.2	12.2	60.1	44.8	54.4	24.3	60.8	6.6	
Short form	88.2	81.0	11.1	60.6	44.4	55.6	25.3	58.3	10.8	

Table 1 continued

Survey form type and sex, race, age, and family income	All types of expense		Health expenses							
	Including insurance premium	Excluding insurance premium	Hospital	Doctor	Dental	Prescription medicine	Optical	Health insurance premium	Other	Population (in thousands)
	Percent of Persons with Out-of-Pocket Expense									
Black[b]										
NMCES		53.2	3.9	38.4	14.3	34.7	5.9		7.1	
FMES—both forms	69.8	55.1	8.6	41.2	21.0	39.8	15.8	45.0	4.3	25,864
Long form	68.9	53.5	7.9	39.4	19.3	39.1	15.1	44.0	1.7	
Short form	70.3	56.3	9.3	42.9	22.4	40.1	16.6	45.7	6.5	
Family Income										
Less than $3,000										
NMCES		61.7	4.9	43.0	22.7	41.5	7.1		14.4	
FMES—both forms	68.2	60.8	8.5	40.9	21.3	44.6	21.9	38.2	7.8	14,743
Long form	71.6	61.8	9.4	42.0	18.8	45.6	21.0	39.6	7.3	
Short form	64.6	59.5	7.7	39.5	24.2	43.1	22.7	36.7	8.4	
$3,000–$4,999										
NMCES		63.3	5.9	46.6	18.3	46.2	7.9		15.2	
FMES—both forms	70.5	62.2	9.9	44.3	19.9	47.3	19.1	46.3	9.7	12,434
Long form	73.0	62.7	9.3	44.1	20.2	48.5	18.6	46.5	6.1	
Short form	68.5	62.0	10.5	45.1	19.9	46.7	19.9	46.5	13.1	
$5,000–$6,999										
NMCES		67.7	5.6	52.8	21.8	50.0	7.0		15.5	
FMES—both forms	78.2	67.9	11.7	50.3	27.3	50.6	23.4	53.8	8.6	11,375
Long form	78.1	64.9	11.5	48.2	27.5	48.5	21.6	53.7	6.8	
Short form	78.5	70.9	11.7	52.3	26.8	52.4	25.5	54.2	10.5	
$7,000–$9,999										
NMCES		70.4	5.7	54.1	26.3	49.1	7.9		14.5	
FMES—both forms	82.0	71.0	12.0	53.2	30.8	48.2	20.9	55.6	7.7	20,650
Long form	80.7	68.3	12.2	50.6	30.0	45.6	19.8	57.1	4.7	
Short form	83.1	73.8	11.9	56.0	31.7	50.8	22.2	53.9	11.1	
$10,000–$14,999										
NMCES		72.4	5.0	56.2	29.7	49.6	8.1		14.1	
FMES—both forms	89.3	79.1	12.4	58.5	40.2	53.9	22.3	60.9	8.4	38,163
Long form	89.9	78.9	13.0	59.7	40.1	54.0	22.9	63.0	5.7	
Short form	88.6	79.1	11.7	57.0	40.2	53.5	21.5	58.6	11.3	
$15,000–$24,999										
NMCES		80.0	4.6	61.1	35.5	52.8	9.5		15.7	
FMES—both forms	92.5	85.0	11.8	65.8	50.5	57.1	24.9	60.7	8.0	60,882
Long form	93.5	86.2	12.4	65.8	51.7	57.3	25.0	62.5	6.6	
Short form	91.5	84.0	11.2	65.9	49.4	57.0	24.7	59.0	9.3	
$25,000 and over										
NMCES		82.0	3.8	61.0	46.2	51.2	11.0		16.2	
FMES—both forms	93.6	86.1	10.8	64.0	58.1	56.9	28.1	64.1	8.4	53,852
Long form	92.6	85.1	10.3	63.7	58.1	55.8	26.5	63.9	5.4	
Short form	94.5	86.9	11.3	64.2	58.0	57.9	29.6	64.2	11.2	

[a]Relative standard errors for estimates from the HIS-FMES may be found in Appendix I of Series 10, No. 122 of Vital and Health Statistics, Data from the National Health Interview Survey. The relative standard error for an estimate from either the short form or the long form is approximately 1.36 times greater than the relative standard error for both forms combined. Relative standard errors for estimates from NMCES may be found in the Technical Notes sections of the Data Preview Series for the National Health Care Expenditures Study, NCHS, 1979.
[b]The "Black" category for NMCES includes some non-Black, non-White persons.

ments made in 1976 into 1977 in FMES, edit and imputation strategies, reporting of payments in FMES for services received prior to 1977, and in NMCES, inclusion of some amounts for which families were liable but which may not have been paid. Also, because of the complexity of NMCES data, estimates for some of the population subgroups shown must still be considered preliminary. Nonetheless, there is sufficient stability in the major out-of-pocket estimates for each service component to permit a general discussion of similarities and differences in the patterns of estimates from the two surveys.

Persons with out-of-pocket expense. While differences between NMCES and FMES estimates of the percent of all persons with any out-of-pocket expense excluding health insurance premiums were small (2.3%, 75.3% in

Table 2

Average out-of-pocket health expenses for persons with such expense by type of expense, according to NMCES and NHIS form type, sex, race, age, and family income, United States 1977[a]

Survey form type and sex, race, age, and family income	All types of expense		Health expense						
	Including insurance premium	Excluding insurance premium	Hospital	Doctor	Dental	Prescription medicine	Optical	Health insurance premium	Other
	Percent of Persons with Out-of-Pocket Expense								
All persons									
NMCES		205	409	101	113	38	72		80
FMES—both forms	336	276	375	122	113	66	81	139	112
Long form	332	275	385	116	117	65	82	135	156
Short form	340	275	358	127	108	67	80	143	87
Under 17 years									
NMCES		124	284	58	117	17	59		48
FMES—both forms	200	160	215	71	89	29	64	93	78
Long form	198	162	231	72	90	28	61	89	137
Short form	201	159	199	71	87	30	66	98	59
Under 6 years									
NMCES		97	339	68	45	18	37		41
FMES—both forms	196	154	227	87	45	31	46	94	70
Long form	187	147	182	90	46	29	49	91	175
Short form	203	159	263	84	42	33	43	98	33
6–16 years									
NMCES		136	234	53	124	17	60		50
FMES—both forms	201	163	205	63	95	28	65	93	82
Long form	203	168	265	62	96	27	62	88	111
Short form	201	160	145	64	94	28	68	98	74
17–44 years									
NMCES		190	395	101	102	26	71		72
FMES—both forms	298	250	354	118	103	46	84	120	73
Long form	287	243	365	110	108	44	86	116	96
Short form	308	257	343	126	99	49	82	124	61
45–64 years									
NMCES		226	510	123	125	55	76		85
FMES—both forms	458	363	439	154	146	96	84	184	104
Long form	457	367	412	149	152	97	85	179	101
Short form	459	359	470	159	139	95	83	189	106
65 years and over									
NMCES		326	390	155	127	78	75		120
FMES—both forms	539	447	569	180	144	127	80	183	204
Long form	548	451	616	162	148	124	82	181	299
Short form	530	432	464	197	137	130	77	187	131
Sex									
Male									
NMCES		175	373	89	109	34	68		83
FMES—both forms	306	252	384	115	110	60	78	136	119
Long form	307	251	402	110	117	58	78	134	148
Short form	305	250	351	119	103	62	79	138	101
Female									
NMCES		230	429	110	116	42	74		79
FMES—both forms	363	297	368	127	115	70	83	142	107
Long form	355	296	371	120	117	70	85	136	162
Short form	371	297	364	134	113	71	81	148	76
Race									
White									
NMCES		206	380	101	114	39	71		81
FMES—both forms	342	275	361	120	113	66	80	142	115
Long form	336	274	365	113	117	65	80	137	160
Short form	348	276	352	127	108	67	80	147	88

Table 2 continued

Survey form type and sex, race, age, and family income	All types of expense		Health expense						
	Including insurance premium	Excluding insurance premium	Hospital	Doctor	Dental	Prescription medicine	Optical	Health insurance premium	Other
	Percent of Persons with Out-of-Pocket Expense								
Black[b]									
NMCES		197	671	104	97	35	78		79
FMES—both forms	270	286	500	141	105	66	93	108	61
Long form	300	305	618	153	106	63	111	109	59
Short form	345	259	371	126	100	68	78	108	62
Income									
Less than $3,000									
NMCES		289	940	134	132	48	82		122
FMES—both forms	362	306	558	154	93	88	79	169	93
Long form	348	315	667	135	87	93	73	144	104
Short form	374	293	392	176	98	80	85	197	82
$3,000–$4,999									
NMCES		240	485	116	111	62	75		89
FMES—both forms	363	308	473	142	79	96	81	167	83
Long form	367	332	712	138	82	86	104	165	89
Short form	360	284	269	144	77	105	72	170	81
$5,000–$6,999									
NMCES		259	519	141	110	57	75		92
FMES—both forms	376	361	588	156	129	90	79	158	101
Long form	356	340	449	133	149	83	79	143	137
Short form	395	361	609	171	108	97	80	173	75
$7,000–$9,999									
NMCES		220	348	118	118	49	69		76
FMES—both forms	357	292	389	134	120	86	91	151	124
Long form	355	294	442	117	116	82	108	150	115
Short form	362	290	316	153	125	91	73	154	133
$10,000–$14,999									
NMCES		210	407	107	109	39	75		102
FMES—both forms	312	253	342	118	95	60	75	133	82
Long form	304	242	326	110	92	60	70	132	110
Short form	321	266	365	128	99	61	81	134	68
$15,000–$24,999									
NMCES		175	304	90	97	31	68		64
FMES—both forms	298	242	253	109	107	52	77	120	79
Long form	296	244	256	111	110	53	76	115	82
Short form	301	240	250	108	104	53	79	124	78
$25,000 and over									
NMCES		198	340	89	127	31	72		72
FMES—both forms	351	279	409	116	126	50	84	138	93
Long form	341	275	423	115	138	50	89	131	127
Short form	358	282	401	117	114	53	80	144	76

[a] Relative standard errors for estimates from the HIS-FMES may be found in Appendix I of Series 10, No. 122 of Vital and Health Statistics, Data from the National Health Interview Survey. The relative standard error for an estimate from either the short form or the long form is approximately 1.36 times greater than the relative standard error for both forms combined. Relative standard errors for estimates from NMCES may be found in the Technical Notes sections of the Data Preview Series for the National Health Care Expenditures Study.

[b] The "Black" category for NMCES includes some non-Black, non-White persons.

NMCES and 77.6% in FMES) (see Table 1) and there was no difference for physician expenses, the NMCES estimate of all persons with out-of-pocket expense for hospital services was more than twice that in FMES (4.7% in contrast to 11.3%).

There also were varying differences for dental, prescription, optical, and other expenses, from about 15% for persons with optical expenses to 2.9% for those with prescription expenses, with NMCES estimates lower than FMES estimates for all except persons with "other" expenses.

These differences and similarities remained roughly the same when age, sex, race, and family income were considered. For example, for each of the comparisons of the percent of persons with physician expenses between NMCES and FMES with respect to these characteristics, there were no statistically significant differences. Similarly, nearly all differences between NMCES and FMES

Table 3
Rank orders of percent of persons with out-of-pocket expense, by type of expense, according to NMCES and NHIS-FMES, within selected socio-demographic characteristics: United States, 1977

Age, sex, race, and family income	Hospital		Doctor		Dental		Prescription medicine		Optical		Other		Total exc. ins.	
	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES
Age														
Under 6 years	2	4	4	4	1	1	3	3	1	1	1	2	2	2
6–16 years	1	1	1	1	5	5	1	1	2	2	2	1	1	1
17–44 years	3	2	2	2	3	4	2	2	3	3	3	3	3	3
45–64 years	4	3	3	3	4	3	4	4	5	5	4	4	4	4
65 years and over	5	5	5	5	2	2	5	5	4	4	5	5	5	5
Sex														
Male	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Female	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Race														
White	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Black	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Family income														
Less than $3,000	3	1	1	1	3	2	1	1	2	3	2	2	1	1
$3,000–$4,999	7	2	2	2	1	1	2	2	3	1	4	7	2	2
$5,000–$6,999	6	4	3	3	2	3	5	4	1	5	5	6	3	3
$7,000–$9,999	5	6	4	4	4	4	3	3	4	2	3	1	4	4
$10,000–$14,999	4	7	5	5	5	5	4	5	5	4	1	5	5	5
$15,000–$24,999	2	5	7	7	6	6	7	7	6	6	6	3	6	6
$25,000 and over	1	3	6	6	7	7	6	6	7	7	7	4	7	7

in the percent of persons with prescription medicine expenses according to the four socio-demographic characteristics were small and usually less than 4%; most were not statistically significant. The only exception was for persons with a family income of $25,000 or more, where the NMCES estimate of persons with out-of-pocket expenses for prescriptions was almost 6% lower.

Also, with the exceptions noted, the rank orders of persons with out-of-pocket expenses were comparable across subgroups in both surveys (Table 3). (Caution should be exercised in comparing rank orders between the surveys however because, although an estimate may have a rank higher or lower than another estimate, the estimates themselves may not be significantly different from each other.) Thus, for total expenses excluding health insurance premiums and for physician, dental, and prescription medicine expenses, the rankings among the four socio-demographic groupings employed (age, sex, race, and family income) were almost identical in both surveys. For example, persons 6 to 26 years of age were least likely and those 65 years and over were most likely to have had an out-of-pocket expense for physician services. By contrast, while rankings for optical, hospital, and "other" expenses were nearly the same for age, sex, and race, there were some dissimilarities between NMCES and FMES according to income. For example, persons with family incomes of $3,000–$4,999 were most likely to have out-of-pocket hospital expenses according to NMCES but second least likely according to FMES.

Average out-of-pocket health expenses. Estimates of average amounts of out-of-pocket expenses also differed between NMCES and FMES but the pattern differed in some instances from that for the percent of persons with expenses. Here, for all expenses excluding premiums (Table 2), the NMCES estimate was lower than the FMES estimate ($205 and $276, respectively) and this was also found for physician, prescription, optical, and "other" expenses. For out-of-pocket physician expenses, the NMCES was 83% of the FMES estimate of $122 and for prescriptions the NMCES estimate was as little as 58% of the FMES estimate of $66. Similarly, the estimate for optical expenses in NMCES was 89% of that in FMES and 71% for "other" expenses. In contrast, the NMCES estimate for out-of-pocket hospital expenses was about 1.1 times the FMES estimate ($409 versus $375). Only the estimates of average out-of-pocket dental expenses were the same.

Again, the direction of differences between survey estimates of average out-of-pocket expenses remained the same among most population subgroups. Hospital expenses were an exception within all population subgroups except race as were dental expenses for persons 6 to 18 and for persons in the less than $3,000 income category.

Despite the differences in estimates of average expenses for all persons, which were particularly noticeable for prescription medicines, the rank orders of out-of-pocket expenses were again similar or identical across groups defined by age, sex, race and income (Table 4). A

Table 4

Rank orders of average out-of-pocket expenses for persons with such expense, by type of expense, according to NMCES and NHIS-FMES, within selected socio-demographic characteristics: United States, 1977

Age, sex, race, and family income	Hospital		Doctor		Dental		Prescription medicine		Optical		Other		Total exc. ins.	
	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES	NMCES	FMES
Age														
Under 6 years	2	2	2	2	1	1	2	2	1	1	1	1	1	1
6–16 years	1	1	1	1	3	2	1	1	2	2	2	3	2	2
17–44 years	4	3	3	3	2	3	3	3	3	4	3	2	3	3
45–64 years	5	4	4	4	4	5	4	4	5	5	4	4	4	4
65 years and over	3	5	5	5	5	4	5	5	4	3	5	5	5	5
Sex														
Male	1	2	1	1	1	1	1	1	1	1	2	2	1	1
Female	2	1	2	2	2	2	2	2	2	2	1	1	2	2
Race														
White	1	1	1	1	2	2	2	2	1	1	2	2	2	1
Black	2	2	2	2	1	1	1	1	2	2	1	1	1	2
Family income														
Less than $3,000	7	6	6	6	7	2	4	5	7	3	7	5	7	5
$3,000–$4,999	5	5	4	5	4	1	7	7	6	5	4	3	5	6
$5,000–$6,999	6	7	7	7	3	7	6	6	5	4	5	6	6	7
$7,000–$9,999	3	3	5	4	5	5	5	4	2	7	3	7	4	4
$10,000–$14,999	4	2	3	3	2	3	3	3	4	1	6	2	3	2
$15,000–$24,999	1	1	2	1	1	4	2	2	1	2	1	1	1	1
$25,000 and over	2	4	1	2	6	6	1	1	3	6	2	4	2	3

possible exception was found within income categories for several service components, although in most instances the ranked estimates themselves were not different by statistically significant margins.

Ratios of differences. To provide a more detailed overview of relative differences between the two surveys, ratios of NMCES estimates to FMES estimates were calculated (Table 5). For the percent of persons with out-of-pocket expenses for dental and optical services, these ratios were .83 and .38 respectively; the ratio for hospital services was .42 and for other health services it was 1.9. In other words, NMCES estimates were substantially lower than FMES estimates for the percent of persons with out-of-pocket hospital and optical expenses and substantially higher for "other" expenses. As mentioned before, these patterns of relative differences in survey estimates remained fairly stable across population subgroups. Exceptions were found among some age groups and for dental expenses according to income. For instance, the ratio between estimates for all persons with hospital expenses was .42, .23 for persons less than 17 years old but .67 for persons 65 years or older. Nearly all ratios by race, sex, and income were statistically indistinguishable.

The relative differences between NMCES and FMES estimates of average amount of expense (also shown in Table 5) are likewise maintained across most population subgroups. Statistically significant differences in these sociodemographic subgroups were found for hospital expenses by age, race, and income and for dental expense by age and income.

Comparison of estimates from the FMES short and long forms

Estimates. There were no differences between the short and long forms in the reported percent of persons with expense for any type of expense except for "other" expenses (Table 1). Although there were other numerically large differences for some population subgroups estimates, the estimated sampling errors were too large to infer real differences. The short form estimate for the "other" expense component was 67% higher than the long form estimate (10.2% and 6.1%, respectively). The listing of examples under types of expenses on the long form may have led respondents to report more expenses under the specific types of expenses rather than putting them into the residual "other" category.

Short and long form total estimates of average out-of-pocket expenses likewise did not differ for all types of expenses combined, both including and excluding health insurance premiums (Table 2), as well as for hospital, dental, prescription, and optical expenses and health insurance premiums. Estimates did differ for physician and for "other" expenses. The short form estimate of physician expenses was 9% higher than the long form estimate ($127 and $116, respectively). The short form estimate for average "other" expenses was about half that for the long form ($87 versus $156).

Table 5
Ratios of NMCES to NHIS-FMES percent of persons with out-of-pocket health expenses and average expenses for persons with such expense by type of expense, according to sex, race, age, and family income: United States, 1977

Age, sex, race, and family income	Hospital		Doctor		Dental		Prescription medicine		Optical		Other		Total exc. ins.	
	%	$	%	$	%	$	%	$	%	$	%	$	%	$
All persons	.42	1.09	.98	.83	.83	1.00	.95	.58	.38	.89	1.87	.71	.97	.74
Under 17 years	.23	1.32	.97	.82	.82	1.31	.93	.59	.36	.92	2.05	.62	.99	.78
Under 6 years	.25	1.49	1.00	.78	.65	1.00	.99	.58	.10	.80	1.62	.59	1.03	.63
6–16 years	.21	1.14	.95	.84	.84	1.31	.90	.61	.37	.92	2.24	.61	.98	.83
17–44 years	.42	1.12	.96	.86	.82	.99	.93	.57	.40	.85	2.10	.99	.96	.76
45–64 years	.46	1.16	1.00	.80	.85	.86	.96	.57	.41	.90	1.74	.82	.96	.62
65 years and over	.67	.69	1.05	.86	.88	.88	1.01	.61	.34	.94	1.67	.59	.99	.73
Sex														
Male	.35	.97	.95	.77	.80	.99	.91	.57	.36	.87	1.75	.70	.95	.69
Female	.48	1.17	1.00	.87	.86	1.01	.97	.60	.40	.89	1.97	.74	.99	.77
Race														
White	.41	1.05	.98	.84	.84	1.01	.95	.59	.39	.89	1.90	.70	.97	.75
Black	.45	1.34	.93	.74	.68	.92	.87	.53	.37	.84	1.65	1.29	.97	.69
Family income														
Less than $3,000	.58	1.68	1.05	.87	1.07	1.42	.93	.55	.32	1.04	1.85	1.31	1.01	.94
$3,000–$4,999	.60	1.02	1.05	.82	.92	1.40	.98	.65	.41	.93	1.57	1.07	1.02	.78
$5,000–$6,999	.40	.88	1.05	.90	.80	.85	.99	.63	.30	.95	1.80	.91	1.00	.72
$7,000–$9,999	.48	.89	1.02	.88	.85	.98	1.02	.57	.38	.76	1.88	.61	.99	.75
$10,000–$14,999	.40	1.19	.96	.91	.74	1.15	.92	.65	.36	1.00	1.68	1.24	.92	.83
$15,000–$24,999	.39	1.20	.93	.83	.70	.91	.92	.60	.38	.88	1.96	.81	.94	.72
$25,000 and over	.35	.83	.95	.77	.80	1.01	.90	.62	.39	.86	1.93	.77	.95	.71

Differences for population subgroups in estimates of average expenses were found for hospital expenses of blacks and persons in the $3,000–$4,999 income group. The long form estimate for blacks for average hospital out-of-pocket expenses was 67% higher ($618 versus $371) and the long form estimate for the $3,000–$4,999 income group was 2.6 times the short form estimate. It is interesting to note that the long form estimate for blacks for physician expenses was 21% higher than the short form estimate, although this difference is not statistically significant.

The discrepancy of estimates of average physician expenses may be related to differences in the questionnaire instructions. Whereas both the long and short form instructions include amounts paid to doctors while a patient is in the hospital, the long form does not specifically use the word "surgeon," while the short form does. In addition, in the short form the major thrust of the instruction for doctor expenses was to include physician fees whereas this was not the central idea on the long form.

The larger estimate of the long form for other expenses for all persons may be related to the reporting of nursing home expenses. The long form instructions were to include such expenses under "other" whereas the short form instructions made no specific mention of a category in which nursing home expenses should be reported.

Selected methodological characteristics. Table 6 shows a comparison of short and long form response rates, completion statuses, respondent use of records, head of household participation in form completion and interviewer completion of form.

Supplement response rates. Overall the short form had a higher response rate than the long form (88.6% and

Table 6
Selected methodological characteristics of the 1978 FMES by form type (In percent)

	Long form		*Short form*
Supplement response	87.2	(.3)[a]	88.6
Item response			
All types of expense including insurance	76.2	(.5)	77.7
All types of expense excluding insurance	81.2	(.5)	81.3
Hospital	84.0	(.4)	85.7
Doctor	84.2	(.4)	85.7
Dental	84.6	(.4)	86.1
Prescription medicine	83.9	(.4)	85.3
Optical	84.2	(.4)	86.0
Health insurance premium	79.7	(.5)	83.0
Other	83.2	(.4)	84.5
Selected completion statuses			
Forms received before second mailing—pass edit	55.0	(.7)	55.4
Received after first or second mailing—pass edit	66.3	(.7)	69.7
Failed edit	7.3	(.3)	4.3
Form not received—telephone completion	11.7	(.5)	12.2
Use of records			
Referred to records for all amounts	22.7	(.5)	26.0
Referred to records for some amounts	18.3	(.7)	27.3
Did not refer to records	59.0	(.7)	46.7
Participation of the head of the household in filling forms	65.7	(.7)	69.8
Interviewer completed form	2.8	(.1)	1.7

[a]Approximate standard errors for both the long form and the short form are shown in parentheses.

87.2%, respectively). Total completion and total nonresponse rates were not significantly different by form type for any of the population subgroups examined. Subgroups were defined by family income, race of head, education of head, place or residence, geographic region, and size of family.

Item response rate. For all persons, for each item, the short form obtained a higher percent of persons reporting known amounts. However, the differences were statistically significant only for the optical expenses, 86.0% for the short versus 84.2% for the long, and for the health insurance expenses, 83.0% versus 79.7%. For population subgroups, the short form rates as well were usually higher, although not significantly so. Only for persons aged 65 years or older were the short form item response rates lower, with a statistically significant margin observed for health insurance.

Selected completion statuses. The proportion of short forms received after the first or second mailing that passed edit criteria was significantly higher than the proportion of long forms (69.7% versus 66.3%). The proportions received before the second mailing and passing edit were about the same for both form types. The percentage of forms failing edit was significantly higher for the long form (7.3% versus 4.3% of the short forms). Telephone completions were judged undesirable from the standpoint of cost and perceived lower quality of responses. The percentage completed over the telephone was about the same for both forms, 12%.

Use of records. A higher proportion of short form respondents reported referring to records for all amounts (26% versus 23%). In addition, a higher proportion of short form respondents reported referring to records for not all but some amounts (27% versus 18%). Conversely, 59% of long form respondents versus 47% of short form respondents reported not referring to records at all.

Participation of head of household in filling out forms. For 69.8% of short-form households, the head of the household participated in filling out the supplement as compared to 65.7% for the long forms, possibly contributing to the accuracy of reporting.

Interviewer completed form. For 2.8% of the long form households versus 1.7% of short form households, the interviewer filled in the form for the family. Interviewer completion of the FMES adds to the overall costs of the survey. More long form respondents may have requested the interviewer's help upon seeing the lengthy long form. In addition, interviewers may have been more inclined to help respondents in what the interviewers may have perceived as a more difficult form to complete.

Discussion

As expected there were differences between the NMCES and FMES estimates of the percent of persons with out-of-pocket expenses and average expenses for persons with expense that resulted from different approaches to survey design, definition of expense components, reporting period methods of accounting, and editing and imputation procedures.

Another possible explanation for differences in the estimates may be related to reporting of use of services. While it is not possible to examine this issue in FMES, in NMCES if a received service were not reported there would be no chance for the out-of-pocket expenditures for that service to be included in the estimates. The NHIS estimate is considered a more sensitive measure because of the use of a shorter recall period, two weeks versus about three months for NMCES. For this reason a comparison of 1977 NMCES and NHIS use estimates was made (Table 7). While the estimates of the percent of persons with one or more hospital stays in 1977 were statistically the same (11.1% in NMCES and 10.4% in NHIS), the NMCES estimate of dental visits per person for 1977, on the other hand, was significantly lower than the NHIS estimate (1.3 and 1.6, respectively). Fewer reported dental visits in NMCES could have resulted in a lower estimate of percentage of persons with dental expenses. Similarly, the estimated number of physician visits per person for 1977 from NMCES was also significantly lower than the estimate for NHIS (4.0 versus 4.8), although both surveys produced about the same estimate of percentage of persons with out-of-pocket physician expenditures. However, it is believed that some FMES respondents may have reported hospital outpatient and emergency room physician expenses under the hospital expense component, which may have reduced the FMES estimate of percent of persons with out-of-pocket physician expenses.

Since there is no available criterion measurement for evaluating the accuracy of the data from the two surveys the most important criterion is whether or not the policy conclusions would be different from one survey to the other. From this standpoint the most important estimates from these data are the relative burden of out-of-pocket medical expenses on different population subgroups. For this determination, total expenditures not affected by definitional differences between the two surveys were examined. As seen earlier the rank orders according to age, sex, and income were the same for both surveys. Furthermore, the differences in the estimates for total out-of-pocket expenses were not so large as to affect most policy decisions about the burden of health care expenses on relevant population subgroups.

Even with favorable future budget allocations, it is unlikely that a panel health expenditure survey will be repeated in the near future. Thus, refielding of FMES within the next two or three years seems reasonable based on the evidence presented in this paper and on the need for out-of-pocket expenditure data in the face of changes in the government policy relating to public financing of health care.

On every measure of quality of performance examined the FMES short form did as well as or better than the long form. The few differences in estimates produced by the two forms are not problematic. For these reasons, given the lower costs of fielding the short form, this form is recommended for the next out-of-pocket expenditures data collection effort.

A few changes in the short form would be likely to improve the accuracy of the estimates. First is the inclusion of an additional expense item for hospital expenses for other than overnight stays, including outpatient and emergency room fees but still excluding physicians' fees. The hospital question would be changed to include expenses only for overnight stays. Another concern relates to nursing home expenses. Only the FMES long form has the instruction explicitly stating that such expenses should be placed in the "other" expenses category; the short form may need to have a similar instruction to prevent such expenses from being included under hospital expense. Another issue is the accounting period used. For the short form the accounting period parallels that used for reporting income and deductions for income tax purposes and should result in negative amounts for expenses for some persons. Amounts received in 1977 from insurance payments for care received in 1976 could be greater than amounts paid in 1977. However, the forms design and data processing procedures did not take this possibility into account. Consideration should be given to including a separate questionnaire item for amounts received from insurance during the calendar year for services received before the start of the year.

Table 7
Selected measures of health care use by NMCES and NHIS: United States, 1977

	NMCES		NHIS	
Hospitalization				
Number of discharges per 100 persons per year	15.0	(.4)	14.0	(.2)[a]
Average length of stay in days	7.0	—[b]	7.8	(.3)
Percent of persons with one hospital episode or more	11.1	(.3)	10.4	(.1)
Dental visits				
Number per person per year	1.3	(.04)	1.6	(.03)
Percent of persons with visits in past year	41.1	(.5)	49.7	(.2)
Physician visits				
Number per person per year	4.0	(.07)	4.8	(.05)
Percent of persons with visits in past year	72.8	(.4)	75.1	(.2)

[a]One standard error of estimate shown in parentheses.
[b]Not available.

Implications for future research

This paper is based on the premise that differences between NMCES and FMES estimates of persons with

out-of-pocket health care expenses and of average out-of-pocket expenses are related to definitional differences, methods of constructing sample weights, and methods of data editing and cleaning, because both surveys sampled the same population. The effects of alternative strategies with respect to these procedures must therefore be assessed to permit a judgment of the relative effectiveness and efficiency of the two surveys. First are the effects of alternative editing strategies in FMES with regard to an allocation of a zero dollar value to a missing entry. (As explained earlier, the current FMES editing rules exclude such cases from the estimation procedures.) Alternative post-stratification techniques in FMES should also be tested for supplement nonresponse, because such techniques are assumed to affect survey estimates due to the variation in supplement nonresponse rates for certain subgroups of the population. In addition, NMCES expenditures should be reclassified in order to reduce categorical differences in the two surveys, particularly with regard to hospital, physician, and optical expenses. Also, the out-of-pocket health insurance premium data collected in FMES should be compared with the NMCES household and verification data.

Comparisons of out-of-pocket expense estimates from the 1980 National Medical Care Utilization Expenditure Survey (NMCUES) should be instructive. NMCUES, jointly sponsored by the NCHS and the Health Care Financing Administration, employed a similar panel survey design as NMCES, had the same target population, and was based on concepts of out-of-pocket expenses comparable to those in both NMCES and NHIS-FMES. Thus, although NMCUES did not provide record check studies of providers and third-party payers other than for Medicaid and Medicare, comparisons of the estimates from the three surveys should provide useful insights for planners of future health expense surveys.

Discussion: Methodological issues in the measurement of health policy outcomes *and* A comparison of estimates of out-of-pocket expenditures for health services

Lois A. Monteiro, Brown University

We have been presented with very interesting and meaty papers to begin this methods conference. The basic questions that these two papers address are first, how to get the most accurate measure of certain phenomena (Medicaid participation and medical expenditures are today's examples) and, second, whether simpler and easier to obtain methods can be used equally well as more complicated measures and if so, how much will be lost in accuracy. Given that as participants here we all have a vested interest not only in accuracy but also in those intellectually challenging complicated measures that we find most appealing, the question for this discussion almost becomes: Does the emperor wear clothes? Both the papers come to somewhat the same conclusion, that measures of *perceptions* of these phenomena are not completely accurate; that is, they do not equal the more "objective" measure of the same phenomena, but that this much inaccuracy may be tolerable for almost all occasions. The conclusion is in a sense that the simpler, easier method may not be so bad after all. This is an especially welcome word in an era of fiscal cutback! Parenthetically, one can wonder whether six or eight years ago the conclusion drawn might not have been that even a relatively small gain in accuracy would be worth the expense—but so be it.

Let me begin my comments in the order of presentation, with the Kletke paper first. It seems to me that the authors have been very thorough in presenting and in ruling out the possible sources of response error for the differences between perceived participation in Medicaid and actual payment by Medicaid for these pediatricians. But I think they have left out at least one possibility that may help to explain it. The self-reported estimates were obtained from an interview with the physician in which he or she was asked what percent of patients were paid for by Medicaid. I would suggest that the overreports of participation may be due to the subjects giving a positively valued response to the question—that this is a value-laden question, that participation is "socially good," and that the physicians could make themselves look like good guys by increasing their estimates. This might be especially true if the rest of the interview or its purpose had to do with physician income. (The authors don't tell us the context in which the question was asked except that it was a Medicaid participation study.) Clearly, the doctors who didn't participate—who already had made a decision about it—were accurate in that none of them had a Medicaid-paid patient. While my suggestion may seem far-fetched, it might be looked at, since other, more logical sources for overestimation, such as delayed payments and administrative bother, did not prove to be the answer.

My second point is to look again at the amount of overstatement—half of the physicians overstated by 5%. That seems to me to be pretty close for the kind of estimate requested in the interview, a "guesstimate," and is comparable to the difference reported in the second paper, of 6.5% between the FMES and NCMES estimates of persons with out-of-pocket expenses for hospitals (when much greater accuracy might be expected).

The thorough regression analysis presented in the paper makes it clear that the self-reports and the objective reports both have the same relationship to a wide range of possible variables and make the case for the use of the self-reports. What the authors don't tell us is the relative cost and relative difficulty of getting the objective estimates—cost not only in funding dollars but also in respondent burden (of patient-physician encounter records) especially when one considers that the self-report question on estimated participation might be done by mail or telephone instead of by personal interview.

Lastly, one serendipitous finding from this study points out the problem in accurately measuring physician manpower. Of the 1,457 pediatricians selected for the sample from the AMA masterfile only 879 (or 60%) met the criteria of eligibility of (1) being in practice 20 hours per week, (2) practicing in the same community for one year, (3) not in a group of 10 or more doctors. In this case the simpler measure of physician manpower (the masterfile) is clearly not accurate enough to predict manpower.

Let us turn now to the Poe and Waldron paper on the measures of out-of-pocket expenditures. Again the authors are very thorough in searching out and discussing the sources of the response errors. Expenditures are a more complex and more difficult phenomena to measure, and that is reflected in the sophistication of the surveys that collected the data and in the number and variety of comparisons in the paper. Its detail is admirable, and it too reaches the conclusion that "the differences in estimates for total out-of-pocket expenses were not so large as to result in different policy conclusions." The less costly, even if slightly less accurate, measure of self-reported expenditures on a short form will do for most purposes. The authors do not give us figures on the relative cost of these studies in dollars nor the response burden costs of a panel study, although they do note that the short form is most cost efficient. The size of the difference in cost would, I am sure, be very large.

The authors focus on the accuracy of the different measures in predicting the percent of persons who had some out-of-pocket expense, rather than looking at the relative accuracy of the techniques in predicting the amount spent, although they give some information on the amounts in Table 3. Perhaps we should have some discussion of the relative need for accurate dollar amount figures and of the relative merits of the different measuring techniques if it is dollars that are important. It may be that the more costly collection effort is necessary to get accurate dollar figures.

With regard to the question of how frequently surveys of out-of-pocket estimates should be conducted, the authors do not point out that the percent of persons paying out of pocket is related to insurance coverage which is related to employment. The current high rate of unemployment nationally can be expected to have an effect on insurance coverage, use of services, and out-of-pocket expenses that might argue for a repeat of the FMES questions earlier than the two to three years that are suggested as reasonable if we want to learn more of the impact of the economic downturn on health care spending.

In summary I found both papers to be provocative and interesting in that they both contradicted what might have been expected—that more is not necessarily better.

Open discussion: session 1

Introduction

This session consisted of four papers. The unifying theme of these papers is the importance of measuring, explicitly, reliability and/or validity. Reliability refers to the reproducibility of results; validity has to do with the extent to which measurements reflect reality. Explicit indicators of reliability and validity are necessary for (1) judging the quality of data; (2) adjusting the analyses of the particular survey in which the indicators were developed; and (3) using these indicators to guide data collection and analysis in subsequent surveys.

The first paper, by Andrews, dealt with a generalized model for dealing with random and correlated errors. It was tested on data from six surveys.

The remaining three papers dealt with specific comparisons of alternative data sources. Brorsson considered the effects of interviewer variability in the Swedish survey of living conditions emphasizing the influence of the sex of the interviewer. Kletke compared physicians' overall assessments of the number of Medicaid patients in their practice with the results of a patient record check of those practices. Poe contrasted estimates of family out-of-pocket expenditures for health services from the continuing Health Interview Survey-Family Medical Expense Supplement (HIS-FMES) with those from the National Medical Care Expenditure Survey (NMCES)—a panel study explicitly designed to collect expenditure data.

General discussion of the Andrews and Brorsson papers

Andrews paper. Andrews responded to Singer's formal comments by emphasizing that his work did, indeed, concentrate on correlated errors in surveys that could be measured from the survey data themselves. He did not deal with bias estimates which depend on external criteria.

Sirken complained that the terminology used by social scientists differs from common usage of statisticians; communication is thus confused. He proposed that a long session or possibly a conference dealing with this problem would be helpful. To explicate his concerns, he noted that as a statistician he deals with rates. The denominators (e.g., people in different age groups) may be incorrect but perfectly correlated. Andrews agreed that his correlated-error approach would not uncover this problem but reemphasized the importance of correlated error in survey analysis.

Greenberg noted that Andrews's correlated-error approach would discover errors of a bivariate nature but not errors of the multivariate type, which would remain as biases. Andrews conceded that this was probably correct.

Sudman asked for clarification concerning the impact of the number of response categories. Andrews reemphasized that more categories appear to do better than three-answer categories. Andrews reemphasized that more categories were better than fewer, except that dichotomous response categories appear to do better than three-answer categories.

Groves asked about the effect of question positioning in the questionnaire on the correlated-error coefficient. Andrews noted that items placed toward the middle of the questionnaire had the smallest error coefficient. Groves indicated that he was particularly interested in the impact of alternative ordering of items—for example, differential sequencing in a scale or index. Andrews did not address this issue in his study.

Sudman inquired about the seeming unimportance of the particular topics covered on the correlated-error term. He wondered if the importance of a topic might increase with a broader range of topics, such as alcoholism. Andrews agreed this might well be the case.

Brorsson paper. Axelrod inquired about the implications of the results for selection of interviewers. Should male interviewers be excluded? In general, should emphasis be on selection or interviewer training? Brorsson felt that male interviewers should not be excluded and that emphasis should be placed on interviewer training. He noted that he certainly would not make a radical change in the sex composition of his interviewers in the middle of a data-collection session.

Verbrugge expressed concern about possible discrimination against male interviewers. She sees the interviewing field as one which should be opened to males as part of a general affirmative action movement.

Greenberg asked about the possible influence of interviewer motivation, whether this might explain some of the differences noted between the sexes. Brorsson replied that he did not have data in his study on this point, but that it was his impression that male interviewers in Sweden were more likely to see interviewing as a transitory activity between longer term jobs. Conversely, women were more likely to be committed to interviewing as a long-term career.

Rouse suggested that the sex results on interviewers depend on the topics covered. She felt that, generally, same-sex, same-race combinations tend to elicit more responses. However, it was suggested that this expectation is not always borne out. For example, some studies of contraceptive practices and gynecological treatment of women did not produce differential results according to sex of interviewer.

Sirken stated that the standardization approach employed by Brorsson is only one of the possible approaches to dealing with differential respondent characteristics. Another, which he favors, is to randomize respondents among interviewers. Given that this latter approach was not used in the Swedish study, Sirken asked Brorsson how he would suggest other researchers use his results. Brorsson replied that his findings should be replicated in other settings before being used to make decisions about interviewer selection.

General discussion of the Kletke and Poe papers

Kletke paper. Kletke responded to Singer's comments by agreeing that perceived appropriate social response might account for physicians' reporting that they had more Medicaid patients than their records suggest. He thought this might be particularly true for pediatricians, who tend to see themselves as more socially responsible than other kinds of physicians. He noted that in considering differences between physician self-reports and records, physicians' overestimates were likely to be more important at the extremes of the distribution. That is, they will be greater if doctors have either very few or very many Medicaid patients than if the proportion is somewhere in between. He said that he would like to look at the relative expense of collecting information directly from doctors as compared to using their records. However, this was impossible because of the multiple purposes of the study. Finally, it would be difficult to compare the relative expense of the two types of data since the aggregated patient record data are available for only 60% of the pediatricians who provided self-reports.

In discussing Kletke's paper, Kovar reiterated the point made in the paper that the units of analyses were not the same for the datasets compared. The doctors counted patients, whereas the records were based on visits. This difference could account for some of the discrepancy in the findings. She also emphasized that poor children on Medicaid would be underrepresented in this study of pediatricians in private practice.

Axelrod inquired if any analysis had been carried out according to sex of respondents. Kletke answered that it had not.

Greenberg wondered if physicians would know many of the details requested, suggesting that these details should have been asked of the business office rather than of the doctor. Kletke agreed and indicated that the physician's clerical staff probably did provide much of the information.

Poe paper. As co-author of the Poe paper, Walden responded to Singer. He noted that in a study with many purposes, like the National Medical Care Expenditure Study, it is very difficult to cost out particular parts of the study, such as comparing record checks with direct questions of physicians.

Sudman asked if the Consumer Expenditure Survey could be used to provide medical care expenditure estimates for the nation. Walden responded that this has been attempted but that the study does not seem ideally suited for that purpose.

Horowitz expressed concern that the recall period in the Health Interview Survey was one year. He asked why the recall period could not be shorter, as it is for use and morbidity information. Poe responded that there is an accounting problem due to the time lags between time of service, billing for service, and reimbursement by third parties. She also pointed out that reporting is sometimes tied to income tax records, which are not useful for short recall checks. Fuchsberg further suggested that deductibles cause a major problem for reporting out-of-pocket costs for short recall intervals because people do not know at the time of service how much of the total bill they will ultimately be responsible for during a deductible period. The problem is complicated when the deductible applies to the experience of the total family rather than of an individual.

Bonham questioned the ratio comparisons of NMCES to NHIS-FMES in Table 3. He noted that while the total expenditures ratios were reasonably close to 1 (.97 for percent of persons with out-of-pocket expenditures and .78 for average expenses for persons with expenses), the differences by type of expenditure were much greater. He particularly noted the greatly divergent ratios for hospital services, .42 for percent of persons and 1.40 for average expenses. He suggested that either the studies were not measuring the same things or a large error existed in one of them.

Walden agreed that these results were disturbing, but noted that the main problem was with hospital services; the differences for other services were much less. Poe said that one reason HIS-FMES might show more people being hospitalized is that some hospital outpatient ambulatory visits were counted as admissions. This would, in turn, lower the apparent; relative average expense per person of hospital inpatient services. Another difference between the studies is that HIS-FMES does not include the medical expenses of persons who died during the year preceding the interview whereas NMCES does.

SESSION 2:
Telephone survey methodology

Chair: Robert Groves, Survey Research Center, University of Michigan

Recorder: Morris Axelrod, Department of Sociology, Arizona State University

Estimating and adjusting for nonphone noncoverage bias using Center for Health Administration Studies data*

Martha J. Banks, Center for Health Administration Studies, University of Chicago

Ronald M. Andersen, Center for Health Administration Studies, University of Chicago

In the last decade, telephone survey techniques have come to be considered seriously when designing a high quality survey. Reasons for this include the rising cost of field work (especially travel costs), a trend toward lower response rates in personal interview surveys (often due to respondents' fears of allowing strangers into their homes), and the recognition that most population groups have a fairly high phone coverage rate.

There are a number of ways in which differences between results from telephone and face-to-face interviewing can be studied. Most of them require that special methodological studies be conducted. For example, a personal interview sample survey and a telephone sample survey can be conducted concurrently, using the same study questions. A comparison of the results allows for an examination of many aspects of in-person/phone differences. For example, see Groves and Kahn (1979) for a discussion of results from three national samples—a nonclustered phone sample, a clustered phone sample, and a clustered area sample. Comparison of results from THIS (Telephone Health Interview Survey) with those of HIS are found in Monsees and Massey (1979b). Jordan et al. (1979) present data from concurrently conducted personal and telephone portions of the Los Angeles Health Survey and compare results. Studies using the National Crime Survey and the Current Medicare Survey, both panel studies, allow assessment of differences in follow-up contact results (after an initial personal visit) by whether the follow-up was done in person or by phone (where possible). See Bushery et al. (1978) for further information.

However, another type of examination of differences between those with and without telephones can be conducted by performing secondary analyses on existing data from face-to-face surveys. During the course of many surveys done in person, interviewers ask respondents for their home telephone number so that the interview can be verified. Therefore a variable can be constructed which indicates whether the person has a home phone. Comparisons of results for those with and without home telephones should give some idea of the extent of and the effect of omitting the nonphone population in conducting a similar survey by telephone.

The data in this paper are from a national survey conducted during 1976 for the Center for Health Administration Studies. This project studied access to medical care in the United States. As part of this effort, Black Southerners living outside of SMSAs and those of Spanish heritage living in the Southwest were oversampled at about 3.4 to 1. Altogether, 7,787 persons in 5,432 families were interviewed. The overall response rate was 85%. For further information about the study, see Aday et al. (1980), especially Appendix A.

Table 1 presents the phone coverage for persons and families as obtained in the CHAS 1976 survey. The data suggest that, overall, about 10.1% of all families and 9.3% of all persons had no home phone in 1976.[1] Among the variables given in Table 1, financial status variables are the best predictors of telephone coverage. (Two measures of financial status are shown. Besides family income, a poverty status variable is provided which compares the family's income to the poverty level cut-off for that family. Therefore the poverty status variable takes family size into account and thus perhaps better measures the family's ability to pay for various good and services.) Family income and poverty status have the largest eta's among the ordinal variables (.235 and .232 respectively, with phone coverage as the dependent variable) and the highest uncertainty coefficient among either nominal or ordinal variables (.084 and .074 respectively). All the other variables in Table 1 have uncertainty coefficients between .001 and .052.

Besides the fact that phone coverage is positively correlated with income, it appears that groups with low phone coverage are Southerners (especially Southern Blacks living outside SMSAs), Southwestern Hispanics, persons whose family head was under 25, those divorced, separated, or never married, and those in one-person or seven-or-more-person families. People tend to have high phone coverage if they live in the Northeast or North Central, are non-Hispanic Whites, are 35 or older, or are members of families containing three or more adults.

The results are similar to those obtained in other data: Groves and Kahn (1979), Table 6.1; Monsees and Massey (1979b), Tables 1, 2, and 3; and Thornberry and Massey (1978). The only meaningful difference between their results and the CHAS 1976 data is that we report somewhat lower phone coverage in the West.

It should be noted that the phone coverage estimates obtained from a sample survey may differ from the true percents due to sampling error and to bias resulting from imputing a few responses, because nonrespon-

* The authors wish to thank the following CHAS staff for their assistance: Timothy Champney, Christopher Lyttle, Valerie Pape, Joyce Van Grondelle, and Tanya Winard.

Table 1
Phone coverage for families and persons, CHAS 1976, in percent

Characteristic	Families			Persons			
	Phone	Non-phone	Percent of U.S.	Phone[a]		Non-phone[a]	Percent of U.S.
Region							
Northeast	93.8	6.2	22.0	95.0	(0.8)	5.0	22.4
North central	95.5	4.5	22.9	95.6	(0.6)	4.4	30.6
South	82.3	17.7	32.1	83.4	(0.9)	16.6	32.7
West	89.7	10.3	16.0	90.1	(1.0)	9.9	14.4
Residence							
SMSA central city	88.7	11.3	28.1	90.3	(0.9)	9.7	25.6
SMSA other	92.8	7.2	35.4	93.4	(0.7)	6.6	36.7
NonSMSA urban	88.0	12.0	12.0	88.8	(1.3)	11.2	11.7
Rural nonfarm	86.6	13.4	19.0	86.4	(1.2)	13.6	20.1
Rural farm	93.5	6.5	5.4	94.2	(1.6)	5.8	5.9
Race							
Spanish heritage, southwest	75.4	24.6	3.2	73.8	(4.9)	26.2	4.1
Other white	92.0	8.0	85.5	92.9	(0.5)	7.1	83.8
NonSMSA southern black	63.2	36.8	2.1	60.6	(3.9)	39.4	2.4
Other nonwhite	81.5	18.5	9.2	86.3	(1.9)	13.7	9.6
Age							
0–5	NA	NA	NA	85.5	(1.6)	14.5	9.3
6–17	NA	NA	NA	90.3	(1.0)	9.7	24.6
18–34	NA	NA	NA	87.8	(1.0)	12.2	25.0
35–54	NA	NA	NA	93.1	(0.9)	6.9	21.8
55–64	NA	NA	NA	95.0	(1.1)	5.0	9.4
65 plus	NA	NA	NA	94.3	(1.0)	5.7	10.0
Age of head							
Under 25	74.9	25.1	9.1	74.2	(2.2)	25.8	6.3
25–34	87.2	12.8	20.2	88.0	(1.0)	12.0	21.7
35–44	90.9	9.1	17.7	91.0	(0.9)	9.0	26.5
45–54	93.4	6.6	18.1	94.0	(0.8)	6.0	21.0
55–64	93.1	6.9	15.6	94.2	(1.0)	5.8	12.7
65 plus	93.1	6.9	19.3	94.2	(1.0)	5.8	11.8
Sex of head							
Male	90.4	9.6	77.1	91.5	(0.5)	8.5	85.2
Female	88.4	11.6	22.9	86.1	(1.2)	13.9	14.8

dents may have different phone coverage than do respondents, or because respondents may misreport phone coverage—either underreporting because they do not want to be bothered further or overreporting because having a phone perhaps is more socially acceptable than not having one. However the coverage estimates should help give a fairly good picture of nonphone noncoverage bias, which is the intent of this paper.

Although information about how socioeconomic groups differ in their phone coverage is interesting, it should be kept in mind that these differences may or may not affect the differences between the phone population and the entire population in terms of the health characteristics of principal interest. The magnitude of observed health differences between the phone population and the entire population depend on the magnitude of health differences between the phone and nonphone population as well as on the coverage rates. It may be inappropriate to use coverage rates as proxies for noncoverage bias measures. (For example, one population subgroup may have quite a bit lower phone coverage than do other population groups, but have little or no more nonphone noncoverage bias.) Therefore we should, where possible, directly assess the impact of noncoverage rather than use coverage rates as proxies.

Table 2 is the first of several tables showing the effect on health care estimates when those without phones are omitted. The table gives estimates of the percent of the population who have contacted a doctor during the preceding year. Based on all persons, 76.7% of the population contacted a doctor. The figure for those in telephone households is 77.6%. The ratio of the two, .988 (given in column 5), is significant at the 5.0—standard error level.

Examining the fifth column of Table 2 shows that there are no population groups given in which the ratio of the total estimate to the phone estimate is significantly greater than 1.000. All of the ratios are either about 1.000 or significantly below it. The distribution of significance levels is as follows:

Significance level	Ratio below 1.000	Ratio above 1.000
Less than 1 standard error	12	5
1 to 1.6 standard errors	11	-
1.6 to 2 standard errors	8	-
2 to 3 standard errors	12	-
3 standard errors or above	8	-

Table 1 continued

Characteristic	Families			Persons			
	Phone	Non-phone	Percent of U.S.	Phone[a]		Non-phone[a]	Percent of U.S.
Marital status of head							
Married	92.6	7.4	67.2	92.3	(0.5)	7.7	80.8
Widowed	91.3	8.7	12.6	92.0	(1.3)	8.0	6.8
Divorced	85.0	15.0	7.6	82.6	(2.2)	17.4	5.4
Separated	74.1	25.9	4.9	73.7	(3.0)	26.3	3.5
Never married	79.3	20.7	7.8	80.2	(2.6)	19.8	3.5
Family size							
1	83.8	16.2	22.2	83.8	(1.5)	16.2	7.6
2	92.2	7.8	28.6	92.2	(0.9)	7.8	19.5
3	89.9	10.1	16.6	89.9	(1.1)	10.1	17.0
4	93.7	6.3	14.9	93.7	(0.8)	6.3	20.3
5	93.2	6.8	9.4	93.2	(1.1)	6.8	16.0
6	92.5	7.5	4.3	92.5	(1.6)	7.5	8.9
7 or more	83.7	16.3	4.0	83.2	(2.2)	16.8	10.7
Adults in family							
1	83.7	16.3	28.0	82.9	(1.2)	17.1	13.9
2	91.4	8.6	53.1	90.6	(0.6)	9.4	55.9
3	95.4	4.6	13.1	95.0	(0.9)	5.0	18.6
4 or more	94.8	5.2	5.8	93.6	(1.4)	6.4	11.6
Family income							
Less than $3,000	75.8	24.2	8.2	71.8	(2.4)	28.2	5.3
$3,000–$4,999	79.9	20.1	11.4	74.9	(1.9)	25.1	8.9
$5,000–$6,999	85.7	14.3	11.2	83.1	(1.7)	16.9	9.4
$7,000–$9,999	87.5	12.5	13.5	87.3	(1.4)	12.7	12.6
$10,000–$14,999	93.3	6.7	24.1	93.9	(0.8)	6.1	25.5
$15,000–$24,999	97.3	2.7	21.5	97.5	(0.5)	2.5	25.9
$25,000 or more	97.2	2.8	10.0	98.5	(0.6)	1.5	12.4
Poverty status							
Below poverty	74.4	25.6	13.5	71.3	(1.6)	28.7	14.5
100%–125% poverty	83.9	16.1	6.8	86.4	(1.9)	13.6	6.7
125%–200% poverty	89.1	10.9	19.4	90.8	(1.0)	9.2	20.1
200%–300% poverty	92.0	8.0	22.3	94.3	(0.8)	5.7	24.0
300%–400% poverty	94.5	5.5	16.8	96.4	(0.7)	3.6	16.6
400% or more poverty	96.7	3.3	21.2	97.7	(0.6)	2.3	18.1
Total	89.9	10.1	100.0	90.7	(0.4)	9.3	100.0

[a]Numbers in parentheses are the standard error estimates for both the phone and the nonphone populations.

Therefore using data from only the phone population would tend to overstate the percent seeing or speaking with a physician.[2]

On the other hand, comparing population groups using data for only those with phones would result in conclusions nearly identical to those based on comparing population groups using data for all persons. Both groups of data show that those in the Northeast are most apt to contact a doctor and those in the South (especially nonSMSA Southern Blacks) are the least likely to. Both datasets indicate that those in SMSAs are more likely to see or talk to a doctor than are those living outside SMSAs, as are preschool children and the divorced. Contacting a doctor is positively correlated with the financial status of the family, as both the phone data and the total data show. Therefore, while a dataset based on only the phone population may overstate the percent contacting a doctor within the year, estimates of differences between population subgroups may contain little bias.

Tables 3 through 8 present the same conclusions for six other health care variables. Table 3 indicates that telephone data overstate the mean number of physician visits. Table 4 shows that the percent with a regular source of care is higher for those with phones, and Table 5 shows that a smaller proportion of those with phones report themselves to be in poor health. The percent hospitalized is about the same for both groups (Table 6), but a higher percentage of people with phones have health insurance (Table 7) and see a dentist during the year (Table 8) than does the population as a whole. Therefore, the phone population is somewhat more advantaged in health care than is the entire population, and those without phones are quite a bit less advantaged than are those with home telephones.

There are several ways in which these results might be used. When analyzing telephone data, a researcher might merely keep in mind the fact that the entire population might be a bit more disadvantaged than the data suggest. This approach would be most appropriate when working with sample sizes small enough that the bias would comprise only a small part of total error. (For example, in Table 2 the bias between 77.6% and 76.7% contacting a doctor contributes 63% of the mean square error. With a sample one-fifth as large, the variance would increase fivefold, so the bias would contribute only 26% of the mean square error. With a sample of 100, the bias would be only 11% of the mean square

Table 2
Percent contacting a doctor during the year, by phone coverage: CHAS, 1976[a]

Characteristic	Percent contacting a doctor during the year			Ratios of the percents	
	Phone population	Nonphone population	Total population	Nonphone population to phone population	Total population to phone population
Region					
Northeast	81.4 (1.4)	74.8 (6.4)	81.1 (1.4)	.918 (.081)	.996 (.004)
North central	76.6 (1.3)	83.5 (5.3)	76.9 (1.3)	1.089 (.071)	1.004 (.003)
South	74.8 (1.2)	66.4 (2.5)	73.4 (1.1)	.888 (.036)	.981 (.006)
West	79.2 (1.5)	52.0 (4.7)	76.5 (1.5)	.656 (.060)	.966 (.006)
Residence					
SMSA central city	78.1 (1.3)	69.1 (3.8)	77.2 (1.3)	.885 (.051)	.989 (.005)
SMSA other	79.9 (1.1)	67.5 (4.2)	79.1 (1.1)	.845 (.054)	.990 (.004)
NonSMSA urban	75.1 (1.9)	68.7 (4.2)	74.4 (1.7)	.915 (.061)	.990 (.007)
Rural nonfarm	76.0 (1.6)	66.5 (3.7)	74.7 (1.5)	.875 (.052)	.983 (.007)
Rural farm	70.6 (3.2)	63.3 (10.6)	70.2 (3.1)	.896 (.155)	.984 (.009)
Race					
Spanish heritage, southwest	72.0 (5.7)	44.0 (11.0)	64.7 (5.3)	.612 (.160)	.898 (.042)
Other white	78.0 (0.8)	72.0 (2.9)	77.5 (0.8)	.924 (.039)	.995 (.003)
NonSMSA southern black	69.6 (4.7)	58.2 (6.5)	65.1 (3.8)	.827 (.110)	.936 (.043)
Other nonwhite	77.5 (3.8)	74.4 (8.4)	77.0 (3.4)	.960 (.118)	.994 (.016)
Age					
0–5	89.3 (1.6)	76.1 (4.5)	87.4 (1.5)	.852 (.053)	.978 (.008)
6–17	71.6 (1.6)	53.9 (4.8)	69.9 (1.5)	.753 (.069)	.976 (.007)
18–34	79.0 (1.3)	73.1 (3.2)	78.3 (1.2)	.925 (.043)	.991 (.005)
35–54	76.0 (1.5)	65.9 (4.8)	75.3 (1.5)	.868 (.065)	.991 (.004)
55–64	79.6 (2.1)	81.7 (6.5)	79.7 (2.0)	1.027 (.086)	1.001 (.004)
65 plus	79.9 (1.9)	69.5 (6.7)	79.3 (1.8)	.870 (.086)	.993 (.005)
Age of head					
Under 25	82.5 (2.3)	80.9 (3.6)	82.1 (2.0)	.981 (.052)	.995 (.013)
25–34	82.4 (1.3)	69.1 (3.7)	80.8 (1.2)	.838 (.046)	.980 (.006)
35–44	75.0 (1.5)	54.4 (4.8)	73.2 (1.4)	.725 (.065)	.975 (.006)
45–54	75.3 (1.6)	70.9 (5.2)	75.0 (1.5)	.943 (.072)	.997 (.004)
55–64	78.3 (1.9)	74.0 (6.2)	78.1 (1.8)	.945 (.082)	.997 (.005)
65 plus	76.1 (1.9)	64.2 (6.2)	75.4 (1.8)	.844 (.084)	.991 (.005)
Sex of head					
Male	77.2 (0.8)	63.7 (2.4)	76.0 (0.7)	.825 (.032)	.985 (.003)
Female	80.2 (1.5)	81.6 (3.1)	80.4 (1.3)	1.019 (.043)	1.003 (.006)
Marital status of head					
Married	77.5 (0.8)	64.8 (2.6)	76.5 (0.8)	.835 (.035)	.987 (.003)
Widowed	75.1 (2.2)	69.2 (6.1)	74.6 (2.1)	.922 (.085)	.994 (.007)
Divorced	84.1 (2.3)	79.4 (5.7)	83.3 (2.2)	.944 (.072)	.990 (.013)
Separated	75.6 (3.6)	73.5 (5.2)	75.1 (3.0)	.972 (.083)	.993 (.022)
Never married	77.1 (3.1)	69.0 (6.1)	75.5 (2.8)	.895 (.087)	.979 (.017)
Family size					
1	78.7 (1.9)	68.7 (4.4)	77.1 (1.7)	.874 (.059)	.980 (.010)
2	79.6 (1.5)	74.4 (4.5)	79.2 (1.4)	.935 (.059)	.995 (.005)
3	79.0 (1.5)	74.1 (4.1)	78.5 (1.4)	.938 (.055)	.994 (.006)
4	82.4 (1.4)	68.2 (5.0)	81.5 (1.3)	.827 (.062)	.989 (.004)
5	77.7 (1.9)	67.5 (6.2)	77.0 (1.8)	.869 (.082)	.991 (.006)
6	71.0 (2.9)	73.2 (6.9)	71.2 (2.7)	1.031 (.105)	1.002 (.008)
7 or more	65.9 (3.1)	52.7 (6.2)	63.7 (2.8)	.801 (.102)	.967 (.017)
Adults in family					
1	81.7 (1.4)	74.1 (3.1)	80.4 (1.2)	.906 (.041)	.984 (.007)
2	78.7 (0.9)	66.2 (2.7)	77.5 (0.9)	.842 (.036)	.985 (.003)
3	75.0 (1.8)	65.3 (6.6)	74.5 (1.8)	.871 (.090)	.994 (.005)
4 or more	72.4 (2.6)	60.3 (9.4)	71.7 (2.5)	.832 (.134)	.989 (.009)

error.) When planning a very large sample, the estimates given in Tables 2 through 8 can be used to determine whether a telephone frame should be used in conjunction with an area frame, as discussed for HIS in Casady et al. (1981).

Another approach to dealing with the noncoverage bias is to use the ratios given in the tables as adjustment factors to be applied to another dataset based on telephone households.[3] The assumption would be that the ratios remain fairly stable over time and from one dataset to another. It probably would be wise to check out these assumptions on data from a recurring survey such as HIS. A limitation of this adjustment method is that it can be performed only on variables for which the relationship between estimates for the phone population and the nonphone population or total population is

Table 2 continued

Characteristic	Percent contacting a doctor during the year: Phone population	Nonphone population	Total population	Ratios of the percents: Nonphone population to phone population	Total population to phone population
Family income					
Less than $3,000	71.3 (2.9)	69.0 (4.4)	70.6 (2.4)	.968 (.073)	.991 (.021)
$3,000–$4,999	77.2 (2.2)	69.0 (3.9)	75.1 (1.9)	.894 (.056)	.973 (.014)
$5,000–$6,999	76.5 (2.1)	63.4 (4.7)	74.3 (2.0)	.829 (.065)	.971 (.011)
$7,000–$9,999	75.4 (1.9)	61.8 (4.9)	73.7 (1.8)	.819 (.069)	.977 (.009)
$10,000–$14,999	76.3 (1.4)	75.5 (4.8)	76.3 (1.3)	.988 (.066)	.999 (.004)
$15,000–$24,999	79.4 (1.4)	62.7 (9.1)	79.0 (1.3)	.789 (.116)	.995 (.003)
$25,000 or more	81.0 (2.0)	80.8 (12.5)	81.0 (2.0)	.997 (.157)	1.000 (.002)
Poverty status					
Below poverty	72.4 (1.9)	66.8 (2.9)	70.7 (1.6)	.923 (.048)	.978 (.014)
100%–125% poverty	74.9 (2.7)	62.4 (6.0)	73.2 (2.5)	.834 (.086)	.977 (.012)
125%–200% poverty	74.3 (1.6)	66.7 (4.3)	73.6 (1.5)	.898 (.061)	.991 (.006)
200%–300% poverty	77.2 (1.4)	70.5 (5.1)	76.8 (1.4)	.914 (.069)	.995 (.004)
300%–400%	79.7 (1.6)	73.5 (7.8)	79.5 (1.6)	.922 (.099)	.997 (.004)
400% or more poverty	83.4 (1.4)	74.7 (9.4)	83.2 (1.4)	.895 (.114)	.998 (.003)
Total	77.6 (0.7)	67.7 (2.0)	76.7 (0.6)	.872 (.026)	.988 (.002)

[a]Numbers in parentheses are the standard error estimates.

Table 3
Mean number of physician visits during the year, by phone coverage; CHAS 1976[a]

Characteristic	Phone population	Total population	Ratio, total to phone
Region			
Northeast	4.2 (0.3)	4.1 (0.2)	.992 (.010)
North central	4.0 (0.3)	4.1 (0.2)	1.017 (.015)
South	3.9 (0.2)	3.8 (0.2)	.976 (.017)
West	4.9 (0.4)	4.7 (0.3)	.957 (.010)
Residence			
SMSA central city	4.7 (0.3)	4.6 (0.3)	.975 (.011)
SMSA other	4.2 (0.2)	4.1 (0.2)	.988 (.009)
NonSMSA	4.0 (0.4)	3.9 (0.3)	.989 (.016)
Rural nonfarm	3.8 (0.3)	3.8 (0.2)	1.003 (.028)
Rural farm	2.7 (0.3)	2.7 (0.3)	.992 (.013)
Race			
Spanish heritage, southwest	4.1 (1.0)	3.5 (0.8)	.870 (.074)
Other white	4.1 (0.1)	4.1 (0.1)	.994 (.008)
NonSMSA southern black	3.4 (0.8)	3.1 (0.5)	.927 (.110)
Other nonwhite	4.4 (0.8)	4.4 (0.7)	.999 (.051)
Age			
0–5	4.5 (0.3)	4.5 (0.3)	.999 (.036)
6–17	2.4 (0.1)	2.4 (0.1)	.984 (.015)
18–34	4.3 (0.3)	4.3 (0.2)	.984 (.013)
35–54	4.4 (0.3)	4.3 (0.3)	.995 (.012)
55–64	4.9 (0.4)	4.9 (0.4)	.995 (.012)
65 plus	6.1 (0.4)	6.0 (0.4)	.986 (.009)
Age of head			
Under 25	4.4 (0.4)	4.4 (0.4)	1.012 (.057)
25–34	4.2 (0.2)	4.1 (0.2)	.986 (.017)
35–44	3.3 (0.2)	3.2 (0.2)	.972 (.011)
45–54	3.8 (0.3)	3.8 (0.3)	1.003 (.014)
55–65	4.7 (0.4)	4.6 (0.4)	.986 (.011)
65 plus	5.7 (0.5)	5.6 (0.4)	.985 (.008)
Sex of head			
Male	3.9 (0.1)	3.9 (0.1)	.984 (.008)
Female	5.2 (0.4)	5.1 (0.3)	.991 (.017)
Marital status of head			
Married	3.9 (0.1)	3.9 (0.1)	.987 (.008)
Widowed	5.3 (0.5)	5.2 (0.4)	.972 (.012)
Divorced	5.6 (0.8)	5.3 (0.7)	.949 (.030)
Separated	4.0 (0.7)	4.3 (0.6)	1.079 (.084)
Never married	5.0 (0.7)	4.8 0.6)	.956 (.044)
Family size			
1	5.6 (0.5)	5.3 (0.4)	.954 (.021)
2	5.0 (0.3)	4.9 (0.3)	.986 (.011)
3	4.7 (0.4)	4.7 (0.3)	.993 (.015)
4	4.0 (0.2)	4.1 (0.2)	1.009 (.021)
5	3.4 (0.2)	3.3 (0.2)	.975 (.010)
6	3.1 (0.4)	3.0 (0.4)	.981 (.013)
7 or more	2.7 (0.4)	2.7 (0.3)	.991 (.036)
Adults in family			
1	5.2 (0.3)	5.2 (0.3)	.995 (.026)
2	4.1 (0.2)	4.0 (0.1)	.981 (.008)
3	4.1 (0.4)	4.0 (0.3)	.978 (.005)
4 or more	3.2 (0.4)	3.2 (0.3)	.985 (.015)
Family income			
Less than $3,000	4.6 (0.5)	4.5 (0.4)	.993 (.050)
$3,000–$4,999	5.1 (0.5)	4.9 (0.4)	.952 (.042)
$5,000–$6,999	4.8 (0.5)	4.5 (0.4)	.951 (.022)
$7,000–$9,999	4.2 (0.3)	4.1 (0.3)	.967 (.023)
$10,000–$14,999	4.0 (0.2)	3.9 (0.2)	.988 (.010)
$15,000–$24,999	3.9 (0.2)	3.8 (0.2)	.988 (.003)
$25,000 or more	3.7 (0.4)	3.7 (0.4)	.998 (.005)
Poverty status			
Below poverty	3.9 (0.3)	3.9 (0.3)	.993 (.040)
100%–125% poverty	4.1 (0.4)	4.0 (0.4)	.985 (.031)
125%–200% poverty	4.4 (0.3)	4.3 (0.3)	.983 (.015)
200%–300% poverty	4.1 (0.2)	4.1 (0.2)	.987 (.010)
300%–400% poverty	4.0 (0.3)	4.0 (0.3)	.996 (.009)
400% or more poverty	4.1 (0.3)	4.1 (0.3)	.989 (.003)
Total	4.1 (0.1)	4.1 (0.1)	.987 (.007)

[a]The numbers in parentheses are the standard error estimates.

Table 4
Percent with a regular source of care, by phone coverage; CHAS 1976[a]

Characteristic	Phone population		Total population		Ratio, total to phone	
Region						
Northeast	87.1	(1.2)	86.5	(1.2)	.994	(.004)
North central	90.3	(0.9)	90.2	(0.9)	.999	(.002)
South	89.4	(0.9)	87.9	(0.8)	.983	(.004)
West	87.7	(1.2)	85.1	(1.2)	.970	(.005)
Residence						
SMSA central city	86.2	(1.1)	85.1	(1.1)	.986	(.004)
SMSA other	88.2	(0.9)	86.9	(0.9)	.985	(.003)
NonSMSA urban	91.3	(1.3)	90.2	(1.2)	.988	(.005)
Rural nonfarm	91.0	(1.1)	90.5	(1.0)	.994	(.004)
Rural farm	93.4	(1.7)	93.4	(1.7)	1.000	(.004)
Race						
Spanish heritage, southwest	89.5	(3.9)	82.9	(4.1)	.926	(.032)
Other white	89.3	(0.6)	88.5	(0.6)	.991	(.002)
NonSMSA southern black	91.8	(3.4)	89.8	(2.8)	.978	(.024)
Other nonwhite	84.1	(3.3)	84.1	(3.0)	1.000	(.013)
Age						
0–5	95.7	(1.0)	94.5	(1.1)	.988	(.006)
6–17	92.1	(0.9)	90.6	(0.9)	.984	(.004)
18–34	80.9	(1.3)	79.9	(1.2)	.989	(.005)
35–54	88.1	(1.2)	87.5	(1.1)	.993	(.003)
55–64	93.7	(1.3)	93.5	(1.2)	.998	(.003)
65 plus	91.5	(1.3)	90.7	(1.3)	.991	(.004)
Age of head						
Under 25	77.9	(2.6)	78.5	(2.1)	1.008	(.015)
25–34	87.4	(1.1)	85.8	(1.1)	.981	(.005)
35–44	89.0	(1.1)	87.8	(1.1)	.987	(.004)
45–54	89.4	(1.1)	89.1	(1.1)	.997	(.003)
55–65	93.4	(1.1)	93.2	(1.1)	.997	(.003)
65 plus	90.2	(1.3)	89.3	(1.3)	.990	(.004)
Sex of head						
Male	89.2	(0.6)	88.0	(0.6)	.987	(.002)
Female	87.5	(1.2)	87.5	(1.1)	1.001	(.005)
Marital status of head						
Married	89.6	(0.6)	88.8	(0.6)	.991	(.002)
Widowed	88.1	(1.7)	87.1	(1.6)	.989	(.005)
Divorced	89.3	(2.0)	88.7	(1.8)	.993	(.010)
Separated	82.9	(3.2)	80.7	(2.7)	.974	(.019)
Never married	76.3	(3.1)	75.2	(2.8)	.985	(.017)
Family size						
1	83.2	(1.7)	80.6	(1.6)	.969	(.009)
2	89.3	(1.1)	88.5	(1.1)	.991	(.004)
3	88.7	(1.2)	88.1	(1.1)	.993	(.004)
4	88.5	(1.1)	87.9	(1.1)	.993	(.003)
5	93.0	(1.1)	92.2	(1.1)	.991	(.004)
6	87.1	(2.1)	86.8	(2.0)	.997	(.005)
7 or more	88.1	(2.1)	86.3	(2.0)	.980	(.011)
Adults in family						
1	86.6	(1.2)	85.0	(1.1)	.982	(.006)
2	89.2	(0.7)	88.0	(0.7)	.987	(.003)
3	90.2	(1.3)	89.7	(1.2)	.995	(.003)
4 or more	88.1	(1.9)	88.0	(1.8)	.999	(.005)
Family income						
Less than $3,000	87.3	(2.2)	86.9	(1.8)	.995	(.013)
$3,000–$4,999	86.4	(1.8)	82.2	(1.7)	.950	(.012)
$5,000–$6,999	89.4	(1.5)	89.0	(1.4)	.996	(.007)
$7,000–$9,999	88.0	(1.4)	86.7	(1.4)	.984	(.006)
$10,000–$14,999	88.9	(1.0)	87.9	(1.0)	.988	(.003)
$15,000–$24,999	89.9	(1.0)	89.9	(1.0)	1.000	(.002)
$25,000 or more	89.0	(1.6)	88.7	(1.6)	.997	(.002)
Poverty status						
Below poverty	86.7	(1.5)	85.2	(1.3)	.983	(.009)
100%–125% poverty	87.0	(2.1)	84.9	(2.0)	.976	(.009)
125%–200% poverty	90.2	(1.1)	88.8	(1.1)	.985	(.004)
200%–300% poverty	90.1	(1.0)	89.4	(1.0)	.992	(.003)
300%–400% poverty	89.4	(1.2)	89.0	(1.2)	.995	(.003)
400% or more poverty	87.4	(1.3)	87.2	(1.3)	.997	(.002)
Total	88.9	(0.5)	89.9	(0.5)	.989	(.002)

[a]The numbers in parentheses are the standard error estimates.

Table 5
Percent in poor health, by phone coverage; CHAS 1976[a]

Characteristic	Phone population		Total population		Ratio, total to phone	
Region						
Northeast	3.2	(0.7)	3.2	(0.6)	1.002	(.043)
North central	2.6	(0.5)	2.8	(0.5)	1.072	(.064)
South	5.4	(0.6)	5.9	(0.6)	1.085	(.053)
West	3.3	(0.7)	3.3	(0.6)	.983	(.049)
Residence						
SMSA central city	4.1	(0.6)	4.0	(0.6)	.985	(.038)
SMSA other	3.1	(0.5)	3.3	(0.5)	1.048	(.047)
NonSMSA urban	4.2	(0.9)	4.4	(0.8)	1.049	(.067)
Rural nonfarm	3.5	(0.7)	4.4	(0.7)	1.236	(.115)
Rural farm	5.5	(1.6)	6.2	(1.6)	1.117	(.099)
Race						
Spanish heritage, southwest	2.7	(2.1)	2.5	(1.7)	.942	(.346)
Other white	3.5	(0.4)	3.8	(0.3)	1.077	(.037)
NonSMSA southern black	9.9	(3.7)	9.9	(2.7)	.995	(.214)
Other nonwhite	4.8	(1.9)	4.8	(1.8)	1.005	(.134)
Age						
0–5	1.1	(0.5)	1.5	(0.6)	1.350	(.353)
6–17	0.9	(0.3)	1.0	(0.3)	1.140	(.178)
18–34	1.3	(0.4)	1.3	(0.3)	1.012	(.087)
35–54	3.4	(0.7)	4.1	(0.7)	1.182	(.082)
55–64	10.1	(1.6)	0.9	(1.6)	1.084	(.043)
65 plus	12.6	(1.6)	3.4	(1.5)	1.065	(.033)
Age of head						
Under 25	0.9	(0.6)	0.9	(0.5)	.992	(.290)
25–34	1.6	(0.4)	1.6	(0.4)	1.047	(.100)
35–44	1.8	(0.5)	2.1	(0.5)	1.176	(.127)
45–54	2.6	(0.6)	3.1	(0.6)	1.177	(.095)
55–64	7.7	(1.2)	8.4	(1.2)	1.085	(.048)
65 plus	10.2	(1.3)	0.8	(1.3)	1.063	(.034)
Sex of head						
Male	3.3	(0.3)	3.5	(0.3)	1.068	(.033)
Female	6.1	(0.9)	6.5	(0.8)	1.065	(.060)
Marital status of head						
Married	3.2	(0.3)	3.4	(0.3)	1.055	(.033)
Widowed	9.2	(1.5)	9.9	(1.4)	1.080	(.051)
Divorced	3.9	(1.2)	4.4	(1.2)	1.129	(.181)
Separated	5.5	(1.9)	5.5	(1.6)	.995	(.159)
Never married	2.8	(1.2)	4.0	(1.3)	1.437	(.386)
Family size						
1	6.5	(1.1)	7.5	(1.1)	1.147	(.094)
2	8.0	(1.0)	8.1	(0.9)	1.005	(.030)
3	3.2	(0.7)	3.5	(0.6)	1.089	(.080)
4	1.6	(0.4)	1.6	(0.4)	1.017	(.065)
5	2.2	(0.7)	2.6	(0.7)	1.180	(.133)
6	2.5	(1.0)	2.8	(1.0)	1.106	(.132)
7 or more	2.0	(0.9)	2.4	(0.9)	1.194	(.270)
Adults in family						
1	5.2	(0.8)	5.9	(0.7)	1.149	(.085)
2	3.5	(0.4)	3.7	(0.4)	1.079	(.043)
3	4.1	(0.8)	4.1	(0.8)	1.004	(.036)
4 or more	2.7	(0.9)	2.6	(0.9)	.974	(.059)
Family income						
Less than $3,000	13.9	(2.2)	4.2	(1.9)	1.020	(.085)
$3,000–$4,999	9.3	(1.5)	9.2	(1.3)	.989	(.075)
$5,000–$6,999	5.7	(1.2)	5.7	(1.0)	1.003	(.076)
$7,000–$9,999	4.0	(0.9)	4.0	(0.8)	1.002	(.070)
$10,000–$14,999	3.0	(0.6)	2.9	(0.5)	.975	(.030)
$15,000–$24,999	1.7	(0.4)	1.7	(0.4)	.994	(.032)
$25,000 or more	1.4	(0.6)	1.4	(0.6)	.985	(.001)
Poverty status						
Below poverty	9.1	(1.2)	9.3	(1.0)	1.025	(.073)
100%–125% poverty	4.1	(1.2)	5.0	(1.2)	1.222	(.168)
125%–200% poverty	4.2	(0.7)	4.2	(0.7)	.993	(.042)
200%–300% poverty	3.3	(0.6)	3.2	(0.6)	.972	(.026)
300%–400% poverty	2.4	(0.6)	2.3	(0.6)	.968	(.014)
400% or more poverty	1.7	(0.5)	1.7	(0.5)	1.005	(.042)
Total	3.7	(0.3)	4.0	(0.3)	1.073	(.030)

[a]The numbers in parentheses are the standard error estimates.

Table 6
Percent hospitalized in the year, by phone coverage; CHAS 1976[a]

Characteristic	Phone population		Total population		Ratio, total to phone	
Region						
Northeast	11.0	(1.2)	11.1	(1.1)	1.010	(.024)
North central	10.8	(1.0)	10.9	(1.0)	1.016	(.021)
South	11.6	(0.9)	11.5	(0.8)	.990	(.026)
West	12.0	(1.2)	11.5	(1.1)	.950	(.019)
Residence						
SMSA central city	12.9	(1.1)	12.5	(1.0)	.971	(.019)
SMSA other	10.1	(0.8)	10.1	(0.8)	.999	(.018)
NonSMSA	12.6	(1.5)	12.5	(1.3)	.992	(.028)
Rural nonfarm	11.3	(1.2)	11.6	(1.1)	1.025	(.036)
Rural farm	8.3	(1.9)	8.3	(1.8)	1.001	(.045)
Race						
Spanish heritage, southwest	13.0	(4.3)	11.6	(3.5)	.889	(.128)
Other white	11.0	(0.6)	11.0	(0.6)	.998	(.013)
NonSMSA southern black	10.4	(3.8)	9.9	(2.7)	.953	(.189)
Other nonwhite	12.8	(3.0)	13.2	(2.8)	1.035	(.085)
Age						
0–5	9.4	(1.6)	10.3	(1.5)	1.092	(.077)
6–17	5.2	(0.8)	5.4	(0.7)	1.025	(.048)
18–34	11.8	(1.1)	12.0	(1.0)	1.014	(.028)
35–54	12.0	(1.2)	11.7	(1.1)	.976	(.016)
55–64	14.2	(1.8)	14.7	(1.8)	1.028	(.027)
65 plus	20.9	(1.9)	20.1	(1.8)	.961	(.010)
Age of head						
Under 25	14.4	(2.2)	14.6	(1.8)	1.017	(.072)
25–34	10.8	(1.1)	10.8	(1.0)	1.000	(.031)
35–44	8.3	(0.9)	8.3	(0.9)	1.001	(.031)
45–54	9.7	(1.1)	9.6	(1.0)	.999	(.022)
55–64	13.4	(1.6)	13.5	(1.5)	1.005	(.023)
65 plus	17.5	(1.7)	17.0	(1.6)	.971	(.013)
Sex of head						
Male	10.6	(0.6)	10.6	(0.5)	.996	(.012)
Female	14.9	(1.3)	14.8	(1.2)	.990	(.028)
Marital status of head						
Married	10.6	(0.6)	10.6	(0.6)	.996	(.013)
Widowed	15.8	(1.9)	15.2	(1.7)	.964	(.019)
Divorced	14.2	(2.2)	14.8	(2.0)	1.042	(.074)
Separated	11.7	(2.7)	11.7	(2.2)	.998	(.103)
Never married	12.0	(2.4)	11.6	(2.1)	.964	(.071)
Family size						
1	15.7	(1.7)	14.9	(1.5)	.949	(.032)
2	14.7	(1.3)	14.2	(1.2)	.965	(.015)
3	11.5	(1.2)	11.3	(1.1)	.988	(.026)
4	9.2	(1.0)	9.6	(1.0)	1.036	(.028)
5	9.8	(1.3)	10.3	(1.3)	1.051	(.038)
6	8.5	(1.8)	8.1	(1.6)	.952	(.024)
7 or more	9.6	(1.9)	9.8	(1.7)	1.109	(.077)
Adults in family						
1	16.5	(1.3)	16.0	(1.1)	.969	(.027)
2	11.0	(0.7)	10.9	(0.6)	.992	(.016)
3	9.5	(1.3)	9.6	(1.2)	1.007	(.023)
4 or more	9.6	(1.7)	9.6	(1.6)	.999	(.039)
Family income						
Less than $3,000	14.0	(2.2)	13.1	(1.8)	.937	(.068)
$3,000–$4,999	15.2	(1.9)	14.5	(1.6)	.956	(.053)
$5,000–$6,999	13.5	(1.7)	13.2	(1.5)	.979	(.043)
$7,000–$9,999	11.4	(1.4)	11.1	(1.3)	.977	(.035)
$10,000–$14,999	11.2	(1.0)	11.3	(1.0)	1.008	(.021)
$15,000–$24,999	11.1	(1.1)	10.9	(1.0)	.982	(.008)
$25,000 or more	7.1	(1.3)	7.2	(1.3)	1.009	(.022)
Poverty status						
Below poverty	13.8	(1.5)	13.3	(1.2)	.964	(.050)
100%–125% poverty	10.5	(1.9)	10.8	(1.7)	1.034	(.063)
125%–200% poverty	10.7	(1.1)	10.5	(1.0)	.987	(.024)
200%–300% poverty	12.4	(1.1)	12.3	(1.1)	.989	(.016)
300%–400% poverty	10.7	(1.2)	10.7	(1.2)	1.002	(.019)
400% or more poverty	9.6	(1.1)	9.5	(1.1)	.990	(.012)
Total	11.2	(0.5)	11.2	(0.5)	.997	(.011)

[a]The numbers in parentheses are the standard error estimates.

Table 7
Percent with health insurance, by phone coverage; CHAS 1976[a]

Characteristic	Phone population		Total population		Ratio, total to phone	
Region						
Northeast	92.9	(1.0)	92.4	(1.0)	.995	(.003)
North central	93.8	(0.8)	93.3	(0.8)	.995	(.003)
South	88.0	(0.9)	84.3	(0.9)	.958	(.005)
West	88.7	(1.2)	85.9	(1.2)	.969	(.005)
Residence						
SMSA central city	89.0	(1.0)	87.2	(1.0)	.981	(.004)
SMSA other	92.5	(0.7)	90.6	(0.8)	.979	(.003)
NonSMSA urban	92.8	(1.2)	91.6	(1.1)	.987	(.004)
Rural nonfarm	91.4	(1.1)	88.4	(1.1)	.968	(.005)
Rural farm	87.1	(2.4)	85.2	(2.4)	.978	(.007)
Race						
Spanish heritage, southwest	72.8	(5.7)	66.9	(5.2)	.919	(.042)
Other white	92.5	(0.5)	90.9	(0.5)	.983	(.002)
NonSMSA southern black	87.0	(4.2)	81.7	(3.5)	.939	(.031)
Other nonwhite	85.6	(3.2)	84.7	(2.9)	.989	(.013)
Age						
0–5	88.4	(1.7)	86.2	(1.6)	.976	(.008)
6–17	91.9	(0.9)	88.5	(1.0)	.963	(.005)
18–34	87.3	(1.1)	85.6	(1.1)	.980	(.005)
35–54	91.9	(1.0)	89.9	(1.0)	.979	(.004)
55–64	89.8	(1.6)	89.3	(1.6)	.995	(.004)
65 plus	100.0	(0.0)	100.0	(0.0)	1.000	(.000)
Age of head						
Under 25	82.6	(2.3)	82.6	(1.9)	1.000	(.013)
25–34	92.2	(0.9)	89.3	(1.0)	.968	(.005)
35–44	90.5	(1.0)	87.6	(1.1)	.968	(.005)
45–54	92.1	(1.0)	90.7	(1.0)	.985	(.003)
55–65	88.1	(1.5)	86.8	(1.5)	.985	(.004)
65 plus	95.5	(0.9)	95.2	(0.9)	.996	(.002)
Sex of head						
Male	91.9	(0.5)	89.6	(0.5)	.975	(.002)
Female	86.4	(1.3)	86.2	(1.2)	.997	(.005)
Marital status of head						
Married	92.1	(0.5)	90.0	(0.5)	.978	(.002)
Widowed	87.2	(1.7)	87.0	(1.6)	.997	(.005)
Divorced	88.8	(2.0)	87.5	(1.9)	.986	(.011)
Separated	82.9	(3.2)	81.2	(2.7)	.979	(.018)
Never married	85.8	(2.6)	83.3	(2.4)	.971	(.014)
Family size						
1	91.5	(1.3)	89.0	(1.3)	.972	(.007)
2	93.2	(0.9)	92.2	(0.9)	.990	(.003)
3	91.2	(1.1)	89.7	(1.1)	.984	(.004)
4	92.6	(0.9)	90.7	(1.0)	.980	(.004)
5	92.2	(1.2)	90.4	(1.2)	.981	(.005)
6	91.7	(1.8)	90.8	(1.7)	.989	(.005)
7 or more	80.9	(2.6)	76.2	(2.5)	.943	(.013)
Adults in family						
1	91.8	(1.0)	89.3	(1.0)	.972	(.006)
2	92.0	(0.6)	89.7	(0.6)	.975	(.003)
3	91.8	(1.2)	90.6	(1.2)	.986	(.004)
4 or more	84.9	(2.1)	83.6	(2.1)	.984	(.007)
Family income						
Less than $3,000	83.0	(2.4)	79.3	(2.2)	.956	(.016)
$3,000–$4,999	83.8	(1.9)	78.3	(1.8)	.935	(.013)
$5,000–$6,999	81.2	(2.0)	77.6	(1.9)	.955	(.010)
$7,000–$9,999	85.6	(1.6)	84.2	(1.5)	.983	(.007)
$10,000–$14,999	92.9	(0.8)	92.4	(0.8)	.994	(.003)
$15,000–$24,999	94.8	(0.7)	94.2	(0.8)	.994	(.002)
$25,000 or more	97.5	(0.8)	97.3	(0.8)	.998	(.002)
Poverty status						
Below poverty	76.8	(1.8)	73.3	(1.6)	.954	(.013)
100%–125% poverty	86.3	(2.1)	83.3	(2.1)	.965	(.010)
125%–200% poverty	88.8	(1.2)	87.2	(1.1)	.982	(.004)
200%–300% poverty	92.7	(0.9)	91.7	(0.9)	.989	(.003)
300%–400% poverty	94.5	(0.9)	94.1	(0.9)	.996	(.003)
400% or more poverty	98.2	(0.5)	97.9	(0.5)	.998	(.002)
Total	91.1	(0.5)	89.1	(0.5)	.978	(.002)

[a]The numbers in parentheses are the standard error estimates.

Table 8
Percent seeing a dentist during the year, by phone coverage; CHAS 1976[a]

Characteristic	Phone population		Total population		Ratio, total to phone	
Region						
Northeast	58.0	(1.8)	56.8	(1.8)	.980	(.006)
North central	52.5	(1.6)	51.6	(1.5)	.983	(.005)
South	43.3	(1.4)	40.1	(1.2)	.927	(.009)
West	52.4	(1.9)	49.5	(1.7)	.944	(.008)
Residence						
SMSA central city	52.3	(1.6)	50.5	(1.5)	.967	(.008)
SMSA other	54.3	(1.4)	52.2	(1.3)	.962	(.005)
NonSMSA urban	46.6	(2.2)	43.6	(2.0)	.935	(.009)
Rural nonfarm	46.9	(1.9)	44.0	(1.7)	.938	(.010)
Rural farm	46.1	(3.5)	44.8	(3.3)	.971	(.012)
Race						
Spanish heritage, southwest	36.1	(6.1)	31.1	(5.1)	.860	(.064)
Other white	53.1	(1.0)	51.6	(0.9)	.971	(.004)
NonSMSA southern black	23.1	(5.3)	17.8	(3.5)	.769	(.074)
Other nonwhite	41.3	(4.4)	38.9	(4.0)	.943	(.029)
Age						
0–5	26.9	(2.5)	24.2	(2.2)	.902	(.018)
6–17	65.9	(1.6)	61.5	(1.6)	.933	(.006)
18–34	54.0	(1.6)	51.9	(1.5)	.961	(.008)
35–54	50.1	(1.8)	48.4	(1.7)	.966	(.006)
55–64	45.1	(2.6)	44.9	(2.5)	.996	(.010)
65 plus	34.0	(2.2)	33.2	(2.1)	.974	(.010)
Age of head						
Under 25	42.4	(3.1)	39.9	(2.5)	.941	(.030)
25–34	51.4	(1.7)	48.3	(1.6)	.940	(.008)
35–44	58.8	(1.7)	55.3	(1.6)	.942	(.006)
45–54	53.8	(1.8)	52.3	(1.8)	.972	(.006)
55–65	46.8	(2.3)	46.0	(2.2)	.983	(.008)
65 plus	36.4	(2.1)	35.4	(2.0)	.974	(.008)
Sex of head						
Male	51.8	(0.9)	49.5	(0.9)	.956	(.004)
Female	45.9	(1.8)	43.9	(1.7)	.957	(.012)
Marital status of head						
Married	52.6	(1.0)	50.5	(0.9)	.959	(.003)
Widowed	36.7	(2.5)	35.7	(2.3)	.974	(.013)
Divorced	49.2	(3.2)	45.7	(2.9)	.929	(.024)
Separated	47.0	(4.2)	40.3	(3.4)	.857	(.029)
Never married	47.7	(3.7)	47.0	(3.2)	.984	(.030)
Family size						
1	47.3	(2.3)	45.5	(2.0)	.963	(.017)
2	45.8	(1.8)	45.4	(1.7)	.990	(.009)
3	48.6	(1.9)	46.4	(1.7)	.955	(.009)
4	55.3	(1.8)	53.3	(1.7)	.964	(.005)
5	53.1	(2.2)	50.6	(2.1)	.953	(.006)
6	53.8	(3.2)	53.2	(2.9)	.971	(.011)
7 or more	52.0	(3.3)	45.4	(2.9)	.874	(.013)
Adults in family						
1	50.3	(1.7)	47.8	(1.6)	.949	(.012)
2	50.8	(1.1)	48.2	(1.0)	.949	(.005)
3	52.2	(2.1)	50.8	(2.0)	.973	(.006)
4 or more	50.8	(2.9)	49.1	(2.8)	.965	(.010)
Family income						
Less than $3,000	32.2	(3.0)	28.6	(2.4)	.888	(.036)
$3,000–$4,999	37.9	(2.5)	34.1	(2.1)	.899	(.026)
$5,000–$6,999	36.4	(2.4)	34.8	(2.1)	.955	(.021)
$7,000–$9,999	40.5	(2.2)	38.1	(2.0)	.941	(.014)
$10,000–$14,999	50.7	(1.6)	49.5	(1.6)	.978	(.006)
$15,000–$24,999	58.0	(1.7)	57.4	(1.6)	.990	(.004)
$25,000 or more	68.8	(2.4)	69.1	(2.3)	1.004	(.003)
Poverty status						
Below poverty	34.3	(2.1)	29.8	(1.6)	.869	(.022)
100%–125% poverty	41.2	(3.1)	38.6	(2.7)	.937	(.018)
125%–200% poverty	44.9	(1.8)	43.3	(1.7)	.966	(.009)
200%–300% poverty	50.9	(1.7)	49.8	(1.6)	.979	(.006)
300%–400% poverty	55.8	(2.0)	55.3	(1.9)	.991	(.006)
400% or more poverty	65.8	(1.8)	65.8	(1.8)	.999	(.004)
Total	51.0	(0.8)	8.7	(0.8)	.955	(.003)

[a]The numbers in parentheses are the standard error estimates.

Table 9
Ratio of total to phone estimates when the phone estimates are standardized[a]; CHAS, 1976

Adjustment cetegory	Health care variable						
	Percent contecting a doctor in the year	*Mean doctor visits in the year*	*Percent with a regular source of care*	*Percent in poor health*	*Percent hospitelized in the year*	*Percent with health insurance*	*Percent seeing a dentist in the year*
Unadjusted	.988	.987	.989	1.073	.997	.978	.955
Adjusted by							
Region	.989	.989	.989	1.048	.997	.980	.958
Residence	.989	.988	.989	1.074	.996	.978	.956
Race	.989	.987	.989	1.062	.995	.981	.959
Age	.987	.990	.989	1.124	1.012	.979	.956
Age of head	.987	.988	.991	1.116	.998	.979	.957
Sex of head	.988	.985	.989	1.068	.995	.979	.955
Marital status of head	.988	.986	.990	1.072	.996	.979	.956
Family size	.989	.988	.989	1.087	1.001	.979	.956
Adults in family	.987	.983	.987	1.062	.992	.978	.955
Family income	.990	.980	.990	.991	.986	.983	.966
Poverty status	.991	.989	.990	1.006	.990	.985	.968

[a]To the total population's distribution on various socioeconomic variables.

known. For other variables, adjustment would be problematical.

An alternative approach is to weight the telephone data in some way so that the entire population is approximated more closely. The usual way of doing this is to find a demographic variable (or variable combination) to use as the adjustment variable. However, since we have seen in Tables 2 through 8 that those without phones are more health disadvantaged than are those with phones, it is unlikely that such an adjustment will be completely satisfactory. This is shown more clearly in Table 9, where each health variable in Tables 2 through 8 is standardized by the demographic, social, and financial variables. The poverty variable provides the most reduction in bias (followed closely by family income), but the adjusted phone data still understate the health problems of the total population.

The next step in this research will be to identify a variable that can be used in an adjustment to minimize the average bias over a whole range of health variables. We suspect that this variable will be a constructed variable and might include health variables as well as socioeconomic variables.

In sum, we have examined differences between the phone and nonphone populations over a range of health and health services characteristics using data from a national household survey. The nonphone population was consistently more disadvantaged, as indicated by poorer health and lower use of health services. However, the relationship between socioeconomic variables and health characteristics tended to be similar for the phone and nonphone populations. Thus, estimates of health parameters for the population as a whole based on the phone population may be biased, but generalizations about the effects of socioeconomic variables on health will be fairly accurate. After considering alternative ways that health services researchers might take the noncoverage bias into account, we conclude that a general weighting system has advantages but that an optimal weighting procedure that would minimize bias of the estimates of a broad range of health variables is yet to be devised.

Footnotes

[1] Persons first were asked for their telephone number, then for its location. Those who reported a phone located outside the household were classified as not having a home phone.

[2] Notice also that there is not a consistent relationship between phone coverage rates and the ratios between estimates. For example, persons in families whose head is under 25 have quite low phone coverge, only about 74%. However, the ratio between the total and the phone population certainly does not suggest a larger noncoverage bias for this group than for persons in families whose heads are 25 or older.

[3] The user of this technique can use the ratio of the total to phone data or, if estimates of phone noncoverage are known to differ from those given in the CHAS 1976 data, the ratio of the nonphone to phone data can be used.

A comparison of the telephone and personal interview modes for conducting local household health surveys*

Richard A. Kulka, Research Triangle Institute

Michael F. Weeks, Research Triangle Institute

Judith T. Lessler, Research Triangle Institute

Roy W. Whitmore, Research Triangle Institute

Introduction

As noted at the third conference in this series on the current state of the art of survey procedures for health surveys (Greenberg et al., 1981), the demand for subnational health data has increased exponentially during the past decade. This is due primarily to federal promotion of health planning, regulation, and evaluation at state and local levels through the establishment of State Health Planning and Development Agencies (SHPDAs) and Health Systems Agencies (HSAs), respectively. In response to these evolving needs and demands for subnational health survey data, the National Center for Health Statistics (NCHS) has undertaken several initiatives including the provision of technical assistance to federal, state. and local agencies interested in carrying out their own surveys and the systematic evaluation of alternative survey methods that could be used to meet the data needs of these agencies (Massey, 1978).

Of particular significance among the latter activities are recent efforts by the NCHS to provide a thorough evaluation of the comparability of telephone and personal interviews as mechanisms for the collection of health interview data, since the use of presumably more cost efficient telephone interview methodology (or mail survey techniques) as an alternative to personal interviews is generally regarded as essential to the systematic and widespread use of health surveys in state and local planning and evaluation (e.g., Aday, Sellers, and Andersen, 1981; Greenberg et al., 1981). In fact, a major impetus for the establishment of a timely, effective, and flexible telephone health-interview capability within NCHS was "to evaluate a methodology which state and local areas could possibly implement for the collection of data" (Massey, 1978: 590).

Fundamental to that development effort has been a concern with how the telephone interview can best be used by the National Health Interview Survey (NHIS) on a continuing basis (Burnham and Massey, 1980) and with a professed need for basic methodological research to determine the conditions or circumstances under which the telephone or personal interview, or some combination of the two, represents the optimal approach for the collection of health survey data (Massey, 1978). A number of recent methodological studies conducted by NCHS have indeed provided some reason for optimism about the viability of telephone surveys for collecting health and health-related information (e.g., Cannell, Groves, and Miller, 1981; Cannell, Thornberry, and Fuchsberg, 1981; Fitti, 1979; Massey, Barker, and Moss, 1979; Massey, Barker, and Hsiung, 1981; Monsees and Massey, 1979a; Thornberry and Massey, 1978).

Nevertheless, few regard the evidence garnered to date conclusive. In particular, systematic comparisons of telephone and personal interview data collected in state or local health surveys, such as those recently reported by Jordan, Marcus, and Reeder (1980), are clearly required as valuable supplements to national comparisons if this research is to have maximum relevance to the conduct of local health surveys (cf. Burnham and Massey, 1980). For example, while recent NCHS studies (e.g., Massey et al., 1979; Cannell, Groves, and Miller, 1981) suggest that quite acceptable response rates (i.e., greater than 80%) can be obtained in *national* telephone health surveys, such rates may not be as readily obtainable in *local* health surveys conducted by telephone (e.g., Jordan et al., 1980), except perhaps with personal interview follow-up of nonrespondents (e.g., Siemiatycki, 1979) or in telephone reinterviews of persons initially interviewed in person (e.g., Aneshensel, Frerichs, Clark, and Yokopenic, 1982a). Alternatively, response rates in local telephone surveys conducted by a well-known and respected local or regional survey organization (e.g., the Survey Research Laboratory at the University of Illinois) may actually be *better* than those obtainable in a national survey (cf. Sudman and Ferber, 1974).

Similarly, while it is uniformly acknowledged that telephone interviews are usually less expensive than face-to-face interviews (e.g., Cannell and Fowler, 1977; Quinn, Gutek, and Walsh, 1980), such costs tend to vary considerably depending on the nature, location, and type of survey, as well as on the particular organization conducting the survey. Since the major impetus for proposing the use of telephone interviewing in local health surveys is their alleged cost effectiveness, once again systematic comparisons of the relative costs of actually implementing the two survey modes for the collection of health data in several local areas, rather than relying on indirect national cost comparisons, are clearly desirable.

Finally, of paramount concern to researchers considering the adoption of telephone interview methods for

* This project was supported with funding from the Department of Health and Human Services under contract number 233-80-2055, directed by F. William Stewart of the National Center for Health Statistics. The contents of this paper do not necessarily reflect the views or policies of the Department of Health and Human Services.

health surveys at any level are potential differences in the reliability or quality of responses obtained by a telephone survey in comparison with those gathered in face-to-face interviews, an issue which has clearly attracted the greatest amount of research attention to date (e.g., Bushery, Cowan, and Murphy, 1978; Groves and Kahn, 1979; Jordan et al., 1980; Klecka and Tuchfarber, 1978; Locander, Sudmann, and Bradburn, 1976; Massey et al., 1979; Quinn et al., 1980; Rogers, 1976; Siemiatycki, 1979; Woltman, Turner and Bushery, 1980). While most recent reviews of the literature conclude that these differences are neither large enough nor systematic enough to suggest that one of these two modes of data collection is consistently superior to the other (see, however, Singer, 1979), evidence from studies involving the collection of health interview data specificially is, in our view, quite mixed (e.g., Massey et al., 1979; Jordan et al., 1980; Yaffe, Shapiro, Fuchsberg, Rohde, and Corpeño, 1978; Siemiatycki, 1979; Hochstim, 1967; Aneshensel et al., 1982a; Bushery et al., 1978; Cannell, Groves, and Miller, 1981).

In short, major concerns raised by NCHS and others with regard to the feasibility of employing telephone interview methods as an alternative to personal interviews in health surveys merit systematic and broadbased examination in the context of local area surveys if efforts to promote the use of this mode of gathering information for planning at the local and state level are to be successful. In this paper we report the preliminary results of such an evaluation, conducted within a broader study of several different survey methods that might be used to meet the data needs of state and local health agencies.

Specifically, in recognition of needs by Health Systems Agencies and other planning agencies for local data, NCHS contracted with Research Triangle Institute (RTI) to conduct a methodological study to evaluate the feasibility of implementing local surveys at the HSA level. These surveys would collect data similar to those obtained in four national surveys conducted by the Center: (a) the National Health Interview Survey (NHIS); (b) The National Ambulatory Medical Care Survey (NAMCS); (c) the Hospital Discharge Survey (HDS), and (d) the Health and Nutrition Examination Survey (HANES). As a result, between February and August 1981, RTI conducted the Community Health Information Policy Study (CHIPS) in the service area of the Florida Gulf Health Systems Agency (FGHSA) encompassing four counties in the Tampa Bay area (Hillsborough, Manatee, Pasco, and Pinellas). The CHIPS consisted of: (1) a Health Interview Survey (HIS), patterned after the national study; (2) a Household Follow-Up Survey (HFUS), involving record checks with medical providers reported by HIS respondents; and local versions of (3) the NAMCS and (4) the HDS.[1] Based on data collected in two of these surveys—the HIS and HFUS—this paper describes the basic design of a field experiment on telephone and in-person interviewing conducted as part of the comprehensive CHIPS evaluation of data collection methods potentially applicable to health surveys in local areas. It also provides a comparison of these two interview modes with respect to response rates, potential for nonresponse bias, costs, response differences, and accuracy of reporting.

Methods

The health interview survey. The CHIPS HIS sampling design consisted of three distinct frames: an area frame, a telephone frame, and a list of persons eligible for Medicaid. The area frame included the entire four-county area and the area sample consisted of 439 housing units in 104 noncompact clusters, allocated equally to the four counties and to two SES strata within each county. The telephone frame included all of the possible telephone numbers in each of the 168 area code-prefix combinations serving the FGHSA area.[2] The telephone frame was stratified by area code-prefix combination, each of which was identified as primarily serving one of the four counties. An equal probability sample of 1,318 four-digit suffixes was generated without replacement within strata and allocated in such a way as to yield expected contacts with the same number of eligible households in each county. The list frame was used to ensure adequate representation of the indigent population in the FGHSA area, a subpopulation of particular concern in health planning. A sample of 280 cases was selected from the list of Medicaid eligibles, 70 from each of the four counties, with each county subsample subsequently allocated equally at random to the field and telephone interview modes. Thus, the overall intent of the sampling design was to allocate the entire HIS data collection effort equally between personal and telephone interviews.

The HIS was conducted in the FGHSA service area during the 13-week period from February 2 through May 3, 1981. From a data collection standpoint, the HIS actually consisted of four distinct components: an area-frame personal interview survey, a list-frame personal interview survey, a random-digit-dial (RDD) telephone survey, and a list-frame telephone survey. A staff of eight RTI field interviewers conducted the first two surveys, while six telephone interviewers in RTI's in-house Telephone Survey Department worked simultaneously on the two telephone surveys. The field interviewers also assisted with the list frame telephone survey by attempting to follow up and interview in person cases that could not be located or contacted by telephone.

The same questionnaire was used by both the field and telephone interviewers. It contained 76 items and covered a wide variety of health-related topics. An adult member of the household served as respondent and provided information for all family members.[3] If a household contained unrelated persons, separate interviews were conducted with each family unit represented.

Since the HFUS was to follow the HIS and would

involve record checks with medical providers reported by a sample of HIS participants, an effort was made during the HIS to obtain "permission" forms from persons who reported ambulatory care visits or hospital stays within a specified reference period (the preceding 12 months for hospital stays and ambulatory visits for chronic conditions; the preceding two weeks for ambulatory visits for an acute condition). One permission form was to be secured for each provider to authorize the release to RTI of medical records data on the named individual for the past 12 months. In the case of personal interviews, permission forms were secured primarily at the conclusion of the interview. For the telephone interviews, the respondent was asked at the end of the HIS if he/she would agree to complete (or arrange to have completed) the necessary permission forms by mail. If the respondent was willing, permission forms were mailed to him or her, along with a cover letter and return envelope. Persons who did not return the forms promptly were followed up by telephone and encouraged to do so. Those who were subsequently selected into the HFUS sample and who had still not returned their permission forms were followed up in person by the field interviewers.

The follow-up survey of medical providers. The purpose of the Household Follow-Up Survey was to assess the accuracy of data collected from HIS households with regard to ambulatory care visits and hospital stays, including a comparison of accuracy of reporting by mode of interview. The sampling frame for the HFUS included all persons in HIS respondent households (1) with one or more reported ambulatory care visits and/or hospital stays, and (2) for whom necessary permission forms had either been obtained or promised. The "promised" category included HIS telephone households where the respondent had agreed to return the necessary permission forms by mail but who had not yet done so at the time the HFUS sample was drawn. All persons in the HFUS frame with one or more reported hospital stays were automatically included in the sample. Some sampling, however, was done for those persons with one or more ambulatory care visits and no hospital stays.

The HFUS was conducted during June through August, 1981. For ambulatory care visits, the survey methodology involved an initial mail phase with telephone follow-ups of nonrespondents. Overall, a total of 398 unique patient/provider combinations were identified from the ambulatory care visits reported by persons in the HFUS sample. Of these, 73 had to be excluded from the survey for lack of a permission form (HIS telephone households where the promised permission forms were not obtained). Of the remaining 325, completed abstraction forms were received for 278, yielding an abstraction form completion rate of 86%.

For the hospital stay component of the HFUS, field staff were used to complete the abstraction forms rather than a mail/telephone methodology. A total of 207 unique patient/hospital combinations were identified, of which 32 had to be excluded for lack of a permission form. However, abstraction forms were completed for all of the remaining 175.

Results

Response rate comparisons. Table 1 shows the distribution of sample cases by final result category for the four HIS survey components, along with two response rate calculations for each. The two methods of calculating response rates reflect the problem posed by "indeterminate" cases—those whose eligibility status could not be determined. The lower bound response rate assumes that all indeterminates were eligible for interview and is derived from the fraction:

$$\frac{\text{interviews completed}}{\text{interviews completed + noninterviews + indeterminates}}$$

The upper bound response rate, on the other hand, assumes that all indeterminate cases were ineligible and therefore excludes them from the denominator of the response rate fraction. For most surveys with indeterminate cases, the truth no doubt lies somewhere in between these two extremes, thereby arguing for the use of both response rates in combination to calculate a "confidence" range for the actual rate. In keeping with standard protocol, cases confirmed to be ineligible for interview are excluded from the base in both methods.

The response rate for the area-frame survey was 88% under either method, since there were no cases of indeterminate eligibility. For the field list-frame survey, the response rate range was 84%–90%, reflecting 9 "unable to locate" cases, where the address obtained from the Medicaid data file was inaccurate or incomplete and field tracing efforts were unsuccessful in locating the sample member. In contrast, the RDD telephone survey achieved a response rate of 62%–70%, the range reflecting 65 "ring, no answer" cases (i.e., the telephone number is called at least eight times, rings normally each time, but is never answered). Finally, the response rate for the telephone list frame survey was 65%–85%, with the 29 "unable-to-locate" cases responsible for the relatively wide range.

A comparison of response rates for the area-frame personal interview survey (88%) and the RDD telephone survey (62%–70%) is of particular interest, of course, since these two survey components are quite comparable, having been conducted in the same area, at the same time, by the same organization, using the same questionnaire.[4] Looking first at the two rates in isolation, they would appear to be consistent with results obtained in similar surveys. The area-frame response rate, for example, is close to that achieved by RTI in other recent area-frame household health surveys, although it is probably

Table 1
Final results and response rates for the four CHIPS HIS survey components

Final result category	Area frame		Field list frame		RDD telephone frame		Telephone list frame[e]	
	Number	%	Number	%	Number	%	Number	%
I. Interviews completed	351	80	114	79	352	27	83	59
II. Noninterviews								
Refused	37	8	7	5	126	10	9	6
Breakoff (R. terminated interview prematurely)	1	—[a]	0	0	18	1	3	2
No eligible R. at home after repeated calls	5	1	3	2	4	—[a]	1	1
Temporarily absent (out of area until after deadline)	4	1	0	0	2	—[a]	1	1
Other	1	—[a]	3	2	4	—[a]	1	1
Total	48	10	13	9	154	11	15	11
III. Ineligible cases								
Vacant	25	6	N/A	N/A	N/A	N/A	N/A	N/A
Not a housing unit (e.g., merged, demolished, used solely for nonresidential purposes)	7	2	N/A	N/A	N/A	N/A	N/A	N/A
No member of household in area at least 4 weeks during 1981	1	—[a]	7	5	8	1	10	7
Nonworking, working nonresidential, and other ineligible phone numbers	N/A	N/A	N/A	N/A	743[c]	56	N/A	N/A
Other	7	2	1	1	0	0	3	2
Total	40	10	8	6	751	57	13	9
IV. Indeterminate cases								
Ring, no answers	N/A	N/A	N/A	N/A	65	5	N/A	N/A
Unable to locate	N/A	N/A	9	6	N/A	N/A	29	21
Total	0	0	9	6	65	5	29	21
TOTALS	439	100	144[b]	100	1,322[d]	100	140	100
Response rates								
Lower bound estimate (I/I + II + IV)	—	88	—	84	—	62	—	65
Upper bound estimate (I/I + II)	—	88	—	90	—	70	—	85

[a]Less than 0.5%.
[b]Includes 140 cases originally selected plus 4 secondary reporting units discovered during data collection.
[c]Breakdown of ineligible numbers: nonworking = 613; working nonresidential = 79; and double wrong connection = 21.
[d]Includes 1,318 random numbers originally assigned plus 4 secondary reporting units discovered during data collection.
[e]Includes field follow-up efforts on 56 of the 65 cases that could not be located by telephone. The field staff successfully traced 36 of these and interviewed 27.

near the upper end of the range for more general household surveys conducted by nongovernmental agencies (e.g., Marquis, 1979; Sudman, 1976b: Steeh, 1981). The RDD telephone response rate is also in line with other household telephone surveys conducted by RTI and with response and refusal rates reported in the literature (e.g., Dillman, Gallegos, and Frey, 1976; Lucas and Adams, 1977).

With regard to differences between the two modes, considerable variation is found in the literature with respect to *comparisons* of personal interview and telephone survey response rates obtained in identical or similar studies. For example, in a comparison of a national area-frame survey and two national RDD surveys, Groves and Kahn (1979) reported a response rate of 74% for the personal interviews and an overall response rate of 59%–70% (using the bounded approach described above) for the two telephone surveys. Siemiatycki (1979) compared mixed-mode strategies in a 1974 health survey in Montreal with a similar area-frame survey conducted in 1971–72 and obtained a 74% response rate for the telephone component (before mail and in-person follow-ups) compared with the 84% personal interview response rate achieved in the earlier study. Hochstim (1967) also compared mixed-mode strategies in two household studies conducted in Alameda County, California, and obtained comparative personal/telephone interview response rates for the initial contact mode of 90% vs. 72%, respectively, for one study and 89% vs. 79% for the other. A more recent study conducted in the Los Angeles area by Jordan and his colleagues (1980) achieved telephone and personal interview response rates of 49% and 64%, respectively. In each of these comparisons, the personal interview mode achieved a somewhat higher response rate than the telephone mode, although, except for one of Hochstim's comparisons, differences between the two are somewhat less pronounced than that observed in the CHIPS comparison.

A comparison of the CHIPS field and telephone *list*-frame survey response rates is confounded somewhat by substantial differences in the problems experienced under the two modes in locating sample members and by the fact that most of the unable-to-locate telephone cases were sent to the field interviewers for follow-up. In spite of these confounding factors, however, a comparison can

Table 2
Socio-demographic characteristics of the telephone and personal interview respondents

Characteristic	Telephone			In-person					
				Telephone households			All households		
		N			N			N	
	%[a]	Unweighted	Weighted (in thousands)	%[a]	Unweighted	Weighted (in thousands)	%[a]	Unweighted	Weighted (in thousands)
Sex									
Male	48.6	444	559	45.7	323	531	45.3	393	627
Female	51.4	471	592	54.3	393	631	54.7	482	757
Age									
0–14	21.4	186	289	20.8	171	241	20.9	207	289
15–24	16.0	135	171	11.3	94	131	12.3	127	171
25–44	25.4	207	315	21.8	168	254	22.8	204	315
45–64	18.8	184	312	23.0	146	267	22.5	172	312
65 and over	18.4	203	298	23.1	137	268	21.5	165	298
Education									
Grade school (0–8)	27.1	240	297	27.4	218	311	28.0	275	377
Some high school (9–11)	14.4	123	158	13.6	101	154	13.8	133	186
High school graduate (12)	28.2	267	309	33.3	224	378	33.3	267	449
Some college (13–15)	16.5	142	180	14.4	94	163	14.2	108	191
College graduate (16+)	13.8	97	150	11.3	57	128	10.7	66	144
Race									
White	86.1	819	992	84.0	564	976	81.2	675	1,125
Nonwhite	13.9	96	160	16.0	152	186	18.8	200	260
Hispanic origin									
Yes	6.2	43	69	8.5	37	98	7.4	45	102
No	93.8	844	1,048	91.5	679	1,063	92.6	830	1,282
Family income									
Less than $3,000	2.9	19	28	3.3	32	36	3.9	51	50
$3,000–$4,999	5.1	40	50	7.1	59	76	9.0	88	117
$5,000–$6,999	5.3	47	53	8.1	61	88	8.9	83	115
$7,000–$9,999	10.7	83	106	17.7	123	192	20.1	160	261
$10,000–$14,999	19.5	171	193	17.1	110	186	16.5	135	214
$15,000–$24,999	27.7	211	274	29.9	191	325	27.0	208	350
$25,000 and over	28.9	199	286	16.8	92	183	14.6	96	190

[a]Based on frequencies weighted to account for different probabilities of selection.

be made with regard to the response rates achieved by the two modes for those sample members who were contacted. As noted in Table 1, the field interviewers contacted 127 eligible sample members in the field list-frame sample and interviewed 114, for an upper bound response rate of 90%. The telephone interviewers, on the other hand, contacted 75 sample members in the telephone list-frame sample and interviewed 56, for an upper bound response rate of 75%. Once again, then, we find that the personal interview mode was superior to the telephone mode with respect to response rate and by approximately the same margin found in our comparison of upper bound response rates for the area-frame/RDD surveys. Overall then, allowing for some idiosyncracies in the particular procedures employed in the CHIPS surveys and known variations in response rates by locale and/or different survey organizations, the response rate ranges observed in the FGHSA study do not appear to be unreasonable estimates of what one might expect in telephone and in-person health surveys conducted in other local areas.

Potential for nonresponse bias. In spite of its typicality, the substantial difference in response rates obtained in the telephone and personal interview samples of the CHIPS HIS survey raises the possibility of important differences in nonresponse bias in the two datasets. One way of assessing this potential for differential bias due to nonresponse in the two surveys is to compare respondents to each survey mode in terms of their basic sociodemographic characteristics, as shown in Table 2. Since the sample of all personal interview respondents includes some persons from households that do not have telephones and would thereby not fall within the telephone sampling frame, characteristics of respondents interviewed in person are presented separately for households in the personal interview survey which have a telephone and for all households interviewed face-to-face. While differences in characteristics of telephone and personal interview respondents in general may reflect undercoverage bias due to the exclusion of households without telephones from the telephone frame, differences between telephone respondents and per-

sonal interview respondents in telephone households are more clearly indicative of differences in nonresponse bias.

A comparison of telephone respondents and the *entire* group of in-person respondents reveals few striking differences in sociodemographic characteristics, with the exception of family income. In general, the telephone respondents tend to be younger, better educated, and more likely white than their in-person counterparts. While these differences are generally consistent with prior research (cf. Groves and Kahn, 1979; Cannell, Groves, and Miller, 1981; Aneshensel et al., 1982a), they are relatively small and could result from the exclusion of nontelephone households from the telephone sampling frame. With respect to family income, however, the considerably higher incomes reported by telephone respondents (especially in the highest category) represent a substantial difference, suggestive of an added influence from nonresponse bias in addition to undercoverage effects.

This conclusion is supported by a comparison of the characteristics of the telephone group and in-person respondents living in telephone households. Theoretically, these two groups represent comparable samples from the same household frame, since the percentages in Table 2 have been weighted to account for different probabilities of selection. Nevertheless, the exclusion of nontelephone households from the personal interview group has little impact on the characteristics of this group, and the differences noted in earlier comparisons generally persist, providing a further indication of a differential influence of nonresponse bias in the telephone and in-person interview samples.

Cost comparisons. It is difficult to evaluate comparative cost data reported in the literature for personal and telephone interviews because of the numerous variables involved, including differences in study specifications, variations in the survey components included, dissimilar methods of recovering indirect costs, and differences in start-up costs across organizations. Comparisons of dollar amounts across time are also confounded by the effects of inflation. In spite of these problems, however, it is clear from the literature that telephone interviews are in general substantially less expensive to conduct than in-person interviews. Some examples of the ratios of telephone to personal costs found by other researchers include 27% and 33% (Hochstim, 1967);[5] 29% (Tuchfarber and Klecka, 1976); 40% (Coombs and Freedman, 1964); 43% (Groves and Kahn, 1979); 44% (Lucas and Adams, 1977; Siemiatycki,1979). Thus, the literature suggests that the cost of a telephone survey generally ranges from about one-fourth to one-half the cost of a comparable personal interview survey (cf. Quinn et al., 1980).

Table 3 provides our estimates of the comparative costs for selected survey components of the CHIPS HIS personal and telephone surveys. The personal interview costs include both the area- and list-frame surveys conducted in person, as well as the field costs associated with the 56 telephone list-frame cases sent to the field for follow-up. Telephone costs include the RDD telephone list-frame surveys, exclusive of the field costs associated with the latter. Unfortunately, it was not feasible to compile separate costs for each of the four surveys, as was done for response rates in the previous section. The survey components shown in Table 3—sampling, interviewer recruitment, training, and data collection and quality control—are those that were most sensitive to cost variations by mode. Other survey components (instrument development, preparation of manuals and forms, in-house processing of the survey data, analysis, and overall technical management) have been excluded from the comparison since they were essentially the same for both modes.

Table 3
Comparison of estimated direct costs for sampling and data collection components of the HIS telephone and personal interview surveys[a]

	Survey component	*Telephone interviews*	*Personal interviews*
I.	Sampling		
	Sampling and survey staff salaries	$ 2,704	$ 8,711
	Listing salaries, mileage, and expenses	0	3,514
	Survey staff travel	0	493
	Miscellaneous expenses	56	112
	Total	$ 2,760	$12,830
II.	Interviewer recruitment		
	Survey staff salaries	$ 0	$ 400
	Survey staff travel	0	380
	Total	$ 0	$ 780
III.	Training		
	Survey staff salaries	$ 215	$ 368
	Telephone supervisor salaries	295	0
	Interviewer salaries	638	1,179
	Survey staff travel	0	894
	Interviewer mileage and expenses	0	455
	Miscellaneous expenses	0	219
	Total	$ 1,148	$ 3,115
IV.	Data collection and quality control		
	Survey staff salaries	$ 1,434	$ 3,373
	Telephone supervisor salaries	796	0
	Interviewer salaries	4,384	8,769
	Interviewer mileage and expenses	0	6,249
	Telephone charges	3,609	323
	Postage and shipping	0	989
	Survey staff travel	0	625
	Total	$10,223	$20,328
	Overall total	$14,131	$37,053
	Per interview cost	$34.63[b]	$ 75.31[b]

[a]Excludes overhead and other indirect costs. Direct costs are estimated where exact figures are not available.
[b]Based on a total of 492 personal interviews and 408 telephone interviews.

The considerable difference in sampling costs between the two modes reflects the additional work involved in selecting an area-frame sample vis-à-vis an RDD sample. Clusters must be selected, field listing materials (segment sketches, listing sheets, and maps) prepared, listers recruited and trained, the listing operation conducted and checked, and the final household sample selected. For CHIPS, additional sampling labor was also required because of the problems involved in using dated (1970) census information. Although the relationship of area-frame sampling costs to RDD sampling costs may vary somewhat across organizations, depending on such factors as the availability of prelisted segments and trained listers, area-frame sampling is nevertheless inherently more expensive than RDD sampling.

Interviewer recruitment costs were incurred only for the personal interview mode, since RTI has a fully staffed in-house telephone survey unit. RTI had three "regular" field interviewers in the Tampa area available to work on the HIS; however, it was necessary to recruit five additional interviewers, and the costs shown for this category reflect the effort expended in developing leads and travelling to Tampa to interview applicants.

The eight field interviewers were trained by in-house survey staff in a 2½-day group session held in the Tampa area, while the six telephone interviewers and two supervisors attended a similar session held at RTI. Cost differences here are basically attributable to the travel time and training expenses (room rental, refreshments, etc.) involved in conducting the field training program.

The cost variations in the data collection and quality control category reflect several inherent differences between the two modes. As expected, the field effort required more supervision than the telephone operation, since the field interviewers were working out of their homes while the telephone interviewers were all located at work stations in a single room on the RTI campus. The field interviewers were supervised by an in-house survey staff member,[6] while the telephone interviewers were supervised by two of the supervisors in the Telephone Survey Department. However, a survey staff member also monitored the telephone operation, and we have included this labor in the comparison since this was essentially a supervisory activity. Differences in interviewer salaries reflect the travel time involved in conducting the personal interviews, while the interviewer mileage and expenses, postage and shipping, and survey staff travel are also indicative of other inherent differences between the two modes. Partially offsetting these field expenses are the telephone WATS charges, although the personal interview survey also incurred some telephone charges due to communications with the field staff.

As shown in Table 3, the average cost per interview for the personal mode (N = 492) was $75.31, compared to $34.63 for the telephone mode (N = 408), yielding a telephone-to-personal interview cost ratio of 46%. However, this comparison is confounded somewhat by three factors. First, interviewer assignments for the area-frame sample housing units were randomized at the cluster level as part of a methodological study of interviewer variance. As a result, interviewer mileage and travel time were larger than they otherwise would have been if geographic proximity to the interviewer's residence had been considered in making field assignments. Second, the field interviewers had to make some postinterview callbacks to obtain permission forms from household members with ambulatory care visits or hospital stays who were not at home at the time of the interview. (However, this was not a large cost item, since callbacks were combined whenever possible with other field-work travel and some permission forms were mailed to the interviewers' homes.) Finally, the per interview costs are skewed somewhat in the other direction by the fact that field interviewers completed more interviews than their telephone counterparts, thus reducing the effect of fixed costs (sampling, recruitment, training, and survey staff travel) on the mean per interview cost. Making adjustments for the estimated effect of all these factors (considering geographic proximity in making area-frame field assignments, no callbacks to obtain permission forms, and field interviewers completing the same number of interviews as telephone interviewers), we estimate that the telephone-to-personal interview cost ratio would have been approximately 43% rather than the 46% reported above. With or without these adjustments, however, the HIS comparison of personal and telephone interview costs is clearly consistent with previous reports in the literature; it probably provides a reasonable estimate of the magnitude of cost difference one might expect if health surveys were to be conducted by telephone or in person in another local health planning area.

Response differences. Up to this point in our analysis of the CHIPS telephone-personal interview comparison, we have essentially corroborated prior research suggesting that telephone interviews in an HSA area can be conducted at a substantially lower cost than face-to-face interviews, but likely at the expense of lower response rates and a corresponding possibility of greater nonresponse bias. Generally, however, the ultimate question in such comparisons is the extent to which such sacrifices result in data of comparable quality to that obtained in face-to-face interviews. Most studies of this issue rely on comparisons of response distributions obtained by these two survey modes (as opposed to direct assessments of data quality), and, while many such studies report few differences between the aggregate figures based on telephone and personal interviewing procedures, results of such comparisons derived from recent studies involving the collection of health interview data are somewhat less consistent, as noted earlier.

Comparisons of responses to selected health measures by telephone and personal interview respondents

(the latter presented both for all respondents and only those that have a telephone) are presented in Table 4. As indicated, data presented in the table are weighted only to account for different probabilities of selection (due, for example, to oversampling of certain groups or multiple telephone ownership) and not for differences in nonresponse or undercoverage. Among the few comparisons presented, only one striking difference is apparent: While only 38% of all personal interview respondents were reported as having at least one dental visit during the past 12 months, 48% of all telephone respondents were so reported. This finding is consistent with that reported by Cannell, Groves, and Miller (1981) for dental visits during a two-week period, based on national data, but other comparisons provided in Table 4, including those variables most directly comparable to those used in their analysis (e.g., disability days, doctor visits) provide little evidence of a general tendency for telephone respondents to report more health events than respondents interviewed in-person. On the contrary, in the only other statistically reliable comparison presented, a higher proportion of respondents interviewed in person (7.7%) than by telephone (3.8%) are reported as having at least one functional impairment related to self-care.

Thus, while some differences are apparent in this table, they are less consistent than those observed in the NCHS national comparison (Cannell, Groves, and Miller, 1981). Moreover, because of this lack of consistency, we are inclined to attribute the differences between telephone and personal interview respondents by income and age (see Table 2), respectively, rather than to differences in mode per se. Social class discrepancies in the use of dental care are well documented in the literature (e.g., Anderson and Andersen, 1972; Wilson and White, 1977), as is the greater prevalence of functional disability among the elderly (e.g., Busse and Pfeiffer, 1977; Kane and Kane, 1981).

Overall then, while our comparison of responses to selected health measures by telephone and personal interview respondents does indeed reveal some differences, such differences are inconsistent and do not appear to reflect any obvious trend suggesting the superiority of one mode over the other. Rather, they appear to result from sociodemographic differences between telephone and personal interview respondents—differences which, as noted previously, may be enhanced by a greater influence of nonresponse on the telephone survey data.[7]

Table 4
Telephone personal interview comparisons: percentages[a] of persons with selected health-related characteristics

Health event or behavior	Telephone interview	In-person interview: Telephone households	In-person interview: All households
Disability day in past two weeks	13.3	14.0	13.5
Doctor visit in past two weeks	9.7	8.0	7.8
Dental visit in past twelve months	48.3	41.2	38.5
Hospitalization in past twelve months	10.1	12.4	11.9
Chronic condition in past twelve months	52.6	51.0	49.7
Functional impairment—self-care	3.8	7.1	7.7
Functional impairment—instrumental	3.0	4.3	4.0
Approximate N's Unweighted	896	717	875
Weighted (in thousands)	1,381	1,158	1,149

[a] Adjusted for unequal probabilities of selection only.

Accuracy of reporting. Distributional comparisons by interview mode such as those just presented are valuable in isolating significant differences in responses to telephone and personal interviews. However, the detection of such differences offers only limited evidence for the contention that telephone interviewing yields more or less reliable or valid data than face-to-face interviewing, since these comparisons do not involve a direct validity criterion for judging which of two divergent estimates is the more accurate. Generally, in health service methodological research, it is assumed that the larger of two estimates is the more accurate (e.g., Kovar and Wright, 1973; Cannell, Oksenberg, and Converse 1977b), since studies comparing respondent reports with medical records generally reveal consistent underreporting, while evidence of overreporting is rare (e.g., Cannell et al., 1965; Cannell and Fowler, 1965). The basic comparisons by mode of data collection just described were implicitly based on this widely accepted assumption that higher reporting is generally indicative of "better" or more accurate reporting, although the data presented permit no clear conclusion with regard to the relative accuracy of reporting obtained by the two modes. It should be noted, however, that even where significant differences between these two modes are *not* found, one data collection mode may still conceivably provide more accurate data than the other.

Because a sample of ambulatory care visits and all hospital visits reported in the CHIPS Health Interview Survey were subject to follow-up verification by comparison to medical provider records, the relative accuracy of self-reported condition and use data obtained by the two survey modes can be assessed more directly by reference to these medical record data. Such record-check data are "imperfect," in that they are derived from a "mentioned provider verification" validation design which tends to find overreporting to be greater than underreporting

Table 5
Extent of agreement on ambulatory care visits by person/provider pairs and type of condition

Result	Acute conditions		Chronic conditions		All conditions	
	Telephone interview	In-person interview	Telephone interview	In-person interview	Telephone interview	In-person interview
Exact or partial agreement	71.4	46.9	52.8	45.8	56.3	46.0
Lack of agreement	10.7	20.4	30.9	37.8	27.2	35.0
Loss to follow-up[a]	17.9	32.7	16.3	16.4	16.5	19.0
Total	100%	100%	100%	100%	100%	100%
N	28	49	123	262	151	311

[a]Includes out of scope provider, visits to a dead, retired, or noncooperating provider, and other cases where reconciliation data were not obtained.

(Marquis, 1978). Other studies involving telephone versus personal interview comparisons on health characteristics, which have had such data available, were similarly compelled to rely on less than perfect record-check methods (e.g., Hochstim, 1967; Siemiatycki, 1979; Yaffe et al., 1978), and even data biased toward over-reporting provide useful evidence regarding the relative accuracy of telephone and personal interviews as methods for the collection of health interview data.

Ambulatory care visits. Procedures for determining the accuracy of reported ambulatory care visits involved comparisons made on a patient-provider basis, with all visits to one provider considered at the same time, although assessments of agreement for chronic and acute conditions varied somewhat. For chronic conditions, agreement between the provider and respondent was examined for each provider/condition combination reported by the household respondent. All such conditions reported for that provider were first fully enumerated, then a form completed by the provider was examined to see if he or she reported at least one visit during the specified 12-month reference period for each of the same conditions. Exact agreement was assigned for any condition reported by the household respondent when that condition was also reported by the provider for at least one visit during the reference period, while partial agreement was inferred when at least one visit reported by the provider during the reference period was in the same broad category of conditions as the one reported by the respondent. The comparison of acute conditions followed a similar process except that agreement between the provider and the respondent was examined for each combination of reasons/conditions and dates of visits reported in the HIS. Exact agreement was assigned for a visit when the respondent and the provider reported the same date and the same reason/condition for the visit, while partial agreement was inferred if the respondent and the provider agreed on the reason/condition but not the date, or vice versa.

The results of these comparisons are provided in Table 5, separately for acute and chronic conditions and then for all conditions combined. In each case, the proportion of patient-provider combinations for which either exact or partial agreement was observed is higher among respondents interviewed by telephone than among those interviewed in person, suggesting a greater accuracy of reporting of ambulatory care visits by telephone respondents. However, as noted in our previous descriptions of the HFUS methodology, provider follow-up was not possible for cases where a permission form was not obtained from the respondent, and our success in obtaining permission forms was considerably greater for respondents interviewed in their homes than for those interviewed by telephone. Specifically, we were able to obtain permission forms for 82% of personal interview respondents from whom they were requested, but for only 53% of the telephone interview respondents, primarily due to the logistics of obtaining forms from them by mail.

Given this large difference in our success in obtaining permission forms, it is possible that the differences in agreement observed in Table 5 are more a function of differences by mode in the types of follow-up sample members for whom permission forms were available rather than of mode per se. Indeed, as shown in Table 6, the telephone and personal interview follow-up respondents for whom permission forms were available are quite different in their basic social characteristics. In particular, the personal interview respondents on which this provider follow-up study is based are younger and have lower incomes than those interviewed by telephone, both differences being striking enough to possibly account for the differences observed in Table 5.

However, when the same comparisons are examined in Table 7 *within* each of these sociodemographic subgroups, the observed trend of higher levels of agreement among telephone respondents generally persists. Within each age and income group, in particular, one observes a higher proportion of agreement for telephone respondents. Thus, in spite of substantial differences between the telephone and personal interview follow-up samples, observed differences in accuracy of reporting are apparently not accounted for by these

Table 6
Socio-demographic characteristics of persons in the ambulatory care follow-up sample from whom permission forms were obtained (by mode of interview)

Characteristic	Mode of interview			
	Telephone		In-person	
	%	N	%	N
Sex				
Male	32.5	25	35.0	55
Female	67.5	52	65.0	102
Age				
Under 35	29.9	23	47.1	74
35–64	40.2	31	31.9	50
65 and over	29.9	23	21.0	33
Race				
White	20.8	16	28.0	44
Nonwhite	79.2	61	72.0	113
Family income				
Less than $7,000	48.1	37	71.3	112
$7,000 or more	44.2	34	26.2	41
Refused	7.7	6	2.5	4

sociodemographic differences.

While it is therefore tempting to interpret these differences as greater accuracy of reporting by telephone respondents, there are, of course, other potential differences not entirely assessed by these characteristics that may account for these results. For example, it may well be that persons who are conscientious enough to return a signed permission form by mail are more meticulous about detail in general, therefore more accurate in reporting their ambulatory care visits. And "conscientiousness" is not a trait necessarily related to age or socioeconomic status.

Hospitalizations. As noted earlier, procedures for determining the accuracy of hospital stays were somewhat different than those for ambulatory care visits. Information on hospital visits reported by household respondents were compared with data abstracted from hospital records and agreement codes assigned as specified in Table 8. In addition to specifying agreement or lack of agreement on the fact of hospitalization at the time reported by the respondent, varying levels of agreement or lack thereof on the specific condition(s) reported were coded.

Overall, consistent with the analyses presented for ambulatory care visits, a slightly higher proportion of agreement on *either* the fact of *or* reason for stay is evident for hospitalizations reported by telephone (88%) than for those reported in personal interviews (79%). Moreover, *exact* agreement on both the fact and condition of stay is more frequent for telephone-reported hospitalizations (35% versus 28%). However, when the two other categories reflecting at least partial agreement on condition are also considered, there is very little difference between respondents to the two modes in their tendencies to provide hospitalization information consistent with hospital records on both of these dimensions, and the extent to which hospital visits are overreported is essentially the same within these two groups.

Nevertheless, within this general pattern there are some additional interesting trends. Lack of agreement on condition is almost twice as high for hospitalizations reported in person than by telephone (15% versus 8%), and a high proportion of personal interview respondents failed to provide information sufficient for a match to be made at all.[8] In contrast, telephone respondents were much more likely not to report a condition at all. Less than 8% of personal interview respondents failed to specify the condition for which they were hospitalized, and nearly two-thirds of these unspecified cases were found to be *sensitive* conditions (e.g., the patient said "don't know," while the hospital reported a genital or urinary tract condition). On the other hand, almost one-quarter of the telephone respondents in the follow-up study said that they could not specify the conditions associated with their hospital stay, and only about one-quarter of these turned out to be sensitive conditions.

Thus, at least with regard to these conditions, it would appear that respondents interviewed in person make a somewhat greater effort to report more fully than do telephone respondents. However, considering the fact that personal interview respondents manifest a higher proportion of clear disagreements between their reports of such conditions and what is indicated in hospital records, this greater effort apparently results in more reporting errors. Thus, although telephone respondents are substantially less likely to actually report

Table 7
Extent of agreement on ambulatory care visits by person/provider pairs according to mode, by selected socio-demographic characteristics

Characteristic	Mode of interview			
	Telephone		In-person	
	% Agree	N	% Agree	N
Sex				
Male	64.7	51	45.2	93
Female	52.0	100	46.3	218
Age				
Under 35	63.2	38	43.8	137
35–64	51.3	76	47.9	117
65 and over	59.5	37	47.4	57
Race				
White	60.0	120	44.6	233
Nonwhite	43.3	30	50.6	77
Family income				
Less than $7,000	56.9	72	42.7	220
$7,000 or more	56.2	73	53.6	84
Refused	50.0	6	57.1	7

Table 8
Results of follow-up patient reported hospitalizations by mode of interview

Result of comparison	Mode of interview: Telephone N	Telephone % of total visits	In-person N	In-person % of total visits
I. Agreement on stay and condition				
Exact agreement on primary condition[a]	27	35.0	39	27.9
Agreement on hospital-related secondary[b] condition	5	6.5	13	9.3
General agreement (same general condition or body system	9	11.7	23	16.4
Total	41	53.2	75	53.6
II. Agreement on stay but not on condition				
Patient reported vague symptoms or condition *consistent* with hospital's report	2	2.6	4	2.8
Patient did not specify a condition	19	24.7	11	7.9
Lack of agreement on condition[c]	6	7.8	21	15.0
Total	27	35.1	36	25.7
III. Lack of agreement on stay				
Patient overreporting[d]	8	10.4	18	12.9
Patient data insufficient to allow match	1	1.3	10	7.1
Total[e]	9	11.7	29[f]	20.7
Total visits reported	77	100.0	140	100.0
Total number of patients reporting	61	—	107	—

[a]Same three-digit IDC-9 code or one very similar in meaning.
[b]Condition reported by patient listed by hospital as secondary condition.
[c]Widely divergent conditions reported by patient and hospital.
[d]Patient reported visit not in hospital's records.
[e]Excludes a few cases of underreporting where visits in hospital record were not reported by patient.
[f]Includes one obvious hospital error.

the condition(s) for which they have been hospitalized, a higher proportion of the conditions they do report are accurate.

However, as in the case of ambulatory care visits, we were able to acquire permission forms from a substantially higher proportion of persons who reported hospitalizations in personal interviews (82%) than by telephone (55%). As a result, personal interview respondents included in the hospitalization follow-up are less affluent, somewhat younger, and more likely to be male than telephone interview respondents in these comparisons. Unfortunately, the results presented in Table 8 have not yet been analyzed by sex, age, or income to examine potential effects of these differences in patient characteristics on the hospitalization follow-up results.

Summary and conclusions

In this paper we have described a methodological comparison and evaluation of the personal and telephone interview modes as a means of collecting health data from households. Results were analyzed from the standpoint of five key criteria: response rates, potential evidence of nonresponse bias, costs, response differences, and accuracy of reporting. The findings are summarized below.

The personal interview mode was clearly superior to the telephone mode with respect to response rate. The area-frame survey produced an 88% response rate compared to a 62%–70% response rate for the RDD telephone survey (with the range reflective of the "ring, no answer" numbers), while in the list-frame survey of Medicaid eligibles, the personal mode yielded a 90% response rate among contacted sample members compared to the telephone mode's 75%. These results are consistent with other personal/telephone comparisons reported in the literature, although the degree of difference between the two modes is somewhat greater than that found in most of the previous comparisons.

The relative potential for achieving an adequate response rate with the two modes is an issue, of course, because of concerns about bias due to nonresponse. Indeed, our comparison of the sociodemographic characteristics of personal and telephone respondents in the CHIPS HIS study did reveal some differences, with telephone respondents tending to be younger, better educated, more likely white, and having higher family incomes than personal interview respondents. Although most of these differences were relatively minor and are consistent with prior research, the magnitude of the income difference, with telephone respondents reporting considerably higher incomes, may indicate that the lower level of response achieved by telephone resulted in some differences in nonresponse bias between the two modes.

Our cost comparison of the two modes estimated a telephone-to-personal cost ratio of approximately 43%.

This finding is consistent with previous cost comparisons reported in the literature, which indicate that the cost of a telephone survey can generally range from about one-fourth to one-half the cost of a comparable personal interview survey.

In our comparison of actual responses obtained by the two survey modes, we observed some differences in reported health-related characteristics, but these variations were not very consistent, and likely reflect differences in the sociodemographic characteristics of the telephone and personal interview respondents, rather than mode differences per se. Thus, from the limited number of response comparisons presented in this paper, it is not possible to conclude that there were any important mode effects in the reporting of health variables.

The relative accuracy of respondent reporting of ambulatory visits and hospital stays was measured through follow-up record checks with providers. While the telephone respondents appear to be somewhat more accurate in their reporting, this difference could be the result of our disparate success in securing permission forms for the release of medical record information. One interesting finding was that personal interview respondents appeared to make a greater effort to report a reason or condition for hospital stays, which may help to account for their higher inaccuracy rate.

Of the five criteria evaluated in this paper, the response rate and cost comparisons are clearly the most compelling, while further research is called for in the areas of nonresponse bias, response differences, and accuracy of reporting. In particular, additional record-check studies using two-directional designs (Marquis, 1978) that directly assess the relative accuracy of reporting of ambulatory care visits and hospitalizations by mode of interview would clearly be desirable. Given the limited resources available to local health planners, however, the telephone mode or a mixed-mode approach (e.g., Siemiatycki, 1979) may well be the wave of the future. Thus, the task confronting health researchers is to continue investigations into these other critical areas and find ways to measure and if necessary (and possible) adjust for these effects.

Footnotes

[1] A local version of the HANES was ruled out during the planning stage due to cost considerations.

[2] A total of five such prefixes were excluded—three that served commercial numbers exclusively and two with fewer than 1% of their numbers located in the sample area.

[3] In the field all adult members of the household who were present at the time of the interviewer's visit were invited to participate in a group interview, although feedback from the field interviewers indicates that such group interviews were a rarity.

[4] It should be recalled, however, that the low-SES domain was oversampled in the area-frame survey while no oversampling was involved in the RDD telephone survey.

[5] Prior reports in the literature of cost ratios presented by Hochstim are confounded somewhat by the fact that some of the telephone interview strategy interviews were conducted in-person or by mail. To avoid this confounding and derive comparisons comparable to the others presented here, these percentages are based on direct data collection costs for an initial telephone wave from this mixed-mode strategy in relation to an initial personal interview wave.

[6] While RTI frequently uses an off-site field supervisor in area-frame surveys, this option was not available to us on the HIS since we did not have a supervisor based in Tampa. Nevertheless, it is unlikely that the use of an in-house supervisor has greatly skewed the cost comparison, since the two supervisory methods are comparable in terms of direct costs.

[7] In some analyses not presented, two types of postsurvey adjustments were made in an effort to reduce any observed differences by mode due to nonresponse and undercoverage: (1) a weighting class adjustment using four county strata as weighting classes in the telephone frame and eight county-by-socioeconomic-status strata in the area frame; and (2) a poststratification adjustment (based on population estimates for the area provided by the University of Florida) using eight age-by-county strata. Neither adjustment had any significant impact on either the percentage estimates for each mode or the magnitude of differences between modes as presented in Table 4. However, this lack of effect was most likely due to the small number of weighting classes and poststrata used in these adjustments rather than to the influence of a true mode effect independent of nonresponse and undercoverage biases.

[8] Some of these visits were outside the reference period and thereby not abstracted. Additional analysis is currently underway to determine the number of cases for which this is true.

Nonparticipation in telephone follow-up interviews*

Alfred C. Marcus, UCLA-Jonsson Cancer Center

Carol W. Telesky, UCLA-Jonsson Cancer Center

Introduction

Survey research methodology has recently rediscovered the telephone, and survey researchers in the health professions have been at the forefront of this movement. For example, in 1975 the National Center for Health Services Research and the National Center for Health Statistics sponsored the first conference on health survey research methodology. As noted by Reeder (1977): "The advantages presented had a decided edge over the disadvantages. It was the consensus that, with proper strategy telephone interviews can be efficient" (p.3).[1]

With the recent emergence of telephone interviewing, panel surveys (i.e., surveys in which people are reinterviewed over time) have become more feasible economically. However, a serious bias can occur in panel surveys when people fail to complete the follow-up interviews. Perhaps the most prominent study of this problem in morbidity surveys was reported over a decade ago by Bright (1967 and 1969). She found that nonwhites were somewhat harder to locate for face-to-face follow-up interviews, as were people 20–29 years of age, the lower socioeconomic groups, divorced or separated people, and those who made a residential move. The findings reported by Bright involved a face-to-face follow-up of respondents after a hiatus of at least five years. In this paper we have examined the same problem in a health survey that made use of repeated telephone interviews for approximately one year.

Methodology

Sample. The data to be reported were collected during a one-year study as part of the 1976–1977 Los Angeles Health Survey (LAHS). The sample design for the LAHS was a three-stage random probability sample of Los Angeles County that was developed by the Institute for Social Science Research (ISSR) at UCLA (Sumner, 1976). From October through December 1976, demographic and health data were obtained from 1,210 respondents during the initial face-to-face interviews. The initial interview was then followed by seven telephone reinterviews at approximately six to eight week intervals and a final face-to-face interview at the end of the panel year.

Independent variables. In the analyses reported below, a number of variables were used to predict nonparticipation in the follow-up telephone interviews. Included were standard sociodemographic measures, a general SES score for the household from Duncan's socioeconomic index (a composite ranking of education and occupation), total family income, number of children in the household, and number of adults in the household. In addition, several health related measures were used to predict nonparticipation, including respondent's self-rated health status (1 = excellent, 2 = good, 3 = fair, 4 = poor), number of chronic conditions, self-reported tendency to delay seeking care (1 = long, 2 = moderate, 3 = no delay), and the number of restricted activity days and acute illness episodes reported for the two months preceding the first interview.

Results

Subgroup differences in the number of completed interviews. Table 1 reports the distribution of completed interviews in the 1976–1977 LAHS panel survey. As shown, approximately 45% of the original sample completed all follow-up interviews, while another 18.5% missed only one interview. Although nearly three out of four respondents completed six or more interviews, 20% of the original sample completed less than half the interviews.

Table 1
Percent distribution of number of completed interviews

Number of completed interviews	*Percent of sample (N = 1,210)*
8 (all)	45.2
7	18.5
6	8.8
5	4.6
4	3.0
3	2.8
2	2.6
1	4.7
0 (none)	9.8

$\bar{X}$ = 5.8, S.D. = 2.7, Md = 7.24

Although these completion rates are lower than one might hope for, the key issue is whether the number of completed interviews is a random phenomenon or whether it systematically varies by population subgroup. If the rate of noncompleted interviews represents random error, then it should not have a great impact on comparative analyses.

Table 2 reports the main findings from our analysis of

* Data reported in this paper were collected pursuant to grant number 5-R18-CA-18451, "Processes in Health Behavior and Cancer Control," awarded by the National Cancer Institute to the late Leo G. Reeder.

subgroup differences in participation levels. The first column of Table 2 reports bivariate correlations between the predictor variables and the number of completed interviews. The first group of predictors represent the health related variables obtained at the initial face-to-face interview. As shown, only one of these correlations was statistically significant; people who began the survey rating their health as fair or poor were significantly less likely to participate in the follow-up interviews than people who rated their health as good or excellent.

Table 2
Health status and socio-demographic variables as predictors of number of completed interviews

	Number of completed interviews	
	Zero-order correlations	*Regression coefficients (standardized)*
Subjective health status	−.14***	−.14***
Restricted activity days (2 months)	−.05	−.01
Number of acute episodes (2 months)	−.01	−.01
Number of chronic conditions	−.04	.03
Tendency to delay	.05	.05
		R^2 = .02
Education (years)	.15***	.02
Age	.06*	.11***
Sex	−.03	.05
SEI (socioeconomic index)	.18***	.04
Total family income	.23***	.17***
Marital status	.10***	.04
Ethnicity (white vs.other)	.14***	.04
Number of children in household	−.09***	−.05
Number of adults in household	.04	−.01
Employment status of respondent	.13***	.09**
		R^2 = .07
	Total	R^2 = .09

*p ≤ .05.
**p ≤ .01.
***p ≤ .001.

The second group of predictor variables represents a standard list of sociodemographic characteristics. Although most of these correlations are significant, only one of the variables (total family income) has a correlation exceeding .20. Thus, there was a moderate tendency for middle- and upper-income people to participate more frequently in the follow-up interviews. Other subgroups that were more likely to complete the follow-up interviews included people with a high school education or above, older respondents, married respondents, White/Anglos, people with no children in the household, and respondents employed full time.

In the second column of Table 2 the standardized regression coefficients are presented for each of the health-related variables. These coefficients were obtained from a two-step hierarchical regression in which the health variables were entered first, followed by the sociodemographic variables. As shown, subjective health status remained significantly related to the number of completed interviews. However, the combined effects of all health-related variables explained only 2% of the variance in the number of completed interviews. Among the sociodemographic variables only age, income, and employment status were related to the number of completed interviews, and the combined effects explained approximately 7% of the variance for a total R^2 of 9%

The above analyses were repeated using several different versions of the dependent variable (i.e., the number of completed interviews). In one analysis, the subgroup that never missed an interview (N = 547) was compared to the subgroup that missed at least one reinterview (N = 663). The results were virtually identical to those reported above. That is, self-reported health status was again related to nonparticipation ($\beta = -.16$; $P<.001$), as were age ($\beta = .08$; $P<.05$), and total family income ($\beta = .10$; $P<.001$). Additionally, married respondents were more likely to have a perfect reinterviewing record than nonmarried respondents ($\beta = .08$; $P<.05$).

In yet another analysis, the subgroup that was successfully reinterviewed at the last interview (N = 903) was compared to the subgroup that was not reinterviewed at the last interview (N = 307). As before, the significant predictors included self-rated health status ($\beta = -.14$; $P<.001$), age ($\beta = .12$; $P<.001$), total family income ($\beta = .13$; $P<.01$), employment status ($\beta = .11$; $P<.01$), and marital status ($\beta = .08$; $P<.05$). Thus, in all three analyses the people least likely to be reinterviewed included younger respondents, people with lower total family incomes, and those rating their overall health as fair or poor. In two of the three analyses nonmarried and nonemployed respondents were also less likely to be reinterviewed.

Despite the statistically significant differences between those who were successfully reinterviewed and those who were not, the impact of these differences on the characteristics of the sample remain quite modest. This is illustrated most clearly in Table 3, where we compare the original sample of 1210 with the subsample of 903 that were successfully reinterviewed at the final interview. As shown, the percentage differences between the two samples are uniformly small, never exceeding 5% for any given subgroup. Moreover, those subgroups showing relatively high rates of attrition (see last column of Table 3) tend to be relatively small in absolute numbers, thereby reducing their impact on the sample characteristics at the final interview. For example, although people rating their health as poor were much less likely to be reinterviewed (attrition = 41%), they constitute only about 5% of the total sample. Thus, while certain subgroups were indeed less likely to be reinterviewed during the study year, this did not seem to greatly affect the sample at the final interview.

Subgroup differences over time. In Figure 1 the completion rates have been reported for each of the eight follow-up interviews. As shown, there was a precipitous 20% decline in the rate of participation between the

Table 3
Comparison of original sample (N = 1210) with sample at final interview (N = 903)

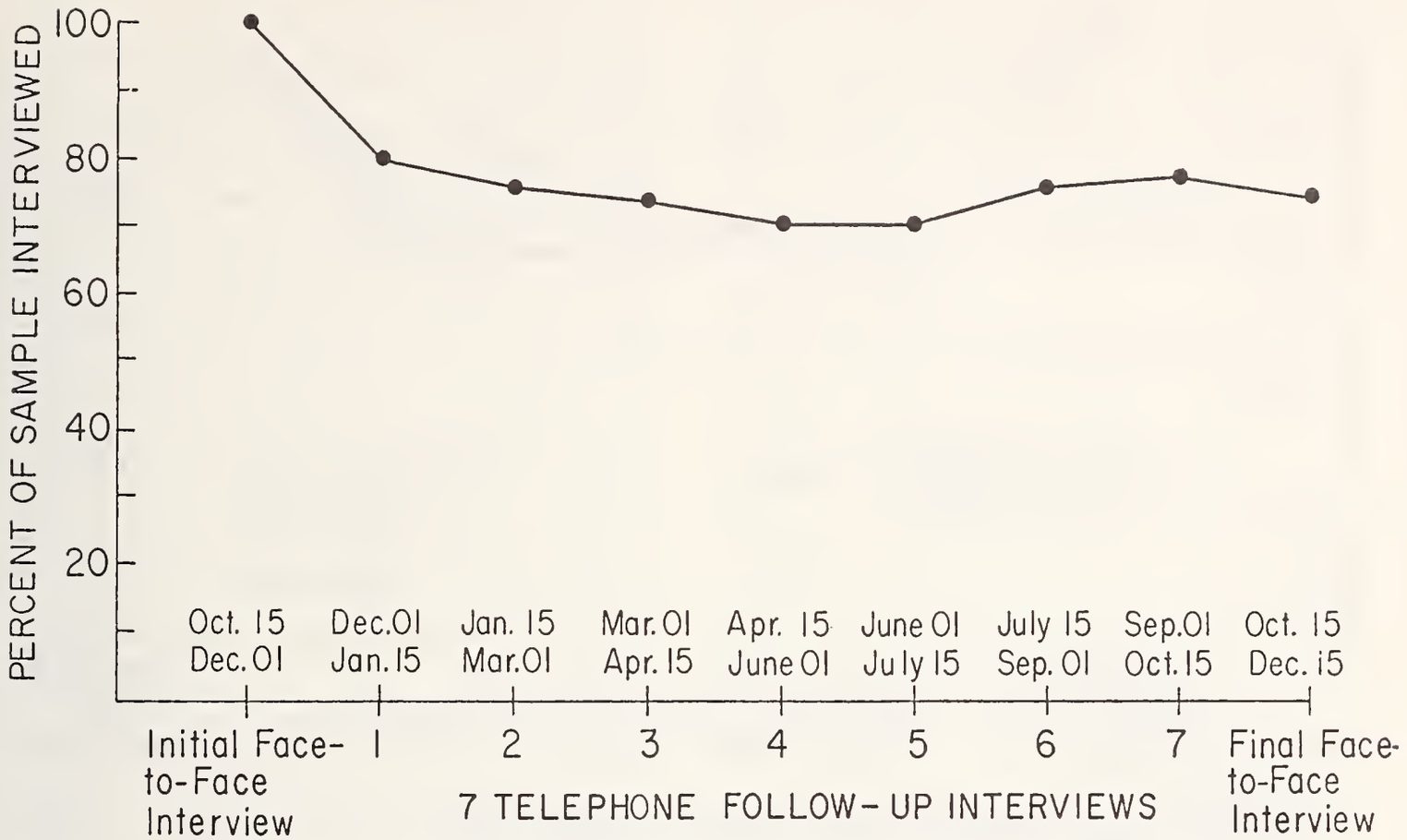
Variables	Original sample N	Original sample %	Sample at final interview N	Sample at final interview %	Attrition rate %
Subjective Health Status					
Excellent	434	(35.9)	346	(38.3)	20.3
Good	504	(41.7)	376	(41.6)	25.4
Fair	206	(17.0)	142	(15.7)	31.1
Poor	66	(5.5)	39	(4.3)	40.9
Restricted days (past two months)					
0	796	(65.8)	608	(67.3)	23.7
1	79	(6.5)	53	(5.9)	32.9
2	74	(6.1)	53	(5.9)	28.4
3+	261	(21.6)	189	(20.9)	27.6
Acute problems (past two months)					
0	855	(70.7)	643	(71.2)	24.8
1	298	(24.6)	216	(23.9)	27.6
2	48	(4.0)	38	(4.2)	20.9
3+	9	(0.7)	6	(0.7)	33.3
Chronic problems					
0	586	(48.4)	435	(48.2)	25.8
1	365	(30.2)	284	(31.5)	22.2
2	174	(14.4)	124	(13.7)	28.8
3+	85	(7.0)	60	(6.6)	29.4
Age					
18–29	356	(29.4)	248	(27.5)	30.4
30–39	247	(20.4)	186	(20.6)	24.7
40–49	180	(14.9)	138	(15.3)	23.4
50–59	181	(15.0)	142	(15.7)	21.6
60+	246	(20.3)	189	(20.9)	23.2
Sex					
Male	526	(43.5)	402	(44.5)	23.6
Female	684	(56.5)	501	(55.5)	26.8
Total family income					
<$5,000	188	(15.5)	120	(13.3)	36.2
$5,000–$9,999	291	(24.0)	190	(21.0)	34.7
$10,000–$13,999	220	(18.2)	168	(18.6)	23.6
$14,000–$19,999	222	(18.3)	177	(19.6)	20.3
$20,000–$24,999	117	(9.7)	92	(10.2)	21.4
$25,000+	172	(14.2)	156	(17.3)	9.3
Employment status					
Nonemployed	463	(38.3)	313	(34.7)	32.4
Part-time	107	(8.8)	79	(8.7)	26.2
Full-time	640	(52.9)	511	(56.6)	20.2
Education					
Grade school	167	(13.8)	111	(12.3)	33.6
Some high school	163	(13.5)	107	(11.8)	34.4
High school	299	(24.7)	225	(24.9)	24.8
Some college	332	(27.4)	256	(28.3)	22.9
College	112	(9.3)	87	(9.6)	22.4
Post college	136	(11.2)	116	(12.8)	14.8
Ethnicity					
White/Anglo	819	(67.7)	633	(70.1)	22.8
Hispanic	207	(17.1)	146	(16.2)	29.5
Black	125	(10.3)	85	(9.4)	32.0
Other	59	(4.9)	39	(4.3)	33.9
Marital Status					
Married	670	(55.4)	531	(58.8)	20.8
Not married	540	(44.6)	372	(41.2)	31.2
Children in household					
0	721	(59.6)	542	(60.0)	24.9
1	181	(15.0)	142	(15.7)	21.6
2+	308	(25.5)	219	(24.3)	28.9
Adults in household					
1	394	(32.6)	281	(31.1)	28.7
2	668	(55.2)	517	(57.3)	22.6
3+	148	(12.2)	105	(11.6)	29.1

initial face-to-face interview and the first telephone callback interview. This pattern is consistent with other longitudinal surveys which show that losses due to sample attrition are much more severe at the beginning of the survey. After this initial 20% loss, the decline in participation continued through the fifth reinterview, but was much less severe. For example, at the second callback there was an additional decline of 5%, while the proportion of completed interviews dropped another 6% between the second and fourth callbacks. At the fourth callback, however, the proportion of completed interviews leveled off at approximately 70%.

At this juncture in the survey, ISSR implemented additional field procedures to increase the number of completed interviews. Following standard survey procedure, ISSR mailed personalized letters to all respondents who had refused to continue in the survey. The purpose of this letter was to thank respondents for their participation and to remind them of the importance of their continued participation. It was hoped that this letter might persuade some of the "soft" refusals to continue in the study. However, only one of 90 refusals was converted by this letter.

Disconnected telephone numbers were by far the most frequent reason for long-term nonparticipation, and this usually signalled a residential move. Respondents who had moved were mailed a personalized letter with forwarding addresses requested from the postal service. For respondents with a forwarding address, a second letter was then mailed asking them to continue in the survey. For people who had moved with no forwarding address, the Department of Motor Vehicles was contacted for a change of address. Addresses identified in this fashion were used to mail these people a personalized letter soliciting their continued participation. Of the 174 people who had disconnected telephone numbers, 29 (or 16.6%) were brought back into the study. An additional 25 respondents had per-

Figure 1
Percent of sample interviewed at each callback in the 1976–1977 Los Angeles health survey



sistently low rates of participation for other reasons (e.g., temporarily incapable, not available for interview). These people were also mailed personalized letters and seven were reactivated.

The impact of this effort on rates of participation is clearly evident in Figure 1. That is, at the sixth callback we see a modest 6% increase in the rate of completed interviews, half of which is attributable to the personalized respondent letters (N = 37). Additionally, at this point in the survey, interviewers were given a $2.00 bonus for each person they reinterviewed who had a poor record of participation (but had not officially dropped out of the study). Thus, much of the remaining increase in participation levels at the sixth callback may be due to this financial incentive offered to the interviewers. As shown in Figure 1, the increase in the proportion of completed interviews at the sixth callback continued through the seventh callback and showed only a very slight decline at the last interview.

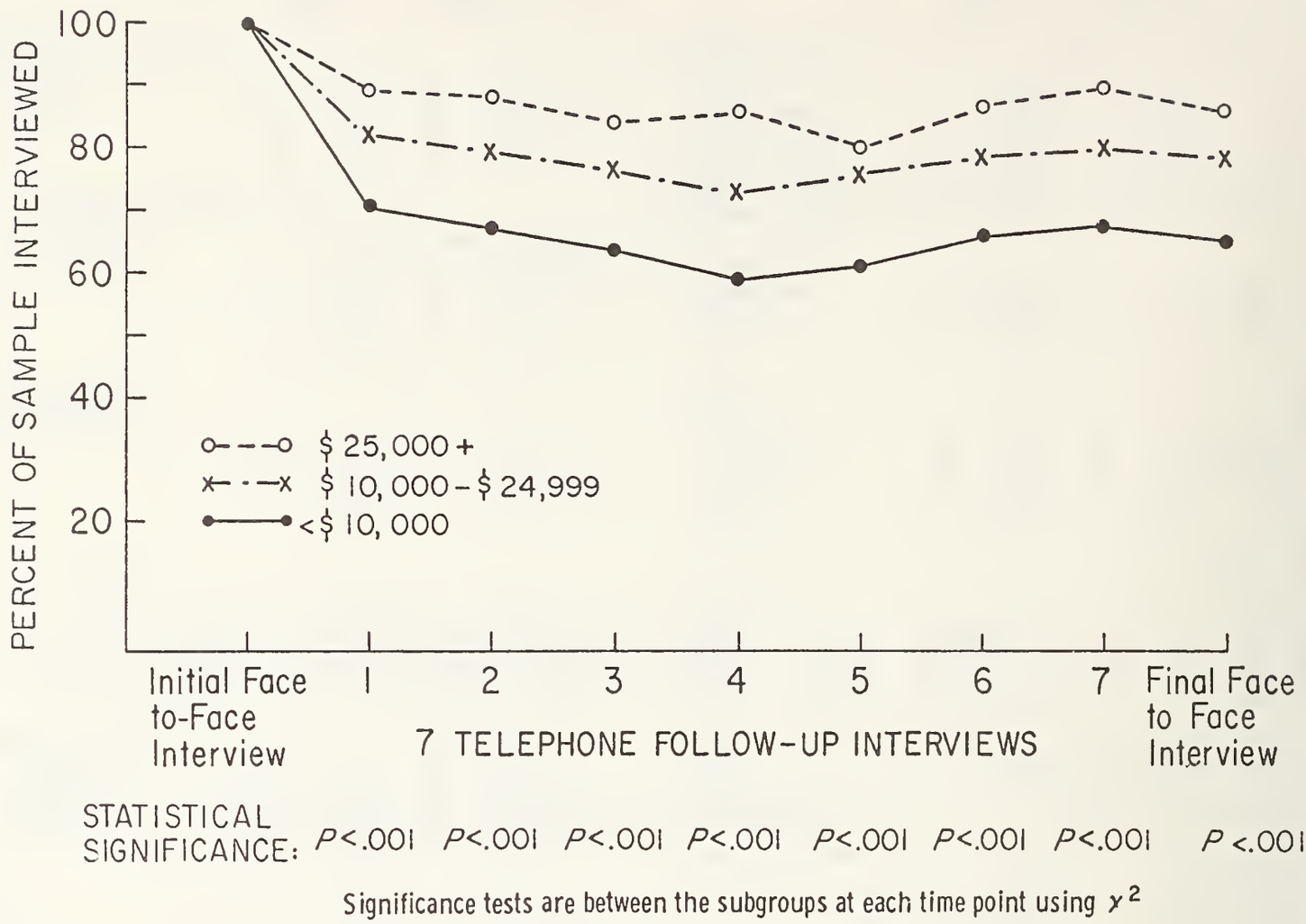
As noted earlier, two of the best predictors of participation in the follow-up interviews were total family income and self-rated health status. In Figures 2 and 3 completion rates have been reported for subgroups that differed on both of these variables. The pattern that was characteristic of the entire sample is reflected in all three income groups shown in Figure 2. That is, beginning at the first follow-up interview, we see a sharp decline in completion rates that level off at the fourth or fifth callback and then increase again peaking at the seventh callback. Perhaps the most interesting aspect of Figure 2 is the consistency of the income differences, with the upper-income group showing a consistently higher rate of participation than the middle-income group, which in turn has a higher completion rate than the lower-income group.

In Figure 3 we see a slightly different pattern. People who began the survey rating their health as fair or poor were consistently less likely to complete the follow-up interviews. However, these people did not show as sharp an increase between the fourth and seventh callbacks as the other subgroups. In fact, of all the subgroups examined in the analysis, this was the only group that showed less than a 5% increase between the fourth and seventh callbacks.

Discussion

We were encouraged to find that nonparticipation was not strongly related to the health variables obtained during the initial interview; only 2% of the variance in

Figure 2
Percent of sample interviewed at each callback by three levels of total family income



Significance tests are between the subgroups at each time point using χ^2

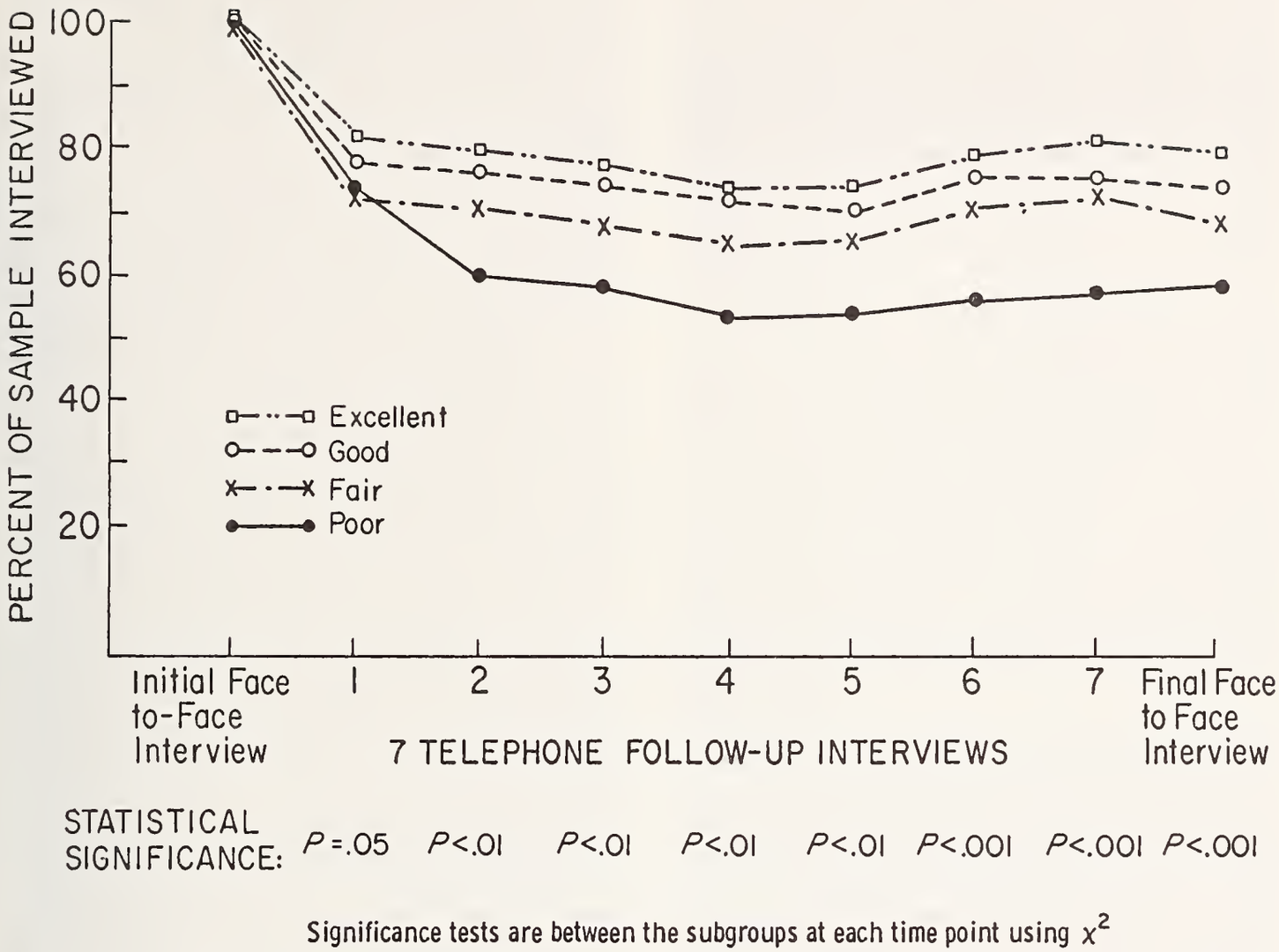
the number of completed interviews could be explained by health-related variables, and most of this variance was accounted for by a subjective rating of health status.

Although nonparticipation in the follow-up interviews was unrelated to initial reports of morbidity or disability, there were important sociodemographic differences. In particular, people who had lower levels of participation were more likely to be younger and have lower family incomes. These findings are consistent with Bright's (1967 and 1969) analysis and concur with findings from other types of panel surveys, including surveys of voting and economic behavior. For example, in an early panel study of economic attitudes, Sobel (1959) found that panel losses were highest in large metropolitan areas, among people under 25 years of age or 65 and over, and among people with lower incomes. More recently, Lansing and Wolfe (1971) reviewed four panel surveys conducted by the Institute for Social Research and found that sample mortality was more likely to occur among younger adults, people 65 years of age and over and people living in the central cities. While Lansing and Wolfe found few differences in sample mortality by racial/ethnic subgroup, there was a slight tendency for lower-income households to participate less in the panel surveys.

Most longitudinal surveys show the largest decline in participation immediately following the initial interview, and our survey was no different in this regard. Following the initial interview there was a 20% decline in the rate of participation that was never recouped. However, between the fifth and seventh callbacks there was a modest 8% increase in rates of participation that could be attributed in large part to personalized respondent letters and a monetary bonus offered to interviewers. The personalized letter technique was particularly useful in locating and reactivating respondents who had disconnected telephone numbers—by far the most common reason for nonparticipation in the 1976–1977 LAHS.

In summary, then, these findings suggest that nonparticipation in telephone follow-up interviews could be a problem in health surveys—particularly among younger respondents and the lower socioeconomic groups, and among people who rate their overall health as poor. Bright has shown that the loss of such people is due more to residential mobility than to refusal to participate. Consequently, the methods used to locate missing respondents represent important components of the survey design. Traditional methods for locating missing respondents include the use of the postal service and telephone companies, searches of public records (e.g.,

Figure 3
Percent of sample interviewed at each callback by self-rated health status



Significance tests are between the subgroups at each time point using x^2

marriage licenses, Department of Motor Vehicles, etc.), and actual community searches in which people are sent into the community seeking leads from neighbors, storekeepers, bartenders, etc. (Bright, 1967; Eckland, 1968). Since residential mobility is most likely to be local, it is essential that protocols be developed for tracking mobile cases within the same community. The letter technique described above is one example of an effective protocol for tracking such cases, providing interviewers with financial incentives may be another technique worth considering. There is some evidence to suggest that a search of public records is not always cost effective (Crider, Willits, and Bealer, 1971) while intensive community follow-up appears particularly useful in tracking younger respondents and the lower socioeconomic classes (Bright, 1967)—two subgroups which had significantly higher rates of sample attrition in this study.

Given the almost immediate loss of 20% of the sample following the first interview, researchers that rely on repeated telephone interviews would be well advised to consider strategies for reducing sample attrition *before* fieldwork begins. All of the methods described above are implemented after the fact. To minimize attrition at the outset, respondents should feel that their continued participation is critical to the success of the study. Additionally, the interviewer should obtain from the respondents enough "anchor points" (Crider et al., 1971) to help in subsequent follow-up activities. These anchor or reference points range from the names and addresses of relatives, friends, co-workers and employers, to the respondent's high school, birthdate (as opposed to age in number of years), and social security number. In regard to the costs associated with respondent follow-up, it should be noted that the various search procedures described above need not be applied to all missing respondents. Kalsbeek and Lessler (1978) have shown, for example, that standard imputation techniques (i.e., statistically derived estimates for missing data) can be combined with an intensive *partial* follow-up to reduce the biases caused by sample attrition.

Footnote

[1] Until recently, the main problem in telephone surveys has been a socioeconomic bias among households with telephones. Given that 95% of the households now have telephones, socioeconomic biases are much less of a problem today. In addition, the development of "random digit dialing" techniques have circumvented sampling problems caused by unlisted or new telephone numbers. Although telephone surveys may have somewhat higher rates of refusals and missing data, the magnitude of these differences has generally been modest (Aneshensel, Frerichs, Clark and Yokopenic, 1982b; Groves, 1979b; Jordan, Marcus and Reeder, 1980; Siemiatycki, 1979; Yaffee, Shapiro, Fuchsberg et al., 1978). Moreover, in reports of sociodemographic characteristics and rates of morbidity, the two modes of interviewing have generally been found comparable. However, there is some evidence to suggest that telephone surveys may yield somewhat lower counts of doctor visits than face-to-face interviews (Siemiatycki, 1979; Yaffee et al., 1978). This potential problem should be manageable, given proper questionnaire development and interviewer training. Siemiatycki (1979) has suggested that reports of doctor visits might be enhanced in telephone surveys if respondents are encouraged to take their time when answering questions. The use of "memory aid" devices in health surveys is also intended to obviate this problem (Marcus, 1982).

A comparison of telephone and personal interviews in the health interview survey*

Peter V. Miller, Survey Research Center, The University of Michigan

Introduction

Telephone surveys evoke a notably ambivalent reaction among survey practitioners. Cost savings, advances in sampling and interviewing methods, and new technology (computer-assisted interviewing) make the telephone a preferred survey medium for many researchers. But, for others, these apparent advantages are outweighed by problems of nonresponse and noncoverage and by perceived shortcomings in data quality.

The reports of previous biennial conferences reflect the mix of positive and negative beliefs about telephone surveys. At Airlie House in 1975, much of the discussion about using the telephone in surveys focused on "mixed mode" designs, in which telephone interviews were used to supplement personal interviews or in which phone contacts were made in follow-up waves of panel studies. The telephone was viewed by participants at that meeting as a useful tool in that supplementary role. It was reported that telephone follow-up interviews compared favorably in several surveys with in-person recontacts in both response rate and reliability across waves. Further, conferees guessed that there might be lower between-interviewer variance when interviewers used the telephone for contacting respondents. But, there was also discussion of real and potential problems with telephone surveys. The subject of "single mode" telephone surveys was dominated by discussion of undercoverage in telephone samples. Discussion of telephone interview dynamics was of the "potential problem" variety. Conference participants speculated on limitations on question types and interview subject matter in phone contacts, debated the effect of lack of visual aids for questions, discussed the optimal length for phone interviews, and expressed the feeling that special interviewing techniques might be required for the telephone (but could not identify what they might be).

On the whole, the Airlie House meeting produced a positive evaluation of telephone contacts, particularly in mixed-mode designs. The difficulties identified were largely "potential" ones, and conferees concluded that there were "few obvious limits on the utility" of the telephone for surveys.

At Reston in 1979, the discussion of telephone surveys by Biennial Conference participants shifted focus from "mixed mode" designs to surveys conducted solely by telephone. At the same time, the discussion became more pessimistic. After Bob Groves discussed design features of a CATI system (Groves, 1979a), Garth Taylor reminded us that there were still no rules on what is a good telephone interviewer and that there are big differences in response rates between in-person surveys and "cold" telephone interviews (one where no previous contact is made in person or by letter)(Taylor, 1979). Reporting the results of a telephone-personal comparison, Jordan, Marcus and Reeder noted that their telephone respondents had more missing data on income and showed more evidence of response error, (Jordan et al., 1979). In her review of the literature comparing telephone and personal interview surveys, Eleanor Singer concluded that the telephone produces lower quality data as judged on a number of criteria including response rates, item nonresponse, response error, and respondent attitudes toward the interview experience (Singer, 1979).

To sum up, the first Biennial Conference produced a generally positive evaluation of telephone surveys, while noting many unanswered questions about them. The third such meeting had markedly more negative comments about phone surveys and raised many of the same unanswered questions. This confusing state of affairs was analyzed by Eleanor Singer at Reston, and she aptly called the telephone interview a "blackbox" because so little was understood about *how* it produced different results from personal interviews (Singer, 1979).

The problem which participants at previous meetings confronted in trying to make sense of research on telephone interviewing is that there is no generally accepted definition of what a telephone interview is. When people note the "promise" of telephone surveys, they are operating under one definition; when they emphasize the *problems* with telephone interviewing they are dealing with another idea of what that involves. For example, the positive evaluation given to telephone surveys at the Airlie House meeting was based largely on a judgment about telephone interviews in mixed mode designs. The Jordan, Marcus, and Reeder paper at Reston, however, compared "cold" telephone interviews with personal interviews and showed unfavorable results for phone contacts (Jordan et al., 1979). Similarly, when participants at Airlie House expressed the hope that telephone surveys might reduce interviewer variance, they had in mind closely monitored telephone interviewing (perhaps done with CATI). In most of the studies Singer reviewed

---

* The contents of this publication do not necessarily reflect the views or policies of the National Center for Health Statistics. Survey Research Center staff having major responsibility for this research includes: Charles Cannell, Robert Groves, M. Lou Magilavy, Nancy Mathiowetz, and Peter Miller.

at Reston, however, the phone interviewing was not conducted at a central location with careful monitoring, but rather was done from home phones by interviewers who were used to conducting personal interviews.

Thus, it appears that the hopes and fears about telephone surveys spring largely from different definitions of what a telephone interview is. As Singer noted, there is no theory by which to predict differences between telephone and personal interviews. The differences appear or they don't and we are often in a post hoc explanatory mode in seeking to identify the reasons for the findings. Without careful specification of the nature of a telephone interview, there is no fruitful way to study its components or make sense of its product.[1]

SRC-HIS telephone experiment

The study discussed here explored a number of dimensions of telephone interviews in order to get a glimpse of the workings of Singer's "blackbox." The study involved a national comparison of telephone and personal interview procedures for the Health Interview Survey. In the fourth quarter of 1979, the Survey Research Center conducted a national RDD telephone survey of some 4,300 families—yielding data on 8,200 people 17 years old and older—using a modified HIS questionnaire. At the same time, Bureau of the Census interviewers contacted a separate national sample of families—yielding 19,800 reports for individuals 17 years old and older—as part of the ongoing HIS. This paper compares data from the telephone and face-to-face interview samples, as do all mode comparisons. But, in addition, we will look in detail at two telephone survey components: interviewing techniques and computer-assisted telephone interviewing (CATI). These telephone survey components are two of the facets of the "blackbox," examination of which may help us to understand telephone-personal interview differences.

The overall telephone study design is summarized in Table 1. The cells contain the number of persons for whom health data were collected within each treatment. I will focus on the interviewing procedure and CATI-nonCATI comparisons here, but let us first look at all of the design facets. A comparison of the results for different respondent rules not discussed here are presented in a technical report on the study submitted for publication by NCHS.

Interviewing procedure experiment. The absence of visual cues in telephone interviews requires reconsideration of appropriate interviewing techniques. In personal interviews, interviewers often communicate understanding and approval of responses in nonverbal gestures. In addition, visual aids are commonly employed in face-to-face contacts to illustrate response tasks. Methods of communicating response tasks and of acknowledging responses (or the need for more information) should probably be systematically employed in telephone interviews. Further, there are indications that telephone respondents may be less motivated to participate in the interview than are those who are contacted personally. Telephone response rates are typically lower than those achieved by personal contacts. Responses to open questions appear to be truncated by telephone respondents. The speed of interviewer-respondent interaction tends to be faster, and telephone respondents often report they would prefer to be interviewed in person (see, e.g., Groves and Kahn, 1979). These findings suggest the need to motivate telephone respondents to participate conscientiously in the survey.

We have developed standardized procedures in face-to-face interviews employing some characteristics which seem suited to addressing the problems of telephone contacts. The techniques—instructions, feedback, and commitment—are intended to inform respondents about the response tasks, to communicate that they have performed them adequately, and to motivate them to take the interview seriously and expend conscientious and diligent effort in responding (Cannell, Miller, and Oksenberg, 1981). We employed these techniques in an experimental group in the telephone survey, comparing the results of this treatment to a control interviewing procedure which was designed to mirror the techniques used by Bureau of the Census interviewers in administering the interview in the personal interview survey.

CATI experiment. The increasing use of the telephone in survey research has been paralleled by research and development on computer-assisted telephone interviewing systems (CATI). With a CATI system, interviewers use video display terminals that present questions and permit the interviewer to enter responses. The computer performs checks on whether responses, as entered by the interviewer, are valid codes, and moves the interviewer from question to question according to a pro-

Table 1
Number of persons with interview data by experimental groups in the SRC telephone survey[1]

Mode	*Random respondent rule*			*Telephone answerer respondent rule*			*Total*
	Experimental interviewing behavior	*Control interviewing behavior*	*Total*	*Experimental interviewing behavior*	*Control interviewing behavior*	*Total*	
CATI	837	933	1,770	964	937	1,901	3,671
NonCATI	1,009	1,095	2,104	1,064	1,162	2,226	4,330
Total	1,846	2,028	3,874	2,028	2,099	4,127	8,001

[1]209 individuals in households where the random respondent could not be interviewed but where some other family member responded have been removed from this table.

grammed logic. This technology has promised to advance survey data collection by offering greater flexibility in questionnaire construction, greater control over interviewer behavior, faster production of data files for analysis, and lower costs (since coding, keypunching and data cleaning are reduced or eliminated altogether). In the SRC-NCHS Study, interviewers conducted a random half-sample of the telephone interviews using a CATI system and the other half using paper questionnaires. We observe that fewer CATI than nonCATI interviews were taken (see Table 1). This difference is due to technical difficulties with the CATI system early in the study.

Respondent selection experiment. In the ongoing HIS, all members of sampled households who are at home when the interviewer calls are interviewed in person. Parents always respond for children 16 years old and younger, and some family member at home responds for other absent adult family members. It was not clear whether this procedure was desirable or practicable on the telephone. In order to investigate the effects of alternative respondent selection rules, half of the households in the telephone sample were assigned to a "random respondent" group and the other half to a "knowledgeable phone answerer" rule.

In interviews with the first half-sample, adults in each family were listed and one was randomly selected to answer questions concerning his or her own health and that of other family members living in the household. In households assigned to the "knowledgeable phone answerer" rule, any person 19 years old or older who answered the phone and was capable of responding for him- or herself and other family members was used as family informant.

Neither respondent rule sought to interview each individual in a family separately. The knowledgeable phone answerer rule is closer to the HIS procedure than the random respondent rule, since an available adult serves as a proxy respondent for others in the family. In contrast to HIS, however, no attempt was made to speak with other members of the family even if they were at home at the time of the interview. Thus, there is a single self-respondent per family. In the random respondent rule, the self-respondents so selected comprise a probability sample of adults in telephone households. Thus, for analytic purposes we can contrast statistics based on data from all family members (many of whom did not report for themselves) with those based only on randomly selected adults (most of whom were self respondents). Table 1 shows that there were more "phone answerer" cases than randomly selected. This difference is due to refusals or inability to contact the randomly selected respondent.

The design was intended to be balanced over eight different combinations of the three experimental treatments. The actual number of person reports varies across the eight groups, reflecting a higher nonresponse rate in the random respondent rule and difficulties with the CATI system in the first month of the project (which required a reassignment of cases from the CATI group to the paper and pencil questionnaire group). Overall, about 8,200 person reports were obtained but about 200 of those required substitution of another family informant for a randomly selected informant who had refused.

Summary and qualifications. By constructing the experimental design described above, we sought to measure the effects of different components of telephone surveys so as to decompose telephone-personal differences for better understanding and policy guidance. At the same time, there are aspects of telephone and personal interviews which we could not measure for control. We also applied some controls which limit the inferences we can make from this study.

Because no experimental variations were made in the HIS personal interview survey, we cannot identify the effects of individual features of the face-to-face procedure in the same way as the telephone survey. Therefore, while we may be able to identify some of the factors underlying telephone-personal interview survey differences, questions will remain about what features of the personal interview procedure might have produced the differences.

Additionally, we had to make alterations in the HIS protocol which may have had an effect on the data. The alterations were made to adapt the questionnaire for telephone use and were not manipulated experimentally, so we cannot gauge their effects. The HIS is structured to accomodate a group format for the interview. In some sections of the questionnaire, questions are asked of or about each member of the family before a new section is begun. This structure is well suited to the situation when the interviewer is able to gather the family together and involve them in the interview. On the telephone, however, it is difficult to maintain this sort of flow in the questionnaire, since one is dealing with a single respondent. It is then necessary to restructure the HIS questionnaire to ask each section separately about each person, making sure of the focus of the questions. This involves making a decision about how best to stimulate the respondent's memory. Should one focus on the *event* (bed-day, doctor visit, hospitalization, etc.) as the HIS does, or organize the interview by the *person*, and ask about each individual's health events in turn? For the telephone survey, we chose to ask questions (except demographic) about each person individually and to follow those with questions from the condition, doctor visit, and hospitalization pages for each person.

The flow of the interview on the phone, in summary, consisted of asking the respondent all of the questions concerning his or her own health. Next, all of these questions about the next listed eligible person were asked, and this procedure was followed for all eligible persons. After this was completed, we asked about con-

ditions, doctor visits, and hospitalizations for the respondent first and then for the next listed person, and so on.

The rationale is that focusing longer on each person will lead to more careful consideration of that person's health history by the family respondent. We have separated the person-pages from the condition, doctor visit, and hospitalization sections for the same reason that they are separated in the personal HIS. We did not want to discourage reporting of health experiences by "teaching" the respondent that each time he or she reports something a series of follow-up questions will be asked.

Using the telephone to collect information generally obviates the use of visual aids such as calendars as used in NHIS. Attempts must be made to compensate for their absence. Tests were included in the pretests to see whether respondents had calendars available and were willing to use one of their own calendars for the interview. About half of the respondents did use a calendar in the pretests, and we continued to use this procedure. In addition, the reference dates were repeated frequently. These procedures and changes in the flow of the interview may have had unmeasureable effects on the data.

A final questionnaire alteration made for the telephone interview was our omission of certain questions from the standard HIS protocol. We did not ask the "chronic condition list" items (Q.32), and we also omitted "supplement" sections (e.g., the home-care page, immunization page, and residential mobility page). These alterations were made due to financial constraints which limited the length of the telephone interviews. We have analyzed the data in a manner designed to minimize the effects of the condition list omission on estimates of other HIS "core" items. Our comparisons use only the data from the NHIS interview prior to questions in which the lists were administered.

Finally, the telephone-personal comparison, again due to financial limitations, was confined to comparisons of *adult* reports (those 17 years and older). Information on children was not collected on the telephone due to the increased length of the interview that this procedure would have required.

Telephone interviewer training. We conclude this introduction to the SRC-NCHS mode comparisons by describing the administration of the SRC telephone survey, including interviewer hiring, training and monitoring, and the implementation of experimental treatments.

Thirty-five interviewers were hired for this study. Ten left before the interviewing was completed. Of those who remained, 7 were male and 18 were female. Nearly all had at least some college training. About half were between 20 and 25 years of age. All were new to interviewing, except two who had some minor short-time interviewing experience. We wanted interviewers without previous experience so that they could more easily be trained in new procedures.

Interviewer training consisted of three segments: (1) training in interviewing techniques and use of the questionnaires and procedures, (2) training in CATI (computer terminal) operations, (3) interviewing practice. The agenda followed the steps listed below:

The first two days of training were devoted primarily to instruction on techniques and questionnaire content. Included were demonstration and role-played interviews. Lectures were kept to a minimum, with heavy trainee participation in discussion and role playing. The goal was to inform the interviewers on what was to be done and how it was to be done, then to practice under supervision with continual feedback.

Sampling procedures were introduced on the third day. Additional role playing was included. On day four, interviewers were introduced to the computer terminal operations. The first three hours were demonstrations and practice only in terminal techniques. The remainder of the day was spent role-playing interviews, entering answers into the terminal. The fifth day was spent in practice interviewing, calling first acquaintances and then strangers, using CATI.

The next three days were spent in closely supervised practice interviews with strangers. At the conclusion of this period, most interviewers were judged competent to begin production interviewing. A few were given one or two more days of practice prior to regular interviewing.

In addition to the formal training sessions, several methods were used to update and review information with interviewers during the course of the study: (a) written memoranda on changes, corrections, or problem areas; (b) meetings with interviewers to review administrative procedures and discuss interviewing techniques. The latter included role-playing introductions and sharing successful refusal conversion techniques; (c) study manager or supervisors working with individual interviewers on specific problems, using discussions, monitoring, role-playing, additional study, practicing with a tape recorder, or any combination of these. Centralized telephone interviewing provides the opportunity for close monitoring of interviewing with immediate feedback to interviewers to correct errors. The need for this is clear.

There are three questions that must be considered in any system that evaluates the effectiveness of an interviewer. First, does the interviewer know what constitutes an adequate performance? Second, is the interviewer sufficiently skilled to behave in the correct manner? Third, is the interviewer motivated to perform correctly and adequately? Knowledge of correct behavior is, of course, a major component of the interviewer's training. The principles and techniques that are specified during interviewer training are by definition the "correct" behaviors, so that evaluations of a performance may differ in some respects from one staff to another, depending on the principles of interviewing that each one teaches or stresses. A monitoring system should focus on the major tasks that are taught during training, identify

each one, and evaluate the interviewers' performance of them. For this study we developed a monitoring system that involves the coding of interviewer behavior. Monitors listen to the interview and code the interviewer activity as it occurs. The major purpose of monitoring is to identify interview errors for supervisors' use in improving interviewing. It is also used in training to help to identify and correct errors. Table 2 summarizes findings from monitoring interviewers during the study.

Table 2
Mean proportion of various interviewer behavior by question type

Behavior	Mean proportion across questions[1] Closed N = 6905	Restricted open N = 2985	Open N = 330
A. Question delivery			
1. Correct reading	.87	.89	.60
2. Minor changes	.08	.08	.08
3. Major changes	.05	.03	.32
B. Evaluation of question reading			
1. Correct (pace, clarity of speech)	.94	.93	.94
2. Fast Pace	.03	.03	.00
3. Unclear speech	.03	.04	.06
C. Probing and defining activities			
1. Proportion of questions probed	.03	.12	.09
a. Correct probing	.75	.74	.77
b. Incorrect probing	.25	.26	.23
2. Proportion of questions with definitions	.02	.04	.03
a. Correct	.81	.87	.91
b. Incorrect or inappropriate	.19	.14	.09

[1]Of the 153 different questions monitored: 69 were classified as closed, 37 were classified as restricted open, and 57 were classified as open.
Note: The Ns report the number of observations of each question type.

The table clearly shows that, overall, interviewers delivered questions clearly and exactly as worded. Open questions presented the most problems for interviews—these questions were seldom asked (as reflected in the relatively small N) and can also be classified as questions which were burdensome to both the interviewer and respondent. Few questions (less than 9% of all observed questions) required the interviewer to define terms or probe for more information. The experimental interviewing techniques, which provide the respondent with information to adequately perform the interviewing task, reduce the interviewer's need to use probes.

Assignment of sample cases to interviewers. Associated with the coding of interviewer behavior is the measurement of interviewer variance. This approach seeks to describe the extent to which respondents' reports of health events tend to vary depending on which interviewer obtained the report.

To measure interviewer variance it is necessary to randomly assign respondents to interviewers. While this is usually not financially possible for personal interview surveys, it is quite feasible in a centralized telephone facility. We employed an interpenetrated design required to assess interviewer variance in this study.

Interviewers employed on this study conducted interviews using all of the experimental manipulation described in Table 1. That is, there were no "specialists" in the control interviewing procedure, or the CATI technique, questionnaire administration procedure, and respondent selection rule by the sample coversheet, so interviewers did not select the procedure they were to perform. Moreover, the allocation of work was accomplished in such a way that interviewers did not perform the techniques in any particular order. As mentioned above, we also monitored interviewers throughout the study to be certain that they continued to maintain operational distinctions between the treatments (e.g., that they did not use "experimental" interviewing techniques in "control" interview households or *vice versa*). For the CATI-paper questionnaire comparison, interviewers worked on the automated system during alternate weeks, one week on CATI, one week on paper questionnaires. The CATI and nonCATI interviews shared a common component—a "family folder"—in which interviewers kept track of family members and their conditions, doctor visits and hospitalizations.

In the first month of the study, difficulties with the CATI system led us to collect more interviews with paper questionnaires than had been planned. (See Table 1.)

In the case of 94 families in the random respondent selection rule, the selected respondent could not be interviewed. In those situations (covering some 213 people) we substituted another family member to respond for the family. Those individuals are not included in Table 1 nor in subsequent analyses presented in this report.

Telephone-personal interview differences

Response rates. The response rate obtained by the Census Bureau on the regular Health Interview Survey usually lies between 95% and 97%. Sample addresses where no contact was made are included in the denominator for HIS if they were judged to be occupied housing units. The total response rate for our telephone interview survey was about 80% (see Table 3). The latter rate is the ratio of the number of families having complete and partial interviews with at least one family member to the total eligible number of sample telephone numbers. It is, thus, a family-level response rate. All sample working household numbers that were never answered are included in the base of the response rate. A person-level response rate differs from the family-level rate only because of cases where data were not obtained on all eligible persons in the household. The person-level response rate for the telephone survey is 79.5%. The response rate on HIS is similar to that of other surveys conducted by the Bureau of the Census that permit proxy interviews within a household; the response rate for the telephone survey is higher than that obtained by most telephone surveys conducted by the Survey Research Center. The

higher than usual response rate, we believe, was attributable to a variety of characteristics of the project: to the legitimacy of the Public Health Service as a health survey sponsor, to the topic of health events, to lengthy training of the interviewers, to continual monitoring, and to high morale.

Table 3
Response rate information for the SRC telephone survey

I. Family level statistics		
Disposition category	*Number of families*	*Proportion of all eligible families*
Interviewed families	2184	.796
Partially completed families	55	.020
Family refusals	373	.136
Other noninterviews	132	.048
Nonsample		
Nonworking	1021	
Other	508	

II. Person level statistics		
Disposition category	*Number of persons*	*Proportion of total estimated eligible persons*
With interview data	8,210	.795
Without interview data		
Refusal (estimated)	1,579	.153
Other noninterview (estimated)	532	.052

[1]Households without complete enumerations were estimated to contain on the average 1.86 eligible persons.

Response differences. Hundreds of different variables were measured by HIS and also obtained as part of the telephone survey. The analysis of this paper concentrates on several measures that are standard dependent variables in HIS analysis. There are four categories of statistics presented in Table 4. First, since reports of a health condition are given for only a small proportion of the population, the percentages that have reported at least one event of various types are presented. This category is separated into measures that asked about the last two weeks and those that asked about the last 12 months. The third type of statistic are percentages of persons classified into the modal category of variables whose response distributions are more dispersed than those in the first two categories. The last class of statistics contains means for some of the variables that are counts of events. Each of the statistics is presented for the total telephone sample, the total HIS sample, and for that part of the HIS sample who had telephones.

We can see in Table 4 that, in general, the telephone interviews produced more reporting of health events than did the personal HIS interviews. The magnitude and meaning of the differences are matters of debate. It appears that the telephone and personal interview data are not very far apart, but the seemingly minor differences in percentages reflect large health differences in the population. Further, a common assumption in the reporting of health events is that "more is better," but there is some reason to doubt this reasoning. We will return to this later, and only point out now that the overall differences identified here do not automatically favor the personal interview mode, as did a variety of other studies reviewed by Singer at the last survey methods conference (Singer, 1979). The telephone data appear to be as good as and maybe even better than the personal interview findings.

Table 4
Percentages of persons in various response categories on health measures for the health interview survey and the SRC telephone survey

Statistic	*Total phone sample (7)*	*HIS telephone households (8)*	*Total HIS (9)*
Percentage with at least one: (two week recall)			
Bed days	8.7	7.7	7.8
Cut down days	9.8*	7.1	7.0
Work loss	7.6*	4.5	4.5
Dental visit	7.1*	5.3	5.2
Doctor visit	17.5*	13.6	13.5
Acute condition	16.3	NA	NA
Percentage with at least one: (one year recall)			
Limitation of activity	23.9*	18.7	18.9
Doctor visits	73.5	73.5	73.3
Health status (% excellent)	41.5*	44.0	43.3
Hospitalizations	13.0	13.3	12.5
Chronic conditions	32.3	NA	NA
Percentage in modal category			
12-month bed days (none)	46.0*	53.9	53.7
Time since last dental visit (2 wks.–6 mos.)	33.7*	31.3	30.3
Time since last doctor visit (2 wks.–6 mos.)	39.2*	43.5	43.5
Means (per 100 persons per quarter)			
Bed days	189.8	208.0	216.5
Work loss days	192.4	111.5	111.2
Dental visits	59.2	40.9	41.0
Doctor visits	166.4	124.8	126.8
Acute conditions	119.0	68.3	75.4

*Statistically significant difference between HIS estimate and SRC telephone survey estimate.

Exploring the telephone survey "blackbox"

Now that we have seen the overall differences between the telephone and personal interview surveys, it is time to look at some components of the telephone mode to try to understand the differences. One of the difficulties with telephone-personal interview survey comparisons is that the style of interviewing may differ across the modes and confound interpretations of differences between them. We designed this study to provide an independent reading on interviewing effects by experimentally manipulating two interviewing treatments in the telephone survey.

Description of experimental treatments. We randomly assigned the telephone sample to one of two interviewing treatments. The first, which we will call "control," featured techniques based on an analysis of census interviewers' techniques in the HIS. We sought to constrain the interviewer behavior in this treatment to be equivalent to census procedures as we understood them, based on our observations at census interviewer training sessions and an analysis of tapes of mock HIS interviews taken by census interviewers. Our intent in designing this treatment was to standardize interviewer behavior as much as possible across the telephone and personal modes so that differences between SRC and census interviewers' questioning style would not be confounded with the effects of the mode of communication in the telephone-personal comparison.

The "control" procedure restricted interviewer-to-respondent communication to asking questions printed in the questionnaire and to using probes and introductory statements at interviewer discretion. For comparison to this procedure, the other half of the sample was interviewed using some experimental techniques. These techniques were based on earlier research in face-to-face interviews. They included three experimental procedures: commitment, instructions, and feedback.

Commitment. It is important that respondents understand that the interview is an important undertaking and that some effort will be needed to perform response tasks adequately. If respondents are motivated to perform well, they may be less likely to treat the interview as a game or to rush through it. More careful thought is likely to produce better reporting. One technique we have used in personal interviews to help motivate respondents is "commitment."

In personal interviews we sought commitment by having the respondents sign a statement in which they promised to work hard to give accurate and complete information. In telephone interviews, as in this study, the commitment statement is read to respondents and they are asked to indicate verbal agreement. If the respondent fails to agree with the statement, the interviews are terminated. In practice, virtually no respondents who are asked to commit themselves refuse to do so. The commitment statement used in this study was:

> This research is authorized by the Public Health Service Act. It's important for the Public Health Service to get exact details on every question, even on those which may seem unimportant to you. This may take extra effort. Are you willing to think carefully about each question in order to give accurate information?

If the respondent agreed, the following statement was read:

> For our part, we will keep all information you give confidential. Of course, the interview is voluntary. Should we come to any question which you do not want to answer, just let me know and we'll move on to the next one.

Instructions. Besides attempting to motivate respondents through commitment, we tried to orient respondents to the interview by the use of instructions on the purpose and goal of questions, and on how to go about answering them. Respondents typically pick up cues on what is expected of them only incidentally through interaction with the interviewer. Attempting to teach respondents what is expected of them through such indirect action is frequently ineffective. We attempted to communicate desirable behavior by including specific instructions at various points in the questionnaire for the interviewer to read.

Researchers concerned with task performance have identified two main functions of instructions, first, to clarify the goal toward which the performance is directed and, second, to clarify specific tasks required to achieve the goal. In the interview this first type clarifies the goal of the interview by informing the respondent what is expected of him or her: to give accurate and complete answers to all questions.

In this study, these general goals were articulated by including performance instructions preceding the questions as well as in the commitment statement above. The second type of instruction details specifically how the respondent should go about producing accurate answers on individual questions, and what level of accuracy is required. Examples of specific question instructions include: "This is sometimes hard to remember, so please take your time." "For this question, we'd like to get as exact a number as possible."

Feedback. The instructions procedure is designed to clarify general and specific goals of the interview and also to motivate better performance. Instructions are not complete, however, without communication to respondents on how well they have carried out the response task. Thus, the third experimental technique we employed was *feedback*.

We came to the idea of programming feedback in interviews after an analysis of personal interview interactions. This research demonstrated that much of the interaction that takes place in face-to-face and telephone surveys is not limited to the strict asking and answering of questions (Cannell, Lawson, and Hauser, 1975). The findings led us to focus on the two-way process, or chaining of behaviors between interviewer and respondent, rather than on the separate activity of each.

In this view of the communication, the way in which interviewers react to respondents' earlier answers is an important determinant of their behavior in later questions. Interviewers' reactions constitute a feedback to respondents that can influence their behavior in general and the accuracy and completeness of the reported information in particular. Like commitment and instructions, feedback reactions can be both informative and motivational in quality. They tell respondents when they have fulfilled task requirements, and they serve as reinforcers capable of shaping subsequent behavior.

Following our practice, feedback statements were designed into the questionnaire in the experimental interviewing treatment. In general, feedback statements were made contingent on "good" performance, and both negative and positive feedback statements were used. For example, interviewers estimated the length of time that the respondents took to think over answers to some of the questions which required respondents to search their memories. Respondents who took less than about three seconds before replying negatively to a question about reducing activities in the recent past because of illness or injury were read the following: "You answered that quickly. Are there any days you might have overlooked?" Positive feedbacks, on the other hand, were used to indicate to the respondent that the answer given fulfilled the goals of the question. For example: "I see." "This is the kind of exact answer we need." "That's useful information." "Thank you. This is helpful."

In summary, commitment, instructions, and feedback are three procedures which we have used in several studies in an effort to improve reporting. The techniques become part of a "script" for the interview, which interviewers are trained to use in a standardized manner. In this way, we seek to reduce between-interviewer variability in the use of techniques, as well to communicate more productively with respondents.

These techniques, singly and in combination, have been shown to improve reporting in face-to-face and telephone interview surveys on health and mass media use (see Cannel, Oksenberg, and Converse, 1977a; Miller and Cannell, 1977; and Cannell, Miller, and Oksenberg, 1981). We anticipated that employing the procedures in this study would improve reporting of health variables in the HIS.

Effects of experimental interviewing techniques. Table 2 displays the overall effects of the experimental techniques. As is characteristic of the health variables, only a small proportion of the population reported affirmatively to questions asking for incidents of illness and health-care use during the previous two weeks. Larger numbers of respondents reported health events and experiences for the previous year. Table 5 shows the percentage of the sample for whom one or more illnesses or health behaviors were reported or, for variables that are not counts of health events, the percentage in the modal category for particular variables.

Nearly all of the health events—bed-days, work-loss days, doctor visits for the past two weeks and for 12 months, etc.—were reported more frequently by the experimental group. The majority of the differences are significant at the 5% level. Nonsignificant differences were found for reporting of medical and dental visits within the previous two weeks and for ratings of subjective health status. In addition to more health events and behaviors, the experimental group reported a higher level of limitation of activities (largely non-major activity limitations).

Acute and chronic conditions were also reported more frequently in the experimental group. These findings suggest that the experimental techniques sensitized respondents to health problems, making the difficulties more salient and enhancing the respondents' tendency to perceive themselves in poorer health; or it may be that the techniques make it easier to *admit* to poor health, an undesirable admission for some respondents.

Table 5
Reporting of health events in telephone survey by experimental interview treatment

Statistic	*Control form*	*Experimental form*
Percentage with at least one: (two week recall)		
Bed days	7.3	10.0
Work loss days	6.3	8.8
Cut down days	8.4	11.5
Dental visit	6.8	7.4
Doctor visit	17.4	17.5
Acute conditions	14.9	17.7
Percentage with at least one: (one year recall)		
Chronic conditions	29.2	35.8
Limitation of activity	20.4	27.6
Doctor visits	72.6	74.5
Health status (% excellent)	42.1	41.3
Hospitalizations	13.4	12.5
Statistic percentage in modal category:		
12-month bed days (none)	48.1	43.6
Time since last dental visit (2 wks–6 months)	37.3	37.1
Time since last doctor visit (2 wks–6 months)	39.2	39.2

During the coding process each reported condition was rated on two scales one for *seriousness* and the other for the potential *embarrassment* it might cause to report it.

Both more serious and more embarrassing episodes were reported under the experimental conditions. Of the total conditions reported, approximately one-third in each experimental group were classified as serious and 17% to 18% were rated as embarrassing. The increased reporting in the experimental group was not accounted for simply by increased reporting of less serious or embarrassing conditions, but appears to reflect an overall increase in condition reporting.

The data in Table 5 illustrate one reason why the telephone survey produced higher estimates of health events than did the HIS personal interviews—the experimental interviewing techniques we employed for half of the telephone sample households elicited substantially higher reporting for the dependent variables than did the control techniques which were modelled after census interviewers' style. How do we interpret these results? We must make assumptions about the

direction of the reporting errors for the health variables in order to make an interpretation. The predominant assumption among researchers in the field, we pointed out earlier, is that health events are underreported. If one believes this, the object of data collection techniques should be to increase reporting on the health measures. We have seen that the experimental interviewing procedures do tend to produce higher reports of illness and health-care use than do comparison procedures. Therefore, we might claim that the experimental techniques produce *better* reporting than the control procedure which was modelled on the current Census HIS techniques.

We must acknowledge, however, that there is not unequivocal acceptance of the underreporting hypothesis; there are some cogent criticisms of the evidence on which it is based. In an analysis of hospitalization record check studies, Marquis pointed out that the finding of underreporting of hospitalization episodes is common to *retrospective* record check studies—those which select respondents from hospitalization records and interview them to see if they report the event. He notes that the only error which is possible to discover in such studies is underreporting, since people who were known not to be in the hospital are never contacted. He suggests, therefore that the "underreporting" uncovered in such record check studies might well be random error. If this argument is correct, techniques designed on the assumption of an underreporting bias in the measures may actually produce overreporting on the health variables.

Another argument which supports the possibility of overreporting involves the notion of "forward telescoping." Since the health events mentioned in the analyses above often require respondents to report things which they experienced during particular time intervals before the interview, it may be that those who received the experimental interviewing treatment tended to "move" events they experienced forward in time so as to place them within the reference period we set up in the questionnaire. So, for example, respondents in the experimental group might have reported more two-week cutdown days because they were motivated to report *some* health experiences, and they reported things which had actually happened before the two-week period as having occurred during the reference period.

We cannot entirely rule out the possibility that the experimental interviewing treatments produced overreporting. A previous study using the procedures, however, found that they tended to reduce both underreporting and overreporting. I reported at the Reston conference that the experimental procedures, administered to a sample of women in a study of mass media use, elicited *more* reports of television watching and x-rated movie attendance and *fewer* reports of book reading. If one accepts the hypothesis that the former two behaviors are likely to be underreported and that the latter one is likely to be overstated, then there is some evidence that the interviewing procedures can reduce reporting biases in both directions.

We can also present some data from the present study which indirectly bear on the issue. At the beginning of the telephone interview, we suggested to respondents that they might find it easier to report health events if they had a calendar handy for reference. Approximately 75% of the 4,400 family respondents indicated that they had a calendar ready for use. Since these individuals may have been less likely to "telescope" health events into the reference periods set up in the interview, we wanted to see how reported calendar usage related to health reporting, to demographic characteristics of the reporter, and to experimental interviewing treatments. If we found that those saying they used calendars reported fewer health events, we would suspect that the "more is better" hypothesis is not accurate. Further, if we found that there were substantial differences between experimental interviewing treatments in reported calendar use, we would be obliged to see whether the experimental effects were explained or specified by this variable.

To summarize the findings of this analysis, we discovered that reported calendar use was generally unrelated to the family income of the reporter, or to education, sex, or age. (Women were slightly more likely to report using a calendar.) There was no difference between experimental interviewing treatments in reported calendar use. Finally, for several selected health events, there were small or no differences in reporting between those who said they used a calendar and those who did not. The differences, however, tended to favor the "more is better" hypothesis, since those who reported using a calendar reported slightly more health events. Again, these analyses only suggest that the experimental interviewing treatment produced better reporting. A study with external validating records would be helpful to sort out the interviewing treatment differences.

CATI and nonCATI questionnaires. Another part of the research design for this telephone survey randomly assigned to half-samples the mode of the questionnaire: Half were assigned to typical paper and pencil questionnaires, half were assigned to a questionnaire programmed into a computer-assisted telephone interviewing (CATI) system. The random assignment was made on a sample number basis; thus, all of the families and persons in the same household were given the same treatment. At different times each interviewer used both modes of asking the questions; interviewers alternated conducting CATI or nonCATI interviewing, changing methods each week.

This part of the research design was only partially fulfilled because of CATI hardware problems that developed in the first month of interviewing. During that time, instead of using CATI on a random half-sample, only 36% of the interviews were taken using the computer. Because the problems occurred during a relatively short period of time, the balance of CATI and nonCATI

interviews for some interviewers was more affected than for others. The first weeks of work on the study processed sample cases in the first of three replicate samples. Because of this we can separate the sample cases affected by hardware problems from those in the other two replicate groups without risking compounding of differences between mode with other differences between the CATI and nonCATI groups.

The CATI questionnaire was designed to replicate as closely as possible the form of the paper and pencil questionnaire. The inherent differences in the two procedures, however, required some adjustment of the paper and pencil version to maximize the comparability with the CATI version. The complexity of the questionnaire coupled with the limitations of the SRC-CATI system at the time of our implementation also resulted in some adaptations of the questionnaire unique to the CATI instrument.

To understand fully the nature of the differences between the CATI and nonCATI questionnaires, we will review the flow of the interview and the associated tasks of the interviewer. The questionnaire collected information on all members of a family; data are collected both through self-reports and proxy reports (a family member reporting for someone else in the same family). A set of core questions, known as the "person section," is asked for each family member. Depending on the information obtained, further "supplements" are completed. These supplements are used to collect more detailed information on conditions, doctor visits, and hospitalization.

These interview complexities required the following capabilities in the CATI system:

(1) collection of core information for each member of a family (person sections)
(2) collection of supplemental information for only those family members with health events requiring further questioning
(3) ability to collect a varying number of these supplement sections per person
(4) assisting the interviewer in identifying the current referent person and the current questionnaire segment

These requirements in the nonCATI format were accommodated through the use of multiple-booklet questionnaires. A separate booklet was used for the person section and for each of the three types of supplements. Booklets were added to the case as needed to complete the questioning. Identifying information (case number, referent person, and interviewer number) was recorded on the cover of each booklet used during an interview. Thus, after completing the person section for the respondent, the interviewer could select the next appropriate booklet from stacks in the interviewing station, record necessary identifying information, and proceed with the interview.

The CATI instrument design closely paralleled the flow of the nonCATI questionnaire in its movement between the person and supplement sections. At the end of the first person section the next screen presented the available options for continuing the interview. The interviewer entered the desired section to complete next and the information needed to identify the person being referred to in the questions. At the end of each section (the equivalent of a booklet in the nonCATI version), the interviewer was returned to this same screen.

As an aid to the interviewer, information concerning the referent person and the relevant section of the interview was displayed at the top of each CATI screen. For example, if the interviewer was collecting information on the third doctor visit for the second person in the family the display would show:

PERSON # = 2 DOC VISIT = 3

The SRC-CATI system described here provided the researcher with a number of controls designed to reduce interviewer error. Each CATI screen (which usually was equivalent to one question) had both a text and numeric field where responses could be recorded. A text field was always available to the interviewer to record probes and comments. The researcher determined whether to include a numeric field, and if so, whether to initialize the cursor on the screen to the numeric or text field. Interviewers could move between the two fields with one keystroke. When a numeric response was required, a list of valid responses was also programmed by the researcher. If an interviewer failed to enter a valid response, an invalid response message would appear at the bottom of the screen with a blinking cursor indicating to the interviewer that a new response had to be entered. In addition to checking for valid responses, the system enforced proper branching to the next question contingent on previous answers.

Performance characteristics for CATI and nonCATI interviews. The overall family level response rate for sample cases assigned to CATI treatment was 78.7% versus 81.5% on nonCATI interviews. The lower overall response rate for the CATI sample cases was due to the technical problems with the computer system in the first month of the study. During that period, some refusal cases begun on CATI were converted using the paper-and-pencil technique when the system was having trouble. It is safe to conclude that CATI in fact had no large effect on response rates.

The number of interviewer hours required to obtain one interview differed between CATI and nonCATI cases. Each case required 52 minutes of interviewer time on the average using CATI, while using paper-and-pencil questionnaires required 46 minutes for each case. The additional time required for the CATI cases may reflect a difficulty interviewers had with using paper coversheets to record calls and household composition but a computer terminal for display of questions and recording of responses. We suspect that an efficient

algorithm for choice of next number to dial and machine documenting of call records would reduce the amount of interviewer time needed per case.

Interviewer reactions. Interviewer reactions to the different types of questionnaires followed predictable lines. The advantages of the on-line system were seen as (1) the ease and reliability of routing through a complex questionnaire and (2) the ease of typing rather than handwriting and paper shuffling. On the other hand, interviewers noted that the response time (the time between a key stroke and the next question display) was often too great, and that, in this particular CATI application, it was difficult to skip around in the questionnaire to locate particular questions for correcting or editing responses. In evaluating the paper-and-pencil questionnaire, the interviewers positively noted their greater sense of control over the interview and quicker questionnaire administration. Overall, however, 75% of the interviewers either preferred the CATI version or said that they like both types equally.

CATI and nonCATI response differences. We did not expect to find notable response differences between the two types of questionnaires. However, since the CATI version eliminated obvious wild codes automatically, by making interviewers re-enter the data if they made a mistake and enforced correct skip patterns, we might have discovered some effect of this in the distributions produced by the two types of questionnaires. Most of the differences were not statistically significant, and they followed no particular pattern. It may be that the advantages of CATI would be more noticeable when the interviewers are not as able, experienced and monitored as they were on this study. Since there were very few skip-pattern and wild-code errors in the paper-and-pencil interviews, the advantage of CATI in eliminating these errors is not obvious in this study.

Interviewer variance. One difference between the CATI and nonCATI interviews which is worthy of note is that the estimates of interviewer variance were lower for interviews taken with the computerized system. Values for our interviewer variance estimates were relatively low compared to personal interview surveys for which such numbers have been calculated, but, despite this, had we collected all of the data by paper and pencil the average design effect for interviewer assignment would have increased by 134% (from Deff = 1.16 to Deff = 2.34). This is due to the effect of (hypothetically) increased interviewer workload, which would inflate the estimates of correlated response deviations associated with interviewers. Thus, while we were unable to discover any notable, consistent response distribution differences between CATI and nonCATI interviews, the machine-directed contacts did show lower interviewer variance.

Conclusions

This study presents one look at telephone-personal interview survey differences and seeks to identify some of the reasons for the findings (in this case, interviewer techniques), as Singer suggested at our last health survey methods conference (Singer, 1979). There remain a number of unanswered questions about communication in telephone interviews, but these results sensitize us to the fact that the mode of the interview may not be as important for determining survey results as are such factors as training, monitoring, and interviewer corps morale. We have seen that telephone surveys do not necessarily produce data of lower quality than those collected in personal interview research and, given some assumptions about reporting error, they may even produce better information.

Another lesson is that, in order for us to make sense of mode differences, we must carefully define the nature of the modes in question. For example, we cannot equate telephone reinterviews done by interviewers without special telephone training and "cold" telephone contacts made by well-trained, closely monitored interviewers working with a computerized questionnaire.

Finally, this study justifies some of the optimism about telephone surveys which feature new technology. While the CATI application used in this study left something to be desired, interviewer variance was reduced over paper-and-pencil interviews and both versions of the questionnaire produced lower interviewer variance estimates than have been seen in personal interview surveys. The fact that we can make such measurements in phone surveys is reason for a positive view about using that medium. We may be able to learn some things about the survey craft in centralized phone facilities that are not practicable (or perhaps even possible) to observe in personal interview surveys. That possibility alone is reason for optimism.

Footnote

[1] The same, of course, is true of personal interviews. Although treated as the standard in telephone-personal comparisons, there is considerable variety in studies employing personal contacts, and this is an additional source of confusion when one reviews the literature.

Discussion: Telephone survey methodology

Norman M. Bradburn, National Opinion Research Center, University of Chicago

We have heard a number of excellent papers investigating possible biases arising from conducting interviews over the telephone instead of face-to-face. My principal reaction to these papers, as well as to others in the developing literature on this subject, is to be surprised that there aren't really any dramatic differences between the two modes of interviewing. The differences that are found are rarely very important or large and in many instances turn out not to be consistent.

Why, then, do we continue to be surprised that there aren't many differences? I suspect that one important reason is that we can't believe our good fortune when see that a method we are reluctantly adopting primarily for economic reasons does not turn out to be inferior. The tenor of a lot of what has been said here today, which has certainly been said more forcefully earlier, is: "Well, good or bad we are going to have to do it because of the economics of it." It's as if all the things that we are forced to do for economic reasons are less desirable than things we think we are doing for noneconomic reasons. One indication of this attitude is seen in the amount of effort put into research on interview mode differences as opposed to research on questionnaire wording, learning effects, reenforcement, and other things that go on in the interviewing process, all of which can produce dramatic differences in responses. I also suspect that one of the reasons that comparisons between open-ended and closed-ended questions are a favorite form of research on questionnaire construction is that this is the one mode of question wording where there are strong economic incentives to move in one direction rather than another.

In any case it seems almost too good to be true that telephone interviewing produces results that are practically no different from face-to-face interviewing, and I continue to be surprised at our good fortune.

When I first read the papers I thought, "Aha! At last we have found a consistent difference." That one consistent finding had to do with the reports of dental visits and was reported across three papers presented here that have data on dental visits. For a brief moment I thought that it was also going to be true for reports of physician visits, but, alas (or hooray, depending on your feelings about telephone interviewing), the findings did not turn out to be consistent. For the rest, the results indicated to me that there are no important or consistent differences between modes.

Of course, things are not all the same between modes. Certain types of questions are more difficult to ask on the telephone than face-to-face such as open-ended questions and questions that require the use of some sort of visual material. There may be some questions that are easier or better to ask on the telephone, such as those that profit from greater anonymity or are unduly influenced by a social desirability bias. As we learn more about where there are real, consistent modal differences, I am confident that we will develop strategies for question wording that will be uniquely adapted to particular modes of interviewing and will eliminate any differences.

One still unresolved question is whether you can conduct interviews on the telephone that are as long as those done in person. One of the things I remember most vividly from the 1975 Airlie House Conference was the way in which views on permissible lengths of telephone interviews changed over the course of the discussion. That conference was an informal discussion in which people recounted their experiences with what was then a much less well-tried technique. They began by saying, "Well, we did fifteen-minute interviews and they didn't seem to be any different from face-to-face interviews, but that's about as long an interview as you can do." Then somebody said, "No, we did twenty-minute interviews and they went okay." Somebody else added, "We did half-hour interviews, but I'm sure you couldn't do anything longer." And then someone said, "No, we did forty-five minutes, but we doubt you can go more than that." And so on and so on. By the time we were finished everybody recognized that their interviews had worked well under many different conditions; yet, still, they were sure there were limits out there that they hadn't quite reached and that they were just beyond the longest interviews they had done. We have heard today from Peter Miller that, although they are not in continuous sessions, Michigan is doing 2½-hour interviews on the telephone. But even in face-to-face interviews it is true that interviews of such length will most likely be done in several sessions.

When we do find a difference between interviewing modes—let us say for the sake of argument that the apparent inconsistency in reports on dental visits is a real difference—how can we decide which method produces the more valid information? If it is information about behavior such as dental or physician visits, we can, of course, make record checks. But if the difference turns out to be about attitudes, then we may be in trouble. We don't have much in the way of theoretical notions about why there should be modal differences, nor do we have a good theory of the interview which might predict which mode ought to produce a particular type of bias. However, two factors were brought out in the papers as con-

trasting explanations for the differences that were observed. The first has to do with coverage, the second with nonresponse. For the nation as a whole, coverage is quite good—on the order of magnitude of about 95% of households. However, coverage can be a real problem if one is concerned about those segments of the population, such as the poor and Hispanics, where coverage is considerably less than that. I also suspect that on the coverage issue we may have reached our highest point and may see a declining proportion of households with telephones in the future. I say this because I think the effect of the court decision to break up AT&T will be to increase dramatically the price of home telephones and that we shall see substantially fewer households with them. It is less clear what will happen to long-distance rates. It may be that increases in telephone costs and decreases in coverage may make telephone interviewing less economically desirable than it now seems. I doubt, however, if the change will be large enough to limit substantially the present cost advantage. The Banks and Anderson paper is particularly interesting in contributing to our understanding of the coverage bias resulting from a telephone survey.

As to response rates, one of the articles of faith in the field is that response rates of telephone interviewing are considerably lower than those of face-to-face interviewing. One of the contributions of the papers today is to demonstrate that response rates on the telephone can be quite high if one puts one's mind to it. The papers, inadvertently perhaps, have also reminded us that response rates in face-to-face interviewing can also be quite high if one puts one's mind to it and backs it up with some money. The point here, I think, is that high completion rates may be somewhat more difficult to accomplish on the telephone, but can be achieved if attention is given to it.

I was very much impressed that there are some data to support something I have always believed. It was mentioned in passing this morning that the expectations of the interviewer about how hard it is going to be to complete an interview are a very significant factor in the completion rate. For a long time I have observed that response rates in surveys at NORC go up and down and up and down regardless of what seem to be national trends. The completion rates seem to have more to do with what the field staff thinks is possible than with anything else. When they are convinced that they can get a 95% to 96% completion rate (which NORC has been doing on the logitudinal study of youth—supposedly the hardest group to track and find), then they get completion rates at that level. This is not true in all studies, obviously. We see that health studies generally get higher completion rates than do more general studies, which may again be a self-fulfilling prophecy, because interviewers believe that health studies are easier to do and respondents are more interested in them than in other kinds of surveys.

Another point that was discussed in different ways in the papers, primarily in the Banks and Anderson paper, relates to the kind of analysis that can be done in a number of studies. One relatively straightforward way of handling the coverage on respsonse problems is to do some sort of demographic adjustment for the lack of coverage of differential response rates. However, these demographic adjustments may not really do the trick if in fact there is an interaction between the characteristics of households that are nonresponsive or not covered by the sampling frame and the health variables we are interested in. If there is an important interaction, the simple demographic adjustment may, in fact, make things worse rather than better. It may increase rather than decrease bias.

Having said that it is possible to get a high response rate even on the telephone, I would ask what is perhaps a heretical question: Are high response rates all to the good? Is it always proper to say that a higher response rate is better than a lower response rate? We take that as axiomatic. However, I do at least want to raise the possibility that it may not be always a good thing. When we make extraordinary attempts to get high completion rates we may be introducing error—more error than we are taking out—by reducing the variance with the high response rate.

Let me differentiate between two problems in getting high response rates: one is getting the people who are hard to locate; the other is converting people who initially refuse to participate. I suspect that the hard-to-finds are similar to their demographic counterparts in the way they respond to questions and that when we do an adjustment for nonresponse we will do okay if we are adjusting just for the people we couldn't find. There are, however, a number of people who end up responding because they have given up trying to fend off the interviewer. Some gung-ho interviewers get their feet in the door and pressure respondents so much that the best strategy for getting rid of them is to go through the interview quickly—in other words, do it but don't work hard. When one is trying to measure things that may require effort, to search memory, locate records, (recalling visits to doctors, health care costs, etc., for example), we may, in fact, introduce more error by pressuring these reluctant respondents into answering our questions than we get rid of by reducing nonresponse. Charlie Cannell published a paper some years ago which showed that when you look at the people you convert from nonrespondents into reluctant respondents, they, in fact, do report less accurately. He found that the total estimate of what is going on was made worse by including reluctant respondents. I think we need to give some serious consideration to this aspect of the response rate problem.

Telephone interviewing is thought to have considerable cost advantages over face-to-face interviewing. It is hard to compare costs across different organizations and even across many studies, and the question of how much savings come from telephone interviewing needs consid-

erably more attention than it has received so far. I was very glad to see the Kulka paper give attention to some aspects of cost. For myself, I think that costs are underestimated.

One of the things I am not sure about is how to account for developmental and maintenance costs. I suspect that most of the cost reports for telephone interviewing do not adequately capitalize or figure into overhead, or however you want to put it, the cost of developing and maintaining a telephone facility that is idle much of the time. Like computers, telephone facilities can be very costly in the aggregate unless you keep them busy all the time. One thing about interviewers is that, on the whole, they are not a large, continuing, running cost. I suspect that as we get more technologically sophisticated—that is, really get to using the computers that are necessary to give us a substantial reduction in costs, get the software developed, maintain a programming staff, factor in lost time because of down time and various other things related to the fixed costs of maintaining a telephone interviewing facility—we will find that the cost differentials between telepehone and personal interviewing will be much narrower than we think now when we look only at the marginal costs of doing a survey. I don't know what the proper accounting rules for these factors should be, but I doubt very much if they are now properly brought into the cost figures reported for telephone versus personal interviewing comparisons.

A couple of minor points: I don't really see why response rates at the local level should be any different from those at the national level, at least for a telephone versus face-to-face comparison. Again, I think it may be more a problem of the areas that one is working in, and you must be very careful when comparing across studies or organizations that are conducting studies in different areas.

A second problem with telephone interviewers was just mentioned in passing. I don't think enough attention has been given to it. This relates to the topic of matching interviewers with respondents in studying interviewer variance. One of the difficulties with telephone interviewing is knowing the characteristics of the respondent so that you can match the respondent with the interviewer. Telephone interviewing allows for much easier random allocation of interviewers to respondents, but makes it harder to do matching, at least through the initial contact and screening phases. In attitude areas, where minority groups may be an important part of the population, the problem of bilingualism or what language you conduct the interview in was mentioned briefly by someone. All these sorts of things are easier to control in a face-to-face field situation where you can make contact in the right way because you have relatively more information about the households before you make the initial contacts. In telephone interviewing this is just another problem that has to be coped with. My impression is that people are developing techniques that handle the problem fairly well, although I think it needs more attention.

Open Discussion: Session 2

The discussion began with a request to Marcus for clarification of a point: If the first and last waves of percentage distributions are approximately the same, and some post-stratification adjustments were done, would the distributions have been biased, that is, lost differentially in some of the subgroups? Marcus replied that they looked at two variables for which they imputed missing data: breast self-examination and smoking. They found that their estimates were within 1% or 2% of the actual findings. There wasn't much gain at all based on trying to make adjustments based on nonresponse.

A question was directed to Banks regarding Table 9 of her paper. Since she was not able to find any single adjustment category to raise the nonresponse category to 1.0, a multivariate analysis was suggested to take into account many of these factors simultaneously, even though no one by itself would bring the rate to 1.0. Banks replied that because of the conference deadlines they were not able to do such an analysis, but that would be the next step.

Another comment about the Banks paper indicated that, especially in Table 2, selected groups such as Hispanics in the Southwest and various age groups were not well represented. It was suggested that telephone surveys may fail to include some of these groups. Banks replied that Table 2 was not a good example to use to judge this problem; other tables showed a closer correspondence.

Fuchsberg commented that researchers have had 40 years of experience administering personal interviews and only about 10 years with the new technology of telephone interviewing. He noted that because telephone interviewing is in its infancy, researchers need to do appropriate experiments to develop methods to increase response rates; they should be able to get better results in the long run. Fuchsberg said that it may not be fair to compare the old and new methodologies at different stages of their developments.

Another participant noted that response rates between first interviews and reinterviews are not that large and that a tremendous amount of resources could be saved if the methodology is perfected.

Marcus commented that he liked the idea of a combination of methods when appropriate. His own preference is that the first contact be face to face and that the interviewers be given special training to use the telephone for subsequent interviews. As methodologies improve, he said, researchers will be able to combine both methods, perhaps using face-to-face for follow-ups on panel surveys and for nontelephone subscribers.

Fuchsberg commented that the costs for mixed mode operations may be higher because there are two sets of administrative structures, two systems for training and supervision, etc. Lois Montiero replied that in her mixed mode interviewing in 1970, it was not more expensive because they used the same interviewers for both methods; the interviewers were part-time and were used only as needed.

Horvitz also expressed concern about cost and components of error, but saw advantages in the many possibilities such as: combining telephone and household interviewing; using address lists geocoded to census blocks; selecting an area sample and interviewing by phone but preparing face-to-face interviews for respondents without telephones; and following up by in-person contact when respondents refuse to be interviewed by telephone. A further advantage of mixed mode is that certain groups like the elderly may be able to report better in person than over the telephone. Mixed mode means that you can change methods in order to more adequately address specific components of the population. However, these decisions about which mode to use should be studied for costs and types of error.

Banks added that another advantage of telephone interviewing is that you can quickly switch interviewers to deal with language or other telephone problems at the time you encouter them.

Czaja raised the question whether a good telephone interviewer has the same characteristics as a good personal interviewer. Marcus replied that IISR (UCLA) uses the same interviewers for both methods, but noted that there is much disagreement on the point. Some think that personal interviewing experience is important for a telephone interviewer; others think that no experience is an advantage.

Kulka was asked what he meant in his paper by agreement of the report with the record and what procedures he used. He replied that, on chronic conditions, agreement was defined very liberally if it was in a given reference period. Partial agreement might reflect only agreement on general physical condition. For hospitalizations, the criterion for agreement was whether there was any record within a 12-month period.

Discussing telephone response rates, Presser commented that in the past six months or so, response rates have improved. They used to be about 70% and are now getting higher—75% or more. This may be due not just to increased experience but also to a change in supervision. Marcus noted similar experiences. Wright commented that when he used the telephone for follow-ups he got a very good response rate—98%—after doing face-to-face household interviews. He provided special training for interviewers on the telephone. Several comments were made to the effect that researchers can soon

expect to do as well with telephone as with personal household interviewing.

Bryan commented on the emphasis placed on household surveys and then asked Marcus about his experience with professionals, such as doctors. Marcus replied that he had no direct experience with professionals but that response rates are generally higher with better educated respondents. He described a recent telephone survey of physicians that got an overall response rate of 77% with two calls. The office manager was contacted first to get as much information as he or she could give about billing, etc., then they talked to the physician only on subjects the physician could answer. Poorer response rates were obtained from GPs, pediatricians, and internists; higher rates from more specialized groups. It was noted that when surveying professionals by telephone, it is useful if the respondent is sent the questionnaire before the call.

Fuchsberg described a series of smoking surveys using random digit dialing (RDD) and noted an average response rate of 78% over several studies; the highest was 89%. He also commented that most refusals occur in the first 15 seconds. If you can use the first words of contact to get the person committed, he or she will usually stay on the phone.

Fowler stated that for health-related studies, he gets in the range of 75% to 80%, with 79% on a drunk-driving survey. Two state surveys (more like polling) were much lower when there was a special topic. Response rates with vague topics were much worse, he noted. Fowler also described an experience with probability samples, selecting addresses, then locating phone numbers from a reverse phone book. Their response on the phone was 60% to 70%. They then picked up the rest in personal visits to homes. Response rates were comparable, and they got rid of all telephone coverage biases that way. He recommended this as a very fast, middle-range way to do a community survey.

Peter Miller was asked whether there are any P values for Table 5 of his paper and which components of an experimental condition have the most impact. He replied that all are significant at the .05 level except doctor visits, health status, and hospitalization. He commented that it is difficult to determine which ones or how they contribute; each component seems to contribute.

In a discussion of record checks, Kulka commented that interviews produce more information than records. The only way to resolve the problem is to examine the total array of error and to do careful studies of total random error. Montiero expressed concern about reports of hospital stays, commenting that this ought to be the easiest thing to measure, but that accurate reporting was a problem. In response, Poe noted that even when "hospital stay" is defined explicitly, people report outpatient visits as hospital stays. Kulka noted that on a cancer study, they matched reports and records and found patients reporting outpatient visits as a hospital stay many times.

SESSION 3:
Studies of survey measurement techniques

Chair: Lu Ann Aday, Center for Health Administration Studies, University of Chicago

Recorder: Maurice Satin, Division of Mental Health Care Systems, Long Island Research Institute

Internal consistency analysis: A method to validate health outcome, function status, and quality-of-life measurement*

John P. Anderson, Department of Community and Family Medicine, University of California at San Diego

James W. Bush, Department of Community and Family Medicine, Health Policy Project, University of California at San Diego

Charles C. Berry, Department of Community and Family Medicine, University of California at San Diego

Introduction

Measurement of the quality of life—the systematic classification and expression of function status, the presence of symptoms and problems, and other possible attributes of personal and social well-being, such as housing, jobs, transportation, interpersonal relations, etc.—is a problem of increasing importance. It encompasses the followup of clinical treatments and other health programs and extends more generally to the assessment of all nonhealth-related aspects of life quality (Campbell et al., 1976; Hill et al., 1973; Wingo and Evans, 1978; Dalkey et al., 1972; Environment Protection Administration, 1973).

More than a decade ago, an analytical framework was proposed for a General Health Policy (resource allocation) Model (GHPM) that demonstrated the need for and central role of a metric scale for combining function status, symptoms, and problems into a comprehensive expression for the health-related quality of life (Fanshel and Bush, 1970). In subsequent research, operational approaches to all components of this comprehensive model have been developed and evaluated (Bush et al., 1972; Bush et al., 1973; Kaplan et al., 1976; Bush et al., 1982).

Until valid measures of the health-related quality of life are available, confidence in all measures of nonfatal outcomes is seriously compromised, and the scientific evaluation of many widespread, expensive, and possibly ineffective therapies is prevented. This study is another step in progressive, cumulative research to develop accurate, reliable, efficient function classification instruments for the QWB scale.

* The authors gratefully acknowledge the research support provided for the initial and follow-up surveys by Grant No. 2R18HS00702 from the National Center for Health Services Research; the collaboration of John Scott, Charles Cannell, Irene Hess, and others of the Survey Research Center, Institute for Social Research, University of Michigan, which conducted the initial survey; and the research assistance of Stephanie Johnson, Sharen Sava Cox, and especially Judy Jamison, for compiling the background data for discrepancy analysis.

Because of the inadequacy of conventional approaches to validation for this problem, we have regularly employed and refined a construct approach to instrument validation we call Internal Consistency Analysis (ICA). This method not only allows the confident detection of individual classification errors, it also exposes the sometimes massive, widespread, and systematic disagreements between putatively identical measurement methods and provides a basis for progressive improvement of the classification instruments.

This paper presents the general principles of Internal Consistency Analysis as it has been applied to two successive field experiments, including prospective evidence from those studies concerning ICA as an approach for improving instrument validity. In addition, we will adapt a set of standards from other disciplines for assessing validity of different instruments and compare these results with traditional correlational standards.

Methods

Prospective internal consistency analysis. Although Internal Consistency Analysis was developed in the context of specific studies, its principles are outlined here in a very general form.

1. Develop two or more sets of questions, or perhaps independent instruments, to categorically measure or classify the same phenomena.
2. Submit both sets of questions to respondents in a randomly counterbalanced design, to test the effect of presentational order from administering the instruments in close proximity.
3. In addition to categorically coded responses, obtain as much ancillary descriptive information about each topic as possible. This can be accomplished in several ways, including (a) standardized open-ended probes to amplify each categorical response, (b) space and training for interviewers to record all such open-ended responses, (c) tape recording all inteviews *in toto*, (d) use of ancillary questions about the general subject to augment the available descriptive information, also with both categorical responses and follow-up probes, (e) training the interviewers to complete extensive thumbnail sketches with standardized elements to capture additional information such as observations of the subject, information from proxies, etc.
4. List the items or questions where identical or similar responses are expected, and tabulate all discrepancies between the classifications for each such item or question.

5. Summarize all the descriptive data on each discrepancy from all sources in a convenient standardized format.
6. Have several investigators review the data jointly, including primary or original materials on each error, to minimize individual bias in interpretations.
7. Establish a consensus on the actual state of the subject, the instrument in error, and the direction of the error (false positives vs. false negatives).
8. Establish a residual category for undefinable states, to avoid assumptions where the available evidence is adequate for a confident classification.
9. Using descriptive information, including the respondent's own words in answering categorical questions and descriptive probes, evaluate the probable cause for each error and develop a convenient classification for similar and closely related causes, also with a residual category for errors with insufficient evidence for clear assignment.
10. Devise solutions to the problems detected, including revised instruments, training methods, coding procedures, etc., and consolidate the proposed solutions in a new interview schedule.
11. Repeat the study and the internal consistency analysis (steps 1 through 10) from above.
12. Tabulate the changes in the error rates for comparable categories to determine whether improvements actually occur, i.e. whether discrepancies have decreased and more accurate responses have been obtained on one or all instruments.

Surveys. Two household interview surveys were conducted as one part of the overall effort to develop and test multiple operational components for the General Health Policy Model, including the Quality of Well-Being (QWB) scale (Patrick et al., 1972; Kaplan et al., 1976).

In the initial survey, which had an 80% response rate, each person's daily state of functioning was classified on three scales: (1) a five-step mobility (MOB) scale, (2) a four-step physical activity (PAC) scale, (3) a five-step social activity (SAC) scale.

The functioning of 1,324 subjects (866 respondents, 89 intentionally selected dysfunctional persons, and 369 randomly selected children) was assessed on each of the three scales using two different instrument: (1) *a self-administered form,* where respondents selected the number of the one step on each scale that best described themselves (and the subjects for whom they were reporting) on the single day before the interview (yesterday); and (2) *an interviewer-administered form,* where respondents answered a series of direct questions and follow-up probes about their own (and other subjects') functioning, also for the single day before the interview.

In general, the interviewer form first asked whether the subject actually performed a specified activity, and then probed to determine whether the reasons for non-performance were related to health. The responses were then logically mapped into the scale steps. Figure 1 presents the definitions of the steps on each scale for the two forms. Both forms were administered to all respondents in a randomly counterbalanced experimental design to test for crossover effects from the order of presentation.

In the followup survey, conducted with the same respondents about the same persons one year later, the redesigned self-administered and interviewer-administered forms were used to again classify the functioning of the population.

Based on the results of the initial ICA, the two instruments were altered in the following ways: (1) Elimination of all capacity wording from both the self-administered and interviewer forms, casting all questions and descriptive items strictly in the performance mode; (2) Altering the response rules on the self-administered form, by instructing the individuals to respond affirmatively to all items describing their current functioning. Using this information, each scale classification was made following the same rules as for the interviewer form.

In the followup, information was gathered on 694 of the initial survey respondents, who also provided proxy information on 292 of the selected children and 77 of the dysfunctional persons, for a total analyzable sample of 1,063 subjects. The relative frequency and distribution of the disagreements in function classification between the two methods provides the central empirical focus of this paper.

Conventional variable analysis. Standard statistical analyses related to the discrepant classifications include the following: (1) demographica variables: to determine if the discrepancies could be substantively linked to any standard demographic characteristics (e.g., age, sex, education, ethnicity, income, etc.); and (2) the order of presentation: to determine if the occurrence of discrepancies could be substantively linked to the experimental variation of the forms, e.g., whether the errors were predominant in one mode of administration; and (3) interviewer-respondent interaction characteristics: using the interaction codes developed by Cannell et al. (1975) on tape recordings of the interviews to see if interviewer or respondent behavior in the interaction could be responsible for the production of discrepant classifications in any substantial way.

Data for internal consistency analysis. We synthesized descriptive data from all available sources to construct an internally consistent actual state for all subjects with discrepancies. These sources included: (1) narrative responses to open-ended probes following each categorical question and probe on the interviewer form; (2) responses to a standardized, comprehensive list of symptom/problem complexes and, in the followup, to a standard list of chronic conditions, plus narrative comments on each response recorded by the interviewer; (3)

Figure 1. Comparison of interviewer and self-administered scale step definitions

Scale Step	Mobility Scale: Direct Questions (Logical Results)	Mobility Scale: Flash Cards (Exact Words)
5.	Drove without help -or- Did not drive without special equipment, but traveled without help -or- Did not drive or travel freely (not health-related), but did go outside without help -or- Did not drive or travel freely, or go outside (not health-related).	"Able" to both drive and use public transportation (bus, train, etc.) without help. (For a child: able to travel as usual for age.)
4.	Did not drive or travel freely (health-related), went outside without help -or- Did not drive or travel freely (health-related), did not go outside without help (not health-related).	"Could" go outside without help, but could not drive and/or could not use public transportation without help from another person. (For a child: needed more help to travel than usual for age.)
3.	Did not drive or travel freely (not health-related), did not go outside or needed help (health-related) -or- Did not drive or travel freely (health-related), did not go outside or needed help (health-related).	"Needed" help to go outside, or stayed inside all day for health reasons.
2.	In hospital, but not in special unit.	In a hospital, nursing home, mental hospital, home for the retarded, etc., as a patient.
1.	In hospital, in special unit.	In a special unit of a hospital, such as an operating or recovery room, intensive care unit, incubator, isolation ward, etc., for any part of a day.

Scale Step	Physical Activity Scale: Direct Questions (Logical Results)	Physical Activity Scale: Flash Cards (Exact Words)
4.	Not in wheelchair, not in bed or chair, no limitations above -or- Not in wheelchair, but in bed or chair (not health-related), no limitations on walking.	Walked without any of the physical limitations above
3.	Not in wheelchair, not in bed or chair, one or more limitations on walking -or- Not in wheelchair, but in bed or chair (not health-related), one or more limitations on walking.	Walked with one or more of the following limitations: --used cane, crutches, artificial limbs or braces, etc. --had trouble lifting, stooping, or using stairs or inclines, or --had trouble walking as far or as fast as usual for other persons of the same age.
2.	In wheelchair, but able to move it without help.	Moved own wheelchair without help.
1.	In wheelchair, unable to move without assistance -or- Not in wheelchair, but in bed or chair (health-related).	In a bed or chair for most or all of the day for health reasons.

Scale Step	Social Activity Scale: Direct Questions (Logical Results)	Social Activity Scale: Flash Cards (Exact Words)
5.	Worked, did housework, schoolwork, etc., without limitations, -or- Did not work, do housework, schoolwork, etc. (not health-related) AND Performed other activities without limitations, -or- Did not perform other activities (not health-related).	Could work, do school or housework, and other activities
4.	Limited in other activities (health-related) -or- Limited in other activities (health-related), but did not limit work.	Could work, do school or housework, but limited in amount or kind of other activities, such as clubs, shopping, church, hobbies, games or sports, as usual for age.
3.	Limited in work activities (health-related) -or- Limited in work and other activities (not health-related).	Could work, do school or housework, but limited in amount or kind, or needed special classes or working aids.
2.	Did self-care, did not work (health-related) -or- Did self-care, did not work (not health-related), but could not have worked (health-related) -or- Did self-care, could not have worked (health-related).	Could not work, do school or housework, but could eat, dress, bathe, and toilet without help.
1.	Didn't do one or more self-care, did not work (health-related) -or- Didn't do one or more self-care, did not work (not health-related), but could not have worked (health-related) -or- Didn't do one or more self-care, could not have worked (health-related)	Needed more help than usual for age for eating, dressing, bathing, or using toilet.

responses to categorical questions, and narrative responses to open-ended probes regarding calls and visits to physicians, and/or other health services use; (4) narrative responses to open-ended catch-all questions, asked at the end of both the self-administered form and the total interview, regarding any health-related information that had not, in the respondent's view, been sufficiently described previously; (5) any amplifying, descriptive comments made after single number responses on the self-administered form; (6) separate descriptions by the trained interviewers in a *household* open-ended thumbnail sketch about the situation, any apparent but unreported health problems, the apparent reliability of the informant; and (7) complete tape recordings of all interviews where the respondents permitted taping (over 90%), as suggested by Bucher et al., (1956).

Determination of actual states, form in error, and direction. Research assistants compiled information from all sources for each of the 471 initial and 323 follow-up survey discrepancies for joint review by the authors. This review undertook these tasks: (1) To resolve discrepancies and establish the actual state of functioning on each scale for each individual, as well as the erring form and direction, i.e., whether more or less dysfunction was recorded than actually existed. Precedents for such procedures include using final diagnoses, i.e., a judgmental "best code" based on multiple symptoms, signs and tests, as estimates of true prevalence for comparison with screening test results (e.g., Bernadt et al., 1982). (2) To assess the probable cause of the error to guide new versions of the instruments and procedures of the interviews. Identification of the actual state automatically identified the form and and direction of error. The form in error could be (1) the interviewer form, (2) the self-administered form, (3) both forms, i.e., where the actual state was not matched by either form, or (4) neither form in error, where the data on function were accurately gathered and coded, but a data-entry or programming error occurred.

Appendix 1 illustrates the amounts and kinds of information available for each case, and how inferences were made about the actual states, the direction, and the probable cause of the errors.

Receiver operating (validity) characteristics. Specifying the direction of error—falsely reporting either more or less dysfunction than actually existed—enabled us to employ Receiver Operating Characteristic (ROC) analysis, now commonly used to assess the validity of laboratory values, X-rays, and other diagnostic tests in clinical medicine and epidemiology.

The usual terms for errors when using these characteristics ("false positives" and "false negatives") are confusing, however, when applied to quality-of-life measures, where "positive" is "better." For our purposes, "positives" are cases of dysfunction and "negatives" are fully functional cases. Therefore, we shall employ an adaptation of the usual terminology.

Still further, multiple levels or steps of dysfunction mean that the standard categories in conventional ROC analysis must be augmented to be analytically complete, as in Figure 2. In addition to the two standard error types—called (c) *false function* (dysfunction reported as full function) and (b) *false dysfunction* (full function reported as dysfunction)—we must also distinguish (e) *off step dysfunction* (dysfunction reported as the incorrect step or level of *dysfunction*). Removing the false dysfunction and the off step dysfunction from the total reported dysfunction leaves (a) the *exact step dysfunction* (dysfunction reported in the correct step of dysfunction).

The *exact step dysfunction* is always the numerator in calculations of sensitivity and predictive value dysfunctional for an instrument, while the categories of error (now three instead of two) will be incorporated in the denominators for such calculations as appropriate.

With these terms, the Receiver Operating Characteristics are calculated as follows: (1) *Sensitivity*: the exact-step-dysfunction (a) divided by the total actual dysfunction, which is the sum of (a) the *exact step dysfunction*, plus (e) the *off step dysfunction*, plus (c) *the false function*. (2) *Predictive Value Dysfunctional*: the *exact step dysfunction* (a) divided by total reported dysfunction, which is the sum of (a) the *exact step dysfunction* plus (e) the *off step dysfunction* plus (b) the *false dysfunction*. (3) *Specificity*: the full (correctly reported) function (d) divided by the total actual function; which is the sum of (b) the *false dysfunction* plus (d) the *full (correctly reported) function*. (4) *Predictive Value Functional:* the *full (correctly reported) function* divided by the total reported function, which is the sum of (c) the *false function* plus (d) the *full (correctly reported) function*.

The validity characteristics of the binary Bayesian analysis are useful, but because of the multiple steps and levels to be distinguished by the instruments for precise quality-of-life measurement, the ordinary categories must be augmented to clearly expose all the possible types and sources of error. Conventional correlational

Figure 2
Error types

	Actual dysfunction	Actual function	
Reported dysfunction	(a) Exact step dysfunction (e) Off step dysfunction	(b) False dysfunction	Total reported dysfunction
Reported (full function)	(c) False function	(d) Full function	Total reported (full) function
	Total actual dysfunction	Total actual function	

criteria of validity will also be presented and contrasted with the Bayesian approach.

Probable causes for measurement errors. Each discrepancy/error was assigned to one of a set of "probable cause" categories, on the basis of the most reasonable inference. A general description of each category is given here. Appendix 2 defines the causes more completely and presents the standardized rules for assigning each probable cause.

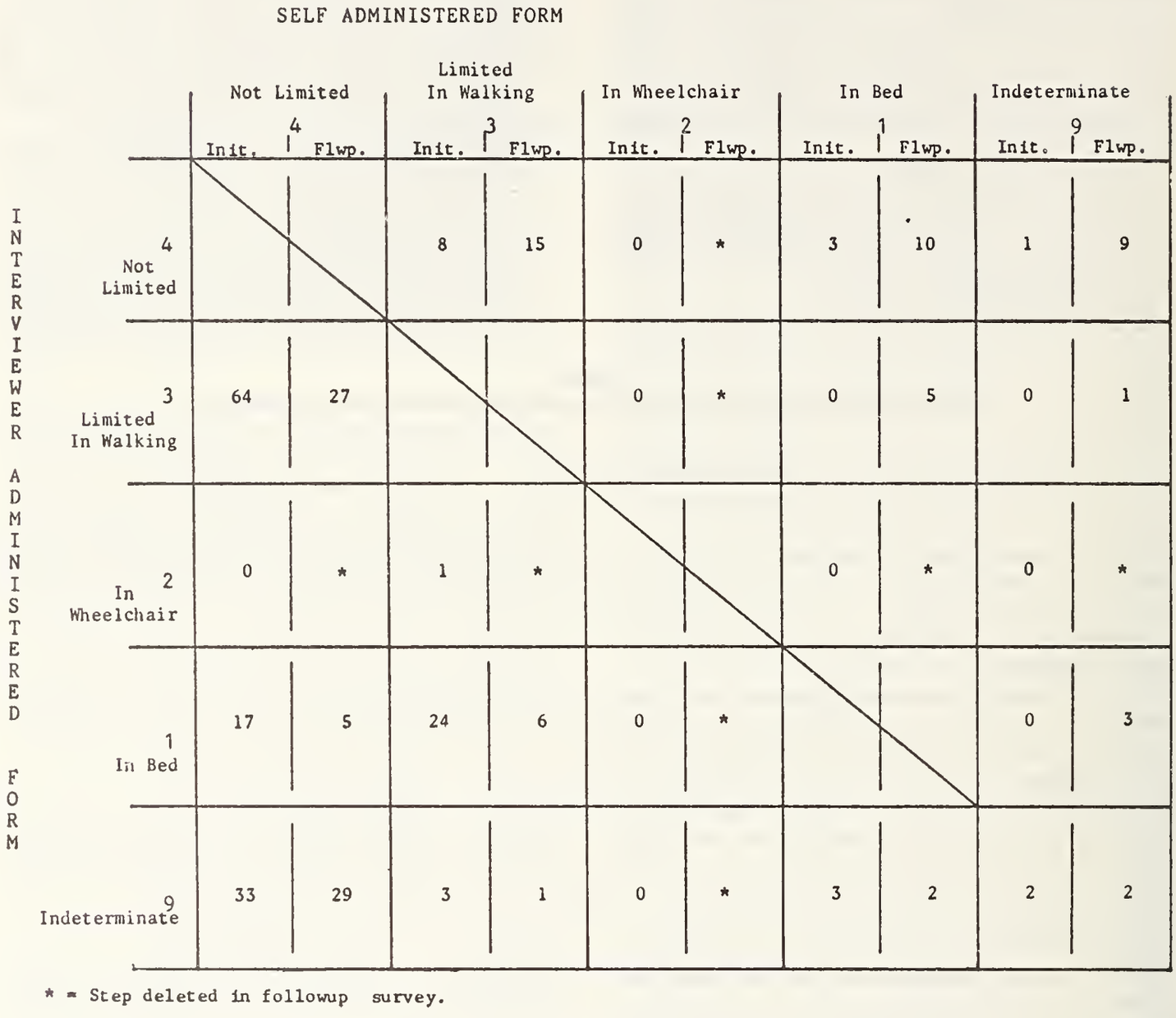
1. *Mechanical Errors* (ME): Errors in the mechanical process or in gathering and processing survey information, such as interviewer omission of selected children, failure to follow a skip pattern, coding, data entry, programming errors, etc.
2. *Comparability Problems* (COMP): Errors produced by inadvertent differences between the self-administered and interviewer forms at specific scale and step intersections, as outlined in Appendix 2.
3. *Respondent Problems* (RP): Errors produced by respondent conditions that prevented appropriate interpretation and processing of information (e.g., retardation, senility, inebriation, etc.), so that the respondent could not be viewed as a reliable informant (usually reported by the interviewer).
4. *"Other" Problems* (Other): Diverse, clearly specifiable, but low frequency circumstances that altered reponses on one or both forms.
5. *Performance/Capacity Errors* (P/C): Errors due to a specific difference in wording, where one form asked about the activities that a subject actually performed yesterday, while the other form asked about

Figure 3
Distribution of initial and follow-up survey discrepancies by step intersection, mobility scale

INTERVIEWER ADMINISTERED FORM	SELF ADMINISTERED FORM: Not Limited (5) Init.	Not Limited (5) Flwp.	Limited in Travel (4) Init.	Limited in Travel (4) Flwp.	In House (3) Init.	In House (3) Flwp.	In Hospital (2) Init.	In Hospital (2) Flwp.	In Hospital, In Special Unit (1) Init.	In Hospital, In Special Unit (1) Flwp.	Indeterminate (9) Init.	Indeterminate (9) Flwp.
5 Not Limited			31	41	8	*	0	9	0	0	2	5
4 Limited In Travel	3	15			6	*	0	0	0	0	0	1
3 In House	9	*	7	*			0	*	0	*	0	*
2 In Hospital	1	0	2	1	0	*			1	0	1	0
1 In Hospital, In Special Unit	2	0	0	0	0	*	1	0			0	0
9 Indeterminate	37	21	2	3	1	*	0	0	0	0	2	3

* = Step deleted in followup survey.

Figure 4
Distribution of initial and follow-up survey discrepancies by step intersection, physical activity scale



what the subject "could," "needed," "required," or "was able" to do on that same day.

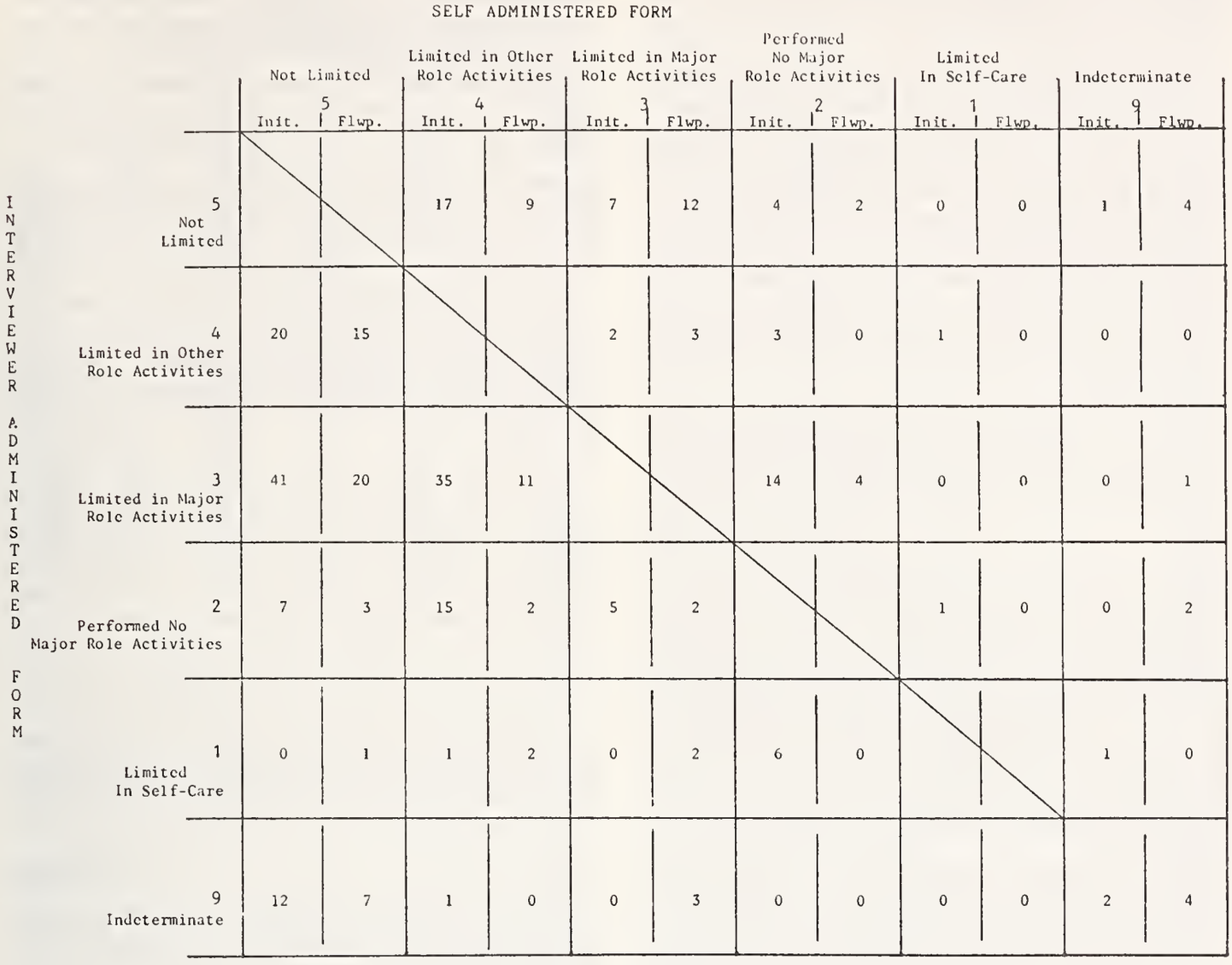
6. *No Apparent Reason* (NAR): A residual code used for all errors where the available evidence was insufficient to identify a specific cause for the misclassification.

Effectiveness tests for internal consistency analysis. Multiple causes for discrepancies provide a test for the effectiveness of ICA through the implied revisions. If error frequencies at scale and step intersections dominated by correctable probable causes improve more than such frequencies at scale and step intersections where the causes were not known or not as open to corrections, the difference can be accepted as measureable positive evidence for the effectiveness of the ICA directed revision. If such differences do not arise, the effectiveness of ICA is unsupported.

Results and analysis

Discrepancies. Administering the two forms to the same population during the initial survey produced different (or *discrepant*) classifications of functioning on 471 occasions. This represents approximately 12% of all 3,972 scale classifications (1,324 subjects × 3 scales). These errors involved 25% of all subjects and over 60% of all persons reporting dysfunction on any one scale. Over 90% of all these errors were on the self-administered form.

Figure 5
Distribution of initial and follow-up survey discrepancies by step intersection, social activity scale



Figures 3, 4, and 5 report the discrepant classifications from the initial survey for all three scales and briefly describe each scale step. To understand the off-diagonal entries, consider Figure 4, which shows that the interviewer form classified 64 separate subjects as being "Limited in Walking" (step 3), while the self-administered form classified all the same persons on the same day as being "Not limited in any way" (step 4). The discrepancies were not evenly distributed among the steps on the scales, with the self-administered instrument systematically indicating higher levels of function than the interviewer mode.

Conventional variable analysis

1. Demographic variables. Analyses of the relative frequency of discrepant classifications by the usual range of demographic variables (e.g., age, sex, ethnicity, education, income, etc.) showed no substantial differences, once the type of sample subject (respondent vs. dysfunctional person vs. selected child) had been controlled.

2. Experimental balance variable. The presentational order of the two forms was associated with discrepancies only on the Physical Activity scale.

When the interviewer form was presented first, 7% of the PAC classifications were discrepant, but when the self-administered form was presented first, 13% of PAC classifications were discrepant. The Mobility and Social Activity scales showed no such differences.

3. Interviewer-respondent interaction analysis. One hundred sixty initial survey households were sampled, stratifying on the basis of respondent ethnicity, educational attainment, and health status, such that interviews involving discrepant classifications were overrepresented in the analysis. Tapes of these interviews were analyzed using codes of recorded interviewer and respondent

behavior, as developed by Cannell et al. (1975). The analysis detected no systematic relation betwen the codable behaviors and discrepancy production, beyond occasional interviewer errors.

Receiver operating (validity) characteristics. Table 1 presents the sensitivity, specificity, and predictive values for the forms. The results for the self-administered form have been limited to the Physical Activity and Social Activity scales, since the experimental Mobility scale was considered inappropriate to include in this analysis.

These values may be interpreted as (1) proportions of the actual state being correctly recorded by instruments (sensitivity and specificity), and (2) proportions of the events as recorded by the instruments that are correctly recorded (predictive value dysfunctional and predictive value functional).

Table 1 reveals that the specificity and predictive value functional for both forms are .94–.99 through all occasions of their use. The fact that these figures remained so uniformly high indicates that both instruments validly record full function, and that improvement in other validity characteristics was not obtained by sacrificing specificity or predictive value functional.

The interviewer form sensitivity values in the initial and follow-up surveys were .90 and .89 respectively, with predictive value dysfunctional figures of .93 and .95. The interviewer form predictive value dysfunctional of .95 means that when the form reports dysfunction in a given step, it does so correctly 95% of the time. The interviewer form characteristics would be considered quite high for almost any laboratory tests (Bernadt et al., 1982).

By contrast with the sustained high levels for the interviewer form validity characteristics, and by contrast with the specificity and predictive value functional of the self-administered form itself, the sensitivity and predictive value dysfunctional for the self-administered form were both much lower initially and changed much more. The *self-administered form* sensitivity improved from .46 to .64 (+.18) and its predictive value dysfunctional improved from .63 to .76.

Table 1 also displays correlations between the two forms, as well as their correlations with the actual state for the initial survey. The self-administered form correlated .89 with the actual state, while the correlation between the interviewer form and the actual state was .92.

The high values for these correlations should be contrasted with the more accurate sensitivity and predictive value dysfunctional results that are proportions of direct observations (not correlations) and that reveal much higher rates of error on the self-administered form.

Distribution of correctable and uncorrectable causes. When compiled and distributed by intersecting steps after the Internal Consistency Analysis, the probable causes displayed a systematic distribution—different causes predominated at different intersecting steps. At the intersection of interviewer form 3 and self-form 4 on the PAC scale, for example, the residual "No Apparent Reason" code accounted for 39 of the 64 errors—a clear majority. Other intersections showed other predominant causes.

Some of the causes of errors were related to specific features of the instruments that were potentially correctable, e.g., Performance/Capacity and Comparability. Other causes indicated problems that we either could not attack, e.g., Respondent Problems, or could not even positively identify, i.e., No Apparent Reason. To prospectively test the validity of ICA objectively, independent of the results of any ICA, we selected intersections containing three or more errors for analysis. These intersections were then divided into three groups: (1) intersections with predominantly (equal to or greater than 50%) correctable causes; (2) intersections with predominantly uncorrectable cause; and (3) the remaining intersection where neither type of cause predominated. In the initial survey, the correctable group had 109 and the uncorrectable group had 160 discrepancies. The two groups together constituted 90% of the relevant discrepancies, since only 27 of the PAC and SAC errors fell outside of the groups.

Table 2 shows the relative frequency of the discrepancies in the initial and follow-up surveys grouped into the correctable and uncorrectable majority categories.

Relating these causes of frequencies to the total actual dysfunction on the PAC and SAC scale in the two surveys more clearly reveals the proportionate changes. The

Table 1
QWB initial and follow-up survey validity characteristics and correlations

	QWB validity characteristics: Self-administered form (PAC & SAC scales)			Interviewer form (all scales)	
	Initial survey	Follow-up Survey	Δ	Initial	Follow-up
Sensitivity	.46	.64	+.18	.90	.89
Predictive value dysfunctional	.63	.76	+.13	.93	.95
Specificity	.99	.99	+.00	.99	.99
Predictive value functional	.94	.96	+.02	.99	.99

	Instrument level of well being (QWB) correlations, initial survey		
	Self form	Interviewer form	Best code
Self form	—	.94	.89
Interviewer form		—	.92
Best code			—

Table 2
Initial survey—follow-up survey test intervals discrepancy frequency comparison by initial survey majority probable cause code

Scale	Step intersections		Initial survey Freq.	Initial survey Prop.	Follow-up survey Freq.	Follow-up survey Prop.
No apparent reason and respondent problems in majority						
PAC	IAF3,	SAF4	64	.123	27	.080
SAC	IAF3,	SAF2	14	.027	4	.012
SAC	IAF3,	SAF5	41	.079	20	.059
SAC	IAF4,	SAF5	20	.038	15	.045
SAC	IAF5,	SAF2	4	.008	2	.006
SAC	IAF5,	SAF4	17	.033	9	.027
		Totals	160		77	
			521	.307	337	.228
			Δ = .079			
			Rel. = .257			
Performance/capacity and comparability in majority						
PAC	IAF1,	SAF3	24	.046	6	.018
PAC	IAF1,	SAF4	17	.033	5	.015
SAC	IAF1,	SAF2	6	.012	0	—
SAC	IAF2,	SAF3	5	.010	2	.006
SAC	IAF2,	SAF4	15	.029	2	.006
SAC	IAF2,	SAF5	7	.013	3	.009
SAC	IAF3,	SAF4	35	.067	11	.033
		Totals	109		29	
			521	.209	337	.086
			Δ = .123			
			Rel. = .589			

correctable discrepancies dropped by 60% (from .25 to .10) from the initial proportion, while the uncorrectable discrepancies dropped only 27% (from .37 to .27).

Discussion

Need for quality-of-well-being scale. Working with a series of colleagues over the past decade and a half, Bush has guided developed of operational components for a General Health Policy Model (GHPM) for evaluation and resource allocation for public issues concerning health (Hopkins, 1969).

In an early conceptual paper, Fanshel and Bush (1970) demonstrated that Well Years are the necessary and final result of applying expected utility (decision) analysis to health treatments and policies.

Well Years result from integrating the level of wellness, or health-related quality of life, over the life expectancy. The level of wellness at particular points or short intervals is governed by the prognoses of the underlying disease or disorder under different treatment (control) variables.

Using Well Years as the output measure, methods have been developed for outcome measurement (Bush and Fanshel, 1970; Fanshel and Bush, 1970), cost-effectiveness (Bush et al., 1972; Ibid., 1973), resource allocation (Chen and Bush, 1975; Chen et al., 1976), medical-care quality assessment (Bush et al, 1975), well-state utility measurements (Patrick et al., 1972; Ibid., 1973, Kaplan et al., 1978; Ibid., 1979; Bush et al., 1983), community health status estimation and program analysis (Chen and Bush, 1975), prognosis estimation (Bush et al., 1971; Berry and Bush, 1978), classifications for levels of function and quality of life (Patrick et al., 1972; Kaplan et al., 1976; Anderson et al., 1977), and evaluation and policy analysis (Kaplan and Bush, 1982).

Because of its central role in everyday professional and public thinking and therefore in the model of health status, extensive field research has been devoted to developing a quality-of-life measure that would meet widely accepted scientific standards of accuracy and reliability. Because of its assumed ease of administration and economy, the self-administered form was tested in our original study to become the primary instrument for collecting well-state information for the QWB scale. The interviewer instrument was designed and included primarily to validate the self-administered form. Previous interviewing research provided little evidence on the differential validity of the two methods, and nothing suggested such a large number of errors as those detected. Their magnitude and pervasiveness were totally unexpected.

Formal statistical approaches. The codes of tape-recorded interviewer and respondent behavior, as developed by Cannell et al. (1975), revealed several instances where interviewers had incorrectly skipped questions or misrecorded answers, but those instances accounted for few of the total number of discrepancies.

Elimination of the conventional demographic variables, of the presentational order of the forms, and of interviewer behavior as primary causes of the errors left the instruments themselves or the respondents' interpretations as the major possible sources of bias. Statistical and other formal procedures by themselves offered little further hope for investigating the reasons behind the errors. For that, we had to explore the meaningful substance of the respondents' answers.

In this regard, the initial tape-recording analyses had not only eliminated interviewer behavior as a primary cause of classification errors, they had also strongly suggested that closer examination of how respondents were answering the questions might identify additional causes and point to other factors contributing to the obvious mistakes. This closer examination proved richly rewarding and led us to use and codify Internal Consistency Analysis as a formal methodology.

Conceptual foundations of internal consistency analysis. Internal Consistency Analysis synthesizes and extends several previous approaches. Reviewing validation study types two decades ago, Sagen et al. (1961) suggested that Internal Consistency could also help verify aggregate analyses, noting that "...in...Consumer Expenditure Studies,...revenue of all types...must...balance with outlays, if the statistics are to be valid" (p. 7). Hyman (1972) later formally suggested comparing responses to differently worded questions to determine their effects (pp. 194 ff). Bucher et al. (1956) found that most re-

spondent verbalizations go unreported by interviewers, suggesting that tapes might reliably capture more of the interaction than written records alone.

If questions are to be used only once, as in much market and political survey research, a deliberately redundant multiple question approach may have merit. If the same instrument is to be used for repeated comparisons, however, testing to assure that the questions are as well understood as possible is more appropriate.

This approach inquires into the substance of the interaction between investigators (via interviewers) and respondents. By contrast with formal or statistical approaches, which tend to accept recorded responses from interviewers as given, ICA makes the respondents interpretations an object of systematic inquiry, to see if their interpretations correspond to the intent of the question. In contrast to aggregate analyses (as cited by Sagen et al.), we have used ICA to validate individual responses which comprise the aggregate. Although unusual in current interview research, this approach merely applies a major and quite conventional social science research method—referred to variously as "verstehen" (interpretive understanding), "field research," or "qualitative methods"—to health-related quality-of-life measurement (Martindale, 1968).

Such techniques are widely accepted as valid scientific approaches to issues involving meaningful interaction and interpersonal communication, as applied, for example, in classic studies of urban dwellers (Liebow, 1967; Whyte, 1943). Interviews about daily functioning provide an excellent focus for verstehen methods because the content is relatively factual and noncontroversial.

Using both types of questions on the same topic—as in double entry bookkeeping—reveals how closed questions alone frequently produce inappropriate responses (although perhaps accurate for the question as understood), while open-ended probes generate additional information to develop instruments that avoid such errors.

When such explanatory information is known, understandings and interpretations are revealed (e.g., responding "I could work" rather than "I did work") that can be compared with the investigators' purpose or intended interpretation. We deliberately created conditions in our studies to maximize the value of such internal consistency checks.

In Internal Consistency Analysis as defined here, however, inferences from these methods are treated only as hypotheses about events in an initial interview. The indicated instrument changes are then evaluated empirically to test the initial hypotheses. Such prospective evidence provides transpersonal objectivity and distinguishes ICA from unconstrained speculation. Internal Consistency Analysis builds on the strengths of both open and closed responses; categorical responses are available for efficient, objective tabulation, while open-ended responses are available for meaning and interpretations. Each method cross-checks the other. In ICA the methods are complementary, not exclusive.

Both self-administered and interviewer forms are methods of receiving information, but the open-ended probes of the interviewer form reveal what the respondents considered in answering each question. The closed responses of the self-administered form rarely reveal what the respondent had in mind or how well that might match the investigators' intent in asking the question (the true meaning of validity). We gathered as much information as possible on similar and related topics (e.g., the experience of symptoms and problems, reasons for calls and visits to physicians, clinics and hospitals, reports of chronic conditions, means of payment, etc.) in addition to direct data on functioning by two different methods.

The approach might be extended in several ways, such as comparing proxy with respondent information about the same subject or situation, or comparing different accounts from (or about) the same (or different) subjects (or respondents) with different lags in time. All discrepancies between all the accounts could then be investigated.

Purposes and results of applying ICA. Each additional capability—clearly identifying errors, identifying actual states, identifying the form and direction of errors, plus reasonable inferences about the causes of most of the errors—is a considerable advance from the situation that existed before the initial survey. At that time, neither evidence nor professional opinion suggested that the self-administered and interviewer methods were not essentially equivalent means of obtaining quality-of-life information.

We approached the problem of inferring error causes by creating a number of probable cause categories, including a residual category, "No Apparent Reason," for cases where evidence on the reason for the error was totally lacking or too ambiguous to permit confident assignment of causes. The frequent use of this residual category indicates the difficulty of attributing the obvious errors to a particular cause. This category was the second most common in the initial survey, and by far the most common in the follow-up. In addition to the difficulty of the task, the frequent use of this category also indicates our reluctance to infer causes without firm evidence.

ICA does, however, reveal a great many of the causes and also permits us to quantify how many errors still have unknown causes. The "No Apparent Reason" category directly expresses the magnitude of this uncertainty about causality, even though the existence, size, and directive of the errors are known. No matter how detailed our contextual information, however, we cannot know the exact sequence of thoughts and events that produced a given response. Thus, ICA can help not only to improve specific methods, it can also expose the inherent limitations of methods that cannot be remedied. From this perspective, the major overall function of ICA

is to improve interviewing methods in general.

Receiver operating (validity) characteristics. With the specificity and predictive value dysfunctional of .94 and above for all instruments in all studies, quality-of-life questionnaires have little difficulty detecting and properly categorizing full function when it occurs. The sensitivity (.90) and predictive value dysfunctional (.93) for the interviewer form are similarly high on all surveys, so detecting and properly classifying f/Idysfunctionf/R is not a major problem for the interviewer form. Many common laboratory tests produce a higher proportion of errors than the interviewer instrument (Bernadt, 1982).

These validity characteristics are more appropriate for assessing categorial quality-of-life measures than conventional correlations. The correlations in Table 5 suggest that the interviewer and self-administered forms are equivalent instruments. But the self-administered form misses over half the actual dysfunction (sensitivity), and when it does indicate the presence of dysfunction, it correctly classifies it only 60% of the time (predictive value dysfunctional). That it correlates with the actual states almost as well as an instrument with 90% sensitivity and 93% predictive value dysfunctional demonstrates glaringly the insensitivity of correlation coefficients as measures of classification accuracy.

As the off-diagonal cases in Figures 3, 4, and 5 show, however, the problem of misclassification is still substantial. A sensitivity of .64 (for the self-administered form) is still far from .89 (for the interviewer from), and a predictive value dysfunctional of .76 is still far from .95. Thus the self-administered form and the interviewer form are not equivalent means for obtaining the same information on life quality. Correlation coefficients of .90 or above in such cases simply demonstrate the ineffectiveness of correlations as indicators of instrument validity.

Accuracy of internal consistency analysis. The major focus of this paper has been on ICA as a comprehensive method for generating new, useful information about questionnaires. Claims to scientific or objective status are conventionally documented in reliability studies by showing that the classification rules are being accurately followed.

Reliability studies are indeed valuable checks on the objectivity of procedures, but a sterner and less frequently encountered test, even in clinical research studies, is prospective demonstration of effectiveness. The number of coding schemes that can present evidence of reliability is many times larger than the number that can demonstrate empirical success.

Table 1 presents the main evidence for the validity of ICA generally and for the probable cause categories. It shows that the differential distribution of majority probable cause codes effectively discriminated between specific scale step intervals concerning the frequency of discrepancies in the follow-up survey. Such results cannot be produced by shifting cases from one probable cause to another. They are not the result of any judgment on our part. Rather, they represent real differences in the frequency of measurement errors that occur at specific scale step intersections, where failure to produce such changes would be impossible to hide. Thus, it represents a very stringent test of the predictive accuracy of the categories and analytic methods.

Improvements in the sensitivity of the physical and social activity scales of the self-administered form between the initial and the follow-up surveys indicate that we had substantial success. A sensitivity increase from .46 to .64 is a 40% improvement in the most critical aspect of accuracy. This is also reflected in a rise in the predictive value dysfunctional from .63 to .76. This demonstrates the substantial power of Internal Consistency Analysis to improve instrument performance and provides strong evidence for the validity of Internal Consistency as implemented (and advocated) here.

Implications for the general health policy model. Even as improved, however, self-administration is still profoundly inferior to interviewer methods. After the follow-up survey, therefore, the self-administered form was abandoned as an inherently biased instrument. Despite its inconvenience, the interviewer form was adopted as the only valid data-gathering method available for the QWB scale; all further instrument development efforts were directed to making it as efficient and as accurate as possible.

One change was the development of an interviewer instrument that gathered performance data over the previous eight consecutive days, thus appropriating for the interviewer form one of the advantages previously possessed only by the self-administered form. The issue is critical, since such measurement errors seriously undermine scientific confidence in all reports about the quality of life. Furthermore, systematic error (lack of sensitivity) in classifying function status outcomes at this level biases all statistical analyses *against* the effectiveness of clinical treatments and health programs using quality-of-life outcome measures.

On such a critical subject as outcome measurement in clinical trials, for example, where hundreds of thousands of dollars have been expended in patient care funds, salaries, and laboratory research, one might even consider the slight additional expense of using two versions of the instruments to obtain as nearly definitive a record as possible of the final outcome variable on which all other analyses ultimately depend.

The purpose of measuring the health-related quality of life is to determine the efficacy of treatments and health programs and to quantify the efficacy so it can be compared to costs for efficiency analyses and resource allocations. The ideal form of an instrument might be a set of self-administered, closed questions that could be completed rapidly by a former patient in follow-up studies. This would be desirable not only for economy, but

for simplicity in using the categorical responses directly as the outcome measure. Such responses give validity, reliability, and efficiency in the coding procedures themselves—highly desirable properties in themselves.

Our studies indicate, however, that we pay an extremely high price for these desirable attributes with regard to the health-related quality of life. The price is too high, in fact, because the responses to closed, self-administered questionnaires are systematically distorted.

One common circumstance is not appropriate for ICA. We must sharply distinguish between internal consistency analysis and open-ended questions used to improve instruments vs. actually evaluating a treatment or program. In unblinded prospective studies, bias may enter the final classification of health outcomes. Revisions and reclassifications open possibilities for manipulation that are highly undesirable. By searching for dysfunction in the untreated group and by not searching so diligently in the treated group, the classification of enough cases might be altered so as to influence the statistical tests. ICA is most helpful therefore in developing and refining instruments to be used without further modification in the context of particular evaluation studies.

Summary and conclusions

We began with an approach to improving quality-of-life measurement, where standard methods using a "gold standard" are not available. We employed an approach that we call Internal Consistency Analysis (ICA) to contribute to the measurement problem, based on similarities between our method and secondary analyses performed in the past. We regard the findings by ICA only as hypotheses about measurement, however; their accuracy must be established empirically in prospective field studies.

The prospective evidence, generally speaking, validated our hypotheses. We hope that this paper will lead other health policy analysts and measurement scientists to extend and apply this type of effort. Not only is it one of the few means available for improving health-related quality-of-life measurement, it should logically and scientifically precede the application of standard statistical techniques that assume, frequently without examination, that the questions as asked are meaningful to the respondents and that the understood meanings correspond to the purposes of the measurement and the study.

Appendix I: Examples of information on discrepancies developed in ICA

ID # EX 01

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	4	5
	Self-Admin. Form:	4	4	5
	Discrepancy:	MOB 5 vs. MOB 4		

(2) Data:
- (2.1) *Age:* 19 *Sex:* Female *Survey Status:* Respondent
- (2.2) *I'er Form:* No limitations (On MOB, Respondent reported in a car without limitations).
- (2.3) *Self-Admin. Form:* Answered 445. MOB, Respondent reported she dislikes driving. I'er: "Did you not drive because of health?" Respondent: "No, I just don't like to." R7 Probe: "Yesterday were you limited in any other way in the activities that are normal for someone your age?" Respondent replied: "No."
- (2.4) *Thumbnail:* I'er believes Respondent accurate except for SAF MOB scale.
- (2.5) *Symptom/Problem Complexes:*
  #2: Pain or discomfort in one or both eyes, such as burning or itching.
  #28: Overweight for age and height.

(3) Conclusions:
- (3.1) *Best Code:* MOB 5.
- (3.2) *Form in Error:* Self-Administered Form
- (3.3) *Probable Cause:* Respondent Problem

ID # EX 02

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	9	4	5
	Self-Admin. Form:	5	4	5
	Discrepancy:	MOB 9 vs. MOB 5		

(2) Data:
- (2.1) *Age:* 31 *Sex:* Male *Survey Status:* Respondent
- (2.2) *I'er Form:* 9 ("information unascertained") on MOB scale because I'er did not ask "leave house" pattern of questions. R says he did not drive because he didn't have anyplace to go (although he reports backing the car out of the driveway for his wife).
  No limitations noted on other IAF questions.
- (2.3) *Self-Admin. Form:* Replied 545 without comment. No limitations on probes.
- (2.4) *Thumbnail:* No relevant information.
- (2.5) *Symptom/Problem Complexes:* None

(3) Conclusions
- (3.1) *Best Code:* R backed car out of driveway, unlikely he stayed in house to do that. MOB 5
- (3.2) *Form in Error:* Interviewer Form
- (3.2) *Probable Cause:* ME

ID # EX 03

		MOB	PAC	SAC
(I) Problem:	Interviewer Form:	2	2	1
	Self-Admin. Form:	2	3	1
	Discrepancy:	PAC 2 vs. PAC 3		

(2) Data:
- (2.1) *Age:* 70 *Sex:* Female *Survey Status:* Dysfunctional
- (2.2) *I'er Form:* PAC 2 because Dys spent most of the day in a wheelchair, but moved it without help from someone else. Dys is physically unable to bathe herself.
- (2.3) *Self-Admin. Form:* PAC 3—R said Dys lives back East and they haven't heard from her in a few weeks. R7 Probe: Dys limited due to broken arm and diabetes.
- (2.4) *Thumbnail:* Nothing relevant. R got very impatient and irritated towards the end of the interview.
- (2.5) *Symptom/Problem Complexes:* #21: One *hand or arm* missing, deformed (crooked), paralyzed (unable to move), or broken (includes wearing artificial limbs or braces).
  Other symptoms: R says diabetes.

(3) Conclusions:
- (3.1) *Best Code:* PAC 2
- (3.2) *Form in Error:* Self-Administered Form
- (3.3) *Probable Cause:* Possible P/C, coded NAR.

ID # EX 04

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	1	3
	Self-Admin. Form:	5	1	5
	Discrepancy:	SAC 3 vs. SAC 5.		

(2) Data:
- (2.1) *Age:* 17 *Sex:* Female *Survey Status:* Selected Child
- (2.2) *I'er Form:* No problems on MOB scale. On PAC, SC spent most of yesterday in bed or chair because "she has dizzy spells from her head injury." On SAC, SC did go to work, but was limited in amount or kind because of "her head, a lot of pressure, tight muscles in neck."
- (2.3) *Self-Admin. Form:* 515 without comment. R7 Probe, "Yes, because of her head. She banged it against the bed headboard."
- (2.4) *Thumbnail:* R and her roommate (SC) seemed healthy and active. SC was in the room while

R answered for her. I'er suspected some of the Symptoms/Problems were exaggerated to tease or frustrate SC.

(2.5) *Symptom/Problem Complexes:* #3: Trouble hearing (includes wearing hearing aid).
#12: Sick or upset stomach, vomiting, or diarrhea (watery bowel movements).
#16: Headache, dizziness or ringing in ears.
#17: Spells of feeling hot, nervous or shaky.
#19: Pain, stiffness, numbness, or discomfort of neck, hands, feet, arms, legs, or several joints.
#31: Trouble learning, remembering, or thinking clearly.
#32: Loss of consciousness such as seizures (fits), fainting, or coma (out cold or knocked out).

(3) Conclusions:
(3.1) *Best Code:* SAC3
(3.2) *Form in Error:* Self-Administered Form
(3.3) *Probable Cause:* NAR

ID # EX 05

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	4	9
	Self-Admin. Form:	5	4	5
	Discrepancy:	SAC 9 vs. SAC 5		

(2) Data:
(2.1) *Age:* 17 *Sex:* Male *Survey Status:* Selected Child
(2.2) *I'er Form:* "9" on SAC because "WERE YOU LIMITED IN ANY WAY IN NONSCHOOL ACTIVITIES YESTERDAY, Such as...?" marked "yes" by I'er, but no follow-up questions asked. From tape, R answered "no," to question, not "yes." No further explanation necessary.

(3) Conclusions:
(3.1) *Best Code:* SAC 5
(3.2) *Form in Error:* Interviewer Form
(3.3) *Probable Cause:* ME

ID # EX 06

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	4	3
	Self-Admin. Form:	5	3	4
	Discrepancy:	PAC 4 vs. PAC 3; SAC 3 vs. SAC 4		

(2) Data:
(2.1) *Age:* 50 *Sex:* Male *Survey Status:* Respondent
(2.2) *I'er Form:* On PAC, R reported spending most of the day yesterday in a bed or chair because he was home with no work to do. When I'er asked "WERE YOU LIMITED YESTERDAY IN THE PHYSICAL ACTIVITY OF WALKING?" R immediately said yes, even before I'er specified limitations. I'er asked "DID YOU HAVE TROUBLE LIFTING, STOOPING, BENDING OVER, OR USING STAIRS OR INCLINES?" R said "No, because I didn't try any of them." "DID YOU HAVE TROUBLE WALKING AS FAR OR AS FAST AS OTHER PEOPLE YOUR AGE?" R replied "Yes...No, but I would have if I'd tried." On SAC, R reported "The limitation in my work yesterday would have been that I can't stand too long at one time, that's the only thing. I can't stand or walk in excess."
(2.3) *Self-Admin. Form:* On PAC, R specified 3 "c" ("trouble walking as far or as fast"). On SAC, R said 4 without comment. R7 Probe, R said "I don't have proper blood circulation in my legs so I can't walk or run. I'm limited in walking and definitely no running."
(2.4) *Thumbnail:* Nothing relevant to add.
(2.5) *Symptom/Problem Complexes:* #11: Cough, wheezing, or shortness of breath
#33: Taking medication or staying on prescribed diet for health reasons.

(3) Conclusions:
(3.1) *Best Code:* PAC 3; SAC 3
(3.2) *Form in Error:* Interviewer Form; Self-Administered Form
(3.3) *Probable Cause:* ME—I'er should have correctly interpreted his comments about "would have been limited if I'd tried."; COMP—looks like he would have been limited in both Major and Other activities.

ID # EX 07

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	4	3
	Self-Admin. Form:	5	4	5
	Discrepancy:	SAC 3 vs. SAC 5		

(2) Data:
(2.1) *Age:* 12 *Sex:* Female *Survey Status:* Selected Child
(2.2) *I'er Form:* No limitations indicated on MOB or PAC sales. On SAC, SC attended school, but is marked as "limited." R said "The only

activity she's restricted in is tumbling. She has a minimal curvature of the spine which is under treatment. She wears a lift in her shoe. It's really improved since she's been wearing the lift."

(2.3) *Self-Admin. Form:* R answered 545 without comment. P7 Probe: "No."

(2.4) *Thumbnail:* Nothing relevant.

(2.5) *Symptom/Problem Complexes:*

#4: Earache, toothache, or pain in jaw.

#5: Sore throat, lips, tongue, gums or stuffy, runny nose.

(3) Conclusions:

(3.1) *Best Code:* SAC 3

(3.2) *Form in Error:* Self-Administered Form

(3.3) *Probable Cause:* probably P/C

ID # EX 08

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	5	3	5
	Self-Admin. Form:	5	4	5
	Discrepancy:	PAC 3 vs. PAC 4		

(2) Data:

(2.1) *Age:* 53 *Sex:* Male *Survey Status:* Respondent

(2.2) *I'er Form:* No limitations on MOB. On PAC regarding "TROUBLE LIFTING, STOOPING...?" R said "I always have trouble lifting because of my back problem. I usually don't do it. I have other people do it."

(2.3) *Self-Admin. Form:* R answered 545 without comment. R7 Probe, "no."

(2.4) *Thumbnail:* R walks with no apparent problem. His left hand is shorter than the other and not completely formed. He did not use it in turning pages of the Casebook. The left side of his face is slightly pulled or drawn, though he made no mention of it. R was getting over a virus attack that had bothered him for the past few weeks—he missed one day of work but it didn't fall within our question period.

(2.5) *Symptom/Problem Complexes:*

#3: Trouble hearing (includes wearing hearing aid).

#6: Several or all permanent teeth missing or crooked.

#16: Headache, dizziness or ringing in ears. (Has sinus condition that causes headaches.)

#18: Weak or deformed (crooked) back. (A back problem due to a tailbone that is a little deformed. R had an accident that might have triggered it.)

#20: One *arm and one leg* deformed (crooked), paralyzed (unable to move), or broken (includes wearing artificial limbs or braces). (Deformed left hand and right club foot—R had surgery as a child and wears a built-up shoe.)

(3) Conclusions:

(3.1) *Best Code:* PAC 3

(3.2) *Form in Error:* Self-Administered Form

(3.3) *Probable Cause:* NAR

ID # EX 09

		MOB	PAC	SAC
(1) Problem:	Interviewer Form:	1	1	4
	Self-Admin. Form:	5	4	5
	Discrepancy:	MOB 1 vs. MOB 5; PAC 1 vs. PAC 4; SAC 4 vs. SAC 5		

(2) Data:

(2.1) *Age:* 14 *Sex:* Male *Survey Status:* Selected Child

(2.2) *I'er Form:* Overall problem is that SC did all his usual activities until 3 PM, when he received a bad cut on one of his fingers. He was taken to the hospital and admitted at 5 PM, had surgery, and stayed overnight.

On MOB, SC was patient in a hospital, and was in a special unit (for surgery). However, he also rode in a car and went outside the house without help, even after his finger was cut—"he could still open the car door and everything." On PAC, SC spent time in bed—"Yes, after 5PM, when he went into the hospital." On SAC, SC went to school as usual and was not limited in school activities. It was after school was over that he became limited in Other Role activities—"That's when he cut his finger. He and a friend were pitching Bowie knives. SC leaned over to pick his up, and just about that time his friend threw his."

(2.3) *Self-Admin. Form:* MOB, "5 until 3PM, then a 1." PAC, "4. Even in the evening, he was still able to walk around." SAC 5, no comments. R7 Probe: "Not until 3 o'clock. Then he couldn't use his right hand. He had a cut on the knuckle of the ring finger. Surgery to repair a tendon that was cut."

(2.4) *Thumbnail:* No thumbnail

(2.5) *Symptom/Problem Complexes:* #19: Pain, stiffness, numbness, or

discomfort of neck, hands, feet, arms, legs, or several joints.

(3) Conclusions:
- (3.1) *Best Code:* MOB 1; PAC 4; SAC 4
- (3.2) *Form in Error:* Self-Administered Form; Interviewer Form; Self-Administered Form
- (3.3) *Probable Cause:* Other; Other; Other. Normally all might be P/C.

Appendix 2: Rules for assignment of measurement errors/discrepant classifications of functioning to "probable cause" categories.

1. *Mechanical Errors:* This category covers errors by research personnel in the gathering and processing of data. Included are cases where
   (a) an interviewer missed questions that should have been asked, e.g., not following a proper skip pattern, or
   (b) an interviewer failed to properly record a response, or
   (c) where a coding, editing, or keypunching error produced an inaccurate or incomplete classification, or
   (d) where a computer programming error produced an inaccurate report of functioning.

2. *Performance/Capacity Errors:* Discrepancies that meet *all* the following conditions:
   (a) did not involve incomplete information,
   (b) (respondent, selected child, or dysfunctional) had some health problem or condition that accounted for limitation in functioning,
   (c) the problem or illness was not so severe as to totally prevent the subject from moving (e.g., paraplegia), *and*
   (d) the respondent *verbalized* the performance/capacity distinction either (i) directly as it applied to the subject, or (ii) indirectly by noting that the subject's activities had been or were being curtailed for reasons related to health.

3. *Comparability Problems:* Among *adult* subjects, this category was used to cover specific scale and step discrepancies where there were known differences between IAF and SAF definitions or possible response patterns. These were the circumstances:
   (a) On the Mobility scale, the IAF did not specify use of public transportation as a requirement for being in the optimal step, while the SAF did. Additionally, for step 4 on the Mobility scale, the SAF specified "could not use public transportation," but did not specify *for health reasons.* Many respondents, having no health problems, appeared to endorse this item because public transportation is simply not easily available in San Diego County.
   (b) On the Social Activity scale, having limitations in Major Role Activities (SAC steps 2 or 3) proved not to be mutually exclusive with having limitations in Other Role Activities (SAC step 4 regarding recreational activities, etc.), though the IAD SAC pattern of questions makes them appear so. Thus, on the self-administered SAC scale, persons having limitations in Major Role Activities (SAC steps 2 or 3) and a limitation in Other Role Activities (SAC step 4) were given the task of reporting one limitation, and one only. This, in conjunction with evidence that illness-related limitations in both Major and Other Role Activities were present, with no reference to the performance/capacity problem, resulted in assignment of an error involving these steps on this scale to the COMP category.

   Among subjects who were infants or children, this category was also applied on any scale when parents, in providing proxy information about their children, appeared to miss our proviso that questions should be interpreted "as usual for age," and respond that "child could not drive" or "infant could not walk," etc., where other evidence suggested that the subject children were not ill and were performing appropriately for their age.

4. *Respondent Problems:* Definite evidence that the respondent had some serious trouble in processing and interpreting information and could not be viewed as a wholly reliable informant, either generally or on some specific topic. This information was usually conveyed by interviewers in their Thumbnail Sketch, although they also indicated in the instrument margins where they thought respondent was unreliable. Included in this category were cases where the respondent was
   (a) suspected of being drunk, or
   (b) suspected or said he/she was on some sedative or tranquilizing drug(s), or
   (c) appeared to be having a severe (non-Spanish) language difficulty, or
   (d) appeared to be either senile or mentally retarded, or
   (e) appeared to be "thinking positively" and denying obvious physical limitations (e.g., using leg braces, crutches, etc.), or
   (f) the respondent refused to answer some questions about him/herself or another person, or, when answering about another subject, replied "Don't know" to some questions.

5. *"Other" Problems:* This was a "catch-all" category for all discrepancies/errors not covered above, i.e., where a cause *could* be positively specified, but where the frequency was too low to warrant a separate category. These were diverse, unusual sets of circumstances. Some could not be prevented by any conceivable changes in the instruments or research design, while

others pointed to potentially recurring sets of problems. Among the circumstances in common were:

(a) an interviewer properly choosing a Selected Child to interview about, then not conducting an interview on the child at all (as opposed to, for example, missing a question or pattern of questions),

(b) an institutionalized person, for whom the respondent was the "closest living relative" (thus making the institutionalized person a proper subject for inclusion in the sample) dying "one week ago yesterday," and thus becoming an impossible interview subject for "yesterday," and

(c) parents, in answering questions about an injured child's activities in school, including Physical Education classes as Other Role Activities, when they should have been included as part of the child's Major Role Activities, and the like.

6. *No Apparent Reason:* Used where the error could not be *positively* assigned to any of the categories. They did, however, have several characteristics in common:
   (a) the self-administered form was the form in error,
   (b) the error appeared on the physical or social activity scales,
   (c) the error involved underreport of dysfunction,
   (d) the subject *did* have some health problem to link to the report of dysfunction,
   (e) the person's problems did not totally immobilize (physically incapacitate) the subject, and
   (f) the error did *not* involve incomplete information.

The errors in this category were *not* selected for this category because they had these characteristics; but once included, they were found to have these characteristics in common.

Health diaries—problems and solutions in study design*

Lois M. Verbrugge, School of Public Health and Institute for Social Research, The University of Michigan

Introduction

Panel data on individuals help scientists measure and understand changes in people's lives. The number of panel studies has grown in the past twenty years, and they are becoming treasured archives of information about the dynamics of social status, health, earnings, and social ties.

Increasingly, scientists are designing longer and more complex panel studies, which enroll respondents for many years or require very detailed information from them. At the same time, researchers are more concerned about how respondents react to panel studies. What factors reduce people's willingness to join the panel, to stay in it, and to provide good quality reports? Does participation in a panel study influence respondents' attitudes and behaviors, especially the ones being studied?

So far, researchers have relied largely on hunches about "respondent burdens" and "conditioning effects" when designing panel studies. They devise field instruments and procedures with large hopes of minimizing those problems but few clues about how to do so.

How do we learn about respondent burden and conditioning effects in panel studies? With plush funds and ample time, we can conduct experimental studies that vary the panel tasks or field procedures. With restricted resources, we can still learn a great deal from completed and ongoing panel studies—by studying response rates and data quality for panel members; by asking respondents about problems, either at data collection points or when they drop out; by looking for unexpected trends in research variables over time. Panel studies that require lifetime membership or continuous personal records (such as diaries) are especially informative. They are extreme designs which probably entail the greatest respondent burdens and greatest risks of conditioning effects.

This paper analyzes a health-diary study which required daily entries by respondents for a six-week period. Respondents had to answer questions about their health every day, not just on days they felt ill or had medical care. Moreover, they had to fill out the diary themselves; no proxy reports were allowed. The study was conducted in a general population sample of white adults, so it provides a good opportunity to see how numerous social and demographic groups respond to a health diary.

Three topics are considered in this paper: sample attrition, task performance, and conditioning effects. The findings produce recommendations about staff activities, diary format, field procedures, and methodological research, with ultimate goals of retaining panel members, improving record quality, and reducing reactivity to the study.

The Health In Detroit Study

The Health In Detroit Study is a survey of white adults (18 years old and up) residing in the Detroit metropolitan area in fall 1978. A multistage probability sample of households in the Detroit SMSA was selected. In each eligible household, one adult was chosen as the study respondent. An initial interview was conducted face-to-face, covering such topics as current health status, health actions in the past year, health attitudes, life-style behaviors, stress and anxiety, social roles and feelings about roles, time constraints, and other sociodemographic information. Following the interview, respondents kept daily health records (DHR) for six weeks. Each day they answered questions about their general health status, symptoms of illness and injury, curative and preventive actions, mood, and unusual events. At the end of the diary period, a termination interview was conducted by telephone, with questions about general health status, changes in health attitudes and behaviors during the diary period, and reactions to the diary task.

The diaries were bound into week-long booklets. Respondents (R) who agreed to keep diaries were given booklets for Weeks 1 and 2 after the initial interview. They received two subsequent mailings (booklets for Weeks 3-4 and 5-6). Respondents mailed in completed booklets each week. Booklets were edited promptly on arrival at the study office. If ambiguities or substantial missing items appeared, the respondent was telephoned for the information. All diary keepers received several routine contacts during the diary period. Each week, they were telephoned or received a postcard, reminding them to mail in completed booklets and asking if they had any special problems. More details about the study design and goals are published elsewhere (Verbrugge and Depner, 1981; Verbrugge, 1979, 1980a).

Appendix 1 shows the daily health record (DHR) for the Detroit study.

* The author thanks Don Camburn, Kathleen Grasso, Yossi Harel, Elizabeth Keogh, and Julie Rubin for their assistance in the analyses. Technical reports prepared by them were especially useful (Camburn, 1980; Grasso, 1980; Keogh, 1980; Keogh and Camburn, 1982).

Sample attrition

Do the same population groups cause attrition at all stages of a health diary study? Which groups are ideal respondents, completing the study exactly as designed?

We will distinguish initial response rates from panel response rates. In most panel studies, an initial interview is conducted with sampled respondents. Interviewed people become the panel which provides more information over time. Analysis of initial response rates is usually difficult (for all kinds of surveys) because little is known about the noninterviewed people. Information can sometimes be gleaned from household screening forms, the sampling frame, or population census data. Analysis of panel attrition is much easier. The initial interview has ample sociodemographic, attitude, and behavior items for comparing panel dropouts with panel members who stay.

The Health In Detroit data allow some limited analysis of initial response rates and extensive analysis of panel response rates. Table 1 shows response rates based on five sample groups: Eligibles (N = 1041, sampled white adults in the Detroit SMSA), Interviewed (N = 714), Agreers (N = 651, interviewed people who agreed to keep daily health records), Beginners (N = 589, interviewed people who began DHRs and submitted one or more booklets), and Completers (N = 492, interviewed people who provided 42 days of diary data).[1]

We have also named some subgroups of Beginners: Dropouts (N = 97) are people who began DHRs but quit the study before providing 42 days of data. Time-gaps (N = 80) are people who began DHRs but skipped some days, so their diaries are not perfectly consecutive.[2] Perfect Cases (N = 450) are the ideal respondents, who provided 42 consecutive days of data. Nonperfect Cases (N = 139) are those who began DHRs but later dropped out or had time gaps, or both. These four subgroups are not mutually exclusive.

Figure 1 shows the study groups named above.

The Diary Panel consists of all Interviewed people (N = 714). Respondents learned about the diary task at the end of the initial interview. Until that point, no respondent behaviors were based on knowledge of the diary task ahead.

Overall response rates for the Health In Detroit Study are as follows: The interview response rate (69%) is not especially high, but it was typical for surveys in the Detroit SMSA in the late 1970's. Most of the interviewed respondents (91%) agreed to keep diaries, but only 82% actually started them. Two-thirds (69%) of the interviewed respondents completed six weeks of daily health records. This is a remarkably high rate for a mail-back diary strategy (see Sudman and LannOm, 1980; Verbrugge, 1980a). Note that the interview response rates uses Eligibles for the denominator, while all panel response rates use Interviewed cases.

1. A stable family situation enhances interview response rates and diary completion rates, especially the latter.

From a household screening form and interviewer observations at sampled addresses, we have some demographic information about Eligibles. Comparisons of initial attrition with panel attrition are therefore possible.

Interview response rates are lowest for: men; elderly people, followed by middle-aged ones; sole adults (no other adults in households); household heads; nonparents (no own-children present in household); and people in middle- and low-income households. Diary completion rates are lowest for: elderly people, followed

Figure 1
Types of respondents for the health in Detroit study

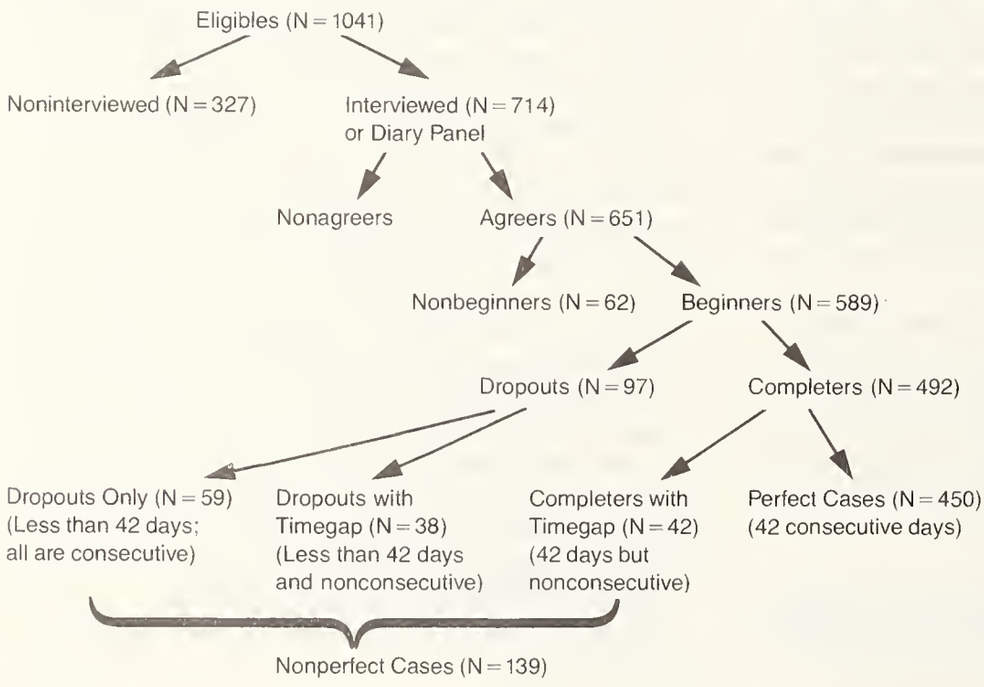


Table 1
Resonse rates by socio-demographic and health characteristics[a]

	Eligible (N)	Interviewed (N)	Interview rate Int'd / Eligible x 100	Rank	Diary agreement rate Agreed / Int'd x 100	Diary beginning rate Began / Int'd x 100	Diary completion rate Completed / Int'd x 100	Rank
Total sample	1041 (960)[b]	714	68.6% (74.4%)		91.2%	82.5%	68.9%	
Gender:								
Male	426	302	70.9		91.4	80.5	67.6	
Female	534	412	77.2		90.5	84.0	69.9	
Age:								
Under 30	251	213	84.9		93.4	82.6	→[c] 63.4	
30–64	532	409	76.9		93.2	85.6	74.3	
65+	177	92	52.0		→ 75.0	68.5	57.6	
Age-Gender:								
Under 30, male	116	98	84.5		87.8	→73.5	→ 53.1	
Under 30, female	135	115	85.2		98.3	90.4	→ 72.2	
30–64, male	241	168	69.7		95.2	85.7	76.8	
30–64, female	291	241	82.8		91.7	85.5	72.6	
65+, male	69	36	52.2		→ 83.3	75.0	63.9	
65+, female	108	56	51.9		→ 69.6	64.3	53.6	
Living arrangement:								
With other adults	808	594	73.5		92.4	84.0	70.4	
Sole adult	172	119	69.2		→ 84.0	75.6	62.2	
Relationship to household head:								
Head	546	390	71.4		90.8	80.3	66.4	
Spouse of head	307	243	79.2		93.0	87.7	76.6	
Other (dependent)	96	81	84.4		87.7	77.8	→58.0	
Presence of own children[d]:								
Own child(ren) present	375	306	81.6		95.8	87.6	74.8	
No own children present	577	407	70.5		87.5	78.9	64.6	
Gender/age/living arrangement:				Rank				Rank
Male, <30, sole adult	17	17	100.0	1[e]	88.2	→70.6	52.9	10[e]
w/other adult	99	81	81.8	4	87.7	74.1	→53.1	9
30–64, sole	28	18	64.3	7	94.4	83.3	66.7	6
w/adult	213	150	70.4	6	95.3	86.0	78.0	1
65+, sole	13	6	46.2	9	→ 83.3	83.3	66.7	6
w/adult	56	30	53.6	8	→ 83.3	73.3	63.3	7
Female, <30, sole adult	14	9	64.3	7	100.0	88.9	77.8	2
w/adult	121	106	87.6	2	98.1	90.6	→ 71.7	5
30–64, sole	44	33	75.0	5	90.9	81.8	72.8	3
w/adult	246	208	84.6	3	91.8	86.1	72.6	4
65+, sole	56	36	64.3	7	→ 66.7	63.9	50.0	11
w/adult	51	19	37.3	10	→ 78.9	68.4	63.2	8
Gender/age/relationship to head/ presence of children[f]:				Rank				Rank
Male, <30, head, none	47	41	87.2	4	85.4	75.6	→ 53.7	13
child	33	28	84.8	8	100.0	→78.6	→57.1	12
other, none	28	22	78.6	12	→77.3	→59.1	50.1	15
child	6	6	100.0	1	83.3	83.3	→50.0	16
30–64, head, none	107	67	62.6	18	91.0	→79.1	71.6	6
child	131	100	76.3	14	98.0	90.0	80.0	3
other, none	2	1	(50.0)[g]	—	(100.0)	(100.0)	(100.0)	—
65+, head, none	57	35	61.4	19	→82.9	74.3	62.9	11
child	3	0	(0.0)	—	—	—	—	—
other, none	3	1	(33.3)	—	(100.0)	(100.0)	(100.0)	—

(continues on next page)

Table 1 continued

	Eligible (N)	Interviewed (N)	Interview rate Int'd / Eligible x 100		Diary agreement rate Agreed / Int'd x 100	Diary beginning rate Began / Int'd x 100	Diary completion rate Completed / Int'd x 100	
Female, <30, head, none	15	11	73.3	15	90.9	81.8	→63.6	9
child	16	15	93.8	2	93.3	→73.3	→53.3	14
spouse, none	23	20	87.0	5	100.0	95.0	85.0	1
child	36	31	86.1	6	100.0	96.8	83.9	2
other, none	35	30	85.7	7	100.0	93.3	→63.3	10
child	5	4	80.0	11	(100.0)	(75.0)	(75.0)	—
30–64, head, none	50	39	78.0	13	92.3	84.6	66.7	8
child	22	18	81.8	9	88.9	83.3	66.7	8
spouse, none	92	74	80.4	10	91.9	85.1	73.0	5
child	110	100	90.9	3	93.0	88.0	77.0	4
other, none	6	4	66.7	16	(75.0)	(75.0)	(50.0)	—
child	3	2	(66.7)	—	(100.0)	(100.0)	(100.0)	—
65+, head, none	57	37	64.9	17	→67.6	62.2	48.6	17
spouse, none	34	16	47.1	20	→81.3	75.0	68.8	7
child	5	1	20.0	22	(100.0)	(100.0)	(100.0)	—
other, none	6	2	33.3	21	(0.0)	(0.0)	(0.0)	—
Household income[h]:								
<$10,000	150	102	68.8		85.3	75.5	53.9	
$10,000–24,999	423	292	69.0		94.2	84.9	74.0	
$25,000+	303	255	84.2		94.9	87.5	73.3	
Marital status:								
Married		467			92.9	85.7	74.3	
Not married		245			88.6	77.1	59.2	
Education:								
8 years or less		51			→66.7	→54.9	47.1	
Some high school		107			88.8	→77.6	64.5	
High school diploma		238			93.7	83.6	66.0	
Some college		199			95.0	87.9	74.4	
College degree or more		116			93.1	87.9	79.3	
Occupation:								
Professional/technical		102			95.1	87.3	78.4	
Administrator/manager		67			92.5	82.1	71.6	
Clerical		120			95.8	91.7	→72.5	
Sales		46			→80.4	→65.2	54.3	
Crafts		148			88.5	→77.0	62.8	
Laborers/operatives/service		77			90.9	83.1	→63.6	
Homemakers (nonemployed women)		138			89.1	82.6	71.7	
Other nonemployed (students, retired, disabled)		13			92.3	84.6	76.9	
Job limitations due to health:								
No job limitations		545			93.6	85.0	71.9	
Limited in kind or amount of work		100			→84.0	→71.0	56.0	
Unable to have a paid job		59			→83.1	79.7	64.4	
Other activity limitations due to health[i]:								
None		510			94.3	86.0	73.9	
1–2		109			→83.5	→71.6	56.9	
3–6		92			84.8	78.3	57.6	
Number of chronic conditions:								
0		66			90.9	81.8	75.8	
1–2		221			92.8	83.3	70.6	
3–5		229			91.7	84.7	69.4	
6–9		134			94.0	83.6	67.9	
10 or more		64			→78.1	70.3	56.3	

Table 1 continued

	Eligible (N)	Interviewed (N)	Interview rate Int'd / Eligible x 100	Diary agreement rate Agreed / Int'd x 100	Diary beginning rate Began / Int'd x 100	Diary completion rate Completed / Int'd x 100
General health status:						
Excellent		256		97.7	89.1	76.7
Good		328		90.9	82.7	69.5
Fair		97		→ 84.5	74.2	56.7
Poor		23		87.0	78.3	→ 52.2
How R usually feels physically (1 = terrible, 10 = wonderful)						
1–5		93		86.0	→ 74.2	→ 53.8
6–7		162		88.3	79.0	64.8
8		224		92.9	87.1	76.3
9–10		231		94.8	84.8	71.4
Health compared to other people same age:						
Better		323		90.7	81.4	70.9
About the same		335		91.6	84.5	69.9
Worse		46		91.3	→ 73.9	→ 45.7
How often sick compared to others same age:						
A lot less		241		92.9	83.4	70.1
Somewhat less		262		89.7	84.7	70.6
Same		151		91.4	82.8	72.2
Somewhat more		41		90.2	→ 73.2	→ 46.3
A lot more		16		87.5	→ 68.9	→ 50.0
How well R takes care of own health:						
Excellent		86		91.9	86.0	75.6
Good		411		92.2	85.4	71.5
Fair		180		89.4	78.9	63.9
Poor		32		93.8	→ 62.5	53.1
How satisfied with own health:						
Very satisfied		359		93.9	85.5	73.8
Somewhat satisfied		250		87.2	79.6	66.4
Somewhat dissatisfied		73		93.1	→ 80.8	→ 61.6
Very dissatisfied		30		90.0	76.7	→ 50.0
In past two weeks, number of days R did not feel well:						
0		355		92.4	84.8	75.2
1–2		167		91.6	83.2	70.1
3–13		140		91.4	82.1	→ 61.4
14		39		87.2	→ 69.2	→ 46.2
In past year, number of days R cut down activities due to illness or injury:						
0		212		88.2	81.6	72.6
1–4		185		95.1	87.0	72.4
5–14		176		94.9	84.1	72.2
15 or more		137		86.1	75.9	→ 54.7
Quality of life in past year: (1 = worst life possible, 10 = best life possible)						
1–5		111		90.1	→ 77.5	63.1
6–7		172		93.0	83.1	68.6
8		181		90.1	80.1	67.4
9–10		245		92.7	87.3	74.3

(continues on next page)

Table 1 continued

	Eligible (N)	Interviewed (N)	Interview rate Int'd / Eligible x 100	Diary agreement rate Agreed / Int'd x 100	Diary beginning rate Began / Int'd x 100	Diary completion rate Completed / Int'd x 100
Stressful life event(s) in past years[l]:						
Yes		507		92.7	84.4	68.8
No		203		87.7	77.8	69.0
How often worn out at the end of the day:						
Every day		47		89.4	→68.1	51.1
Often		116		94.8	87.1	→68.1
Sometimes		304		89.1	81.3	68.8
Rarely		203		94.6	86.2	74.4
Never		40		90.0	85.0	72.5

[a]NAs are excluded in all response rates.
[b]Information about household composition was obtained at 960 households, so the total number of Eligibles for most characteristics is close to that number.
[c]Arrows between columns indicate high attrition from one stage to the next: 15% or more from interview to agreement, 11% or more from agreement to beginining, and 18% or more from beginning to completion.
[d]"Own" means by birth or adoption.
[e]Ranks are shown, with rank 1 for the highest response rate.
[f]A total of 36 categories are possible; those not shown had zero Eligible cases.
[g]If the denominator <5, rates are shown in parentheses.
[h]Interviewed people were asked about household income in the initial interview. For non-interviewed people, interviewers judged the income level by observation and noted it on special non-interview form.
[i]Limitations in housework or chores, free-time activities, mobility, personal care, and physical activities (such as lifting heavy objects).
[j]"People sometimes experience changes in their lives—good things like a raise or a marriage, or bad things like the loss of someone close to them or not getting something they had expected. Has anything like that affected your life in the past year?" Probe if "No:" "Can you think of other kinds of changes, either good or bad, that you have experienced in the past year?" The percents shown include "Yes" to the first question or to the probe. Differentials are similar if we look at only the first question.

by young ones; sole adults; dependents (people who are not head of household or spouse of head), followed by household heads; nonparents; and people in low-income households. There is no gender difference in diary completion rates.

So far, the selective factors look very similar. Important differences surface when we look at life-cycle position as determined by sex, age, relationship to household head, and presence of own children.[3] Interview response rates are especially high for young men and young women in virtually all family situations. Middle-aged mothers (married or not) are also eager respondents. Average response rates appear for middle-aged married fathers and for some groups with uncommon family situations (such as nonmarried mothers living with other adults). Middle-aged men and women without children have lower response rates than those with children. Response rates are low for all elderly groups, especially for elderly women living with their husband or other adults. (Compared to them, elderly women living alone are twice as likely to be interviewed.) In sum, age has the strongest impact on interview response. Family situation and gender are not very important factors for young and elderly adults, but they strongly differentiate middle-aged people. Among people aged 30 to 64,those with a spouse or children or both are more likely to be interviewed; this is especially true for women.

Diary completion rates are different. The most successful completers are young and middle-aged wives and middle-aged men, especially when these groups have children. Average completion rates are achieved by elderly married women, elderly men, and women 30 to 64 who head their own households. The lowest rates appear for young men of all kinds, young nonmarried mothers, and elderly women who live alone. In sum, family situation and gender differentiate response rates for all age groups here. Stable family situations (being married, having children present, or living with other adults) generally enhance diary completion; this is equally true for men and women. The exception is young men, who tend to drop out regardless of their family situation. Age remains an important factor, but less here than for the interview response rates.

Some groups are very enthusiastic about the interview but not about the diary, or vice versa. Groups that are very happy to grant interviews but do not manage to complete the health diaries are young men, young women, and (compared to other elderly people) elderly women living alone. Groups that resist being interviewed but (if they get past that point) keep diaries very well are middle-aged married men and elderly married women.

2. *Among panel members, those who complete the study are socially advantaged and have better health than noncompleters.*

Demographic characteristics that enhance completion have been discussed. Now we consider socioeconomic and health characteristics, based on information in the initial interview.

Completers are more likely to be married, have high education, and have a white-collar job (see Table 1, right-hand column). It is intriguing that among lower white-collar groups, clericals complete the study much better than sales workers. Is the record-keeping task more compatible with their skills and lives? Good general health status, absence of chronic limitations, little recent illness or restricted activity, and a good life in the past

year bode well for completion.[4] People who are not "worn out" at the end of the day also manage to complete diaries better than those who are tired every day. This last result is fascinating; the "worn out" item produces one of the largest differentials in completion rates. Tired people have great trouble adding the diary task to their lives.

3. Among people who begin the diaries, those who keep them perfectly (for six continuous weeks) are very advantaged in social status and health compared to beginners with nonperfect diaries.

Here we compare Perfect Cases with Nonperfect Cases, those who dropped out of the study or had timegaps or both. Perfect Cases have a striking profile. They are more likely to be middle-aged (45–64), married, well educated, and of middle (but not high) income. Their health is better, whether measured in the interview or the diaries. They report better lives and fewer stressful life events in the past year. (See Table 2.) By contrast, Nonperfect Cases tend to be young adults, divorced/separated or never married, or the sole adult in a household. They have more young children present and have poorer health (in the past year and also during the diary period). Nonperfect Cases report more changes in their health habits and in their lives during the six-week diary period. And their recent life events concern marriage and family more often than the events reported by Perfect Cases.

It is worth noting that men and women Beginners are equally likely to produce perfect diaries (42 consecutive days).

Thus, Nonperfect Cases have more difficult and changeful life situations. It is difficult for them to do a task which requires constant record-keeping. Failure to complete the study reflects a poor match of the diary task to their lives, rather than hostility about or boredom with keeping diaries.

4. Refusal to keep diaries is often due to serious health problems, cognitive difficulties, and few social supports at home. By contrast, subjective and short-term problems often explain dropout after agreeing to keep diaries and beginning them.

There are three key attrition points in the panel: refusal to keep diaries, failure to begin them despite agreement, and discontinuation of diaries. The overall attrition rates are 8.6% loss from the initial interview to diary agreement, 8.7% more from agreement to beginning, and 13.6% more from beginning to completion. Reasons for attrition seem to differ at these stages. Some population groups are especially likely to refuse to keep DHRs, others are likely to agree but then do not begin them, and still others are likely to start DHRs but fail to complete the six weeks. Table 1 denotes the groups with especially large attrition at each stage.

Refusal to keep diaries is especially frequent for people with little education (8 years or less, 33.3% loss from

Table 2
Characteristics of perfect and nonperfect cases[a]

Note: All respondents in this table kept DRHs for one or more days. Perfect Cases filled them out for 42 consecutive days. Nonperfect Cases kept them for less than 42 days or had a timegap (skipped day) or both.

	Perfect Cases	Nonperfect Cases	X^{2b}
N	450	139	
Demographic Characteristics (household screener and initial interview)	(Percentage distribution)		
Gender			
Male	41%	41%	NS
Female	59	59	
Age			
18–34	39%	52%	**
35–44	20	22	
45–54	16	7	
55–64	14	9	
65+	11	9	
Marital status			
Married	71%	58%	*
Widowed	6	6	
Divorced/separated	10	15	
Never married	13	20	
No. of adults in household			
1 (respondent only)	19%	27%	NS
2	62	58	
3	12	10	
4 or more	7	4	
No. of R's own children ages 6–12 in household			
0	76%	76%	NS
1	14	9	
2 or more	10	15	
No. of R's own children ages <6 in household			
0	80%	81%	NS
1	15	13	
2 or more	5	7	
Socioeconomic Characteristics (initial interview)			
Education			
Less than high school	18%	23%	*
High school diploma	32	40	
Some college	31	25	
College degree or more	19	12	
Household income			
Less than $10,000	15%	27%	**
$10,000–19,999	24	26	
$20,000–24,999	20	8	
$25,000–34,999	21	21	
$35,000 or more	20	18	

(continues on next page)

Table 2 continued

	Perfect Cases	Nonperfect Cases	X^2[b]
Current work status			
Full-time employed (40 + hours/week)	47%	48%	NS
Part-time employed (1–39 hours/wk)	17	22	
Not employed	36	30	
Health (initial interview)			
Job limitations due to health			
No job limitations	80%	78%	NS
Limited in kind or amount of work	12	14	
Unable to have a paid job	8	8	
Other activity limitations due to health[c]			
None	77%	67%	*
Low	12	18	
Medium	7	7	
High	5	8	
Number of chronic conditions			
0	10%	6%	NS
1–2	31	32	
3–5	33	32	
6–9	18	22	
10 or more	8	7	
General health status			
Excellent	40%	34%	NS
Good	46	45	
Fair	11	16	
Poor	3	4	
In past year, number of days R cut down activities due to illness or injury			
0	31%	24%	NS
1–2	17	14	
3–7	26	25	
8–14	10	12	
15–28	5	9	
29 or more	10	17	
Quality of life in past year (1 = worst life possible, 10 = best life possible)			
1–4	6%	5%	NS
5–6	17	20	
7–8	40	41	
9	18	22	
10	19	13	
Stressful life event(s) in past year[d]			
Yes	71%	80%	**
No	29	20	
Health (daily health records)	(Averages over diary period)[e]		
Physical feeling (1 = terrible, 10 = wonderful)[f]	7.7	7.5	
Symptomatic days	15.6	16.1	
Days of restricted activity	3.4	5.1	
Days with medical or dental care	0.8	1.0	
Number of medications "to treat symptoms bothering you today"	13.5	14.3	
Number of medications "to prevent illness or to become more healthy in general"	25.5	13.6	
Days with unusual event	16.0	15.2	
Recent Changes (termination interview)	(Percentage distribution)		
Any change in eating or sleeping habits in past 6 weeks (diary period)			
Yes	15%	21%	NS (.07)
No	85	79	
Any stressful life events in past 6 weeks[d]			
Yes	48%	50%	NS
No	52	50	

[a] NAs are excluded in percentage distributions.
[b] * is $P < .05$, ** is $P < .01$; *** is $P < .001$. NS means not significant ($P \geq .05$).
[c] An index based on limitations in housework and chores, free-time activities, mobility, personal care, and physical activity (such as lifting heavy objects).
[d] See footnote j of Table 1.
[e] For Nonperfect Cases with less than 42 diary days, information was inflated to a 42-day period. Significance tests are not shown.
[f] This was rated each day by respondents. Each respondent's average for the diary period was computed; those values were averaged to obtain the figures shown here.

interview to agreement) and for elderly women living alone (33.3%). Other socio-demographic factors related to refusal are elderly age in general (25.0%), being a dependent (16.0%), being the sole adult in a household (16.0%), having no own children present (12.5%), low income (14.7%), and sales occupations (19.6%). People with permanent disabilities are reluctant to agree (15%–17% loss). Recent illness and restricted activity also increase refusal but not nearly so strongly as long-term limitations do. In sum, people of elderly ages, with low education, with sparse family situations, and with permanent disabilities tend to refuse to keep health diaries.

Despite agreeing to keep them, young men are unlikely to start the diaries (14.5% loss between agreement and beginning). This is especially true for young married fathers (21.4%), young men living alone (17.6%), and young men living as dependents in a household (18.2%).

Other strong factors are recent illness (not feeling well for the entire past two weeks, 18.0%) and poor self-rated health (12%–21% on various items). To a lesser extent, low education (11.8%), nonmarriage (11.5%), and sales or crafts occupations (15.2%, 11.5%) also increase not starting. But the most startling factor is the "worn out" item: People who say they are worn out every day are least likely to honor their agreement (loss of 21.3%). In sum, groups that do not pick up their pens after agreeing to do so tend to be constantly fatigued people, young men, and those with recent health problems or a negative view of their health.

Some people try to do the diary job but quit during the six weeks. Poor health (especially poor self-rated health status) inhibits continuation; the loss between beginning and completion is often 20% or more. Young men (20.4%) also have trouble, as do women under 64 who head their own households (16.6%–20.0%). Dependent status (19.8%), low income (21.6%), and clerical or laborer occupations (19.2%, 19.5%) are linked to discontinuation; to a lesser extent, nonmarriage and modest education are too. (The late dropout for clericals is a

surprise.) In sum, subjective feelings of poor health status make it hard for people to continue diary-keeping. Young men and female household heads also give up quite readily.

To compare the three stages we see first that elderly age, little education, and physical disabilities discourage agreement. Many of these respondents would have objective problems completing the diary form. Second, people who are perpetually tired and young men tend to say they will keep diaries but then do not. We do not know whether their agreement was genuine or not. Third, poor health (recent or subjective) and young age discourage continuation of diary keeping. Illness disrupts daily routines, including any daily record keeping. Routine may also help explain dropout rates of young adults if their lives are less scheduled than those of older people. Why subjective health status becomes so important is a mystery; we wonder if people who think their health is poor become depressed when keeping health diaries.

One final point about these stages: Differentials are smaller at the beginning (agreement) and at the end (completion). In other words, panel selectivity increases. It is not surprising that we end up with a rather special group of people. The truly surprising feature is the high level of agreement at the beginning stage to keep diaries by a general population.

Summary and recommendations. The results above identify groups that drop out of a health diary study at different points by refusal to be interviewed, refusal to keep health diaries, failure to start them despite agreement, and discontinuation of diary keeping. Selectivity varies at all of these points, but several results stand out.

First, age is a powerful filter in health diary studies. In Detroit, the enthusiasm of young and elderly adults waxes and wanes sharply at various stages. Middle-aged people end up being the best diary keepers, the most likely to complete the study and to do the diary task perfectly. Second, gender is not a strong selection factor in health diary studies. It is far less important than age, education, family situation, and health in the Detroit data. Despite popular beliefs that men will not keep diaries, middle-aged and older men are fine panel members. Only.young men drop out at high rates; we do not know precisely why. Third, people with poor health have trouble at all stages of the diary study. Those with long-term disability tend to refuse the diary task, realizing that they are unable to do it.[5] Later attrition is due to recent illness and subjectively poor health. Short-term disruptions in health and depression about their health make people abandon the diaries. Fourth, stability in life enhances completion of the health diary. Some indicators of stability are moderate or high socioeconomic status, being married and having children (especially both), and having few stressful life events.

The response rates and differentials suggest that interviewed people are fundamentally cooperative and very willing to give further information about themselves. They will probably agree to do many kinds of panel tasks. Refusal and dropout from a diary task occur when people cannot match routine record-keeping with their daily routines.

The implications for staff activities are that staff should focus on getting Agreers to begin the diaries and on helping Beginners continue them. The 9% attrition between agreement and beginning in the Detroit study occurred in a mere one or two days after the interview; it should be almost entirely preventable.[6] Staff should also help diary keepers who are young, who have poor health, low education, or disrupted lives (nonroutine daily lives or recent stressful events). The assistance might be longer training at the outset, discussion of possible problems and negotiated changes in the diary task, more contacts by staff during the diary period, etc. These activities can be overtly designed into the study, rather than ad hoc.

Increasing the diary agreement rate and interview response rate are tougher problems. People who refuse to keep diaries are often truly unable to do the task. We do not know much about the people who refuse to be interviewed. At initial contacts, interviewers may need to focus on convincing people of the study's worth and the importance of their personal information. In sum, interviewers' tactics need to change across stages of the panel study. At the beginning, convincing is critical, later, assisting is the important ingredient.

One qualification is important: The *lower* the interview reponse rate, the *more* selective the diary panel is likely to be, especially for "cooperativeness." Getting reluctant people to be interviewed will entail costs in the panel period, since these respondents will probably need more attention and assistance. Similarly, reducing early attrition in the panel will tax staff later, since the "saved" respondents may keep records poorly. Survey researchers usually think only of the costs of high attrition; but for panel studies, low attrition can have very high costs too.

Task performance

Diary keeping is a clerical task. People vary greatly in clerical skills and enthusiasm—in their ability to keep tidy records, their devotion to detail, and their pleasure in doing routine jobs. How well do diary keepers in a general population do the task?

In the Detroit study, people who agreed to keep DHRs were given careful instructions for filling them out and mailing them. With the interviewers' help, they filled out a DHR for the day of their interview and also looked at the preprinted DHR which had numerous symptoms and health actions filled in. Booklets for Weeks 1 and 2, mailing envelopes, and the two practice forms were placed in an attractive folder and left with respondents. On the folder flaps, lists of possible symptoms and medications were printed, to serve as "cues" to

diary keepers. Interviewers asked respondents to fill in the DHR each day sometime after 8 p.m. (or at the "end of the day" for people with unusual schedules) and to fill it out themselves. Respondents were told that if they ever missed a day, they should fill it in as soon as possible the next day.

Two sources are used to assess diary keepers' performances. First, in the termination interview, respondents discussed how well they followed the instructions. Second, office tallies were made of skipped items and skipped days. These sources are very different; the first involves individual self-reports, the second, aggregated objective indicators.[7]

The main questions are: How much do diary keepers deviate from the task of filling in records *each* day at the *end* of the day by *themselves?* Do some groups have more trouble in this task than others? How frequent are skipped items, and what aspects of the diary's format are related to skips? How common are skipped days, and what population groups are most responsible for them?

1. Most respondents filled out DHRs at the end of the day and did so without anyone's assistance. Many skipped one or more days but usually filled them in later. Women followed the task specifications more closely than men did.

The termination interview was conducted with all Beginners, whether they completed the diary period or not. It included questions about the diary task—how many minutes the DHR usually required; what time of day he or she usually filled it out; if a day was ever missed and what he/she did about it; if anyone ever filled in a DHR for him or her and if so, why.

The majority (68%) of diary keepers spent 1–5 minutes each day on the DHR; only 11% said it normally took more than 10 minutes (Table 3). Virtually all respondents said they filled out DHRs in the evening: 38% at 7–9 p.m., 38% at 10 p.m.–midnight, and 16% at an unspecified evening hour. Few (6%) diary keepers ever relied on someone else's help to fill in DHRs. Typically, the proxy respondent was the spouse. In almost all instances, the respondent told the helper what to write down. The most common reasons for proxy response were language/reading problems and illness. Two-thirds (65%) of the diary keepers said they skipped one or more days sometime during the six weeks, but virtually all of them filled in the information later (13% next morning, 68% "next day," 12% "later"). Only 2% of the people who skipped days said they never filled them in.

Men were less conscientious record keepers than women. They spent less time filling out the DHR each day and were more likely to do it early in the day. More men had someone fill out the DHR for them, usually their wife. Women with proxy responses tended to get help from their children or a telephone interviewer, not their husband. More men skipped days; although they

Table 3
How well diary keepers follow instructions, by gender
(Based on self-reports in the termination interview)

	Total	Men	Women
N	577	235	342
	(Percentage distribution)		
Minutes per day to fill out DHR			
1–4	33%	36%	31%
5	35	33	37
6–9	9	7	11
10	10	11	10
More than 10	11	12	10
Nonspecific answer	1	1	1
Time of day R usually filled it out			
Morning hour	3%	5%	2%
Afternoon hour	4	7	2
7–9 pm	38	40	35
10 pm–midnight	38	33	42
Nonspecific "evening"	16	15	18
If R ever missed a day			
Yes	65%	69%	62%
No	35	31	38
If Yes: (N)	(373)	(163)	(210)
How many days R missed during diary period			
1	19%	14%	23%
2	23	20	25
3–5	44	50	38
6–9	7	5	9
10 or more	8	11	5
What R did about missed days			
Filled in DHR next morning	13%	7%	18%
Filled in DHR "next day"	68	72	66
Filled in DHR "later"	12	14	11
Filled in DHR more than 1 day later	4	4	4
Left it blank	2	2	1
Other answers	0	1	0
If someone else ever filled in DHR for R			
Yes	6%	12%	2%
No	94	88	98
If Yes: (N)	(35)	(27)	(8)
Who filled it in for R			
Spouse	62%	73%	25%
Other person	38	27	75
If R was asked the questions			
Proxy asked R the DHR questions	91%	88%	100%
Proxy filled in DHR alone without asking R the questions	9	12	0
How many days R had someone else fill in DHR			
1–2	15%	12%	25%
3–5	9	12	0
6–9	9	12	0
10–41	18	24	0
All 42	48	40	75

usually filled them in on a later day, they waited longer to do so than women did.

Differentials by age, employment status, marital status, presence and ages of children, socioeconomic items (education, income, occupation), and health were also examined. None of them are as large as the gender differentials just discussed. Age ranks second to gender, with elderly people reporting more minutes per day to fill out DHRs, more skipped days, and more proxy response than other age groups. People with children

Table 4
Item nonresponse in daily health records

	An item series	Main, filter, or interior item[a]	Place on page	Closed or open[b]	Mean percent NA	Percent of RS with any NA
Time of day DHR was filled out		M	Middle	O	5.1%	47%
Q1. How felt physically today?		M	Bottom	C(10)	2.9	36
Q2. Any symptoms today?		M	Top	C	0.0	0
Q2c. Symptoms 1-5: Names[c]		F		O	0.0	0
Q2c. Symptom 1: Seriousness	X	I		C(3)	2.4	16
2: Seriousness		I		C(3)	6.2	19
3: Seriousness		I		C(3)	7.1	16
4: Seriousness		I		C(3)	10.9	26
5: Seriousness	X	I		C(3)	3.6[d]	15
Q3. Any restricted activity for symptoms?		F	Top	C	1.6	8
Q3a. Stayed in bed	X	I		C	6.3[e]	11
Q3b. Cut down household chores		I		C	6.4	11
Q3c. Missed work		I		C	6.4	11
Q3d. Missed schooling		I		C	6.4	11
Q3e. Cut down other activities	X	I		C	6.4	11
Q4. Any medical or dental help for symptoms?		F	Bottom	C	2.0	11
Q4a. Made an appointment	X	I		C	24.4[e]	32
Q4b. Telephoned for advice		I		C	24.4	32
Q4c. Visited doctor or dentist		I		C	24.4	32
Q4d. Admitted to hospital		I		C	24.4	32
Q4e. Other medical/dental help	X	I		C	24.5	32
Q5. Talk to relatives or friends about symptoms?		F	Top	C	1.6	10
Q5a. Spouse	X	I		C	3.2	12
Q5b. Other household member		I		C	3.2	12
Q5c. Other relative		I		C	3.3	13
Q5d. Neighbor, coworker, friend		I		C	3.3	13
Q5e. Other person	X	I		C	3.3	13
Q6. Any preventive medical or dental care?		M	Bottom	C	0.5	11
Q6a. Purpose		F		O	13.5	21
Q7. Any pills, medicines, or treatments?		M	Top	C	0.3	6
Q7a. Drug 1: Name[c]	X	F		O	2.4	8
2: Name		F		O	5.8	10
3: Name		F		O	12.4	15
4: Name		F		O	19.8	22
5: Name	X	F		O	31.8	35
Drug 1: Purpose	X	I		C(4)	2.3	9
2: Purpose		I		C(4)	6.0	11
3: Purpose		I		C(4)	12.6	16
4: Purpose		I		C(4)	21.1	25
5: Purpose	X	I		C(4)	31.7	35
Q8. How spirits were today		M	Top	C(10)	0.7	14
Q9. Any unusual events today?		M	Middle	C	2.2	32
Q9a. What happened?		F		O	8.8	34

[a]Main—Everyone must answer this. Filter—People who say "Yes" to a main item must answer this. Interior—People who say "Yes" to a filter item must answer this.
[b]For closed items, the number of response categories is shown in parentheses.
[c]A maximum of five symptoms and five drugs were coded.
[d]Reports of five symptoms on a day were uncommon (N = 20 days), but details for them are well-reported.
[e]Editing rules caused the uniform percents in this series.

present and with high socioeconomic status skipped more days, but they filled in virtually all of them later. Poor health increased record-keeping time and proxy assistance. One note: People in clerical occupations were "average" in the task performance, with one exception. They report the least assistance from other people (1%).

2. *The less visible an item is on the diary form, the more often it is skipped.*

Rates of item nonresponse are low for the Detroit study, because of extensive staff efforts to minimize it. Following written instructions for editing booklets, staff resolved many problems in the DHRs including certain N.A.s. When information was ambiguous or very incomplete, respondents were usually telephoned and the information was filled in retrospectively. Thus,the rates reported here refer to residual N.A.s, but not the original levels. We shall concentrate on comparisons across items, rather than on absolute levels of item nonresponse.[8]

We begin with a brief scan of the levels. Item nonresponse is usually 2%–6% for DHR items (Table 4). The range is .5%–24%. To get these values, the percent N.A. was computed for each respondent, then the average of those individual percents was computed.

Four aspects of diary format are related to item nonresponse. First, in a series of parallel questions, those near the *end* of the list are more likely to be skipped than those near the beginning. Second, *filter* items (answered only if a prior item is answered positively) are skipped more than main items (which everyone must always answer). Items nested even farther (after "Yes" to a filter item) have additional risk of nonresponse. Third, items at the *bottom* of a page are skipped more often than those at the top. Fourth, *open-response* items (where the respondent must write down information) have more nonresponse than close-response items (where the respondent checks a box). And among the closed items, those with more check boxes tend to have more nonresponse.

A few exceptions to these results appear in Table 4; they are due to editing rules and are discussed in table footnotes. One anomaly merits comment: Q2 and Q6-Q9 are main items; Q3-Q5 are filter items. Ideally, nonsymptomatic people ("No" to Q2) should flip two pages to Q6 and avoid Q3-Q5 entirely. But printing constraints forced Q6 to appear on the same pages as Q3-Q5. Surprising, it has very little N.A. This indicates that an awkward skip pattern is feasible and does not always confuse diary keepers, but it should be avoided whenever possible.

Socio-demographic differentials in item nonresponse are not analyzed because editing and retrospective interviewing have undoubtedly masked them.

3. *Skipped days are most common at the beginning of a booklet and are most frequent for young and elderly adults.*

Although many respondents say they skipped days, the final number of blank days is small in the dataset because respondents filled them in later or were interviewed about them. The remaining skipped days are called "time gaps." Timegap cases are people with one or more skipped days in their final set of DHRs. (They are included in the Nonperfect Cases we discussed earlier; now we look at them separately.) Tables for the results reported here are in a Technical Report (Camburn, 1980, Tables 6,8,11,12).

Altogether 23,526 days of diary data were submitted in the Detroit study. Interspersed among them are 449 blank days. The ratio of blank days to filled days is 1:52 or about 1.9%. These skipped days were produced by 14% (N = 89) of the diary keepers.(Beginners). Timegap respondents tended to be young adults (under 30 years old) and elderly adults (79 or older). Consistent with that, never-married and widowed people were more likely to have time gaps compared to married people. Less-educated and full-time employed people were also more likely to be Timegap cases. There was no gender difference in Timegap status, suggesting that although more men report skipped days, their days were ultimately filled in as well as women's.

Several indicators about the timing and volume of time gaps were examined for Timegap cases: number of timegap days, number of timegap episodes, number of days into the diary period when the first timegap occurred, and percent of all days that are timegaps. Timegap people who were young or elderly, never-married or widowed, fulltime employed, or better-educated produced more timegap days, had more episodes, had their first skipped day sooner, and had a larger percent of skipped days in their diaries than others.

Note that these differentials are *among* Timegap cases, yet they are very similar to the factors that distinguish Timegap from non-Timegap cases. (Education is an exception.) This means that the groups with greatest risk of skipping *any* days at all also generate the *most* skipped days. From a more positive perspective, middle-aged, married people tend to be the most consistent diary keepers and rarely skip days.

Of the 449 timegap days, the majority occurred in Week 3 (N = 102) and Week 5 (N = 104). Gaps tended to peak on Day 1 of a week. Going from one booklet to the next obviously posed trouble for some respondents. Although low motivation is a possible reason, the study procedures also created two problems. First, each booklet had a computer-generated label identifying the dates for that week. The dates were based on the initial interview date and assumed that respondents began keeping DHRs the day after the interview, as requested. But some started on the interview day itself; when they received the Week 3 booklet, they waited a day before beginning it. This problem sometimes "echoed" to the next mailing and appeared for Week 5. Second, a mailing error delayed a large shipment of Weeks 3-4 booklets, so some respondents had no forms when their Week 3 began.

The timing of skipped days varied for age and gender groups. Young men (aged 18-34) produced a large share of the early time gaps (in Weeks 1-4). Women of all ages

were largely responsible for time gaps in the final weeks. Recall that young men also tended to drop out of the study. Apparently, women whose motivations waned often decided to stay in the study but became less thorough near the end of the diary period; young men tended to quit entirely.

Timegap cases who eventually dropped out had their first skipped day much earlier than Timegap cases who completed the study (22 days into the diary period vs. 16 days). This is an important diagnostic result, and it leads to a recommendation in the next section.

Summary and recommendations. Clerical expertise for diary keeping is remarkably high in a general population. People found it easy to fill out DHRs in the evening and without assistance. They did have difficulty doing this for 42 days in a row. But when they missed a day, most filled in the information the next day. Study staff worked energetically to detect and remedy skipped items (N.A.) and skipped days; whenever possible, respondents were telephoned for retrospective information. The result is a dataset of 23,256 days with little nonresponse for items or days.

Some population groups had more trouble than others in doing the diary task as specified. By their own reports, men followed the instructions less closely than women did. Several aspects of the diary format and field procedures also influenced the frequency of skipped items and days.

The basic aim of diary studies is to secure prospective data, i.e., information recorded at the time of events or very soon after. If record keeping is either delayed or done too early (for example, before the evening), events can be omitted. Men tend to err on both sides more than women do. We do not know exactly why men are less conscientious, so recommendations for improving their performance are hard to devise. It is important to note that panel attrition was similar for men and women in Detroit. Thus, men (except young ones) tended to stay in the study as well as women, but they were more casual in record keeping.

The analyses do suggest some technical aspects that help diary keepers. First, the diary format should be easy and pleasurable for respondents. The ideal format would have short series of parallel items, few filter items, few questions per page, and mostly closed-response items. When complexity is necessary (such as a chart or a skip pattern), it should be visually and cognitively clear. Boldface type will help to highlight main items. Good-quality paper, sturdy binding, and professional printing enhance the diary's attractiveness. These aesthetic aspects should not be ignored by researchers. Second, for a short panel period, all diary forms should be given out at one time. For long studies, additional forms should be delivered in person. This avoids the vagaries of mail service and also provides an opportunity for interviewers to question, motivate, and assist respondents. Third, diary forms should be arranged contiguously in a binder and dated by the interviewer and the respondent together. Weekly or monthly segments should be clearly identified and easy to remove, so they can be mailed for editing.

Finally, the time-gap results suggest that staff should monitor time gaps very closely. Diary keepers who skip a single day are "high risks" for dropout. As soon as a missed day appears, a respondent should be contacted and given special assistance and encouragement; this may be critical to keeping them in the study and will also improve their task performance.

Conditioning effects

Panel participation may change respondents' attitudes and behaviors, especially those being investigated. This is a conditioning effect called sensitization. Health researchers believe that sensitization increases symptom awareness, medical care, and self-care. Another conditioning effect, called fatigue, is often suggested. Panel participation may become tiresome, especially for a continuous task like a health diary. Researchers believe that fatigue reduces reports of symptoms and health actions over time. Note that sensitization influences the frequency of events, whereas fatique influences the frequency of reporting events; this is an important distinction.[9]

Typically, researchers study conditioning effects by pooling respondent reports for different periods (e.g., weeks or months), computing overall rates for each period, then asking if the rates show trends over time. If rates rise (either at the beginning of the study or persistently over time), sensitization is assumed to be at work. If rates decline, fatigue is assumed. These are plausible interpretations, but too simplistic; the situation is actually more complex. First, trends are the net result of all conditioning effects. In particular, sensitization and fatigue probably pull in opposite directions, so they compensate each other in the observed trends. Second, other time-related factors besides conditioning effects can cause trends or sudden jumps in rates. Examples are season, holidays, mass media reports, political events, and selective panel attrition.

Because trends contain confounded factors, we cannot easily answer these questions for health diary studies: Do diary keepers become more sensitive to their physical discomforts and begin to report more symptoms—things they would not have felt before the study? Does the persistent focus on health prompt them to have more preventive or curative health care, which they would have otherwise delayed or not done? Do they tire of daily record keeping and become more cursory in their reports, omitting some symptoms and behaviors they would have written down early in the study? Which population groups experience the most sensitization and fatigue?

We shall explore these questions, with due caution,

for the Health In Detroit Study. The study offers two kinds of information on conditioning effects. First, in the termination interview, respondents were asked if their health perceptions and behaviors changed during the study, and if they tired of filling out the DHRs. Second, weekly rates for the sample were computed and graphed, and regression lines through them were estimated. Note that the first analysis involves subjective, but explicit, reports about conditioning effects; the second involves subjective, but indirect, indicators. The first analysis uses individual-level data; the second uses aggregated data. It will be interesting to see if results from these very different sources agree.

1. Many respondents say their awareness of symptoms increased during the diary period, but few say their health behaviors changed. Few report fatigue in keeping the records.

Over half (57%) of the diary keepers say they noticed their health problems more during the study than before.[10] Three-fourths (73%) of the sensitized people say this persisted for the entire six-week period; only 6% report it just at the beginning. Only 6% of the diary keepers say they handled health problems differently during the study than before. The most common change was in use of pills and medicines.

Only 19% of the diary keepers report they tired of filling out the daily records. But when asked for details, most say they simply became bored; few say they became careless or incomplete in their reports. Fatigue tended to increase as the diary period lengthened.

These results refer to all Beginners (N=574) who had a termination interview. This includes respondents who completed the study and also those who dropped out. Not surprisingly, Dropouts report fatigue much more often than Completers (32% vs. 16%).[11]

Women report more symptom sensitization than men (54% vs. 45%), but there is no sex difference in health-care changes.[12] Men report fatigue slightly more often than women (20% vs. 18%). (We note this difference only because it is consistent with task performance results.) With regard to age, younger diary keepers report more symptom sensitization and fatigue than older ones.

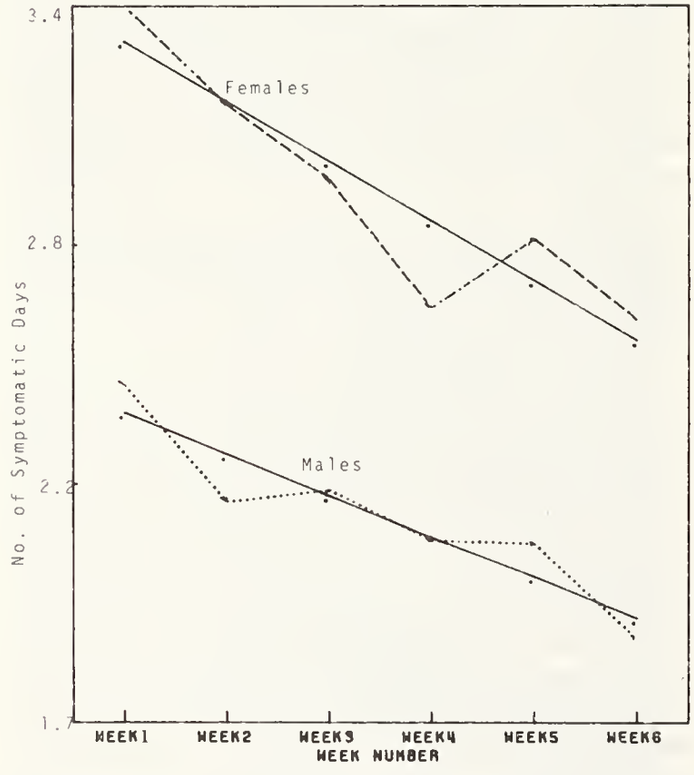
2. Over a six-week period, symptom rates decline and several health behavior rates increase. Panel participation spurred men to take care of their health more than before, through medical care and restricted activity; it spurred more drug use among women.

Figures 2–5 show weekly rates and trends for some DHR variables.[13]

Over the diary period, people report that they feel better physically, have fewer symptomatic days and fewer symptoms. Symptoms that are "not very serious" decline very sharply. ("Very serious" symptoms are essentially constant. Rates for "somewhat serious" symptoms increase. We will discuss this more below.) People report

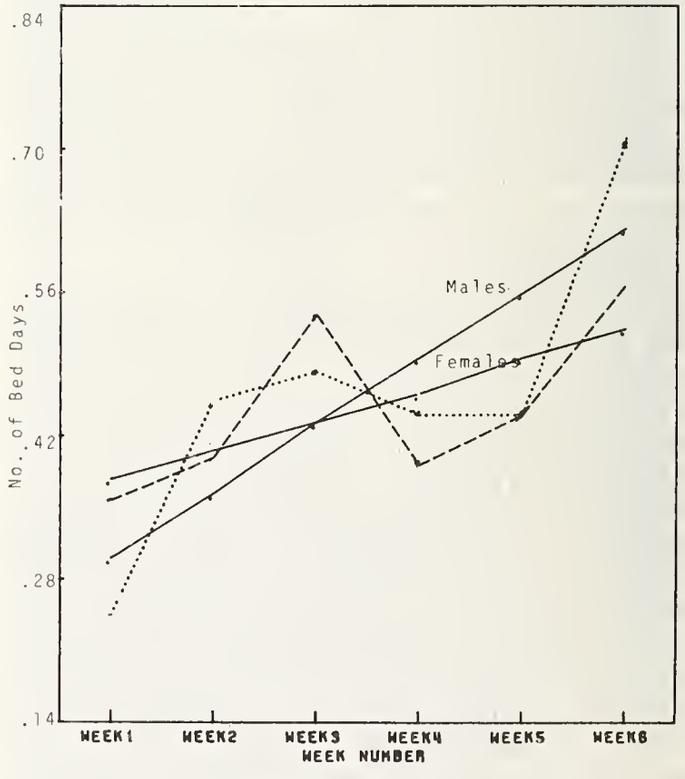
Figure 2
Number of symptomatic days per week, by gender

The six weekly means are connected by dashed lines (females) or dotted lines (males). Unbroken lines show regressions for the weekly means.



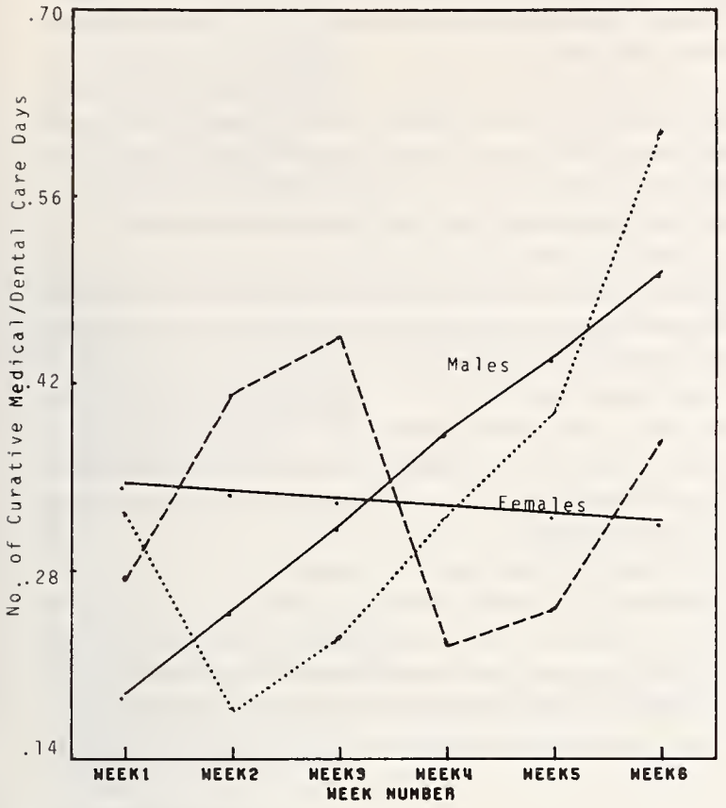
(This figure refers to all diary keepers.)

Figure 3
Number of bed days per week, by gender



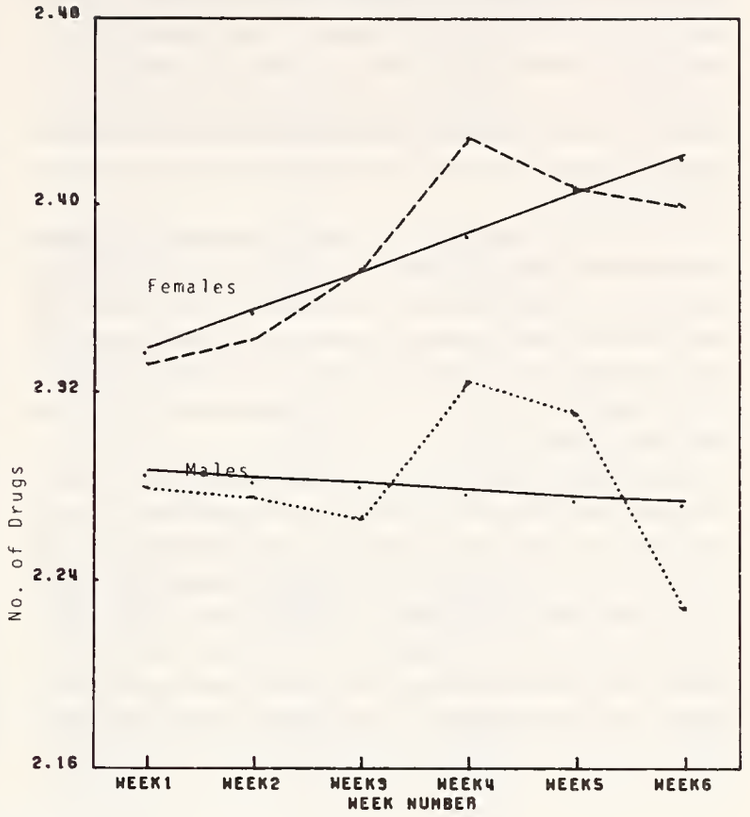
(This figure refers to symptomatic people only.)

Figure 4
Number of curative medical/dental care days per week, by gender



(This figure refers to symptomatic people only.)

Figure 5
Number of drugs per week, by gender



(This figure refers to all diary keepers.)

more positive moods and fewer unusual daily events over time too. The trends for all of these items are quite steady; i.e., the weekly rates move monotonically over time.

Health actions show less consistent trends. Three items increase: bed days, work-loss days, and curative medical/dental care days. Other actions show no obvious trends: lay conversation about symptoms, preventive medical/dental care, drug use, and other types of restricted activity (reduced household chores, days absent from schooling, "other" reduced activities). Compared to the symptom variables, health action variables show much more fluctuation across the weeks; even those with trends are less steady.

The symptom trends appear for both genders, but there are intriguing differences for health actions. Increased bed disability and work loss are much more pronounced for men than for women. In fact, all forms of restricted activity increase for men. And the increase in curative care is due solely to men; it does not appear for women. Men's preventive behavior also increases, for both medical visits and drug use (such as vitamins). By contrast, women's overall restricted activity rates decline over time. (Bed disability and work loss do increase, but all other types of restricted activity decrease.) Curative and preventive medical care decrease. But women do use more curative drugs as the diary period continues. These results suggest that panel participation spurred men to take more preventive and curative care of their health. It influenced women's health behavior less, but probably spurred more drug use.

Does the decline in symptom rates reflect fatigue? Probably not. In the Detroit diary, people marked if they had symptoms or not, and they rated their physical wellbeing on a 1-10 scale. The trends are therefore based on overt statements about feeling fine or not-so-fine. This contrasts with most prior health diary studies, which asked for entries only when symptoms occurred. Blanks were therefore ambiguous; they could mean "no symptoms" or failure to make entries on a symptomatic day (fatigue).

But we have no reason to expect the incidence of symptoms to decline over six weeks, and no reason to expect that people feel better and better. So we must look a little more for conditioning effects. First, respondents may want to please study staff by reporting good health. This "social desirability" effect might increase over time. Second, we cannot totally eliminate fatigue; recall that rates of "not very serious" symptoms dropped sharply, whereas rates for more serious problems were constant or increased. If respondents decided to avoid filling out DHR details by omitting minor symptoms, and this increased over time, it could explain those trends. But those respondents also had to be clever enough to report better physical well being (Q1) and mood (Q8)! Third, sensitization to *good* health might occur. Over time, re-

spondents may have paid less attention to their minor discomforts, actually perceiving them less readily and focusing more on major problems. This change in symptom perception could underlie both the declines in symptom rates and the increases in health action rates. In sum, there is no ready explanation for the symptom trends; we suspect a conditioning effect but cannot pinpoint what it is.

Does the increase in some health behaviors indicate sensitization? Possibly. Women and men apparently are sensitized in different ways. Women increase their drug use, whereas men restrict activities and seek medical care more often. For both sexes, changes in health behavior are modest, no nearly so pronounced as changes in symptom reports.

Other time-related factors can be excluded as explanations. First, the study period was September-December. This is short and mostly in one season (fall). If anything, symptom incidence would increase as winter approaches and rates would rise; but they do not. Seasonal effects are very unlikely in the data. Second, the analysis of panel attrition does show that less-healthy people tend to drop out. This could depress symptom rates over time. To check this, separate trend analyses were performed for Perfect and Nonperfect Cases. The trends stated above appear for *both* groups (Verbrugge, 1980b). Thus, selective sample attrition cannot explain the results here.

Let us blend the Detroit results with other health diary studies. What trends tend to appear in health variables? Table 5 summarizes trends from four studies (Mooney, 1962; Sudman, Wilson, and Ferber, 1974; Sudman and Lannom, 1980; and Health In Detroit).[14]

Looking at month-long segments, downward trends appear for most health indicators from month to month. The declines are greatest from Month 1 to 2, then are more gradual. Looking at changes in a 4–6 week period, illness rates drop 15%–30%. Although specific indicators vary across the studies, all show declines in this range. The declines are largest for minor illnesses. Rates of restricted activity, bed disability, and medical care drop in all studies, except for Detroit which generally shows increases for men and decreases for women. (Is it important that women often kept the diaries for household members in the other studies?) Trends for symptoms are sharper than trends for health actions.

Previous researchers have usually attributed the declines to fatigue. Occasionally, sensitization is mentioned

Table 5
Trends in rates of symptoms and health actions for four health diary studies

	Rate for first segment	*Percentage change*		
	Week 1 (per 100 persons)	From Week 1 to Wk. 2	From Week 1 to Wk. 3	From Week 1 to Wk. 4
California (Mooney, 1962)[a]				
Total illness	12.5	+04%	−01%	−18%
With restricted activity or medical attention	6.4	+02	−03	−19
Without RA or MA	6.0	+07	+02	−18
Acute illness	10.3	+02	−05	−19
With RA or MA	5.3	−04	−08	−21
Without RA or MA	4.8	+08	−00	−17
Chronic illness	2.2	+14	+23	−14
With RA or MA	1.0	+30	+40	−00
Without RA or MA	1.2	−08	+08	−25
	Month 1 (per 100 persons)	From Month 1 to Mo. 2	From Month 1 to Mo. 3	From Month 1 to Mo. 4
Total illness	58	−16%	−34	−43
With restricted activity or medical attention	27	−07	−30	−48
Without RA or MA	31	−23	−39	−42
Acute illness	44	−11	−36	−39
With RA or MA	20	−00	−30	−40
Without RA or MA	24	−21	−42	−37
Chronic illness	14	−29	−29	−57
With RA or MA	7	−29	−29	−57
Without RA or MA	7	−29	−29	−57
Total illness days without restricted activity	178	−25	−32	−44
Acute illness days without RA	113	−19	−34	−49
Chronic illness days without RA	65	−35	−29	−35

Table 5 continued

	Rate for first segment	*Percentage change*		
Total restricted activity days	102	+06	+13	−09
RA days for acute illness	68	+10	−10	−38
RA days for chronic illness	34	−00	+59	+50
Total bed disability days	40	−20	−07	−30
Bed days for acute illness	24	−08	−37	−58
Bed days for chronic illness	16	−31	+31	+06

	Month 1 (per household)	From Month 1 to Mo. 2	From Month 1 to Mo. 3
Marshfield and Chicago (Sudman, Wilson, and Ferber, 1974)[b]			
Days stayed home or felt ill	2.3	−18%	−14%
Acute doctor or hospital visits	1.3	−04	−05
Routine doctor visits	0.7	−27	−34
Medicine purchased or medical bills paid	2.5	−08	−10

	Month 1 (per household)	From Month 1 to Mo. 2	From Month 1 to Mo. 3
Illinois (Sudman and Lannom, 1980)[c]			
Days felt ill but performed usual tasks	8.7	−30%	−31%
Days stayed home from work or school, or unable to do usual activities	4.6	−23	−24
Days spent in bed	3.4	−21	−27
Visits to health professionals	1.7	−16	−15
Visits for hosital care	0.3	−20	−10
Outpatient visits	0.3	−31	−23
Inpatient visits	*	+50	+75
No. of prescription medical supplies obtained	1.1	+01	+03
No. of nonprescription medical supplies obtained	0.7	−03	−10
No. of payments to health care providers	1.0	−12	−08

	Week 1 (per person)	From Week 1 to Wk. 6
Detroit (Health In Detroit Study)		
How R feels physically (1 = terrible, 10 = wonderful)	7.5	+04%
No. of symptomatic days	3.0	−23
No. of symptoms	4.2	−23
No. of very serious symptoms	0.2	−03
No. of somewhat serious symptoms	1.0	+06
No. of not very serious symptoms	2.9	−30
Restricted activity days	0.7	−25
Bed days	0.2	+27
Days cut down household chores or errands	0.6	−15
Work-loss days	0.1	+57
School-loss days	*	+20
Days cut down other planned activities	0.4	−16
No. of restricted activities	1.2	−03
Days with curative medical/dental care	0.1	+16
No. of curative medical/dental care	0.2	+16
Days with preventive medical/dental care	0.3	−18
Days of lay conversation about symptoms	1.4	−22
Days of drug use	3.8	−02
Number of drugs	8.7	+01
Curative drugs	2.3	−05
Maintenance drugs (for asymptomatic conditions)	1.8	+45
Preventive drugs	4.0	−01
Drugs for other reasons	0.6	−06
How R's spirits were today (1 = terrible, 10 = wonderful)	7.4[d]	+04
No. of eventful days	2.9	−15
No. of unusual/special events	3.8	−13

*Less than .1

[a]Based on data in Tables 11 and 34 of Mooney (1962). "With restricted activity or medical attention" means one or both of these actions. "Without restricted activity or medical attention" means neither of them.

[b]Based on data in Table 3.3 of Sudman, Wilson, and Ferber (1974).

[c]Based on data in Tables 3.3, 3.10, 3.14, 3.28-3.31 of Sudman and Lannom (1980).

[d]Average for week.

as the reason for high medical care rates early in the study. Occasionally, seasonality has also been cited for trends.

Summary and recommendations. According to self-reports, symptom sensitization is common (especially for women) and is constant over the diary period. Being in the study does not cause changes in health behavior. Fatigue is rare, and it does not affect how carefully people keep records; it tends to increase over the diary period.

At first, the self-reports seem to contradict the trend analyses, which show *declines* in symptom rates and *increases* in some health behaviors. Typically, we would say that respondents tired of reporting symptoms and that they were sensitized to take better care of their health.

We offer a plausible resolution: Trends in *symptoms* are due mostly to sensitization. Diary keepers did become more aware of health (they thought about it more), but their perceptions of minor problems decreased over the six weeks. More and more, they concentrated on major symptoms and felt minor ones less.[15] Fatigue in keeping records is a secondary factor, and social desirability (to report good health) is tertiary. Trends in *health actions* are more modest, and this is consistent with the self-reports. The trends are probably caused by sensitization, but it spurred different activities for men and women. This resolution is certainly a tentative one, since we did not ask respondents much about conditioning effects and because we cannot separate conditioning effects in the aggregate analysis.

Our recommendations are aimed at theoretical and methodological research, rather than study design. We simply do not know enough about conditioning effects to design studies that minimize them. There are difficult questions to solve first: What kinds of conditioning effects are possible, and how do they affect respondents' real-world behavior and record-keeping behavior? Can conditioning effects be reliably measured by asking respondents about them? How can conditioning effects be measured separately in trend analyses? We suspect that conditioning effects are sizable in panel studies, especially those with continuous records or lifetime membership. But there are few theories, models, or empirical results to demonstrate that or to guide us in study design.

Conclusions

Health diary studies require substantial effort from respondents and make them think about their health constantly during the diary period. Researchers and funding agencies have worried that the respondent burden for these studies is "too great," and that conditioning effects are "large." The Health In Detroit results offer a more sanguine view of participation and performance in health diary studies and a more cautious view of sensitization and fatigue.

Respondent burden (sample attrition and task performance). The Detroit study shows that agreement to keep diaries can be very high in a general population, and that dropout rates can be modest. But there is clear selectivity in the types of people who agree to keep diaries, who actually begin them, and (especially) who submit a complete set. People who are physically or cognitively unable to perform the requested task refuse to keep diaries. By contrast, people who are capable of keeping diaries but have trouble fitting the routine task into their daily lives agree to do the task but then fail to begin, or they start and then quit.

Most Detroit diary keepers had no difficulty filling out DHRs during the evening and without anyone else's help. But doing this every day for six weeks was not easy; a majority of people skipped some days, then filled them in on a later day. Some aspects of diary format and field procedures also caused problems, especially mail delays and machine-dating of booklets. On the positive side, the diary format (Appendix 1) and mail-back procedures were very acceptable to diary keepers. Overall, the study design worked very well indeed. From telephone contacts with diary keepers, we know they enjoyed being in the study and liked the diary form.

In sum, some respondent characteristics and some design characteristics decreased the panel size and the quality of diary data. We have made two kinds of recommendations. Some focus on the task—how to design the diary and field procedures to make the job easier for all respondents. In particular, we recommend that the diary format be attractive and easy to follow and that all diaries be left with the respondent for short studies. Other recommendations focus on communications with respondents—especially those who agree to keep diaries and those who begin them. At these stages, staff can assume that respondents are willing to do the task and that problems will concern fitting it into their daily lives. When problems arise, communications must be prompt; delays will cause incomplete records and dropout. But staff communications can be "proactive" as well as "reactive." The Detroit results identify population groups that are especially likely to leave a health diary study after agreement and after beginning. Special communications aimed at those groups can be actually designed into the study. One general thought about recontacts during the diary period (whether aimed at all respondents or specific groups): Active staff involvement may impress respondents. Seeing staff working energetically may boost a respondent's own enthusiasm for diary keeping. (Consider a design with minimal recontacts; respondents may end up thinking "If they are not doing anything, why should I?")

Rather than presume that diary studies pose high burdens, it may be more useful to presume that people are generally willing to provide data about themselves (see also, Bradburn, 1979). Attrition is often explained by people's inability to do the respondent task or fit it into their daily lives. Certainly, there are some very un-

willing and hostile respondents; they will be difficult to recruit and retain in any kind of panel study, regardless of the topic or respondent task. The point here is to plan staff communications that recruit and retain people "at the margin"—those who will stay or leave, depending on the encouragement and assistance they receive from study staff.

We actually know little about the "social psychology" of panel participation. What factors really make people refuse to join or fail to continue? One way to find out is to ask them. We recommend that panel studies formally interview respondents (including dropouts) at various stages of the study, to learn about their problems and also to listen to their suggestions for improvements.

Conditioning effects. It is very likely that health diary studies change people's thresholds for symptom perception, their evaluations of general health status, and their propensities to take health actions. The Detroit data show some trends in symptoms and actions over time, and many respondents reported being more aware of their health problems. It is also possible that fatigue occurs, so repondents become less careful in keeping records; this is especially likely when daily records are required. The Detroit data suggest that fatigue is a smaller problem than sensitization.

Our interpretations of the Detroit results are cautious. It is frustrating that conditioning effects are inherently tangled in aggregate trends, and we do not really know if respondents can talk about sensitization and fatigue accurately.

Attentive theoretical and methodological work needs to be done on conditioning effects. We need to identify the possible reactions to panel studies, hypothesize how they influence data, and determine how to measure them one by one. In particular, we need to be able to distinguish changes in attitudes and behaviors from changes in record-keeping. Until such work is done, analyses will continue to yield ambiguous results.

The results and recommendations in this paper are most pertinent for health diary studies. What about panel studies with other topics and other strategies? Specific results for sample attrition, task performance, and conditioning effects may differ greatly for them, compared to health diaries. In particular, respondent health may have a weaker selection effect for nonhealth studies and nondiary studies. (Ill people may be more willing to discuss other subjects and more capable of providing interview data.) People with nonroutine daily lives may find it easier to participate in nondiary studies. And conditioning effects may be smaller when data are collected less frequently. Nevertheless, the recommendations about staff communications with panel members, diary format (a self-administered questionnaire), and field procedures probably apply to many panel studies. And the need for research on conditioning effects is universal; knowledge about sensitization and fatigue will help all panel studies.

Footnotes

[1] A few Completers (N = 18) provided more than 42 days of data.

[2] These are actually residual timegaps. Many respondents skipped days but filled them in the next day. Also, when editors found skipped days which were recent, respondents were telephoned for retrospective information, and the timegap was eliminated. If the skipped day was not recent, a respondent was asked to do some extra diary days at the end of their six weeks, to achieve 42 total days of data; this did not eliminate timegap status for them. Finally, some timegaps were not detected until data files were constructed, and timegap status was established at that point.

[3] The household screening form does not indicate marital status of Eligibles. When marital status can be reasonably assumed from relationship to household head, we use the terms "married" and "nonmarried" here.

[4] One anomaly: People who experienced stressful life events in the past year are just as likely to complete the diaries as those who did not have such events. We expected more attrition for the first group. In fact, when we compare Perfect and Nonperfect Cases in the next section, the expected difference appears. So, people with disrupted lives are very willing to join the panel but they have some trouble producing flawless diaries.

[5] Serious health problems are also an important reason for noninterview. In the Detroit study, 9% of the noninterviews are due to the designated respondent's poor mental or physical health.

[6] These people were not interviewed later, so we do not know their reasons for not starting the diaries. We now recognize this is a critical point for attrition—both for staff activities and for followup information.

[7] The study has one other source of information about record-keeping problems. There is a file for each respondent with office records of editing problems and recontacts during the diary period. These can be coded and analyzed to learn about task performance and also staff activities. Such analysis was not budgeted and has not been done.

[8] Analysis of original levels would be feasible. Editors and telephone interviewers used colored pencils for their work; each group used a different color. Respondents used lead pencils or pens. The source of each entry is therefore clearly recorded on the DHRs. (The editor and interviewer entries cover up original N.A.s.) Methodological analyses of item nonresponse were not budgeted, so the source information was not coded.

[9] Other conditioning effects have been proposed; for example, increased knowledge about the topic, crystallization of attitudes, improvement in record-keeping skills.

[10] The sensitization questions are: "Did participating in this study make you notice your health problems more than before?" and "While participating in this study, did you handle your health problems differently than you usually would? For example, were you more likely to visit a doctor, cut down your activities, or take medications?" The fatigue question is: "Sometimes people get a little tired of filling out the daily records and are not as careful or complete as usual. Did anything like this happen to you?"

[11] Dropouts also report more changes in health behavior during the diary period (11% vs. 6%). But they report less symptom sensitization (39% vs. 53%).

[12] Women seem to have more changeful lives in many respects. In the Detroit study, they report more changes in eating and sleeping habits during the diary period, more stressful life events in that time and also in the past year, more unusual daily events during the diary period, and more variability in how they feel from day to day. Greater symptom sensitization is consistent with this profile.

[13] Graphs for the other DHR variables are available on request. Regression statistics for the total sample and the subgroups discussed are also available. Analyses of daily rates yield similar conclusions to the analyses of weekly rates reported here.

[14] No other health diary studies have published rates for segments of their diary period.

[15] If these perception changes did occur, there is a perplexing issue: Was awareness of minor symptoms boosted at the beginning of the study—an early sensitization effect that disappeared later? Or did the typical awareness of minor symptoms change over the study—a sensitization that developed gradually across the six weeks?

DAILY HEALTH RECORD

DAY # 1

An example Health Record is included in your folder

DAY OF THE WEEK _____

DATE _____

TIME _____

1 **How did you feel physically today?** *(Put an "X" in the box which best describes how you felt)*

1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10

TERRIBLE — WONDERFUL

2 **Did you have any symptoms or discomforts today?** *(Fill out the chart below from left to right for each symptom or set of symptoms)*

☐ 1 Yes ➡ ☐ 5 No symptoms or discomforts at all ➡ *Go to Question 6*

SYMPTOM CHART

NUMBERS BELOW CAN BE USED IN Qs 3 AND 4	2a SYMPTOMS & DISCOMFORTS Write symptoms of the same health problem in one box	2b. CAUSE Illness, name the illness Injury, name part of body hurt and type of injury	Not illness or injury, write what you think caused the symptoms.	2c SERIOUSNESS In your opinion, how serious was this condition or set of symptoms today?
1				☐ 1 Very Serious ☐ 3 Somewhat Serious ☐ 5 Not Very Serious
2				☐ 1. Very Serious ☐ 3 Somewhat Serious ☐ 5 Not Very Serious
3				☐ 1 Very Serious ☐ 3 Somewhat Serious ☐ 5. Not Very Serious
4				☐ 1 Very Serious ☐ 3 Somewhat Serious ☐ 5 Not Very Serious
5				☐ 1 Very Serious ☐ 3. Somewhat Serious ☐ 5 Not Very Serious

3 **Because of today's symptoms, did you cut down on the things you usually do?**

☐ 1. Yes ☐ 5 No ➡ *Go to Question 4*

Answer 3a through 3e.			For which symptom, illness, or injury? *(Write in number from symptom chart, question 2)*
3a. Did you stay in bed?	☐ No	☐ Yes ➡	
3b. Did you cut down on household chores or errands?	☐ No	☐ Yes ➡	
3c. Did you miss work?	☐ No	☐ Yes ➡	
3d. Did you miss school?	☐ No	☐ Yes ➡	
3e. Did you cut down on other activities you planned to do (sports, clubs, church, etc.)?	☐ No	☐ Yes ➡	

4 **Did you seek medical or dental help about today's symptoms?**

☐ 1. Yes ☐ 5. No ➡ *Go to Question 5*

Answer 4a through 4e.			For which symptom, illness, or injury? *(Write in number from symptom chart, question 2)*
4a. Did you make an appointment with a doctor or dentist?	☐ No	☐ Yes ➡	
4b. Did you telephone a clinic, doctor's office or dentist's office and get advice?	☐ No	☐ Yes ➡	
4c. Did you visit a clinic, doctor's or dentist's office or an emergency room?	☐ No	☐ Yes ➡	
4d. Did you get admitted to a hospital?	☐ No	☐ Yes ➡	
4e. Was there anything else you did? (IF YES. What was that?)	☐ No	☐ Yes ➡	

5 **Did you talk with anyone else today — such as relatives, neighbors, friends — about your symptoms?**

☐ 1 Yes ☐ 5 No ➡ *Go to Question 6*

5a. Who did you talk with? *(Check all boxes that apply.)*

- ☐ A. Your husband/wife
- ☐ B Other member of your household
- ☐ C. Relative not living with you
- ☐ D Neighbor, coworker, friend (who is not a relative or household member)
- ☐ E. Other person (druggist, minister, priest, etc.) Who did you speak with? _____

6 **Did you visit or telephone a clinic, doctor or dentist's office, or hospital today for any reason besides symptoms you might have had today?**

☐ 1 Yes ☐ 5 No ➡ *Go to Question 7*

6a. For what reasons?

7 **Did you take any pills, medicine, or treatments for your health today — to treat a symptom, prevent illness, or to become more healthy in general?** *(Fill out the chart below from left to right)*

☐ 1. Yes ➡ ☐ 5 No pills, medicine, or treatment taken at all ➡ *Go to Question 8*

	7a PILLS, MEDICINE, TREATMENTS If pills or medicine, write the brand name from the label and the type of drug. Use one box for each pill, medicine or treatment.	7b REASONS FOR TAKING PILLS, MEDICINE, TREATMENTS *(Check all boxes that apply.)* A. To treat symptoms bothering you today B. For other health problem not bothering you today C. To prevent illness or to become more healthy in general D. Other reasons	7c. SYMPTOM OR CONDITION What was the symptom, health problem, or other reason for taking pills, medicine, or treatment?
1		☐ A ☐ B ☐ C ☐ D Go to 7c ➡	
2		☐ A ☐ B ☐ C ☐ D Go to 7c ➡	
3		☐ A ☐ B ☐ C ☐ D Go to 7c ➡	
4		☐ A ☐ B ☐ C ☐ D Go to 7c ➡	
5		☐ A ☐ B ☐ C ☐ D Go to 7c ➡	

8 **How were your spirits today?** *(Put an "X" in the box which best describes how you felt today)*

1 — 2 — 3 — 4 — 5 — 6 — 7 — 8 — 9 — 10

TERRIBLE WONDERFUL

9 **Did anything happen — for better or worse — to make today different than usual?**

☐ 1 Yes ☐ 5. No ➡ *Go to Question 10*

9a. **What happened?** *(Check all boxes that apply.)*

- ☐ A. Birthday, holiday, or special social event
- ☐ B Trip or vacation
- ☐ C. Emergency
- ☐ D. Trouble with family or friends
- ☐ E. Something extra nice with family or friends
- ☐ F. A lot of extra work
- ☐ G Guests
- ☐ H Other: What happened? _____

_____

10 **Any other comments:** _____

_____

_____

_____

_____

_____

REMINDER:

IF you had any symptoms today:
Be sure you answered Questions 1, 2, 3, 4, 5, 6, 7, 8, and 9.

IF you had no symptoms today:
Be sure you answered Questions 1, 2, 6, 7, 8, and 9

Discussion: Internal consistency analysis *and* Health diaries

Judith Kasper, National Center for Health Services Research

The paper by Anderson, Bush, and Berry examines differences in response to a self-administered and interviewer-administered questionnaire about dysfunction or disability in three areas—mobility, physical activity, and social activity. The researchers intended to use the self-administered form for data collection but, in using the interviewer-administered form as a check on the self-administered form, noticed discrepancies between the two. They conclude that in a majority of cases the self-administered form underestimates levels of dysfunction in the population compared to the other form, and they conclude the different wording of the two forms is a major reason. The self-administered form is described as being cast primarily in what the authors call "capacity language"—for example, "Is (the person) limited in his or her work activities or in walking." This is the type of question used by the HIS to assess limitations. The interviewer-administered form is described as containing "performance language"—for example "Does (the person) walk with a cane, have trouble stooping or lifting." Another way of stating this difference, mentioned by Eleanor Singer yesterday, is that one questionnaire (the interviewer-administered form) gives respondents an empirical referent from which to respond while the other does not. Questions which provide such a referent appear to elicit more responses than those that do not (see Frank Andrews's paper), and I think this is the significant difference between "capacity" and "performance" language.

It is clear from this paper that the differences in responses to the two types of questionnaires are considerable, and since the authors are interested in the instrument that identifies the most dysfunction, their preference for the interviewer-administered form is sensible. However, I would like to discuss two issues this paper raised for me. The first has to do with the way a "best code" was arrived at between the two questionnaires. From the examples given in the paper, it seemed that the process always led to choosing the response that identified the most dysfunction, which in turn was most likely to be the interviewer-administered form given the more specific probes this form contained. I'm not really familiar with the "internal consistency" method used to arrive at a best code, but it appears to involve a very detailed case-by-case analysis of responses using taped transcripts of interviews. The method obviously was carefully implemented but the outcome seems predetermined given that reports of more dysfunction are always accepted as more accurate (the few exceptions have to do with cases where the interviewer made some kind of error). An example of what bothers me about this process can be taken from the paper. A 12-year-old child was reported on the self-administered form as not being restricted. In response to a probe on the interviewer-administered form about restrictions in hobbies and sports, the mother said the child could not participate in the sport of tumbling because of a minimal curvature of the spine for which she wore a lift in her shoe. The second response is the "best code" in terms of reporting more disabilities but it is not necessarily the "best code" in terms of which form more accurately assesses the presence of a meaningful disability. One might argue that this child is not "disabled" in any meaningful sense and from that perspective the self-administered form might be more accurate.

The second issue is a related one, concerning what is meant by disability. These two forms were administered to respondents one after the other with the order alternating so that some respondents received the interviewer-administered form first, and others the self-administered form first. The order of the forms did not alter the way in which people responded—more reports of disability for the interviewer-administered form,fewer for the other. I find it interesting that in some cases respondents who first responded positively to the interviewer-administered form with its detailed probes about dysfunction, could then indicate no dysfunction on the self-administered form immediately afterwards. This suggests how complex the concepts of dysfunction or disability are. It also suggests that the self-administered form may be telling us something about the way people think about their level of functioning in the absence of cues as to what others consider disability or reduced levels of functioning. In fact the authors make reference to this issue in their conclusions: "Even with the institution of the above procedures, some respondents still underreport dysfunction. This appears to involve those who have some long-term (disfiguring and/or handicapping) function limitation. which is usually obvious to an observer...." Rather than regard this as a measurement problem and a weakness in questionnaire design or wording, I think the differences in responses to the questionnaires provide some interesting data about differences in perceptions of disability.

Turning to the second paper, we have a self-administered form of the most extreme kind, a health diary which respondents were asked to complete every day for six weeks. In addition, Lois Verbrugge's study was a panel-design study, and my comments have to do with some aspects of her study which relate to my experiences with the National Medical Care Expenditure Survey

(NMCES) and may be characteristic of panel designs in general.

The first issue discussed was panel attrition. The author describes those who dropped out of her study as people with "difficult life situations." Their demographic characteristics were similar to those of persons who dropped out of NMCES. They were more likely to be in poorer health, to have lower income and to be nonwhite than were persons who completed all five rounds of interviews. Yesterday, Al Marcus presented data from a panel design study consisting of an initial personal interview and several telephone follow-ups, and he also pointed out that drop-outs were more likely to have lower income and to be "people who began the survey rating their health as fair or poor." So it seems that despite these very different methods of data collection, certain types of persons, those Lois characterized as having "difficult life situations," may be more likely than others to drop out of panel design surveys. This occurs despite the different levels of respondent burdens in the Verbrugge, Marcus, and NMCES studies.

Panel attrition becomes a problem in panel design surveys, not only because of concern about the representativeness of the remaining population, but because you lose large components of data. In panel designs, researchers tend to ask for more information because more interviews offer more opportunities to ask for information. However, if income data, for example, is not collected until the later interviews, as was the case for NMCES, then information on major data items such as income will be missing for those people who drop out. Lois Verbrugge also observed that the greatest drop in participation in her study occurred early. This experience appears similar to that described in Marcus' paper and to the NMCES experience, where the sharpest decline in participation occurred after the first interview. Lois Verbrugge's comment about gaining a better understanding of the sociology of panel participation is very apt because both in her study and in NMCES, the loyalty of participants once you get past the initial drop off is very impressive and seems to hold up regardless of respondent burden.

The second major issue discussed had to do with the impact of respondents' performance on data quality. Again, the panel design exacerbates problems of nonresponse. Most researchers are used to dealing with some item nonresponse for cross-sectional surveys,but the problem of missing major components of data discussed earlier also occurs. A third type of problem, unique to panel designs, is internal consistency. When people are asked similar questions about insurance coverage for instance across time, their answers may vary, and it becomes difficult when data of this type does not mesh with income or employment data to determine whether these are plausible changes over time or response errors. A more complex level of editing may be required for data from panel design surveys due to these missing data and consistency problems.

Lastly, Lois Verbrugge raised the "conditioning effect" problem noting the decline in reports of symptom counts and symptomatic days over the course of her six-week panel. Although she is not inclined to attribute these declines to respondent fatigue, given other data she has available, this again appears to be a pattern which deserves more study in panel surveys. For example, although this has not been examined in much detail, some decline in reporting of physician visits in the last quarter of NMCES also has been observed.

In summary, I was struck by some of the similarities across these panel design surveys despite very different data collection modes—the health diary, personal, and telephone interviews. The particular methodological advantages and problems of panel design compared to cross-sectional surveys seem deserving of further attention.

A field approach for obtaining physiological measures in surveys of general populations: Response rates, reliability, and costs*

Sonja M. McKinlay, Pawtucket Heart Health Program and Associate Professor of Community Health, Brown University

Diane M. Kipp, Pawtucket Heart Health Program

Patricia Johnson, Pawtucket Heart Health Program

K. Downey, Pawtucket Heart Health Program

R. A. Carleton, Pawtucket Heart Health Program and Professor of Medicine, Brown University

Introduction

With the rapidly increasing cost of large scale epidemiological field work, a likely reduction in funding support for such research, and increasing difficulty in obtaining satisfactory survey response rates from populations already over-burdened by other constituencies (for example, market researchers and pollsters), the need for reliable, cost efficient data collection strategies is paramount. This is particularly relevant in large scale investigations when the collection of complex physiologic data is involved. Ideally, such information as expected response rates and relative costs of different approaches should be available at the planning stage of research in order to inform decision making about optimal data collection strategies. Surprisingly little attention has been devoted to such issues in the epidemiological literature.

Using experience and data from the first of three health surveys associated with the Pawtucket Heart Health Program (PHHP), a community intervention program to prevent cardiovascular disease involving two cities in Rhode Island and Massachusetts, this paper attempts to fill some major gaps in our present methodological knowledge with regard to large health surveys. The first section briefly reviews some reports on the validity and reliability of physiologic measures—particularly blood pressure—using different protocols and in various research settings. The second section describes in detail the field protocol currently employed on the PHHP Health Survey. The third section presents results from ongoing data quality control activities on the validity and reliability of physiologic measures (height, weight, blood pressure, and blood sampling) obtained from a field or household survey approach. The fourth section presents response rates and the major costs of obtaining data in this manner. The fifth section presents an application of these findings and illustrates the potential use of these results in planning future health surveys that involve the collection of physiologic data. The final section is a brief summary and conclusion.

Background

Large scale health surveys in the U.S. requiring physiological measurements have tended to follow the general model used by the Health Examination Survey (HES) and the subsequent Health and Nutrition Examination Surveys (HANES I and II), sponsored by the U.S. National Center for Health Statistics (USNCHS, 1973; Ibid, 1977; Ibid, 1981). Following this approach, interviewers make in-person contact with sampled households and, possibly following a brief interview, arrange appointments for eligible members of the household to attend a centrally located clinic where all physiological measures and procedures are conducted, using fixed, standard equipment. The HES and HANES surveys used a specially designed mobile trailer unit while other more localized surveys and experiments, primarily in the cardiovascular field, have used permanent clinics or temporary centers in church halls, schools, or other available space. See, for example, the Framingham Heart Study (Gordon et al., 1959), the North Karelia Project (Puska et al., 1979), the Western Electric and Chicago Gas Company Studies (Paul et al., 1963). This is the approach currently being used by two other large cardiovascular community demonstrations projects—the Stanford Heart Disease Prevention Program (SHDPP) (Farquhar et al., 1977) and the Minnesota Heart Health Program (MHHP) (Personal Communication, MHHP).

Response rates in such surveys have varied depending on the type of survey, the population, and sampling strategies, and have ranged from 87%, 96%, and 90% for the three cycles of HES, through 70%–75% for HANES I and II, to 62%–74% on SHDPP (Three Cities Project) (Farquhar et al., 1977).

One outstanding and recent exception to this general trend in the survey measurement of physiological variables is the 1978 Canadian Health Survey (Lalonde, 1974), in which all eligible members of sampled households were measured, in the home, for blood pressure, height, weight, and fitness, as well as having blood samples drawn for extensive analysis. This field protocol was completed by a nurse/interviewer team, visiting the household by appointment following an initial house-

* Acknowledgments are due to A. Brescia, A. Howe, J. Correia, M. Victor, and C. Jones, who assisted in the research and analyses for this manuscript, and to the members of the survey team, whose unstinting efforts produced the successful experience reported here. This work is supported by NIH Grant HL23629.

hold interview. The response rate averaged 75% for this national survey (Health and Welfare Canada/Statistics Canada, 1981).

Most of the literature discussing measurement reliability has focused on blood pressure (the only measurement which appears to have been routinely performed at home) and has tended to emphasize methods for obtaining a stable value that accurately screens for hypertension. Personal factors which have been identified as affecting blood pressure are age, sex, race, body mass, history of hypertension, and physical and emotional stressors (Evans and Rose, 1971; Hypertension Detection and Follow-up Project (HDFP), 1978; Heller et al., 1978; Gordon et al., 1976; Gutgesell et al., 1981; Stine et al., 1975). Other external affectors include seasonal changes (Evans and Rose, 1971; Kirkendall et al., 1980; HDFP, 1978), altitude (Kirkendall et al., 1980), and general environmental factors (Heller et al., 1978; Stine et al., 1975; Hawthorne and Smalls, 1980).

Controllable components of variability which have been identified include subject position, the arm selected, and the level of the arm in relation to the heart (Viol et al., 1979; Thulein et al., 1975). Type of machine and cuff size have also been shown to affect measurement (Evans and Rose, 1971; Kirkendall et al., 1980; O'Brien and O'Malley, 1979; Thulein et al., 1975), with the anaeroid sphygmomanometer being particularly inaccurate (Bowman, 1981). Digit preference of observers has been well documented (Kirkendall et al., 1980; HDFP, 1978) and has even been detected with use of the random zero device (Wright and Dore, 1970). Interobserver variability has been noted in many reports, particularly in the recording of the fourth and fifth phases of diastolic pressures (O'Brien and O'Malley, 1979; Rose, 1965; Wilcox, 1961; Eilertsen and Hummerfelt, 1968). The importance of thorough, standardized training is emphasized in most of these studies.

In terms of which of multiple readings one should use, recommendations vary according to the purpose of the measurement. Gordon and coworkers (1976) found that a single casual measurement can be highly predictive of future cardiovascular disease, although it is inadequate to characterize an individual's blood pressure. After a careful analysis of the variance of multiple measurement, taken on the same and different visits to a subject at home by different technicians, Rosner and Polk (1979) concluded that three visits, with two measurements per visit (six readings in all) were required for an accurate assessment of hypertension status. Multiple readings for accurate assessment have also been recommended by other researchers (Soucheh et al., 1979; Shepard, 1981).

Reports on reliability of height and weight measures have been based on fixed, nonportable equipment, generally as recommended by the Center for Disease Control (1980) including the balance beam scale and a rigid, sliding right-angle (usually of wood). Using such equipment, test-retest reliability on adults (Moffat et al., 1980) and children (Sady et al., 1981) was shown to be excellent, with correlation coefficients exceeding 0.95 and 0.90 respectively. No equivalent testing using portable equipment appears to have been carried out.

Field collection of blood samples has been a routine procedure for hospital-based phlebotomy services, with blood samples being immediately returned to the laboratory for centrifuging and storage or analysis. The inclusion of field blood collection in a survey protocol, however, does not appear to have been attempted, at least in the U.S., except as a back-up procedure for missing or hard-to-obtain samples. The Canadian Health Survey did include a field blood sample collection procedure performed by nurses only, which was particularly successful and efficient given logistical complexities of transport in remote areas of the country (Health and Welfare Canada, 1981). Whole blood samples were collected on 90%–95% of adults under 65 years, with lower rates for older adults, and, particularly, for children.[1]

The PHHP health survey protocol

As is clear from the brief review presented above, the most frequently used approach to the collection of physiological data in surveys of general populations is to invite respondents to a central clinic or equivalent setting following a (usually brief) household contact. In the ensuing discussion, this will be termed the "traditional" approach. This must be distinguished from a "field" strategy which involves completing all physiologic measurements and procedures in the household (or other location convenient to the respondent such as the worksite). The approach used in the PHHP Health Survey (1981-82), the results of which are reported in this paper, is an adaptation of the field protocol used successfully in the Canadian Health Survey and is described fully in the *PHHP Health Survey Manual* (Kipp et al., 1981). This section summarizes the strategy.

The PHHP Health Survey was designed with two components—the field and survey center protocols. The field protocol was administered to all selected respondents in a home visit, lasting on average 35 minutes and including an interview of 15 to 20 minutes, height and weight measurement, two blood-pressure determinations, and the drawing of a 30 ml. blood sample. In order to include a submaximal fitness test using an ergometer (Siconolfi et al., 1982), a one-third subsample was asked to complete an additional protocol at a centrally located survey center. This protocol included, primarily, a screening interview to determine eligibility for the test as well as the test itself (for the eligible respondents). This report focuses on the field protocol, specifically the reliability and cost efficiency of the physiological measurements and procedures included in it. The visit to the survey center by the subsample, usually within two weeks of completing the field protocol, provided an excellent opportunity for building into the survey design repeat measurements of height, weight,

and blood pressure (using standard clinic equipment as employed in the "traditional" approach) with which to compare the field measurements.

The survey is conducted in the cities of Pawtucket, Rhode Island, and New Bedford, Massachusetts (about 45 miles apart). To avoid confounding of technician and city differences, the survey staff is divided into two teams which rotate between the two communities every six weeks. A survey center, located centrally in each community, provides a field base for survey technicians as well as the center for the fitness testing. Each of these centers is directed by a registered nurse who, as assistant to the field supervisor, provides immediate supervision of the survey team and conducts the fitness test. Field technicians are mostly college graduates with some community experience who undergo an intensive six-week training program and evaluation directed by the field supervisor, before commencing fieldwork. This training program includes phlebotomy certification (completed with the regular Memorial Hospital phlebotomy service), blood pressure (systolic and fourth and fifth phases diastolic), height and weight measurement technique, interviewing technique, and respondent selection procedures. Technicians are also certified in CPR. The field supervisor coordinates all quality control in the field. This includes monitoring of production rates, feedback from taped interviews, spot checks on dispositions, and full field evaluations.

In both cities, households are randomly selected from available street directories, updated by a block supplement sample (Kish, 1965). Within each sample household, a single respondent is selected from the eligible adults (aged 18 through 64 years at last birthday), using selection tables adapted from those proposed by Kish (1965) and Deming (1960) to approximate a random selection process. To compensate for a slightly higher probability of selecting respondents living alone, some weighting of data may be required in the analysis. Computerized labels with I.D. numbers are generated for all sampled addresses and are attached to the Household Screener instruments, each containing one of twelve respondent selection tables. To minimize travel costs, addresses are assigned to technicians by census tract. Bilingual and trilingual technicians are able to conduct interviews in Portuguese or Spanish (using back-translated instruments and consent forms). Interviews are also conducted in the oral Cape Verdean creole. These are the only languages, apart from English, accommodated by the survey and include the major ethnic groups in the two cities.

The physiological measurements are obtained in the home as follows. A first blood pressure is taken after the respondent has been seated quietly for five minutes, using the right arm positioned at heart level (using the kitchen or dining table for support). This is followed by weight and height measurement, without shoes, in light indoor clothing, using hard floor and vertical surfaces. The remainder of the interview follows (10 to 15 minutes), ending with a second blood pressure in the same seated position. Written consent is then obtained and the blood sample drawn from the antecubital space, using a vacutainer or syringe. Technicians may make only two attempts, one in each arm. If no sample is obtainable because of difficulty getting access to a vein and if the respondent agrees, the nurse supervisor will return at a later date for another attempt (usually requiring puncture of a vein in the wrist or hand). Repeat visits have been required for no more than 4% of respondents and have been markedly reduced in later survey months by the introduction of syringes for technician use.

The equipment employed in the household has been selected for accuracy, durability, and portability. With the exception of the scales everything fits into a sturdy, leather, compartmentalized shoulder bag. Blood pressure is measured using a Baumanometer mercury sphygmomanometer (folding desk top model), and three sizes of cuffs are carried to ensure accurate cuff fit. The manometer is cleaned and recalibrated regularly by the Memorial Hospital's Medical Equipment Laboratory—as frequently as once a week during humid summer months. The scale is battery-powered Heathkit Digital Scale, which weighs 5.4 lbs and, unlike standard spring-operated portable scales, retains its accuracy despite constant handling and transport in the field. Height is measured using a specially designed folding wooden set square and standard carpenter's folding six-feet wooden ruler, with six-inch metal extension. The body and head of the respondent are positioned as recommended by CDC (1980), the base of the set square's position is marked on removable tape placed on the wall by the technician, and the height (from the floor to the mark) is measured using the ruler. Vertical surfaces with baseboards or other protrusions are avoided, as are carpeted horizontal surfaces.

For the subsample completing a fitness test at the survey center, repeat measurements are performed in the following order by the nurse supervisor. Height and weight are first measured without shoes in indoor clothing. The Detecto balance beam is used for recording weight as recommended by CDC. Height is measured using the Detecto scale attachment, removed from the balance beam and attached to the wall for stability. The right-angle is modified with a plexiglass extension to provide a wide flat, rigid surface and thus closely approximate the equipment recommended by CDC. Blood pressure is taken three times, with the respondent seated as for the household protocol, during an interview to screen for fitness test eligibility. The first two readings are taken 10 minutes apart using a random zero manometer, while the third is taken with the standard desk top model mercury manometer approximately 3 minutes later (without adjusting the prior reading for the random zero). This is followed by the fitness test for those eligible. All respondents are offered $10 as reimbursement for expenses on completion of the survey center appointment.

Reliability, test design, and results

In testing for the reliability (and validity) of field measurements, two approaches were employed. The first used the comparison of field and survey center values as described in the previous section, while the second approach involved separately designed and conducted experiments to supplement this information. The first comparison addresses issues of measurement validity as well as reliability—the extent to which measurements taken in the field replicate values obtained in the more traditional clinic setting. To the extent that sources of variation can be controlled in the comparisons, an assessment of reliability can also be made. The experiments address reliability issues exclusively, controlling for major sources of variation in the comparisons.

The set of paired observations produced by replicating measurements within the survey center protocol can be analyzed in various ways. In order to check on validity, correlations, means, and variances can be considered. In addition to these statistics, an analysis of variance provides a more detailed assessment of variance components in the determination of reliability—the extent to which observations are reproducible (Moser and Kalton, 1971). Although analyzed as a balanced block design with one observation per cell, the measurement protocols were not randomly assigned in the order of application (Kempthorne, 1952; Finney, 1960). This was not possible given that (a) the field measurements always preceded the survey center protocol, and (b) the use of the random zero sphygmomanometer had to precede the standard blood pressure reading in the survey center in order to prevent bias in anticipating values. Despite the lack of random assignment, the data were, with one exception, independently obtained. The nurse performing the survey center measurements worked independently of the field technician and generally without knowledge of the values obtained in the household. The only exception to this was the situation in which a respondent was taking antihypertensive medication or other drugs which may affect blood pressure. Data on these medications were obtained in the home and coded by the nurse supervisor as part of the screening procedure for fitness test eligibility.

Although appropriate for validity assessment, these paired data unavoidably included confounded effects. In particular, observer and equipment differences were confounded in all comparisons except for the survey center blood pressure determinations. Differences in setting for the respondent were also included in comparisons (for example, weight gain or loss between observations, clothing differences, variations in respondent apprehension or expectations in the home and in the survey center). Independent experiments were therefore designed to estimate the separate component effects of technicians, equipment and protocols. For convenience, 4×4 Latin Squares were used as the basic design, in two replicates, to provide sufficient degrees of freedom for error estimation (Kempthorne, 1952; Finney, 1960). In all the experiments, subjects (comprising seven women and one man and representing a 20-year age range) were assigned to rows (eight in total). For the height and weight trials, columns represented technicians, and letters represented equipment or protocols. For the blood pressure trial, the basic design was augmented by another set of Latin Squares, orthogonal to the first, to produce a Greco-Latin Square which permitted the inclusion of a third blocking factor (Fisher and Yates, 1953; Finney, 1960). In this design, columns represented the order of measurement on the subject, Latin letters represented equipment as before and Greek letters were assigned to technicians. Using these designs, the contribution of each source of variation could be independently and efficiently assessed.

In determining "treatments," equipment was taken from technicians' bags and labeled as they returned from the field. No further checking or calibration was conducted before the trial in order to simulate, as nearly as possible, normal field conditions, with regular maintenance and calibration of equipment assumed, as described in the Field Manual (Kipp et al., 1981). To minimize the order effect in the blood pressure trial, five minutes seated at rest were ensured between each measurement on each subject using stop watches.

The Latin Squares were randomly selected according to the procedures recommended by Fisher and Yates (1953). In the height trial, with only two measurement procedures, the portable field protocol was assigned to letters A, B while the fixed survey center protocol was assigned letters C, D. Weight protocols comprised three Heathkit Scales (A,B,C) and the Balance Beam (D). The blood pressure trial included four standard manometers used in the field. The random zero sphygmomanometer was not included as (a) an adequate comparison of this equipment with the standard manometer was provided in the paired survey data, and (b) not all the survey technicians were trained in its use.

In completing analyses of variance for both the paired survey measurements and the Latin Square trials, a mixed effects model applied, as respondents/subjects contributed a random effect while protocols were fixed. However, given only one observation per cell, a simple fixed effects model was assumed, after testing for interaction effects in the paired survey data (Scheffe, 1959). Results of this preliminary test of residuals indicated that subject and equipment interaction was statistically insignificant (0.05) for height, weight, and diastolic blood pressure. The statistically significant interaction for systolic blood pressure reflected the tendency to greater discrepancy in successive determinations of high systolic pressure. For the paired survey data, therefore, results should be interpreted conservatively. Given the normal range of blood pressures exhibited among the eight volunteer subjects, interaction effects in the trial are assumed to be negligible.

For clarity, results are presented separately below for

each physiological measurement.

Height. From the paired survey data on 135 respondents, a correlation of 0.98 was obtained. This compares very favorably with previous reports (Moffat et al., 1980; Sady et al., 1978) and indicates that the field measurement, using portable equipment, is not only valid (assuming the survey center equipment produces the more accurate value) but highly reliable. The means presented in Table 1 confirm this reliability. With different nurses completing the survey center measurements in New Bedford and Pawtucket, the results were comparable. The corresponding F value (1,134 df) for protocol differences was negligible (F = 1.63).

Table 1
Mean heights (in inches) and standard errors from survey and trial data

		Protocol	
		Field	*Traditional*
A.	Survey Data		
	Pawtucket (n = 86)	65.36	65.44
	New Bedford (n = 49)	64.53	64.56
	Total (n = 135)	65.06	65.12
	S.E. (n = 135)		

B.	Trial data	A	B	C	D
	(n = 32)	65.59	65.62	65.75	65.75
	S.E.		0.054		

Means calculated from the trial data (Table 1) indicate that almost all the (admittedly small) variability contributed by the height protocols is found in the difference between field and survey center protocols and not in repeat measures using the same protocol—even when administered blind by a different technician. The full analysis of variance for the Latin Square trial, presented in Table 2, confirms the excellent reliability obtained with the portable field equipment.

Table 2
Analysis of variance for height measurement (inches) from a Latin square trial

Source	*d.f.*	*S.S*	*M.S.*	*F.*
Subjects	7	63.11	9.02	398.26**
Protocols	3	0.16	0.05	2.30
Technicians	3	0.10	0.03	1.43
Error	18	0.41	0.02	
Total	31	63.78		

**Significant at the 0.01 level.

Weight. Equivalent results for weight are evident from the means presented in Table 3 and the analysis of variance for the trial data in Table 4. The statistically significant mean square for machines in the trial is almost entirely due to one of the Heathkit scales, which consistently weighed about one pound more, and to the very small residual mean square. The survey results indicate that the protocol effect is indeed negligible, despite the potential for variability in clothing and real weight changes in the respondent between the two measurements, with a correlation coefficient between paired observations of 0.92 (n = 135).

Table 3
Mean weights (pounds) and standard errors from survey and trial data

		Protocol	
		Field	*Traditional*
A.	Survey data		
	Pawtucket (n = 86)	153.79	155.26
	New Bedford (n = 49)	157.68	157.28
	Total (n = 135)	155.20	156.00
	S.E. (n = 135)		

B.	Trial data	Field protocol only			
	(n = 32)	A	B	C	D
	Mean	140.08	140.25	141.35	140.61
	S.E.		0.111		

Table 4
Analysis of variance for weight measurement (pounds) from a Latin square trial

Source	*a.f.*	*S.S.*	*M.S.*	*F.*
Subjects	7	13,140.88	1,877.27	19,155.80**
Protocols	3	7.66	2.55	26.05**
Technicians	3	0.65	0.22	2.21
Error	18	1.77	0.10	
Total	31	13,150.96		

**Significant at the 0.01 level

Blood pressure. Within the survey, several sets of blood pressure measurements were taken. The three used for this analysis are: the second random zero survey center measure (R), and the standard survey center measure (S). Only systolic and fifth phase diastolic values are included in the analysis. (Fourth phase values were recorded optionally, if heard clearly.) The survey data therefore included three blood pressure protocols and the three pair-wise correction coefficients were as follows.

	H v S (n = 135)	H v R (n = 133)	S v R (n = 133)
Systolic	0.74	0.69	0.90
Diastolic	0.53	0.54	0.84

These values, combined with the means and standard errors presented in Table 5, indicate that while the field

protocol for systolic pressure appears to be reasonably accurate and reliable, the same cannot be said for the diastolic (5th phase) reading. This finding is consistent with prior reports (O'Brien and O'Malley, 1979; Rose, 1965; Wilcox, 1961) and reflects the greater difficulty in accurately determining the point of sound disappearance, compared to the first appearance of the sound. The F value for the systolic readings was approximately 1.0 even when the sum of squares was orthogonally partitioned into the two contrasts (H v R+S and R v S), indicating that compared to "between subject" variability, the differences between the three readings were negligible. In contrast, an analysis of variance of the survey data on diastolic pressure yielded a mean square for measurements which was significant at the 1% level, almost all of it due to the differences between field and center determinations. The mean square for subjects was statistically insignificant, underscoring the unreliability of diastolic readings.

Table 5
Mean blood pressures (systolic and fifth phase diastolic) and standard errors from survey and trial data

		Protocol		
		Field	Traditonal	
A.	Survey data		(a) Random Zero	(b) Standard
	Systolic			
	Pawtucket (n = 84)	125.36	125.74	127.57
	New Bedford (n = 49)	126.10	126.61	126.53
	Total (n=133)	125.63	126.06	127.19
	S.E. (n=133)			
	Diastolic			
	Pawtucket (n = 84)	78.38	79.12	82.00
	New Bedford (n = 49)	77.22	84.16	83.31
	Total (n=133)	77.95	80.98	82.51
	S.E. (n=133)			

B. Trial data (n=32)	Field protocol only			
	A	B	C	D
Systolic mean (machine)	116.00	115.00	113.50	114.25
Systolic mean (technician)	113.50	114.25	118.50	112.50
S.E.	2.585			
Diastolic mean (machine)	68.25	73.88	69.38	71.50
Systolic mean (technician)	75.50	72.00	64.50	71.00
S.E.	2.193			

The analysis of the Greco-Latin trial presented in Table 6 provides additional data on the reliability of the field measurement. In particular, equipment does not appear to affect reliability and only for diastolic pressure is the technician effect significant. This is almost entirely due to the very low reading of one technician, who also provided the highest systolic reading. The impact of this one technician in the trial analysis underscores the role of technician variability in the interpretation of survey data.

Blood sampling. The 1978 Canadian Health Survey used a "window" of 16 hours in a portable cooler between collection and centrifuging of blood samples. However, that survey transported whole blood, only assayed total serum cholesterol, and did not include any lipid analysis (Health and Welfare Canada, 1981). Our participating laboratories had already established that such a delay did not affect serum thiocyanate readings and that *separated* serum could be held in a cooler for at least three days without affecting lipid analysis. It remained, therefore, to verify that holding blood samples for a period of up to 16 hours in a cooler in serum separator tubes would not adversely affect cholesterol or lipid values. An assay was therefore designed and conducted. At least 30 mls of blood were drawn from each of eight volunteers and decanted equally into six serum separator tubes (at least 5 mls per tube). After 30 minutes standing at room temperature (the minimum requirement to ensure adequate clotting), the first sample was centrifuged and the resulting serum stored in a cooler before being transported to the laboratory. The remaining five samples were deposited directly in the cooler, to be removed and centrifuged after 2, 4, 6, 8, and 16 hours respectively. All samples were then transported to the laboratory and analyzed in the same batch to control for between batch variability. The resulting total cholesterol and HDL values are presented in Figures 1 and 2, demonstrating clearly the stability of values over 16 hours of storage. Variability of values obtained after different storage times was random and negligible compared to differences between mean subject values.

Response rates and costs

In this health survey, the *response rate* was defined as the number of completed field protocols divided by the number of eligible households. All households which did not exist, which were vacant, which were outside city limits, or in which no one aged 18–64 years lived, were classified as ineligible. For the relatively few houses that were inaccessible or in which no person was ever home, the proportion eligible was estimated. The *hit rate* was defined as the number of completed field protocols divided by the total number of sample households—a statistic which partially reflects the effort required to complete a protocol. In defining these rates, households and respondents can be considered interchangeably, as only one respondent was selected per household (see above).

The distribution of final dispositions (obtained from partial survey data) is presented in Table 7. The hit rate from this table is 46.4%. The lower half of Table 7 presents the distribution of the number of calls (household visits) required in order to obtain a completed field protocol. The difficulty in identifying and scheduling a visit with the designated respondent is illustrated in the comparison with the distribution obtained in earlier surveys, as reported by Cochran (1967). Although the limit on the number of calls was set at five, technicians made additional calls to complete a protocol if the respondent was only identified on the fifth call. Moreover,

Table 6
Analysis of variance for systolic blood pressure measurements from a Greco-Latin square trial

Source	df.	S.S.	M.S.	F.
A. Systolic				
Subjects	7	1,431.87	204.55	3.83*
Machines	3	27.37	9.12	0.17
Technicians	3	167.37	55.79	1.04
Order	3	32.37	10.79	0.20
Error	15	801.89	53.46	
Total	31	2,460.87		
B. Diastolic (fifth phase)				
Subjects	7	1,396.50	199.50	5.19**
Machines	3	147.75	49.25	1.28
Technicians	3	506.00	168.67	4.38*
Order	3	76.75	25.58	0.66
Error	15	577.00	38.47	
Total	31	2,704.00		

*Significant at the 0.05 level
**Significant at the 0.1 level

multiple visits were not limited to complete protocols. The "no one ever home" disposition required five calls to be made (the majority of the "H/H unreachable" dispositions), while ineligibility, refusal, and other nonsamples took an average of nearly two calls to determine (compared to the 2.5 average for completed protocols). Even vacancy was not always determined on the first visit, unless neighbors were available to confirm this disposition.

Table 7
The distributions of final sample dispositions and of responses by call

A. Distribution of Final Dispositions

Disposition	Percent
H/H Nonexistent, vacant, or ineligible.	30.3
H/H Unreachable	4.6
Refusals, Noncompletions for language, etc.	18.7
Completed Field Protocols (Responses)	46.4
TOTAL (n = 2,561)	100.0

B. Distribution of Responses by Call

Call	Percent (this survey)		Percent (prior surveys)*
1	21.7		40.2
2	36.6		34.8
3	20.5		
4	12.1	41.6	25.0
5+	9.0		

*Combined results from surveys completed in the 1950's (Cochran, 1967).

The response rate (preliminary) is 70% and is expected to be higher after follow-up of nonrespondents has been completed. This rate assumes that a (partial) protocol is completed. In isolated cases, physiological data were not collectable (for example, respondents in wheelchairs, collapsed veins). No more than five respondents refused all physiological measurement (about 0.2%) and only 10% refused to have blood drawn (given that they completed the remainder of the protocol). This experience compares well with the 90%–95% blood sample rate (given a completed interview) in the Canada Health Survey. Of the 6% from whom blood was not obtainable, most have been recontacted by the nurse supervisor, and blood samples have been obtained from 50% of these cases.

These hit and response rates, as well as an average of approximately 2.5 visits (calls) per disposition, were combined with technician production rates to yield an estimate of two completed field protocols per person-day of effort. This estimate includes local city travel, routine paperwork, centrifuging, decanting, and storing serum samples, as well as all visits to sample households. In a 7.5 hour workday (excluding meal breaks) this corresponds to 6.0 hours in the field and 1.5 hours in the survey center. The additional labor cost incurred by travel between Pawtucket and New Bedford (25% out of every day of technician effort) is not included in these figures in order to provide generalizable cost estimates.

The labor cost per completed field protocol is, therefore, calculated on the basis of 2/person-day, assuming an average of 240 workdays per year (a 40-hour week, minus vacation and sick leave). This excludes the four to six week training period. A base salary of $12,5000 and 18% fringe benefit is employed for illustrative purposes and will vary for different survey settings.

The remaining cost components included in this analysis are all protocol specific. Fixed equipment costs, which are independent of the number of protocols completed, are not included but amount to approximately $300 per technician. This includes a sturdy leather shoulder bag with zippered compartments, a portable desk top model manometer, stethoscope, three sizes of blood pressure cuffs, carpenter's wooden rule, folding wooden set square, and portable Heathkit scale, portable cooler with test-tube rack and ice packs, and clipboard. Other fixed costs which may be required are liability insurance for all technicians (if not adequately covered by institutional policies) as well as the costs of CPR and phlebotomy training.

Protocol-dependent costs include four major categories—travel, instrument printing, laboratory costs, and consumable supplies. The costs of processing data, once collected, are not included in this discussion as these depend on the type and volume of items as well as available processing systems and will therefore vary widely between survey situations. The four component items included, plus labor costs, cover the major cost items for data collection regardless of site and are easily adapted to local cost structures.

Local travel costs are reimbursed for all field work at the rate of 20¢ per mile and, in the PHHP survey, averages $55/person-month. Given the protocol completion rate of 2/day, this is equivalent to $1.57 per protocol. To reduce travel costs, team members traveling

Figure 1
The effect on total serum cholesterol levels of retaining serum in a separator tube for variable periods, refrigerated, before centrifuging

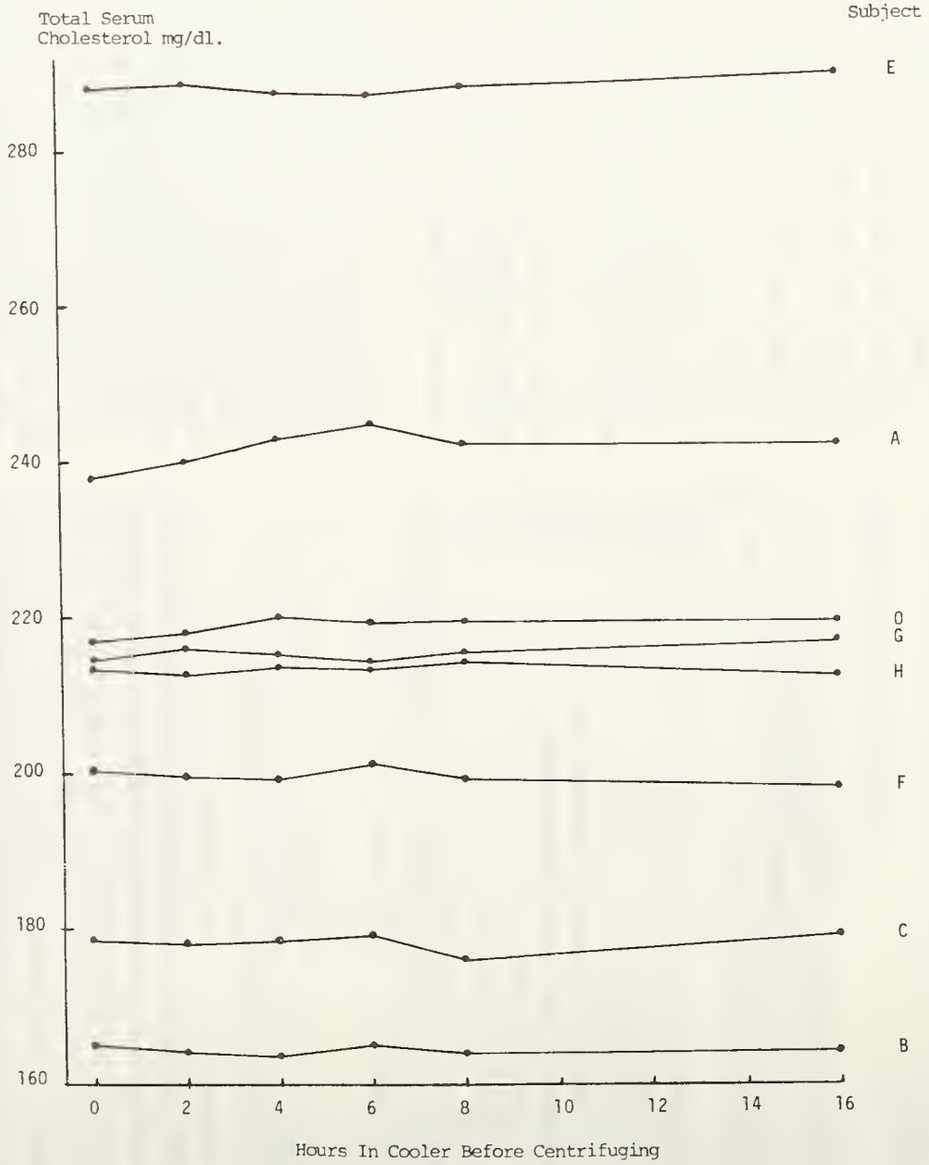
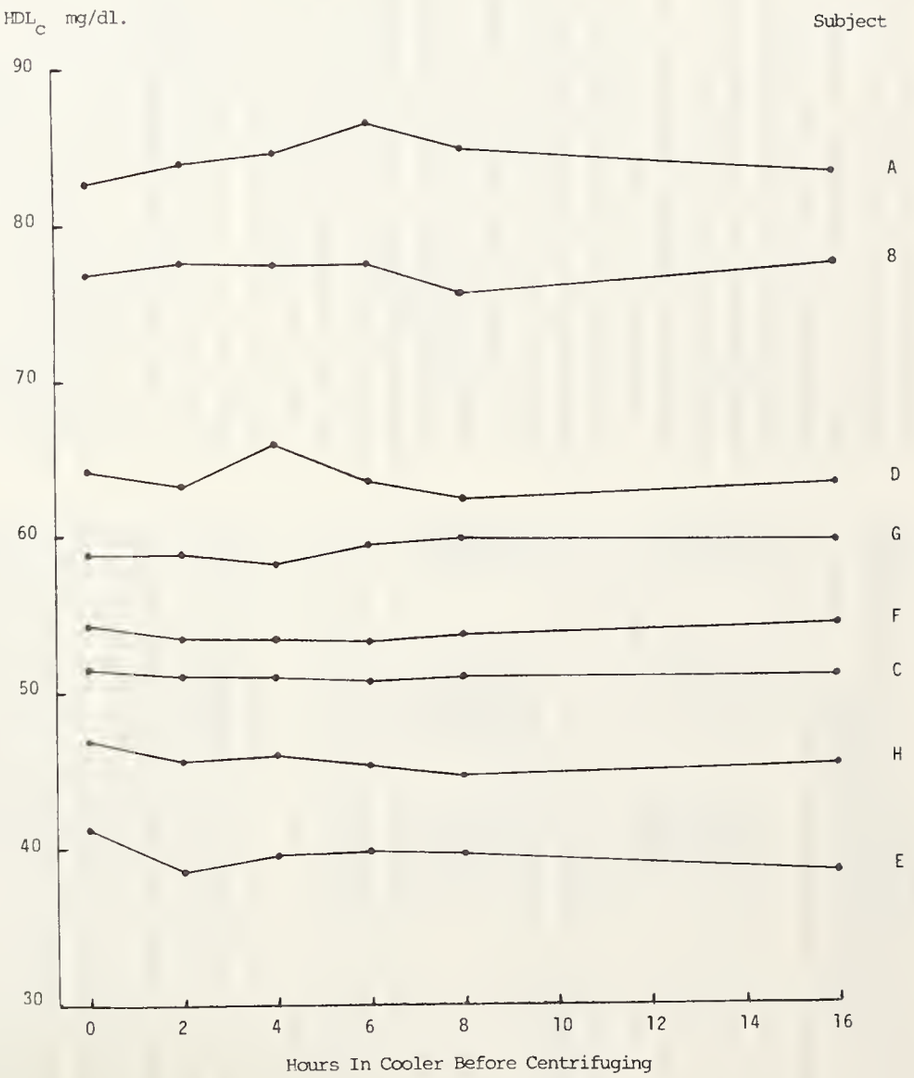


Figure 2
The effect on HDL levels of retaining serum in a separator tube for variable periods, refrigerated, before centrifuging



etween cities used rented cars (at a cost of $250 per car er month). This was not a cost efficient alternative for vithin-city travel and is not included in the cost estimate presented in Table 8. However, where long distance ravel is involved, this option may result in considerable aving compared to mileage reimbursement.

Instrument costs include a reasonable volume 2000–5000 copies) as well as typesetting in two colors. The 75-sheet total includes all the instruments required for a completed protocol (Screener, Interview, Laboratory Forms, Consent Form, and Confidential Record) as well as the Screener required for an additional, noninterview, final disposition.

Laboratory test costs are highly variable depending on the type of cost agreement. Where a laboratory can accommodate a large volume of tests, price may be determined as the cost of labor plus supplies and overhead (as for PHHP). If determined at commercial, single-test rates, the cost per test could be much higher than the $10 quoted here.

Supply costs are based on bulk prices for needles, tubes, vacutainers, sterile wipes, pencils, and other routine consumables.

These cost items are summarized in Table 8 and yield a total of just over $46 per completed protocol. This cost is probably at or near a minimum, given the competitive labor costs in this area of southern New England and the low laboratory costs. The cost of producing and mailing the respondent report of laboratory results is not included here as it is essentially a data-processing cost (reports are generated inexpensively by computer, directly from clean datasets, and postage is covered within hospital indirect costs).

Table 8
Component costs of field protocol

	Item	*Assumptions*	*Cost/Protocol*
1.	Technician Labor	• Annual salary + fringe = 15,340 • Completion rate 2/day for 240 days	31.96
2.	Local Travel	• 20¢/mile • $55/person month	1.57
3.	Instruments (printed)	• 75 sheets (4 sides) @ .28/sheet	2.10
4.	Laboratory Tests (SCN + S.C.)	• $28,000 for up to 3,000 protocols	10.00
5.	Supplies (consumables)	• Vacutainers, needles, sterile wipes, etc.	0.50
	Total		$46.13

Applications

The results presented in this paper have important implications for the design and conduct of future epidemiological field research by providing detailed information on reliability, validity, and cost efficiency of a household survey approach to the collection of physiologic data.

In particular, these results can be applied to future health surveys including physiological measurement and procedures in at least two important ways. First, a precedent is created in demonstrating the feasibility of a field approach to physiological measurement, which could be expanded to, for example, fitness testing (already successfully attempted in the Canadian Health Survey) and urine sampling. Second, data on response rates and costs are available with which to optimally plan future surveys. Specifically, additional pilot and survey center data from the PHHP survey, combined with the results presented in the previous section, provide a useful means for comparing relative costs of the field and traditional approaches.

Before deciding on the field approach described in this paper, a pilot of the traditional approach was conducted in Pawtucket. A brief household contact was used to identify respondents and generate appointments to visit a trailer clinic (modeled on the HANES approach), where the entire protocol including interview, measurement, and blood sampling was conducted. Respondents were not reimbursed but were offered transport and child care as needed. The response rate to the initial household contact was 90% with an average of 3.75 appointments generated per person-day in the field. The response rate at the mobile clinic was 47%, given that a household response was obtained, to yield an overall response rate of 42% (47% of 90%).

This experience contrasts with the response in the survey center for the subsample in the current survey selected for the fitness test. A conditional response rate of 82% was obtained, given completion of the field protocol and payment of $10, to yield an overall response rate of 56% (82% of 70%).

Let us assume, following the pilot experience, that a 90% response rate is obtained for the brief household contact, with 3.75 appointments generated per person-day in the field. Further, assume that respondent payment or other incentive is sufficient to generate a conditional 70% response rate in the clinic, yielding an unconditional response of 63%. Given the experience described above, this is probably an optimistic assumption in the PHHP study population. On the other hand, it will generate a conservative comparison of the costs of a traditional versus field approach.

It remains to estimate the production rate within the clinic setting. To complete the equivalent of the field protocol in a clinic setting would take at least 30 minutes. Eight appointment slots 45 minutes apart would therefore constitute a full person-day of effort in the clinic, allowing time for paper-work and blood handling.

From the survey center experience, the following data on appointments are available.

Appointment	Percent total	Percent kept
First	84	74
Second	14	62
Third	2	33
TOTAL	100	

From these data, an average rate of appointments kept (show rate) of 71.5% can be estimated, which represents 84% of the survey center response rate (71.5%/85%). Assuming this ratio of show rate to response rate is constant,the show rate for a 70% response rate would be 60%. The rate of additional appointments required would be approximately 25%, given 20% additional appointments for a 71.5% show rate. In other words, the 3.75 (average) appointments generated from household contacts per person day would produce an average 4.7 appointments at the clinic, representing approximately 0.6 person-days of effort in the clinic. Moreover, this total of 0.6 person-days of effort will yield, on average, 2.36 complete protocols (63% of 3.75). This is equivalent to a completion rate of approximately 1.5/person-day of technician time (including household contacts and clinic time), compared with 2/person-day for the field approach. This reduced completion rate represents a 33% increase in technician labor per protocol.

In terms of other protocol costs, it is reasonable to assume that travel, printing, laboratory, and supply costs are equivalent for traditional and field approaches. However, the traditional approach must include two other items which could be substantial—respondent payment and overhead costs for maintaining the clinic. It is possible, therefore, that the traditional approach will require up to twice the direct cost of the field protocol.

Moreover, this traditional approach will not, in general, produce as high a response rate. Conservatively, assuming (as in this discussion) that the conditional clinic response is equivalent to the overall response to the field approach, the response rate for the equivalent traditional approach will not exceed 90% of the field response rate. Thus a more expensive traditional approach will yield data of lower inferential value and could not therefore be considered to be a cost-effective alternative.

Summary and conclusion

This paper has presented results on reliability, validity, response rates, and cost which demonstrate the feasibility of a field approach to the collection of physiologic data. In particular, the following has been demonstrated:

1. Physiologic measurement performed under field conditions with portable equipment can be as reliable and as valid as supposedly more closely controlled measurements performed in a central (clinic) location using standard equipment.
2. The response rate of this field approach is an acceptable 70% in lower socioeconomic populations with large Portuguese minorities, and this rate will almost certainly exceed the response which could be anticipated to an equivalent traditional approach.
3. The direct data collection cost for this field approach is less than $50. It does not include or require respondent payment and is presented in a component format amenable to adaptation in other survey environments.
4. The labor cost (the major cost component) is at most 75% of the labor cost required to complete an equivalent protocol using the traditional approach (brief household contact followed by a clinic visit).

Hopefully, this paper will stimulate further work on the development of cost-effective approaches to the measurement of physiological data, which could have wide applications in both survey and field trial research. In particular, it should stimulate closer investigations of relative costs of different approaches and, consequently, more careful planning of large, costly data collection projects with the aim of employing the most cost-effective approach in a given research environment.

Footnote

[1] Unpublished data (1981)—Canada Health Survey, 1978.

Dimensions and correlates of respondent burden: Results of an experimental study

Joanne Frankel, Bureau of Social Science Research, Inc.

Laure M. Sharp, Bureau of Social Science Research, Inc.

Introduction

The research reported here focuses on the topic of respondent burden, defined as the configuration of negative feelings which persons who participate in voluntary personal household interviews may experience. A review of federal guidelines relating to respondent burden (e.g., OMB regulations limiting interviews to an average of 30 minutes) suggest that such policies are based on the assumptions that long interviews are more burdensome than short ones and that repeated interviewing, such as that experienced by persons participating in a survey panel, also leads to an increase in burden. Assumptions about the aggregate respondent burden resulting from the proliferation of survey activities have also been articulated in the private sector, which is concerned about exhausting the good will of the American public.

However, as reported at the Second Health Survey Research Methods Conference, some survey research methodologists (e.g., Bradburn, 1979; Rothwell and Bridge, 1979) contend that the relationship between interview length, number of interviews, and respondent burden is far from clear. There may be other mediating factors, such as the predisposition of the respondent, which affect respondent reactions. At the conference, Rothwell and Bridge also stated that the construct of respondent burden itself is ill-defined, and that, before further research about the correlates of burden is pursued, more definitional and theoretical work needs to be done.

A research project, initiated by the Bureau of Social Science Research in 1978, with funding from the U.S. Department of Housing and Urban Development, provided the opportunity to examine some of these concerns. The project sought to measure the relative importance of three factors identified by Bradburn—time, effort, and repeat administration—on self-perceived respondent burden and to examine demographic and attitudinal factors which may be associated with differences in perceived burden.

The respondent burden study was implemented in two phases with a sample of 500 persons living in surburban Philadelphia. During the first phase, both interview length and the effort required of the respondent to answer certain questions were manipulated, and their effect on perceived burden was assessed. For the second phase of the study, the effort variable was eliminated, and the third manipulated variable—single vs. repeat administration of identical questions over time—was introduced. The "repeat" treatment was applied approximately 10 months after the first data collection cycle by reinterviewing 200 of the original 500 respondents, using either the short (25-minute) or long (75-minute) version of the interview instrument. All versions of this instrument (i.e.,long, short, Time I, and Time II) were based on the Annual Housing Survey and covered topics of presumably moderate salience, such as housing and neighborhood conditions, transportation, and energy.

Respondent burden was measured through observed behavioral indicators and through responses to a self-administered reaction form, which was filled out at the conclusion of the first interview by 300 of the respondents and at the conclusion of the second interview by the 200 members of the panel group. The reaction form was designed to include items which captured those negative attitudes flagged as possible components of respondent burden by other researchers. It contained questions concerning the extent to which the interview was overly time-consuming, boring, difficult, personal, and unimportant. A question asking the respondent to rate the bothersomeness of the survey as compared to other common tasks was also included. Finally, items ascertaining respondent attitudes toward surveys in general, respondent participation in other (earlier) surveys, and respondent attitudes toward the interviewer were also written, since the work of other researchers (e.g., Kahn and Cannell, 1957; Dillman, 1978) had pointed to these factors as important components or determinants of respondent motivation.

The overall findings from both phases of the experiment have been presented in earlier papers (Frankel and Sharp, 1981; Sharp and Frankel, 1981). These findings suggested that respondent burden as measured in this study was a relatively infrequent experience; on a burden index constructed on the basis of answers to the reaction form, only 12% of all participants could be classified as "burdened." The earlier findings also pointed to "burden" differences associated with instrument length, but no discernable effects were found attributable to the effort variable (operationalized by the requirement to search records for financial information pertaining to utility payments, insurance payments, etc.). For this reason, the effort variable was dropped from the analysis presented here, which is based on the design shown in Figure 1. This analysis sought to answer two major research questions:

1. What are the underlying dimensions or components of respondent burden as measured through the reaction form?
2. In explaining the various dimensions of respondent

burden, what are the relative contributions of:
a. interview length;
b. panel participation;
c. respondent demographic characteristics; and
d. past participation in other surveys as reported by respondents?

The first question was explored through factor analysis and the second through regression analysis of the database. This paper reports the results of those procedures.

Figure 1
Study design

	Group 1 25-Minute interview		Group 2 75-Minute interview	
	1A One interview (N=75)	1B Two interviews (N=100)	2A One interview (N=75)	2B Two interviews (N=100)
Phase I				
Interview	☒	☒	☒	☒
Reaction form	☒	☐	☒	☐
Phase II				
Interview	☐	☒	☐	☒
Reaction form	☐	☒	☐	☒

Dimensions of respondent burden: Results of the factor analysis

A description of the 21 variables which were factor-analyzed is provided in Figure 2. Table 1 shows the factor matrix obtained from these variables, and Table 2 presents the percent of total and common factor variance accounted for by each factor. As these tables indicate, seven factors were identified, which considered as a set account for approximately 63% of the total variance in the data. These factors are described below, in order of magnitude of variance explained.

Factor 1. This factor was defined primarily by loadings on six variables: V13, views about the value of the time and effort spent answering the interview questions, with a loading of .72; V6 and V7, ratings of the interest (.71) and importance (.63) of this interview; V1 and V3, views about the general benefits (.52) and interest (.49) of survey participation; and V4, views about the ability of survey participants to affect government decisions (.42). This combination of variables seemed to reveal the presence of an attitude that taps feelings of the interest, importance, and benefits of survey participation. Since respondent burden was defined as negative feelings toward the interview, this factor was named in the negative as well, as "Perceived Uselessness."

The Perceived Uselessness factor suggests that an important component of respondent burden is the extent to which respondents feel that their participation in surveys is of little importance, does not affect decision makers, and does not confer benefits on themselves or others. It is noteworthy that, for this particular group of respondents, this factor was the most important component of respondent burden in that it accounted for the largest percentage of total (24%) and common factor (48%) variance in the data (Table 2).

Factor 2. The second factor loaded on three variables: V10, stated willingness or unwillingness to continue with the interview given a hypothetical opportunity to do so, which had a loading of .85; V9, views about the

Table 1
Varimax rotated factor matrix

			Factors						
			1	2	3	4	5	6	7
V1	Q3A	Survey benefits	.52	.04	−.01	.01	−.02	.05	.03
V2	Q3B	Too many surveys	.17	.11	.46	.02	.11	.15	−.27
V3	Q3C	Surveys are interesting	.49	.12	.11	−.04	−.02	.21	−.05
V4	Q3D	Surveys affect government decisions	.42	.03	.15	−.09	.32	.09	−.20
V5	Q3E	Surveys are too personal	.11	.08	.70	.10	−.02	−.05	.00
V6	Q4	Interest of this interview	.71	.20	.14	.04	.14	.08	.01
V7	Q5	Importance of this interview	.63	.14	.13	−.03	.12	−.10	.04
V8	Q6	The interviewer in this interview	.18	.12	−.00	−.07	.85	−.08	.07
V9	Q7	Length of this interview	.18	.77	.05	.03	.10	.09	.03
V10	Q8	Willingness to continue additional minutes	.28	.85	.16	−.03	.00	.05	−.08
V11	Q9	Difficulty of this interview	.01	.05	.03	.02	−.04	.49	.07
V12	Q10	Accuracy of answers	.08	.04	−.04	−.01	−.07	.20	.31
V13	Q11	Were time and effort well spent	.72	.27	.22	.01	.10	−.14	.14
V14	Q12	Nusiance scale for this interview	.24	.22	.34	.01	.20	.37	.22
V15	Q15	Willingness to be reinterviewed	.30	.28	.27	−.01	−.02	.15	.21
V16	Q16A	Improve surveys: explain use of answers	.10	.02	.04	.84	−.06	.12	−.04
V17	Q16B	Improve surveys: explain confidentiality procedures	−.06	.14	.28	.51	−.01	−.03	.01
V18	Q16C	Improve surveys: use shorter questionnaires	−.01	.04	.14	.12	.17	.03	.32
V19	Q16D	Improve surveys: hire better interviewers	.18	.60	.34	.03	.10	.06	.22
V20	Q16E	Improve surveys: ask less personal questions	.22	.22	.75	.03	.03	.07	.19
V21	Q16F	Improve surveys: ask more open-ended questions	−.07	−.16	−.10	.32	−.00	−.06	.17
		Eigenvalue:	4.67	1.41	1.02	.86	.76	.58	.43

Figure 2
Description of the variables included in the factor analysis

Variable number		Variable label	Description
			Vs1–5 show agreement/disagreement with the following statements:
V1	Q3A	Survey benefits	• Answering surveys is of direct benefit to the people who answer.
V2	Q3B	Too many surveys	• Too many surveys are being conducted these days.
V3	Q3C	Surveys are interesting	• Taking part in surveys can give me a chance to talk about interesting topics.
V4	Q3D	Surveys affect government decisions	• By taking part in surveys, I can affect the government's decisions.
V5	Q3E	Surveys are too personal	• Surveys ask questions that are too personal.
V6	Q4	Interest of this interview	Rating of the interest level of the interview just completed.
V7	Q5	Importance of this interview	Rating of the importance of the interview just completed.
V8	Q6	The interviewer in this interview	Views about the manner of the interviewer conducting interview just completed.
V9	Q7	Length of this interview	Views about the length of the interview just completed.
V10	Q8	Willingness to continue additional minutes	Stated willingness/unwillingness to continue the interview for 15–30 minutes, given hypothetical opportunity to do so.
V11	Q9	Difficulty of this interview	Rating of the difficulty of the interview just completed.
V12	Q10	Accuracy of answers	Views about accuracy of information provided regarding utility bills and household expenses.
V13	Q11	Were time and effort well spent	Views about "how well spent" were time and effort put into answering interview questions.
V14	Q12	Nuisance scale for this interview	Rating of the overall burdensomeness of this interview vis-a-vis other tasks.
V15	Q15	Willingness to be reinterviewed.	Stated willingness/unwillingness to allow the interviewer to return a year hence.
			Vs16–21 show agreement/disagreement with the following suggestions for improving surveys:
V16	Q16A	Improve surveys: explain use of answers	• Explain more about how the answers will be used.
V17	A16B	Improve surveys: explain confidentiality procedures.	• Explain more about how the confidentiality of the answers is protected.
V18	Q16C	Improve surveys: hire better interviewers	• Hire better interviewers.
V19	Q16D	Improve surveys: use shorter questionnaires	• Use shorter questionnaires.
V20	Q16E	Improve surveys: ask less personal questions.	• Ask fewer personal questions.
V21	Q16F	Improve surveys: ask more open-ended questions.	• Give respondents more chance to talk about their ideas and opinions.

length of the interview (was it too short, too long, just right?), with a loading of .77; and V19, views about the use of short questionnaires as a way of improving surveys (.60). This factor, then, clearly tapped respondents' feelings about the time required for survey participation. For this particular group of respondents, the "Time Concerns" factor was the second most important component of respondent burden, accounting for 9% of the total and 14% of the common factor variance in the data.

Factor 3. The third factor, which was termed "Privacy Concerns," accounted for approximately 7% of the total and 10% of the common factor variance. It was defined by loadings on variables representing agreement/disagreement with three statements: asking fewer "personal questions" would be a way of improving surveys (V20, which had a loading of .75); surveys ask questions that are "too personal" (V5, with a loading of .70); and "too many surveys are being conducted these days" (V2,with a loading of .46). In this context, the last statement is interpreted as reflecting the same concern as the others, following the reasoning of NORC researchers (Jones, Sheatsley, and Stinchcombe, 1979) who conducted a study (of reactions toward USDA surveys) from which the "too many" question used in the reaction form was taken. As stated in their report, these authors felt that: "*too many* does not refer so much to the actual number of survey requests as it does to the feeling of some. . .[respondents] that surveys are an unwelcome intrusion into their private lives" (p. 100).

Table 2
Percents of total and common factor variance accounted for by various factors

Factor	Eigen-value	Percent of total variance	Cumulative percent[a]	Percent of common factor variance	Cumulative percent
1	4.67	24.3	24.3	48.0	48.0
2	1.41	9.0	33.4	14.5	62.6
3	1.02	6.9	40.3	10.5	73.1
4	.86	6.4	46.7	8.8	81.9
5	.76	5.9	52.6	7.8	89.7
6	.58	5.7	58.3	5.9	95.6
7	.43	4.9	63.2	4.4	100.0

[a]Percents do not total exactly to cumulative percents due to rounding error.

The three factors just described—Perceived Uselessness, Time Concerns, and Privacy Concerns—were the most important components of respondent burden for this group of respondents. Together these factors accounted for approximately 40% of the total and 73% of

the common factor variance in the data (Table 2).

The remaining factors. Although the final four factors are of less importance from a variance point of view (explaining only about one-quarter of both the total and common factor variance), they are worth examining because they suggest dimensions of respondent burden which are consistent with the theories and findings of other researchers. For example, feelings about the *difficulty* of the interview have been flagged as a possible component of respondent burden in the measurement attempts of other researchers, as well as by survey research theoreticians (Bradburn, 1979, p.49; Dillman, 1978, p.15). This component of respondent burden is suggested by factor 6, which shows a moderate loading on V11, views about the difficulty of the interview (.49), coupled with a lower, but still noticeable loading (.37) on V14, the "overall bothersomeness of this survey" as compared to other social tasks. (The fact that the eigenvalue of this factor is lower than one may be a function of the fact that only one question asking about difficulty per se was included in the reaction form.)

The work of other researchers also suggests that the interviewer is an important determinant of respondent reactions to the interview. This dimension appears to be revealed in factor 5, which loads heavily (.85) on V8, ratings of the interviewer's manner in conducting the interview. Again,the fact that the eigenvalue of this factor is under one may be a function of the limitations of the database, which contained only two questions relating to interviewers (Qs 6 and 16c), as well as a function of the study design which attempted to treat the interviewer as a constant, by using only female and primarily experienced personnel.

Factor 4 showed a high or moderate loading on two variables (V16 and V17), involving agreement or disagreement with the statements that to improve surveys one should: "explain more about how the answers will be used" (.84) and "explain how the confidentiality of the answers is protected" (.51). A lower but still noticeable loading (.32) was seen for V21, involving improving surveys by asking "more open-ended questions." At the risk of reading more into these responses than is warranted, we may place the emphasis on the words "explain" and "open-ended." Given that emphasis, we may conjecture that, in all of these statements, the respondent is asking the interviewer to create an atmosphere in which there is more open exchange between the interview participants, whether it is on the part of the interviewer who is asked to offer more information about the survey or on the part of the respondent who is asking for the opportunity to express his or her ideas in a more open format. If such a interpretation is accurate, it would explain why V17, the request for an explanation of confidentiality procedures, loaded more heavily on factor 4 than on factor 3, Privacy Concerns. Again, however, it must be stressed that the interpretation given here is conjecture.

The final factor, factor 7, did not show any consistent patterning of variables and was therefore uninterpretable.

Correlates of respondent burden: Results of the regression analysis

A regression analysis was carried out to determine the relative contribution of the various manipulated and attribute independent variables in explaining the three major dimensions of respondent burden described above—Perceived Uselessness, Time Concerns, and Privacy Concerns. First, factor scores were computed for all respondents along each of these dimensions, to be used as the criterion variables in the analyses. Three "nuisance" variables—interruptions during the interview, and time of day and day of week of interview administration—were controlled by entering them first into the regression models. The remaining independent variables were then entered by the computer program in a sequence determined by the amount of variance each explained in the burden factor under analysis.

Table 3
Correlation matrix for predictor variables used in the regression analyses

	Interview length	Panel participation	Income	Employment status	Sex	Previous Survey participation	Interview interruptions	Interview time of day	Interview day	Education (L.T. H.S. graduate)	Education (H.S. graduate)	Age (under 40)	Age (40–59)
Interview length	1.000	.069	.029	.076	−.046	−.059	.022	−.073	.035	.009	−.118	.140	.031
Panel participation		1.000	−.051	.008	−.058	−.090	−.035	−.067	−.040	−.024	−.118	−.056	.132
Income			1.000	.159*	.133	.175*	−.024	−.027	−.002	−.403**	−.172*	.192**	.136
Employment status				1.000	.224**	−.007	−.045	.059	.116	−.310**	−.132	.387**	.268**
Sex					1.000	−.185*	−.052	−.055	.132	.019	−.123	−.131	−.029
Previous survey participation						1.000	.020	.045	−.004	−.155*	−.026	.114	.192*
Interview interruptions							1.000	−.038	.104	−.009	.043	−.134	−.088
Interview time of day								1.000	.104	−.075	.079	.068	−.005
Interview day									1.000	−.099	−.037	.035	−.107
Education (L.T. H.S. graduate)										1.000	.669**	−.082	.004
Education (H.S. graduate)											1.000	.058	.107
Age (under 40)												1.000	.385**
Age (40–59)													1.000

Note: Variables were coded as shown in Table 5.
*p.<.05 (2 tailed test).
**p.<.01 (2 tailed test).

Table 4
Correlations between factor scores and predictor variables

	Factor scores		
	Perceived uselessness	*Time concerns*	*Privacy concerns*
Interview length	.134	.442**	.019
Panel participation	−.001	−.076	−.132
Income	.136	.137	−.056
Employment status	.168*	.168*	−.108
Sex	.012	−.088	−.039
Previous survey participation	−.005	.064	−.084
Interview interruptions	.007	.010	.109
Interview time of day	.116	−.083	−.064
Interview day	.177*	.093	.041
Education (L.T. H.S. graduate)	−.109	−.107	.067
Education (H.S. graduate)	−.022	−.037	.047
Age (under 40)	.148*	.102	−.025
Age (40–59)	−.016	−.003	−.083

*p.<.05 (2 tailed test).
**p.<.01 (2 tailed test).

The bivariate correlations among all of the independent variables and covariates used in the regression analysis are reported in Table 3. Significant correlations ($p<.01$) were found between several of these variables, for example *employment status* and sex, education, and age; *income* and education and age; and *previous survey participation**[1] and age. (Several of the other correlations reported, for example between the various education categories, were artifacts of the coding procedure used.)

Identification of these correlations was potentially important, since, at each stage of the forward regression procedure, the effects of previously entered variables are removed from the effects of the new predictor entered at that stage. However, the problem of high correlations among the predictor variables turned out to be moot in this research, because, as reported in Table 4, the correlations between each of these predictors and the criterion measures themselves were generally not significant. There were some important exceptions, such as the significant correlation ($p<.01$) between Time Concerns and interview length, and the significant correlations ($p<.05$) between Perceived Uselessness and several of the nonexperimental variables.

The results obtained from the regression analysis were as might have been predicted from an examination of the bivariate correlations. The findings for factor 1, Perceived Uselessness, are reported in Table 5. Only about 6% of the variance in Perceived Uselessness is explained by the variables of interest. As shown in Part II of the table, no single variable accounted for more than 2% of the explained variance, with most variables accounting for less than 1%.

Much the same can be said about the third component of respondent burden, Privacy Concerns (Table 6). Again, neither the overall nor the variable-specific relationships are significant, and the total amounts of variance explained by the set of independent variables (4%) and by any specific variable (1% or less) are so small that describing the relative contribution of individual variables to this factor is not appropriate.

However, a considerably different picture emerged for factor 2, Time Concerns. As shown in Table 7, 28% of the variance in this component of respondent burden was explained by the variables entered into the regression equation, with 26% accounted for by the independent variables. Further, the overall relationship between Time Concerns and these independent variables is sta-

Table 5
Regression analysis of "perceived uselessness"

I. Source	*Prop. of variance*	*Sum of squares*	*DF*	*Mean squere*	*F*	*Significance*	*Correlation*
Covariates	.04093	5.00376	3				
Independent variables	.06503	7.95021	10	.79502	1.2293	NS	R^2 = .10596
Residual	.89404	109.29866	169	.64674			Adj. R^2 = .03719
Total	1.00000	122.25263	182				

II. Veriebles	*B*	*BETA*	*F*	*Significance*	*Increase in R^2 attributable to variable at step entered*
Employment status	.0880	.1052	1.344	NS	.0206
Interview length	.1037	.1269	2.822	NS	.0158
Income	.0492	.1134	1.773	NS	.0134
Age					
Under 40	.0841	.0835	.853	NS	.0032
40–60	−.1009	−.0883	1.077	NS	.0059
Education					
H.S. graduate	.0760	.0864	.656	NS	.0020
L.T. H.S. graduate	−.0781	−.0662	.312	NS	.0021
Previous survey participation	−.0308	−.0336	.185	NS	.0009
Sex	.0271	.0317	.144	NS	.0009
Panel participation	.0142	.0173	.052	NS	.0003
Constant:	−.2967				

Note: Variables used were as follows: *Covanetes:* Presence of interview interruptions (1 = Yes, −1 = No); interview Time of Day (up to 5:00 p.m. = 1, after 5:00 p.m. = −1); Interview Day (1 = weekend/holiday, −1 = weekday). *Menipuleted Independent Veriebles:* Panel Participation (−1 = No, 1 = Yes); Interview Length (−1 = short, 1 = long). *Attribute Independent Variables:* Income (coded in $5,000 increments); Age (2 vectors for under 40, 40–60, 60+, with members of category coded as 1, and 60+ as −1); Employment Status (1 = employed, −1 not employed); Sex (1 = male, −1 = female); Education (2 vectors for L.T. H.S. graduate, H.S. graduate, and M.T. H.S. graduate); Previous Participation in other (unrelated) Surveys (1 = Yes, −1 = No).

Table 6
Regression analysis of "privacy concerns"

I. Source	Prop. of variance	Sum of squares	DF	Mean square	F	Significance	Correlation
Covariates	.01679	2.00026	3				
Independent variables	.03880	4.62397	9	.51377	.7762	NS	R^2 = .05559
Residual	.94441	112.52754	170	.66193			Adj. R^2 = .01107
Total	1.00000	119.15177	182				

II. Variables	B	BETA	F	Significance	Increase in R^2 attributable to variable at step entered
Panel participation	−.1099	−.1356	3.075	NS	.0151
Employment status	−.0709	−.0858	.979	NS	.0109
Previous survey participation	−.0882	−.0974	1.477	NS	.0094
Sex	−.0433	−.0514	.383	NS	.0022
Income	−.0061	−.0144	.028	NS	.0005
Interview length	.0157	.0194	.064	NS	.0004
Education					
L.T. H.S. graduate	.0314	.0270	.050	NS	.0002
H.S. graduate[a]	−.0106	−.0122	.012	NS	.0001
Age: 40–60[b]	−.0109	−.0097	.014	NS	.0001
Constant:	.0658				

Note: Variables used are coded as shown in Table 5.
[a]For clarity, Ed. H.S. Graduate is grouped under education, although it was actually entered last into the equation.
[b]Tolerance was insufficient for entry of age, under 40 into the equation.

tistically significant ($p<.01$).

The information reported in Part II of Table 7 indicates that the actual length of the interview given to the respondent was the most important variable contributing to the perception of burden along the Time Concerns dimension. Interview length explained almost 19% of the variance in Time Concerns and was related to that factor at the .01 level of significance. Moreover, the relationship was consonant with federal policy concerns. That is, as indicated by the "b" coefficients, participating in the long interview added about .4 (or nearly one-half of a standard deviation) to a respondent's predicted score on the Time Concerns factor.

Less important, but still statistically significant ($p<.05$) in explaining Time Concerns were employment status and sex, each of which contributed about 2% to the explained variance. The "b" coefficients for these variables show that employed persons have a somewhat higher predicted score on the Time Concerns dimension than do persons who are not employed, and that

Table 7
Regression analysis of "time concerns"

I. Source	Prop. of variance	Sum of squares	DF	Mean square	F	Significance	Correlation
Covariates	.01761	2.47027	3				
Independent variables	.26331	36.93752	10	3.69375	6.1884	p<.01	R^2 = .28092
Residual	.71908	100.87280	169	.59688			Adj. R^2 = .22561
Total	1.00000	140.28059	182				

II. Variables	B	BETA	F	Significance	Increase in R^2 attributable to variable at step entered
Interview length	.3958	.4520	44.521	p<.01	.1879
Employment status	.1709	.1907	5.497	p<.05	.0168
Sex	−.1631	−.1783	5.655	p<.05	.0243
Income	.0552	.1188	2.420	NS	.0143
Panel participation	−.0665	−.0757	1.231	NS	.0080
Age					
Under 40	−.0777	−.0720	.788	NS	.0056
40–60[a]	−.0769	−.0628	.678	NS	.0016
Previous survey participation	.0158	.0527	.566	NS	.0024
Education					
H.S. graduate	.0665	.0706	.544	NS	.0016
L.T. H.S. graduate	−.0558	−.0442	.172	NS	.0007
Constant:	−.3542				

Note: Variables used are coded as shown in Table 5.
[a]For clarity, Age 40–60 is grouped under Age, although it was actually entered after Previous Survey Participation.

women have a higher predicted score than do men.

In examining this component of respondent burden, it is also important to note that the Time Concerns factor was not related to the other manipulated variable (panel participation) nor to any of the other attribute variables entered into the regression equation.

Discussion

In interpreting the findings presented above, the following limitations of the study must be kept in mind.

1. The findings were drawn from a study which used personal, voluntary household interviews. Therefore, the findings are not necessarily applicable to other types of data collection activities, such as telephone surveys.
2. The study sample consisted predominantly of respondents who were white, relatively well-educated, and in the middle-income brackets. All of the respondents lived in a suburban area. Therefore, the findings may not be applicable to other populations such as blacks, low-income groups, or residents of inner-cities or rural areas. Moreover, due to the requirements of the statistical procedures used, the analyses reported here were limited to respondents who answered all items in the reaction form. By eliminating respondents who failed to answer one or more of these items, the analysis procedures may have eliminated the most burdened respondents.
3. The treatment interviews dealt with topics of presumably moderate salience, such as housing and energy costs, neighborhood conditions, and transportation. The results might have been different if more (or less) respondent-pertinent topics had been discussed.

However, if they are replicable with other populations or in other settings, the findings summarized above will have implications in a number of areas for researchers and research sponsors.

Alleviating feelings of the uselessness of survey participation. The findings indicate that, for the group of respondents studied here, feelings about the uselessness of survey participation is the single most important component of respondent burden. It is of course possible that this finding may be a function of the instrument used in this research, since the instrument could not possible contain items relating to all possible dimensions of respondent burden, but this finding enjoys strong support in the work of other researchers. In 1978, investigators for Walker Research suggested that feelings about the usefulness, benefits, and importance of survey participation are primary determinants of respondent motivation (Walker Research, Inc., 1978). This suggestion also emerged in the theory developed by Bradburn (1979), which stressed the importance of the *salience* of the task to the respondent. Finally, the finding reported parallels most closely that of a group of NORC researchers (Jones et al., 1979) reported in their study of Dakota farmers and ranchers:

> In their assessment of survey burden, farmers and ranchers are not so much influenced by the number or length or type of surveys as they are by their perception of the quality of the surveys and the effects of surveys upon their lives. Operators who are convinced that surveys produce useful and accurate information that serves primarily their own economic interests tend not to feel burdened by even large numbers of surveys. Those who are not so convinced are likely to feel that even one survey request is too many (p.69).

The importance of this factor suggests that to reduce respondent burden in personal interview surveys, it is important to alleviate feelings among respondents that their participation is unlikely to be beneficial. Determining how to accomplish this goal, however, will not be an easy task. For, as shown in the regression analysis, we do not yet understand very much about the circumstances that create or influence these feelings. The study findings suggest that Perceived Uselessness is not related to interview length, to panel participation, to respondent demographics, or to previous survey participation. Thus, it would appear that Perceived Uselessness is primarily a function of the individual's belief system (for example, denial of the efficacy of individual action) and therefore not subject to remediation by the researcher.

Of course, some effort can be made to change this belief structure by conveying appropriate information; this can be most usefully done on two levels. One is to convey to the public at large the importance and usefulness of the survey method and the likelihood that survey data will, in fact, be used by the research sponsors. This public relations approach was suggested at a recent conference of the American Association for Public Opinion Research (AAPOR) by Corson (1979). This type of effort may be helpful in bringing about changes in attitude over time.

A more immediate opportunity for reducing burden arises when a respondent is actually asked to participate in the survey. Careful and convincing explanation prior to the start of the interview about the importance and utility of the research project should lead to burden reduction. However, as Singer (1978) has pointed out, "conventional survey wisdom advocates keeping the introduction short so as not to lose the respondent's attention" (p.195).

The importance of interview length. Time Concerns emerged as a second component of respondent burden in this research. This factor was indeed related to interview length, with "length" explaining about 19% of its variance. Thus, these findings at least partially confirm the common wisdom expressed through federal policies and in the "hearsay" literature that long interviews are more burdensome than short ones.

However, the Time Concerns factor was found to be a

much less important component of respondent burden than was the Perceived Uselessness dimension discussed above, with the latter accounting for 24% of the total variance in the data as opposed to only 9% for the former. Moreover, the Time Concerns factor did not figure heavily in such important considerations as the respondent's willingness to be reinterviewed, accounting for only 8% of the variance in that variable. (Perceived Uselessness also accounted for only about 9%.) (See Table 1.)

Dealing with privacy concerns. Although less important than the other factors, Privacy Concerns were found to be a component of respondent burden. Again, however, the study carried out here yields no clues concerning the conditions which may alleviate negative feelings in this area among respondents.

One conventional approach which is sometimes used is to inform respondents beforehand of the extent to which the confidentiality of their responses will be maintained. While some support for this technique can be found in the literature concerning response rates (e.g., Hauck and Cox, 1974) and data quality (e.g., Singer, 1978), the effect of such disclosure on respondent burden is not known. And again, Singer's concern about unduly prolonging the introduction to the interview must be taken into account.

The question of survey panels and frequent interviews. Based on this admittedly limited test (one vs. two interviews), panel participation does not appear to be an important contributor to respondent burden. The "panel" variable was not related to any of the dimensions of respondent burden identified through this analysis.

Similarly, participation in previous, unrelated surveys was also found to be unrelated to the various dimensions of respondent burden. Respondents reporting that they had participated in other interviews were not more burdened by the current interview than were other respondents. Thus, previous participation per se was not revealed as a correlate of burden. It is of course conceivable that some portion of those who refused to participate did so because of frequent prior survey participation. Given the fairly high Phase I refusal rates (24%) this possibility cannot be ruled out conclusively. However, interviews with a subgroup of refusers did not confirm this hypothesis; rather, they showed that refusers were primarily motivated by private concerns and had seldom or never been interviewed before. Also, in line with the findings of other panel studies, we experienced a low refusal rate in Phase II (13%), even among respondents in the "long" interview group (15%).

Respondent burden is an important issue for the performance of survey research, and the work we have done so far has only begun to alert us to its many aspects and implications. Much more work is needed before we can develop a comprehensive model of the causes and effects of respondent burden and their consequences for the conduct of survey research. Among high research priorities are the following:

1. We need to see if the burden factors that we have identified emerge in other populations (e.g., low-income groups), in other data collection contexts (e.g., in phone surveys), and perhaps most important, in surveys having higher respondent salience. Thus, a test of the burden associated with health surveys would be especially informative.
2. We need additional research to determine why some respondents are more burdened than others as well as the conditions which might reduce burden for the former.
3. It is also important to know more about the effects of respondent burden on survey quality. The fact that burdened respondents may become refusers seems to us to be less of a problem than the fact that they may become poor respondents who answer inaccurately. The issue of respondent burden and data quality deserves further attention.

Footnote

[1] Previous survey participation is based on a question asking about surveys other than Phase I of the experiment, which is classified under panel participation.

Discussion: A field approach for obtaining physiological measures in surveys of general populations *and* Dimensions and correlates of respondent burden

Wornie L. Reed, Department of Sociology, Washington University (St. Louis)

These two papers advance the discussion of methods of improving data collection in survey research, although they do so along different lines. The paper presented by McKinlay addresses issues concerning the epidemiological approach, and the paper by Frankel and Sharp addresses issues concerning the more general health services research survey. A strong inference that I gathered from the papers is the probable significance of the salience of the research project to potential respondents.

The McKinlay study may represent an important breakthrough in health survey research. Physiologic measures are not usually collected, even though such data may be desirable. From this study it appears that the method of collecting physiologic measures in the field is effective. It is also less expensive than bringing subjects to a central point for testing. Further, the method has a higher response rate than the two-step centralized location method—that is, if these findings are replicated in other studies.

An interesting question about the McKinlay study is why this method works so well. Why is it so efficient, not so much in terms of reliability but as a means of data collection? In spite of the unpleasantness of being stuck by a needle, being weighed, and being measured for height, respondents participated at an adequate rate. I think one reason for this is that the subjects received some immediate benefit—a free health screening examination. Some of the examination results were given immediately, and results of the blood tests were sent to subjects later. In addition, this benefit was received with much less bother than would be involved in going to a central site for testing. With the increasing public attention to matters of health and the increasing costs of physical examinations, it is not difficult to appreciate the importance of these benefits as they may have been perceived by the subjects.

A second issue provoked by the McKinlay paper is the question of the applicability of this type of data collection. Obviously, it would be useful in a similar type of epidemiological study, where the objective would be the search for specific undiagnosed illnesses in a community. On the other hand, it may not be a reasonable substitute for the methods used in the Health and Nutrition Examination Studies (HANES). McKinlay's research team collected measurements of height, weight, blood pressure, and blood sampling; however, the HANES studies conducted more complete examinations, including X-rays, urine analysis, EKGs, and various clinical examinations, and developed disease diagnoses. Thus, it may be difficult or impossible to conduct HANES-type studies in that manner, unless the HANES studies are reduced in scope in the current era of budget trimming.

The first set of findings from the study of respondent burden are reassuring in that they appear to correspond to intuition. I would think that in any list of the negative aspects of participating in a survey interview, potential or actual respondents would include terms approximately equivalent to "the perceived uselessness of the survey," "personal time concerns," and "privacy concerns." On the other hand, intuition may suggest that individuals who refuse to participate in a survey may differ from those who do participate. While they may not differ on data relevant to the substance of the research, they may very well differ on the reason why they did or did not participate. In the absence of good evidence suggesting otherwise, I am inclined to accept this latter view.

The authors have addressed specific issues relative to respondent burden. These are (1) the derivation of factors to measure the dimensions of respondent burden and (2) the assessment of the relative contribution of a set of selected factors to respondent burden. In considering the implications of this research, the question arises as to whether respondent burden is a part of a unidimensional phenomenon that includes, on the one end, the configuration of negative feelings by participants and, on the other end, more deeply held negative feelings by nonparticipants. If the answer to this question is yes, then we may, to some extent, be able to sustain or increase participation in surveys by acting on the results of studies of participants. On the other hand, if the answer to this question is no, the attention to lessening respondent burden by using data collected from participants may be less than fruitful. Our efforts may be misdirected if the configurations of negative feelings toward survey participation held by nonparticipants are from a different dimension than that of participating respondents.

There is evidence that other factors lead some individuals to decline participation in surveys. One such factor is an individual's predisposition at the time of the interview. For example, Cartwright (1967) found that persons who were dissatisfied with their housing situation were less willing to participate in their survey. In my own current survey of older persons in St. Louis, an overwhelming majority of the respondents actually enjoyed the interview; they appreciated the opportunity for social interaction. In fact, interviewers had to be cautioned against spending too much time in the interview.

On the other hand, more than 20% of the potential respondents declined participation. Many expressed the desire to be left alone. Of course, some of this "negative" behavior could be related to the issue of "privacy concerns," as discussed by Frankel and Sharp. Nevertheless, there are indications that reasons for not participating cannot always be inferred from data collected from participants.

As previously stated, the factors derived in this study to measure respondent burden are consistent with general expectations. Consequently, I concur with the implicit assumptions of the authors that these factors measure a significant portion of respondent burden whether the attitude is held by a respondent or a nonrespondent. However, the nonrespondents might differ from respondents on either the importance of the three factors or some additional attitudinal reason for not participating. One type of attitudinal question that may be helpful would be one that measures the salience of the research subject to potential respondents. I wonder if there might be some direct or indirect measure of salience in the substantive data on housing collected in this study.

The importance of the salience of the project for a participant is implied also in the McKinlay paper. Individuals participated at high rates even though there was some bother and discomfort. However, they received a free health screening examination. Thus, the project had high salience because of the benefits received. On the other hand, by the authors' estimates, the respondent burden study which was based on a housing survey had only moderate salience. It may very well be the salience of the project that affects the size and distribution of the factors that measure respondent burden.

Open Discussion: Session 3

The chairperson asked the speakers if they had any responses to the discussants' comments. John Anderson said that in his study "best codes" were not assigned simply on the basis of the Interviewer-Administered Form (AIF). He indicated that instances of overreporting (false positives) on the Self-Administered Form (SAF) and underreporting (false negatives) on the IAF had both occurred. He concluded that contrary to the discussant's suggestion, there was no undue bias in the direction of reporting more dysfunction in arriving at "best codes."

To the criticism that impediments, such as the "lift-in-the-shoe" example, did not represent significant dysfunction, Anderson replied that they had used a "strict constructionist" interpretation of the rules for the scale and felt that limitations of the sort involving inability to engage in gymnastic activities constituted real physical dysfunction.

He indicated that Kasper is correct in noting that this study does not distinguish between objective and subjective physical disfunction, inasmuch as both the SAF and IAF ultimately depend on subjects' perceptions of their physical functioning. Kasper noted that this remained a continuing difficulty in quantifying dysfunction.

Elinson inquired as to whether Verbrugge had examined panel attrition by occupational status. Verbrugge replied that she had not, but thought it was a good idea. She went on to note that, although dropouts seemed to have disordered lives or to have been going through particularly turbulent times, stressful life events, contrary to expectations, were not associated with discontinuance.

Kovar asked J. Anderson how he was going to use these scales. She noted that his paper stated that they were to be employed as measures of program effectiveness and asked if he would give some examples of this. J. Anderson replied that they would be used in pre- post-designs, thereby quantifying program impact on functioning. He pointed to an evaluation of PKU screening in New York State as an instance of effective application of the scale.[1]

J. Anderson wondered if Kovar's objections were based on the Physical Activity Scale (PAS) discussed in the paper or on the characteristics of the other two scales comprising the Quality of Well-Being (QWB) scale. Kovar asked if there was much disagreement between IAF's and SAF's for the other two QWB components. Anderson said that there was more difference between forms on one of the scales and less on the other.

Greenberg then asked about the amount of time that had intervened between form administrations in Anderson's study. Anderson replied that they were no more than an hour apart.

Bryan asked J. Anderson to elaborate further on his use of sensitivity and specificity; he thought one usually sought to optimize one or the other depending on the situation. Anderson responded that sensitivity was "what you're after" (A) and specificity was "what you're not after" (non-A). If you have a good test both will be high. Bryan asked which he would optimize and Anderson responded that he wanted both.

Fuchsberg asked Verbrugge how she got her respondents to participate in so arduous a task. She said she had told prospective respondents of the need to know about fluctuations in day-to-day health in the general population and about self-care. She also pointed out that while respondent motivation was important, so was interviewer motivation. Interviewers, she said, were often inexperienced in recruiting subjects for panel studies and often "hedged" in response to what they perceived to be excessive demands on the respondents. Therefore, careful attention must be paid to what interviewers do at the beginning of these surveys as well as to what can be done to help respondents.

Warnecke asked Verbrugge what the response experience was in her study. She said that about 70% of the study stayed in the full six weeks; 9% refused the initial interview; 9% agreed, but then did not keep diaries; and 9% dropped out in the course of the study.

Bradburn noted the similarity between characteristics of Verbrugge's panel dropouts and those in other studies. He suggested that these characteristics should be taken into account in the initial design of studies. He noted that researchers currently try to keep all the subjects in their studies and improve the reporting on the part of subjects who do not follow protocols. He suggested that an intermediate strategy might focus resources on groups at high risk of dropping out, such as those who report illness at the time of the initial interview.

Bradburn offered some other suggestions for making panel tasks palatable. Ascertaining whether people's degree of life "irregularity" might put them at high risk of dropping out at the beginning of the study might allow the application of special techniques, such as routinizing protocol adherence, to help them stay in the study. He also suggested that long tasks might profitably be broken into smaller ones to minimize respondent burden. Another possibility would be to provide paging devices as reminders of times at which study participants should record relevant information.

Axelrod asked if the age/gender distribution of dropouts could be explained in terms of the number of

outside activities. He also wondered if the decline in symptom reports could be due to seasonality. Verbrugge said the study was spread out over six months and does not reflect seasonal trends. Beed asked if Verbrugge had personal as well as telephone contact with subjects during the course of diary keeping. She stated that they did not as a matter of protocol because of the expense involved.

Axelrod noted that while payment had been shown effective in increasing response, it probably had no bearing on the respondents' concerns about privacy or the perceived uselessness of the study.

Frankel responded that since response depended on a host of completely interrelated factors, it was impossible to predict what impact changing one study design factor might have on overall response rates.

Monteiro noted the inconsistency between salience as an important factor and the finding that the sickest people were the ones most likely to drop out. Reed suggested that salience probably differed for respondents, interviewers, and principal investigators. Monteiro commented that perhaps health surveys were not as salient for the sick as sickness surveys would be.

McKinlay commented that her preliminary results indicated that study refusers were mostly young and gave excuses like they "had just been to the doctor and didn't need an examination." Although all the results are not yet available, she commented that sicker people in her study seemed to participate because of their concern over things like cholesterol level. This may have been due to prior treatment or to unsubstantiated suspicions and to fears of obtaining conventional primary care. She said we may ultimately find that these features of our study offset the tendency for the sick to refuse.

Warnecke noted that in his study of cancer patients, involving interviews of up to three hours, he achieved a 90% response rate. He felt that salience was an extremely important factor and cautioned the group not to accept too readily the findings indicating that sick people do not respond.

Frankel commented that, as opposed to McKinlay's study, the old rather than the young refused to participate in her study.

Kasper said that her earlier comments in the discussion of Verbrugge's paper were meant only to suggest that sicker people would be likely to participate in panel studies, not that they would be less likely to participate in a single cross-sectional interview.

Elinson commented that salience had to be considered in the context of the population samples. Certain issues have high saliency for certain populations and little saliency for general populations.

Sirken asked McKinlay if she was doing analyses of the outliers among her subjects, since these were the people of interest in most epidemiological studies. McKinlay responded that outliers were not the focus of the study. Her concern was with the distribution of coronary risk factors in the population, but that she would ultimately examine outliers.

Sirken noted that the concern with respondent burden in the Frankel study ignored the fact that much of what we do to enhance response rates could be viewed as additional respondent burden. Thus, although we tend to investigate one design factor at a time, when it comes to research we must contend with the interactive effects of multiple factors. The very people added to a study through intense followup efforts may be the ones who later manifest the impact of this burden by dropping out or providing invalid information.

Maurer stated that in the HANES survey most of the subjects agreed to participate in the initial interview. Most of the nonresponse occurred three weeks later when they were asked to come to a central exam center. From the data from the initial interview, HANES staff determined that there were no major differences in medical history between those who agreed to participate initially and those who eventually participated in the study. It was true, however, that there were demographic differences between responders and nonresponders, principally that children were more likely to participate, probably due to parental concerns about their child's health. Maurer further expressed his approval of the results McKinlay had presented.

McKinlay noted that her study was based on the 1978 Canadian Health Survey in which a home fitness test had also been conducted. She went on to indicate that procedures such as drawing blood apparently had little effect on participation rates. Preliminary results indicate that 95% of those participating in the rest of the protocol consented to having blood drawn.

Verbrugge made a final comment reiterating the importance of a "sickness" rather than a "health" oriented presentation of a study to improve a survey's salience for the sick.

Footnote

[1] Bush, J. W., M. M. Chen, and D. L. Patrick, 1973, Analysis of the New York State PKU Screening Program Using a Health Status Index. Report to the State Health Planning Commission, New York, N.Y. NTIS/PB 243-585.

SESSION 4:
Use of records in health survey research

Chair: Daniel Horvitz, Research Triangle Institute

Recorder: Gordon Bonham, Division of Health Interview Statistics, National Center for Health Statistics

Consumer knowledge of health insurance coverage

Daniel C. Walden, National Center for Health Services Research

Constance M. Horgan, National Center for Health Services Research

Gail Lee Cafferata, National Center for Health Services Research

Introduction

How knowledgeable consumers are about their health insurance coverage is an important issue in view of policy recommendations aimed at containing health care costs by fostering competition among medical care providers and insurers in the market for health services. These recommendations place great faith in the consumer's ability to exercise an informed choice among various health insurance options that would reflect the consumer's own risk aversion and tastes for medical services. Proposals which rely on consumer choice among competing health plans or on changes in the tax treatment of health insurance premiums make a basic assumption that consumers either are or can be well informed about insurance coverage (Langwell and Moore, in press).

The issue of consumer knowledge is also of interest because of its potential influence on the use of health services. There is evidence that depth and breadth of insurance coverage is related to the use of health services (Newhouse et al., 1982), and knowledge of health insurance benefits is important if the insured are to use the benefits for which they are covered. Ignorance of particular benefits will lead to the erroneous perception of lack of coverage and may affect the decision to initiate or continue care. Conversely, the incorrect perception of having coverage can result in unanticipated medical expenses. Thus, being over- or under-informed about one's health insurance may distort the use of medical care.

In this paper three aspects of consumer knowledge about health insurance are examined: (1) knowledge about the fact of any coverage; (2) understanding of the amount of the health insurance premium, who pays it, and in what proportion; and (3) knowledge about coverage for specific services. Although findings are also discussed with reference to policy considerations, the major focus of this paper is on methods. We will examine the type of respondent in a household survey who is able correctly to report on his or her health insurance coverage; whether overall this kind of information can be reliably obtained only from insurers; which respondents report that they are not covered when in fact they have employment-related health insurance coverage; and whether respondents are more knowledgeable about particular aspects of their insurance coverage than about others. The findings in this paper are based on a survey of insurers and of employers conducted to verify information on health insurance reported in a household survey.

In general, the literature on knowledge of health insurance suggests that consumers are fairly well informed about certain aspects of their coverage, such as whether or not they have any coverage and whether they have coverage for hospital stays, but that there is a lack of knowledge about other benefits. These findings are consistent across studies, including an early study designed to provide estimates of underreporting and overreporting of insurance in household interviews (USNCHS, 1966). A study of subscribers who joined or changed plans during the open enrollment period at a large university found a lack of detailed knowledge about the extent of coverage and limitations of the plan selected, although the decision to join or change plans was related to characteristics of the plan (Moustafa et al., 1971). Knowledge of the extensiveness of benefits for covered services was found to vary with the complexity of the benefit structure (Marquis, 1981). A study comparing federal employees enrolled in an HMO with those enrolled in a Blue Cross/Blue Shield Plan found that while HMO enrollees were generally more knowledgeable, those with more knowledge in both plans were more likely to make a physician contact and to average more contacts than people with little knowledge of plan benefits, although the relationship between previous use and knowledge was noted (Riedel et al., in press). By contrast, a study using national data found that knowledge of health insurance benefits was not directly related to use of physician ambulatory care, although there was a relationship between knowledge and how the policy was obtained as well as the extent of the employer's contribution (Andersen and Daughety, 1979). Another national study, which did not use the insurance policy to verify self-reported coverage but concentrated on general knowledge, found that overall knowledge regarding health insurance is low. The average proportion of cor-

We gratefully acknowledge the assistance of Pamela Farley, the helpful comments of Marc Berk, Judith Kasper, Louis Rossiter, Gail Wilensky, and Renate Wilson, the careful clerical support of Barbara Bottazzi, John Carrick, and Martha Hartley, National Center for Health Services Research, and the programming support of Jack Bieler, T. J. Jefferson, Karen Pinkston, Jeff Sussman, and Cathryn Zucker, Social and Scientific Systems, Inc. The views in this paper are those of the authors and no endorsement by the National Center for Health Services Research is intended or should be inferred.

rect responses in this survey was between 40% and 65% (Arthur D. Little, Inc., 1980).

The data

The National Medical Care Expenditure Survey (NMCES) of a sample of the civilian, noninstitutionalized population of the United States is the data source for this paper. It provided detailed information on health insurance coverage and on expenditures for and use of health services for the calendar year 1977. NMCES was funded by the National Center for Health Services Research, which cosponsored the survey with the National Center for Health Statistics. Data collection for NMCES was carried out by Research Triangle Institute and its two subcontractors, National Opinion Research Center of the University of Chicago and ABT Associates, Inc. Data processing support was provided by Social and Scientific Systems, Inc.

Data were collected in three separate but complementary survey components.

1. The Household Survey (HS) which collected information from 14,000 randomly selected households each interviewed six times over a 15-month period during 1977 and 1978. During the second round of interviewing, a Health Insurance Supplement was administered to collect information on health insurance plans covering members of the household. In this supplement, the person in the family identified as most knowledgeable about health insurance was asked questions about the premiums and benefits of policies held by family members. (See Appendix B for questions in the Health Insurance Supplement which are the basis for this analysis).
2. The Health Insurance/Employer Survey (HIES) which collected, for each private health insurance plan reported in the household survey, data from employers, insurance carriers or other insuring organizations including information on coverage, premiums, and benefits.
3. The Uninsured Validation Survey (UVS) which collected data from the employers of individuals in the household sample who reported that they were not covered by insurance through their employer. The purpose of this survey was to confirm that these individuals were in fact not covered by employment-related insurance and to obtain information about the benefits and premiums of individuals who were found to be insured.

HIES and UVS respondents were also asked to provide a copy of the policy or certificate describing the subscriber's benefits. Information from these policies, abstracted onto insurance code forms, served as the basis for the record check of insurance coverage as reported in the household survey component of NMCES.

The findings are presented in three parts: health insurance coverage (Tables 1–3), premiums (Tables 4–7), and specified benefits (Tables 8–9). The structure of the analysis was particularly influenced by the work of Andersen and Daughety (1979).

Weighted estimates from the household data and the HIES/UVS data of the population with private health insurance and with coverage for specific benefits by public and private plans are shown in Tables 3 and 8. All other tables are based on unweighted data, and the specific reports of household respondents are compared with those of their insurers. Since there were some missing data in both the HS and HIES/UVS, these comparisons are limited to persons with complete data from whom appropriate weights were not available. In discussing the findings, it is assumed that the information reported in the HIES/UVS is reported without error. (See Appendix A for a discussion of data collection and procedures used for data editing and cleaning).

Findings

Private health insurance coverage. The impact of verification data on the estimates of persons with private health insurance is shown in Tables 1–3. The first two tables are based on unweighted numbers of persons with private health insurance coverage for whom complete data are available from HIES/UVS. Table 3 compares national weighted estimates of persons with private health insurance coverage obtained from the HS with national estimates from the HIES/UVS.

Table 1, which is limited to persons with complete HIES data, shows that most persons covered by private health insurance according to the HS were also reported as covered by at least one verification respondent in HIES (99.9%). By contrast, agreement between the HS and HIES on the specific plan covering household respondents existed only for 77.2%. According to the HIES verification data, 22% of the persons covered by private health insurance according to the HS had coverage by at least one private plan that was not reported in

Table 1
Unweighted comparison of Household Survey (HS) and Health Insurance/Employer Survey (HIES) of persons with private health insurance coverage[a]

	Number of Persons	*Percent of Persons*
Household Survey		
Covered by private health insurance	15,592	100.0
Health Insurance Employer Survey		
All household reported plans covered	12,038	77.2
All household reported plans covered, plus coverage by discovered plans	805	5.2
Some household reported plans covered, no plans discovered, and at least one household reported plan not covered	2,435	15.6
Some household reported plans covered, coverage by some discovered plans, and at least one household reported plan not covered	168	1.1
No household reported plans covered	146	.09

[a]Limited to persons with complete HIES data.
Source: National Center for Health Services Research.

Table 2
Unweighted comparison of Household Survey (HS) and Uninsured Validation Survey (UVS) of persons covered and not covered by private health insurance[a]

Uninsured Validation Survey	Persons in the UVS[b]		Household Survey			
			Covered by private health insurance		Not covered by private health insurance	
	Number	Percent of total UVS	Number	Percent of total UVS	Number	Percent of total UVS
Total	4,409	100.00	3,029	68.7	1,380	31.3
Covered by employer plan	798	18.1	632	14.3	166	3.7
Not covered	3,611	81.9	2,397	54.4	1,214	27.5

[a]Limited to persons with complete UVS data.
[b]Including dependents.
Source: National Center for Health Services Research.

the HS and/or at least one private plan for which they reported coverage but which was found to be not held. Specifically, 15.6% of all persons with private health insurance in the HS had at least one reported plan that in fact provided coverage and at least one reported plan that did not provide coverage, and for about 6% of the persons covered by private health insurance according to the HS, the HIES data indicated that they were also covered by plans that had not been reported in the HS, the so-called discovered plans. For about 17% of the persons covered by private health insurance according to the HS, the HIES data indicated that they were not covered by plans that had been reported in the HS, the so-called rejected plans.

Table 2 shows unweighted comparison of HS and UVS data on persons covered by private health insurance. The persons included in UVS were those who according to the HS were employed but not self-employed and not covered by health insurance obtained through their employer. The UVS respondent confirmed that 81.9% of these persons and their dependents did not have health insurance obtained through the employer of the employed HS member. However, about 68% were covered by private insurance, i.e., obtained other than through the employer as reported in the HS. Only 3.7% of persons for whom the HS data indicated no private health insurance coverage were reported by the UVS respondent to have coverage.

Table 3, showing weighted comparison of HS and HIES/UVS estimates of the population ever covered by private health insurance in 1977, indicates no differences in these estimates, whether for the total population or for groups defined by age, sex, race, and education of household head. There were differences by family income, however. Persons with less than $12,000 income were more likely to be classified as covered by the HS data than by the HIES/UVS data (62.5% and 53.9% respectively). The reverse occurred for persons in the highest income category, where the estimate for the percentage covered in the HS was 89.8% and 95.3% in HIES/UVS.

Premiums. Tables 4–7 compare premium information from HS and HIES/UVS. Again, the person in each household who was identified as being the most knowledgeable about the family's health insurance coverage was asked questions about the source of payment and the amount paid by each source for health insurance premiums. These questions related only to private health insurance coverage and not to public insurance such as Medicare and Medicaid. The same information on premiums was obtained through the HIES/UVS.

This analysis pertains to all plans held by a family. In some families the most knowledgeable person was not a beneficiary of every plan, but the plans are described

Table 3
Weighted comparison of Household Survey (HS) and HIES/UVS estimates of population ever covered by private health insurance in 1977

	Household		HIES/UVS	
	Percent ever covered by private	Percent never covered by private	Percent ever covered by private	Percent never covered by private
Total	78.8	21.2	79.9	20.1
Age in years				
Less than 19	76.9	23.1	78.0	21.9
19–24	73.8	26.1	76.8	23.2
25–54	83.6	16.4	84.2	15.8
55–64	83.2	16.7	84.9	15.0
65 or older	68.9	31.1	68.6	31.3
Sex				
Male	79.4	20.6	80.8	19.2
Female	78.3	21.6	79.0	21.0
Race				
White	81.6	18.4	82.6	17.4
All other	59.5	40.5	61.2	38.8
Education of Head				
Less than 9 years	63.5	36.5	64.8	35.2
9–11 years	72.7	27.3	73.9	26.1
12 years	83.5	16.5	86.6	13.4
13–15 years	86.7	13.3	85.9	14.1
16 years or more	91.1	8.9	91.7	8.3
Family Income				
Less than $12,000	62.5	37.5	53.9	46.1
$12,000–$19,999	85.1	14.8	92.5	7.5
$20,000 or more	89.8	10.2	95.3	4.7

Source: National Center for Health Services Research.

Table 4
Unweighted comparison of Household Survey (HS) and HIES/UVS for knowledge of existence of family out-of-pocket premium expense by most knowledgeable person

	Out-of-pocket premium expense						
HS / HIES/UVS	Yes / Yes (1)	No / No (2)	Yes / No (3)	No / Yes (4)	D.K. / Yes (5)	D.K. / No (6)	Percent correct[a]
	Percent distribution						
All families	52.4	25.7	5.2	15.3	0.6	0.8	78.1
Age in years							
Less than 25	40.4	33.5	6.2	17.7	0.5	1.7	73.9
25–54	44.7	31.6	5.6	16.6	0.6	0.8	76.3
55–64	58.1	21.2	6.5	12.9	0.7	0.6	79.3
65 years or older	74.2	9.6	2.6	12.8	0.4	0.4	83.8
Sex							
Male	49.5	26.2	6.1	16.5	0.7	0.9	75.7
Female	55.8	25.2	4.1	13.9	0.4	0.7	81.0
Race							
White	53.3	24.9	5.0	15.4	0.6	0.8	78.2
All other	43.8	33.3	6.9	15.1	0.6	0.2	77.1
Education of most knowledgeable person							
Less than 9 years	57.0	21.0	7.2	13.7	0.5	0.6	78.0
9–11 years	49.2	29.1	5.0	14.9	0.5	1.3	78.3
12 years	52.4	27.8	3.9	14.5	0.6	0.7	80.2
13–15 years	50.6	24.8	6.1	16.9	0.7	0.9	75.4
16 years or more	52.1	23.8	5.4	17.4	0.6	0.5	75.9
Family income							
Less than $8,000	59.8	22.5	4.1	12.8	0.4	0.3	82.3
$8,000–$13,999	51.7	27.2	6.5	13.5	0.3	0.8	78.9
$14,000–$19,999	50.5	28.6	4.5	15.0	0.6	0.9	79.1
$20,000–$24,999	47.4	25.1	5.9	20.0	1.0	0.6	72.5
$25,000 or more	49.5	25.3	5.3	17.7	0.8	1.4	74.8
Perceived health status							
Excellent	50.2	28.0	4.8	15.5	0.7	0.8	78.2
Good	52.4	24.9	5.5	15.8	0.5	0.9	77.3
Fair	59.4	20.1	5.1	14.7	0.3	0.5	79.5
Poor	57.5	25.7	6.0	9.0	1.2	0.6	83.2
Insurance type from HIES/UVS							
Both group and non-group	84.6	1.4	1.8	11.6	0.4	0.2	86.0
Nongroup only	85.9	1.4	1.7	10.3	0.7	0.0	87.3
Group only—family pays all	62.7	0.0	0.0	36.4	0.9	0.0	62.7
Group only—family pays none	0.0	82.2	15.3	0.0	0.0	2.5	82.2
Group only—family pays some	70.0	0.0	0.0	29.0	1.0	0.0	70.0
Out-of-pocket family premium from HIES/UVS							
Less than $200	34.4	42.9	8.7	12.3	0.4	1.3	77.3
$200–399	79.8	0.0	0.0	19.6	0.6	0.0	79.8
$400 or more	78.8	0.0	0.0	20.2	1.0	0.0	78.8
Total family expense for physician visits							
None	48.0	29.1	5.2	15.9	0.9	0.9	77.1
$1–$99	52.0	26.2	5.0	15.6	0.4	0.8	78.2
$100–$249	54.7	24.6	4.9	14.5	0.7	0.6	79.3
$250 or more	51.5	25.7	5.7	15.7	0.6	0.9	77.2

[a]Corresponds to a Yes/Yes and No/No.
Source: National Center for Health Services Research.

according to the demographic characteristics of the most knowledgeable person. Sources of payment are divided into two categories: out of pocket (family) and other (usually employers but occasionally another source such as a union).

Table 4 shows an agreement rate of 78.1% between the HS and HIES/UVS regarding the existence of any out-of-pocket payments by the family for health insurance premiums. About one-half of most knowledgeable persons in a household correctly state that their families pay some amount of money out of pocket for health insurance, and one-quarter correctly state that their families pay no out-of-pocket amount for health insurance premiums. Thus, for over 20% of families, the most knowledgeable person has an incorrect perception regarding the existence of out-of-pocket premiums. Most of this occurred where the HS indicated that the family did not pay for health insurance premiums and the HIES/UVS reported the opposite. Only 1% of most knowledgeable persons reported not knowing whether their family had any out-of-pocket premium payment.

The data in Table 4 also permit comparison of the percentage of families with out-of-pocket payments for health insurance premiums according to the HS (sum of columns 1 and 3) and according to the HIES/UVS (sum of columns 1, 4, and 5). Based on the HS, it is estimated that 57.6% of families pay out of pocket for health insurance premiums, compared to an estimate of 68.3% in HIES/UVS, a difference of 10.7%.

Higher HIES/UVS estimates for the percentage of families with out-of-pocket premium payments were generally observed across all characteristics of the most knowledgeable persons that were examined, although the relative differences varied within groups defined by these characteristics. The largest ranges were observed among age groups, income categories, and type of insurance. Almost 74% of most knowledgeable persons less than 25 years of age were in agreement with HIES/UVS, compared to almost 84% for persons 65 years of age or older. The range by income was smaller, with approximately 75% of persons from families with incomes over $25,000 correctly reporting whether their family had out-of-pocket payments for premiums, compared to 82.3% for persons from families with incomes less than $8,000. The latter were also slightly more likely to have had an out-of-pocket payment as determined by the HIES/UVS (73%) in contrast to 68% of those from families with incomes greater than $25,000. Not surprisingly, the largest differences were found according to the kind of insurance held by the family. If a family had only nongroup coverage, 87% of the most knowledgeable persons correctly stated that there was an out-of-pocket premium payment. For families with group-only coverage, correct responses were much less likely. Of those with group-only coverage, knowledge was highest for those where the family had no out-of-pocket premium payments (82.2%) and lowest for those where the family paid the entire premium (62.7%).

Table 5
Unweighted comparison of Household Survey (HS) and HIES/UVS for knowledge of amount of total family out-of-pocket premium expense by most knowledgeable person: dollar and percent differences

	Out-of-pocket premium expense					
	HS less than HIES/UVS[a]	*HS equals HIES/UVS*	*HS greater than HIES/UVS*[b]	*HS premium unknown*	*All families*	*All families cumulative frequency*
	Percent of all families					
Dollar difference						
None	1.9	30.1	0.4		32.4	32.4
1–50	12.2		9.7		21.9	54.3
51–100	5.9		4.0		9.9	64.2
101–300	11.3		8.0		19.3	83.5
301–500	4.2		2.7		6.9	90.4
500 or more	5.5		2.2		7.7	98.1
Unknown		30.1		1.8	1.8	99.9
All differences	41.0		27.0	1.8	100.0	
Percent difference						
Less than or equal to 1.0%	2.7	30.1	1.1		33.9	33.9
1.1–10.0%	5.8		5.9		11.7	45.6
10.1–30.0%	7.5		3.4		10.9	56.5
30.1–50.0%	3.8		2.0		5.8	62.3
50.1–75.0%	3.8		1.8		5.6	67.9
75.1 or more	17.4		12.9		30.3	98.2
Unknown				1.8	1.8	100.0
All differences	41.0	30.1	27.7	1.8	100.0	

[a]Includes discovered policies; assumes HS family premium = $0.
[b]Includes rejected policies; assumes HIES family premium = $0.
Source: National Center for Health Services Research.

Differences according to sex, education, and perceived health status were small. Females and those who reported poor health status were slightly more knowledgeable regarding out-of-pocket family premium payments, and those who had more than twelve years of education were slightly less likely to correctly report whether their family had out-of-pocket premium payments. There were no differences according to race nor according to the amount of the out-of-pocket premium as determined by HIES/UVS and the total expenditures for outpatient physician visits by all family members.

A much smaller percentage of HS answers matching HIES/UVS estimates was found with regard to the amount of total family out-of-pocket premium expenses. Only 30% of most knowledgeable persons reported out-of-pocket premium amounts for their families that corresponded to HIES/UVS reports (Table 5). In other words, 41% of persons underestimated the amount that the family paid out-of-pocket for premiums, 27% overestimated this amount, and 2% reported not knowing the amount paid out of pocket for premiums.

Both under- and overreports of premium amounts in the HS were not large for 22% of persons (within $50 of what was reported in the HIES/UVS), but as many as one-third of the respondents reported incorrectly by more than $100. Also, although for 45% of families there was either no discrepancy or a discrepancy of less than 10% between the HS and HIES/UVS reports of out-of-pocket premium payments, the discrepancy was greater than 75% for over 30% of families.

The discrepancies between the HS and the HIES/UVS are more likely to be larger in absolute dollar terms when the HS is the lower of the two estimates. When the discrepancy is greater than 75%, the HS estimate is also more likely to be lower.

Comparisons of HS and HIES/UVS reports of other sources of payment for premium expenses indicated agreement for 73.7% of the families (Table 6). This rate is slightly lower than the 78.1% observed regarding knowledge about the existence of family out-of-pocket premium payments.

Almost half of most knowledgeable persons correctly stated that some source other than their family paid all or part of the health insurance premium, and one-quarter reported that there was no other source of payment. Of the almost 12% of families in which the most knowledgeable person had an incorrect perception regarding the existence of other sources of payment, about 65% stated that there was no other payer although other payers were reported in the HIES/UVS. Almost 15% reported not knowing whether there were any other sources of payment for their health insurance premiums than the family.

Although the percentage of most knowledgeable persons who correctly report on the existence of family out-of-pocket payment for health insurance premiums is similar to the percentage who accurately indicate other payers, the distribution with respect to incorrect responses is different. For knowledge about the existence of out-of-pocket premium payments, 20% were incorrect and 1% did not know; for knowledge of other sources of premium payment, 12% were incorrect and 15% did not know. This suggests that while people are likely to make statements about family out-of-pocket payments for health insurance premiums even when they are incorrect, they tend to admit that they do not know about other sources of payment

Table 6 also yields estimates of the percentage of families with other sources of payment for health insurance premiums according to both the HS (sum of columns 1 and 3) and the HIES/UVS (sum of columns 1, 4, and 5). The report of the most knowledgeable person in the HS yields an estimate of 52.2% of families having other sources of payment for health insurance premiums, while the HIES/UVS yields an estimate of 67.8%, 15.6% higher than the HS estimate. As for out-of-pocket payments, this pattern of higher HIES/UVS estimates appears across all characteristics examined. Knowledge about other sources of payment for premiums varies with age and is more pronounced than for knowledge of out-of-pocket payments. While only 63.3% of most knowledgeable persons less than 25 years of age were in agreement with HIES/UVS, this rate rises to 81.1% for persons 65 years of age or older. The differential is more pronounced for knowledge about other sources of payment for premiums according to insurance type. If a family had only nongroup coverage, the most knowledgeable person was much more likely to state correctly that there was no other premium payer (almost 85% reporting correctly) than in families with group-only coverage, where knowledge is highest for those where the family has no out-of-pocket premium payments (74.7%) and lowest for those where the family pays the entire premium (49.6%).

There were small or negligible differences in knowledge about other sources of premium payments according to sex, education, family income, and health status. Differences by race were more pronounced, with 75% of whites and 61.7% of all others correctly reporting whether their family had other sources of premium payments.

According to the percentage of the premium that was paid by other sources as determined by HIES/UVS, the highest knowledge scores were obtained when the percentage paid by other sources was zero (78%). To the extent that this category contains nongroup plans, for which the entire premium is typically paid out-of-pocket and which in general are associated with higher knowledge scores, this finding is expected. When there is another payment source, as the percentage of the premium that is paid by these sources increases, the knowledge scores also increase.

There were no differences with respect to the percentage of persons who correctly reported the existence of other sources of payment for premiums according to the family's total expenditures for outpatient physician

Table 6
Unweighted comparison of Household Survey (HS) and HIES/UVS for knowledge of existence of other sources of payment (not out-of-pocket) for premium expense of family as reported by most knowledgeable person

	HS / HIES/UVS	Other sources of payments for premium expense: Yes / Yes (1)	No / No (2)	Yes / No (3)	No / Yes (4)	D.K. / Yes (5)	D.K. / No (6)	Percent correct[a]
		Percent distribution						
All families		48.4	25.3	3.8	7.9	11.5	3.1	73.7
Age in years								
Less than 25		42.4	20.9	4.7	11.8	14.0	6.2	63.3
25–54		59.1	13.0	4.1	7.8	13.9	2.0	72.1
55–64		52.1	23.0	3.8	8.1	10.9	2.1	75.1
65 years or older		17.1	64.0	2.7	6.1	4.2	6.0	81.1
Sex								
Male		53.4	19.8	3.6	8.0	12.1	3.1	73.2
Female		42.2	32.1	4.1	7.7	10.8	3.2	74.1
Race								
White		49.1	25.9	3.8	7.2	10.9	3.1	75.0
All other		41.3	20.4	4.4	13.7	16.7	3.6	61.7
Education of most knowledgeable person								
Less than 9 years		32.0	41.5	2.6	9.7	9.3	4.9	73.5
9–11 years		41.3	28.2	5.2	8.8	13.4	3.1	69.5
12 years		52.8	21.1	3.9	7.5	12.3	2.3	73.9
13–15 years		51.4	22.7	2.8	6.9	12.0	4.2	74.1
16 years or more		58.0	18.4	4.4	6.9	10.0	2.2	76.4
Family income								
Less than $8,000		19.8	54.7	4.3	8.6	6.7	6.0	74.5
$8,000–$13,999		47.9	24.8	4.0	9.0	11.6	2.7	72.7
$14,000–$19,999		58.4	14.4	3.9	7.2	13.7	2.4	72.8
$20,000–$24,999		64.5	11.2	2.9	6.9	13.3	1.3	75.7
$25,000 or more		60.1	13.7	3.5	6.9	13.4	2.4	73.8
Perceived health status								
Excellent		53.0	21.5	3.5	7.3	11.8	3.0	74.5
Good		48.3	25.2	4.0	8.2	11.4	3.1	73.5
Fair		37.4	35.2	3.8	8.1	12.3	3.1	72.6
Poor		31.1	39.5	4.2	10.8	8.4	6.0	70.6
Insurance type from HIES/UVS								
Both group and nongroup		37.9	32.9	4.0	12.0	9.8	3.4	70.8
Nongroup only		0.8	83.7	4.8	3.2	0.6	6.9	84.5
Group only—family pays all		0.0	49.6	33.7	0.0	0.0	16.7	49.6
Group only—family pays none		64.8	9.9	0.9	7.9	15.1	1.0	74.7
Group only—family pays some		72.3	0.0	0.0	10.7	17.0	0.0	72.3
Percent total premium paid by other from HIES/UVS								
None		0.1	77.9	11.7	0.4	0.1	9.7	78.0
1–34		62.4	0.0	0.0	22.1	15.5	0.0	62.4
35–66		66.1	0.0	0.0	16.7	17.1	0.0	66.1
67–99		75.3	0.0	0.0	7.6	17.2	0.0	75.4
100		73.5	0.0	0.0	9.4	17.1	0.0	73.5
Total family expense for physician visits								
None		43.7	23.9	6.4	10.7	11.6	3.7	67.6
$1–$99		46.7	27.6	3.6	7.3	11.4	3.4	74.5
$100–$249		49.3	25.3	3.8	6.9	10.7	3.9	74.6
$250 or more		49.6	24.0	3.5	8.4	12.2	2.3	73.6

[a]Corresponds to a Yes/Yes and No/No.
Source: National Center for Health Services Research.

visits, with the exception of families who had no expenditures. Sixty-seven percent of most knowledgeable persons in these families with no expense for physician visits reported correctly, whereas approximately 74% reported correctly in families where there were such expenditures.

Approximately 28% of most knowledgeable persons reported the amount paid by other sources for family premiums that corresponded to the HIES/UVS report (Table 7). This percentage is similar to the percentage who report the out-of-pocket premium amount correctly. Almost 49% underestimated that amount in the HS, 3% overestimated it, and almost 20% reported not knowing the amount that other sources pay. Compared to out-of-pocket premium amounts, household respondents were much more likely to not know the amount paid by other sources and much less likely to not know the amount paid by other sources and much less likely to overestimate this amount. However, reporting of amounts for both out-of-pocket and other sources of premium payment were more likely to be underestimated by the household.

Less than one-third reported the amounts paid by other sources within $100 of what was reported in the HIES/UVS. Almost 28% reported incorrectly by over $500. Thus, the dollar discrepancies between HS and HIES/UVS reports are much smaller for out-of-pocket premiums than for premium amounts paid by other sources. Where the HS respondent underestimated the amount paid by other sources (48.8%), most underestimated this amount by more than 75%. Similar to the findings for out-of-pocket premiums, the discrepancies between the HS and the HIES/UVS with respect to other payers for premiums are more likely to be larger in absolute terms when the HS was the lower of the two estimates.

Selected benefits. In the HS, the most knowledgeable person in the family was asked if family members were covered or not covered for 13 types of benefits by public and/or private plans. In the HIES/UVS, insurance coders indicated if the policy stated that coverage was provided, definitely was not provided, or if it was not possible to determine from the policy if coverage was provided. Table 8 presents weighted national estimates of the population covered for selected benefits based on the HS and on the HIES/UVS. Table 9 compares unweighted estimates of the most knowledgeable person's report of coverage by private and/or public plans with the HIES/UVS data. Data reported by these respondents on other members of the family covered by different plans are excluded, as are respondents for whom complete HIES/UVS data were not available.

For all benefits except routine dental care, ortl odontia, and eye examination for glasses, the weighte d estimates from the HS of the percentage of the popt lation covered were less than those of HIES/UVS (Table 8).

For inpatient hospital benefits, the estimates c f the percentage of the population covered for semi-p ivate accommodations and inpatient surgery were 10% less in the HS than in HIES. The difference for other inpatient physician services was about 15%.

Table 7
Unweighted comparison of Household Survey (HS) and HIES/UVS for knowledge about amount of premium expense paid by other sources by most knowledgeable person: dollar and percent differences

	Other sources of payment for premium expense					
	HS less than HIES/UVS[a]	*HS equals HIES/UVS*	*HS greater than HIES/UVS*[b]	*HS premium unknown*	*All families*	*All families cumulative frequency*
	Percent of all families					
Dollar difference						
0	0.1	28.2	0.0		28.3	28.3
1–50	1.7		0.6		2.3	30.6
51–100	1.5		0.5		2.0	32.6
101–300	9.6		1.0		10.6	43.2
301–500	8.5		0.3		8.8	52.0
500+	27.4		0.5		27.9	79.9
Unknown				19.9	19.9	99.8
All differences	48.8	28.2	2.9	19.9	100.0	
Percent difference						
Less than or equal to 1.0%	0.1	28.2	0.1		28.4	28.4
1.1–10.0%	0.7		0.4		1.1	29.5
10.1–30.0%	1.1		0.6		1.7	31.2
30.1–50.0%	0.9		0.2		1.1	32.3
50.1–75.0%	1.3		0.1		1.4	33.7
75.1 or more	44.7		1.4		46.0	79.7
Unknown				19.9	19.9	99.6
All differences	48.8	28.2	2.8	19.9	100.0	

[a]Includes discovered policies; assumes HS family premium = $0.
[b]Includes rejected policies; assume HIES family premium = $0.
Source: National Center for Health Services Research.

The underestimates of coverage in the HS for ambulatory benefits were large: 65.8% versus 87.3% for X-rays and tests; 42.4% versus 80.5% for physician services; and 37.0% versus 72.7% for prescription drugs. These differences are believed to be related to first-dollar coverage and/or to the size of the deductible. Differences between the two data sources are smaller when the coverage includes first-dollar coverage (data not shown).

Large differences in the weighted estimates between the HS and HIES/UVS were also found for certain types of care that are used by only a small percentage of the population: for example, the estimate for inpatient mental health care was 27.4% from the HS versus 77.3% from HIES/UVS; for ambulatory mental health services, 23.4% versus 72.4%, and for nursing home 16.7% versus 55.9%. The difference in the nursing home estimates, however, may be related to differences between the HS questionnaire which focused on care in nursing homes and the HIES/UVS coding forms which addressed coverage for extended care facilities.

The percentage of the population who do not know if they are covered is substantial and ranges from 7.3% for inpatient surgery to 46.4% for nursing homes. According to the HS, the percentage who do not know if they are covered is highest for the least used services, where it ranges from 38.2% to 46.4%. For the remaining benefits, the percentage of the population who do not know if they are covered is much smaller and ranges from 7.3% to 16.1%.

It should be noted that there is ambiguity in health insurance policies about coverage for some benefits, according to HIES/UVS. About 11% of the population were covered by policies where even carefully trained and experienced coders could not determine whether psychiatric benefits were included. This ambiguity was smallest (2% or less of the population) for ambulatory physician, ambulatory diagnostic, prescription drug, routine dental care, and the three nonpsychiatric hospital benefits.

When the responses of the most knowledgeable persons are compared with the HIES/UVS data on their public and private plans, over 85% of the HS respondents correctly reported coverage for semi-private hospital accommodations and physician inpatient surgery benefits (Table 9). Lower levels of knowledge for less frequently used but also less expensive services obtained on an ambulatory basis were found (53.8% for prescription medicine benefits, 53.9% for ambulatory physician benefits, and 69.6% for X-ray and diagnostic test benefits). The most common error occurred when a lack of coverage was reported but the verification data indicated coverage was in fact provided. HIES information in this respect indicates only that for a given benefit, some coverage is provided by the policy or policies; it makes no assumptions about first-dollar coverage or the payment of deductibles.

The only frequently used services for which high knowledge scores were obtained were routine dental care (77.5%), orthodontia (68.5%), and eye examination for glasses (73.4%). These services are usually not covered under insurance plans.

Benefits for nursing home care and mental health services were also likely to be inaccurately reported. About one-third of most knowledgeable persons reported correctly on coverage for nursing home care. Only 29% of most knowledgeable persons correctly perceived whether their family had coverage for outpatient mental health services. Approximately 32% accurately reported on their coverage for psychiatric hospitalization. The HS respondent was more likely to report not knowing about mental health benefits than any of the

Table 8
Weighted comparison of Household Survey (HS) estimates and HIES/UVS estimates of the population covered for selected benefits by private and public health insurance in 1977

	Household survey			*HIES/UVS*		
Type of benefit	*Percent covered*	*Percent not covered*	*Percent don't know*	*Percent covered*	*Percent not covered*	*Percent don't know*
Semi-private room in hospital	80.2	12.0	7.9	89.3	10.2	0.5
Physician inpatient surgery	81.4	11.3	7.3	90.0	9.5	0.4
Other inpatient physician	76.0	13.0	11.0	89.1	10.2	0.7
Maternity	60.3	23.7	16.1	83.4	14.3	2.3
Eye examination for glasses	19.8	67.2	13.0	15.2	83.1	1.7
Routine dental care	26.5	65.3	8.3	24.9	75.1	0.0
Orthodontia	16.7	69.8	13.5	9.6	83.9	6.4
Ambulatory X-rays and diagnostic tests	65.8	22.9	11.4	87.3	10.8	1.9
Ambulatory physician	42.4	47.7	9.8	80.5	17.4	2.0
Prescription drugs for ambulatory patients	37.0	52.5	10.5	72.7	26.6	0.7
Ambulatory psychiatric or other mental health care	23.4	38.4	38.2	72.4	17.2	10.4
Inpatient mental health	27.4	27.8	44.8	77.3	11.6	11.1
Semi-private nursing home or similar facility	16.7	36.9	46.4	55.9	23.2	20.9

Source: National Center for Health Services Research.

other specified services examined. Thirty-seven percent reported not knowing about outpatient mental health benefits, and almost 45% indicated that they did not know whether they had coverage for inpatient mental health services.

When HS respondents reported incorrectly, they were more likely to report that they were not covered or did not know if they were covered than to report that they were covered when their policy did not in fact include coverage. For example, for prescription medicine benefits, about 37% of most knowledgeable persons reported they were not covered or did not know if they were covered, although their policies included some coverage for such expenditures. Only about 9% reported coverage or responded that they did not know, when their policies provided no coverage for prescription medicines.

The mean number of thirteen selected services for which families actually had coverage according to the HIES/UVS was 8.7. The mean number of correct responses (defined as yes/yes, or no/no and don't know/don't know) was 7.6. There was little variation across the demographic characteristics of most knowledgeable person, insurance type and family health expenditure levels. Of all variables examined, the lowest knowledge score was 6.7 in families that had no expenditures for physician visits and the highest knowledge score was 8.1 in families with incomes greater than $20,000 (data not shown).

Conclusions

Methodological implications. This paper has addressed a number of methodological issues that are important to health survey researchers collecting population data on coverage by private health insurance, health insurance premiums and source of premium payment, and on the types of benefits covered. Two NMCES data sources were compared: the household survey and the HIES/UVS, which provided verification data from the insurers and/or employers of persons in the household survey.

In comparing these two data sources the assumption was made that the data collection process, including the questionnaire design, interviewing procedures, and coding methods did in fact measure coverage, premiums, and benefits without error. This paper does not address the validity of this assumption or other sources of error. The focus here is on one type of nonsampling error—reporting bias (see Kish, 1965, and Andersen et al., 1979, for a discussion of models of total survey error).

Data on whether someone is covered by private health insurance can be obtained accurately from a household survey with the design and methods used in NMCES. Estimates from the verification data were not different from those from the household survey data for the entire population or for subgroups defined by age, sex, and race. Some differences were found by income. Compari-

Table 9
Unweighted comparison of Household Survey (HS) and HIES/UVS for knowledge of most knowledgeable person's own coverage by public and private insurance for selected health services

	HS / HIES/UVS	Yes / Yes	No / No	Yes / No	No / Yes	D.K.[a] / Yes	D.K. / No	Yes / D.K.	No / D.K.	D.K. / D.K.	Correct Response[b]
Semi-private room in hospital		84.9	0.7	2.0	2.1	9.7	0.1	0.4	0.0	0.0	85.6
Physician inpatient surgery		86.8	0.8	1.3	1.4	9.2	0.1	0.3	0.0	0.0	87.6
Other inpatient physician		79.4	1.0	1.5	3.3	13.7	0.4	0.5	0.1	0.1	80.5
Maternity		52.4	2.6	2.3	17.2	21.6	1.5	1.3	0.7	0.4	55.4
Eye examination for glasses		11.1	62.2	8.6	2.0	2.5	12.3	0.6	0.5	0.1	73.4
Routine dental care		16.3	61.2	7.2	4.0	2.1	9.2	0.0	0.0	0.0	77.5
Orthodontia		4.3	62.6	8.6	1.8	1.0	14.5	1.6	3.8	1.6	68.5
Ambulatory X-rays and diagnostic tests		67.6	1.7	1.5	13.7	13.0	0.5	0.9	0.8	0.3	69.6
Ambulatory physician		46.4	7.3	1.7	31.6	10.1	1.0	0.6	1.0	0.2	53.9
Prescription drugs for ambulatory patients		34.1	19.6	4.3	28.9	7.7	4.5	0.3	0.5	0.1	53.8
Ambulatory psychiatric or other mental health care		20.2	4.7	1.4	22.2	37.1	3.6	2.0	4.4	4.4	29.3
Inpatient mental health		25.4	1.8	0.9	15.0	44.7	1.5	2.6	3.0	5.0	32.2
Semi-private nursing home or similar facility		17.1	5.8	1.7	16.0	32.3	7.5	2.4	7.0	10.2	33.1

[a]DK = do not know.
[b]Yes/yes, no/no, and DK/DK.
Source: National Center for Health Services Research.

sons from these data should also be made, however, for other subpopulations that are of particular policy importance.

The UVS data did not substantially affect the estimates of the number of persons covered by private health insurance obtained through employers. In fact, a sizeable proportion of those eligible for UVS were known from the household data to be covered by private health insurance obtained in some other way, most often from the spouse's or parent's employment-related private health insurance.

If, however, more specific data are required, such as the number of plans, the premiums, source of premium payment, or specific benefits covered, data based on the survey design and methods used in the NMCES household survey are likely to involve substantial reporting bias. The comparison made here suggests that the household survey did not obtain data on all the plans covered by household sample. Although the concept of "a private health insurance plan" may have been viewed differently by household survey respondents and HIES/UVS respondents, it is clear that differences between the two data sources in terms of the number of plans covered are not trivial. Some of these differences, of course, may be simply definitional ones, e.g., Is coverage by Blue Cross and Blue Shield coverage by one plan or two?

To the extent that the household survey did not obtain accurate data on all the plans that covered the members of the sample, the differences between the two data sources with respect to knowledge of sources of premium payment, amounts paid by each source, and benefits covered are not surprising. For about 20% of the NMCES households the most knowledgeable person provided information that conflicted with the verification survey with respect to the out-of-pocket payment of premiums. Moreover, the two data sources agreed on the amount of the out-of-pocket premium payment for about one-third of the households. Similar differences were found with respect to the existence of other payers and the amounts paid. There also were substantial differences between the two data sources on the coverage by both public and private insurance for most of the types of health care considered.

Some of these differences are difficult to interpret. For example, it is not possible to determine when household respondents guessed or when they did not know. Also, the coding procedures used for the HIES/UVS data on plan benefits may have forced the coding of "don't know" when another reasonable interpretation would have been that the policy did not provide coverage. In addition, the HIES/UVS data were taken directly from the policy, and no data were obtained on claims payment procedures that may in some cases conflict with the statement of benefits in the policy. Hence, household respondents may have known more about their effective coverage than is implied by the benefits stated in the policy and reported in HIES/UVS.

The analysis of the correlates of differences with respect to characteristics of the household survey respondents was limited and does not provide a basis for selecting households for whom verification data are necessary in order to obtain more precise estimates. Reporting bias was not concentrated among those groups who are generally considered to report more inaccurately. Before recommendations can be made, the correlates should be examined with multivariate statistical techniques, and the question should be approached in light of the total survey design concept that has been discussed in several of the past methodology conferences.

The present findings differ somewhat from those of Andersen and Daughety (1979) with respect to differences between household survey data and verification data on premiums. Andersen and Daughety found that households tended to overestimate their out-of-pocket premium payments, whereas this comparison found the opposite. Since their data collection techniques and methods employed for the 1970 survey were similar to those used in NMCES, differences between the two surveys are not believed to have been of sufficient magnitude to have resulted in the differences between these two sets of findings. Here too, further research is necessary before definitive conclusions can be reached on the collection of health insurance coverage, premium, and benefit data.

Policy implications. The estimates presented here suggest that American consumers are not in every respect knowledgeable about their health insurance coverage. This finding may have implications for the effectiveness of strategies which rely on market forces as the mechanism to slow the growth in health expenditures; however, it should be kept in mind that present consumer knowledge about health insurance does not necessarily reflect how they will act in a more competitive marketplace.

Many competitive approaches depend on the consumer's cost consciousness in purchasing insurance. The assumption is that knowledgeable consumers can choose health insurance plans that are a reflection of their own preferences with respect to risk aversion and tastes for medical care and will in fact choose to buy health insurance with cost-sharing requirements and benefit limitations such that excessive utilization and spending is discouraged.

One type of strategy would encourage the availability of various insurance options for consumers. Currently consumers have little choice with respect to the group health insurance plans which are typically available through employers. Only 18% of the subscribers in employment-related group plans were offered more than one option in 1977 (Farley and Wilensky, 1982). Of course, consumers are free to purchase nongroup coverage but are unlikely to do so when they have a group option because nongroup plans are more expensive and offer fewer benefits. One could argue that consumers,

who now have for the most part only one realistic option, have too low a level of understanding about their health insurance to be able to choose wisely among whatever options might become available in the future. On the other hand, at present, perhaps consumers really do not need to be well informed about their health insurance. If only one plan is available and there are no options, they will take what is offered and perhaps become better informed when they use or consider using a particular service. Evidence presented in this paper suggests that consumers who purchase their insurance on a nongroup basis (i.e., those who have presumably chosen among several plans) have a higher level of knowledge, at least with respect to premiums, than those with group insurance. Thus, when consumers are given options and have to make choices, they may become better informed.

Another type of strategy focuses on eliminating the present exclusion of employer-paid health insurance premiums from employee taxable income. This paper suggests that the impact of this approach may come as a surprise to many American consumers who greatly underestimate the contribution which their employers make to their total premium payment. It should be kept in mind that, though on an annual basis the employee contribution may be on the order of several hundred dollars, the impact of this amount may be less if it is a deduction from salary and wages divided over several pay periods.

Another finding of this paper is that in some situations, trained coders were not able to determine if an insurance policy covered certain services. This suggests that to the extent that policies are ambiguous and confusing, health insurance policies should be written in a more understandable way.

On a more general level, policy makers developing strategies aimed at influencing insurance purchases must be aware of these findings. We currently know very little about how consumers will behave in a more competitive marketplace; if these policies are to work, American consumers must be knowledgeable about their health insurance coverage, benefits, and premiums.

Appendix A

Health insurance data. The household survey data are based on information provided by household respondents during the first five interviews. The survey reference period was January 1 to December 31, 1977. For the interview instruments see Bonham and Corder (1981). Respondents were asked if anyone in the family currently was covered by any of the following types of insurance: Medicare Part A and Part B; Medicaid; Civilian Health and Medical Programs of the Uniformed Services (CHAMPUS) or Civilian Health and Medical Program of the Veterans Administration (CHAMPVA); private insurance for hospital, dental, or physician services.

Questions were also asked about policies that pay supplemental cash benefits only and policies that cover only dread diseases, such as cancer; however, these are not considered health insurance coverage in this paper. Eligibility for direct provision of health services from the Veterans Administration and through such programs as neighborhood health clinics or migrant worker programs are likewise not considered health insurance coverage.

During the second and subsequent interviews, the family was asked about any change in their public or private insurance coverage including the names of plans added or dropped and the names of family members added to or dropped from existing or new insurance. The data on insurance coverage have been edited for consistency with other information about insurance contained elsewhere in the survey. These edits include the following: periods of nonresponse and noneligibility were adjusted according to previous insurance response so as not to artificially create changes in insurance coverage due to periods of nonresponse or noneligibility; inconsistencies between summary reports and household reports were resolved; edit rules were established whereby ambulatory visits paid by public or private insurance were used to establish corresponding insurance coverage; sources of payment for visits reported in rounds of no insurance for persons otherwise reporting insurance were used as a mechanism of adjustment.

Information supplied by the respondent was amended in each round through the household summary update process, which allowed the respondent to correct or add to the information provided in previous interviews. Trained interviewers then updated a computer-generated summary of health insurance coverage previously reported by the respondent. In the fifth interview, respondents reviewed with the interviewer each reported coverage shown in the household summary.

The Health Insurance Employer Survey (HIES) was one of several surveys comprising the National Medical Care Expenditure Survey (NMCES) which focused on the employers and insurance companies of individuals included in the household study. Two types of instruments were sent to these respondents, corresponding to two different groups in the household sample. Members of the household sample who were identified as subscribers to a private insurance policy were asked to sign a Health Insurance Permission Form (HIPF) that was subsequently sent to the employer, union, insurance company, or other organization through which they had obtained their insurance. The purpose of the HIPF was to verify coverage and supplement information obtained from the household with respect to health insurance benefits and premiums. Persons who reported that they were not covered by an employment-related health insurance plan, but who were employed and not self-employed, were asked to sign the Uninsured Validation Permission Form so that their employers could also be

contacted through a second form that was used in the HIES, the Uninsured Validation Survey Questionnaire (UVSQ). The Uninsured Validation Survey (UVS), a component of the HIES, was designed to confirm that these survey participants were not in fact covered by employment-related health insurance and, if they were covered, to obtain data on premiums, on source of premium payment, and on health insurance benefits.

HIPF respondents were only asked to confirm the insurance coverage of the primary subscriber on each private insurance policy reported by the NMCES household. Whether other members of the family were also insured under the policy was determined from verification of the primary subscriber's insured status and whether the plan was an individual, couple, or family policy as reported by the HIPF respondent. The linking of nonsubscribers to the verification of coverage and other information about the policy that was provided by HIPF respondents was based on the particular insurance plans reported for each member of the family during the household survey. The names of these insurance companies or plans had previously been coded using a seven-digit coding system specifically developed for the NMCES survey. Nonsubscribers for whom coverage was reported from a particular insurer were linked to the HIPF responses for the household's family or couple plan from that insurer.

Plans not reported by the household, which were discovered when an HIPF respondent reported additional coverage for the primary subscriber not shown on the form or when a UVS respondent reported that an individual was in fact insured, could not be linked to nonsubscribers on this basis. All discovered plans involving couple or family coverage were expanded to the appropriate nonsubscribers by one of two means. First, if the subscriber for the discovered plan was also the subscriber on a nonindividual plan that had been reported by the household, all individuals who had been linked to the reported plan were linked to the discovered plan. Second, if this rule could not be applied, then coverage for nonsubscribers of discovered family and couple plans was assigned on the basis of family relationships. Family plans held by subscribers under 65 years of age were linked to their children who were either under 21 or unmarried college students who were not primary subscribers of their own plans. Family and couple plans held by a married subscriber were linked to the person's spouse. Along with the linkages established from household-reported insurance plans, these linkages to an HIES response for a primary subscriber were treated as HIES responses for nonsubscribers.

The dependents of any employed but not self-employed individual, who was eligible for the UVS, could potentially have been covered by a nonindividual policy discovered for that person in the UVS. Since coverage for discovered UVS plans were assigned to dependents through the rules involving family relationships described above, these rules were also used to identify spouses and children who were eligible for UVS not through their own employment but through someone else in the family. For "primary UVS eligibles" who were under 65, their children who were under 21 or unmarried, uninsured college students were also considered to be eligible. For each married primary eligible, the person's spouse was also ascribed eligibility for UVS.

(Appendix B follows on next page.)

	CODE ONE	
B	Medicare	01 *(Next plan)*
	CHAMPUS/CHAMPVA.	02 *(Next plan)*
	Indian Health Service. .	03 *(Next plan)*
	Any other plan	04 *(13)*

As you know, the amount of money paid for insurance is called a premium. The next few questions are about premiums.

13. Not counting any amount that may be paid by any other source, what is the premium you or your family pay for (PLAN)?

$_____ *(14)*

None . . . . . 00 *(15)*

Don't know . . . . . 94 *(15)*

14. Is that per week, per month, per quarter, per year, or something else?

Per week 01

Every two weeks. 02

Per month. 03

Every two months 04

Per quarter (3 months) 05

Every six months 06

Per year 07

Other (SPECIFY)_____ 08

Don't know 94

15. (In addition to the premiums that you or your family pay) Does any other source pay all or part of the premium for this insurance?

Yes. . . . . 01 *(A)*

No . . . . . 02 *(Next plan)*

Don't know . . . . . 94 *(Next plan)*

A. Who else pays all or part of the premiums for this insurance? ENTER BELOW. Anyone else?	FOR EACH SOURCE, ASK B & C: B. How much money does (SOURCE) pay toward the premiums of this plan?	C. Is this per week, per month, per quarter, per year or something else?
SOURCE	AMOUNT	ENTER CODE FOR TIME INTERVAL FROM Q. 14.
	$	per
	$	per
	$	per

ASK Q's. 1 THROUGH 15 FOR EACH INSURANCE PLAN. AFTER ALL PLANS HAVE BEEN COVERED, GO TO PART II OF THIS INSURANCE SUPPLEMENT.

PART II

REFER TO TABLE ON BACK OF INSURANCE SUPPLEMENT. ASK PART II ABOUT EACH UNIQUE COMBINATION OF PLANS.

RECORD THE NAMES AND ID LETTERS OF THE PLANS COMBINED FOR PART II. CIRCLE THE PERSON NUMBER(S) OF REPORTING UNIT MEMBERS COVERED BY THIS COMBINATION OF PLANS.

The following questions concern the coverage provided by (NAME PLANS).

1. Thinking about all of these plans together, do any of the plans cover any part of the costs for . . .

ASK "A" FOR ITEMS 1 THROUGH 15 BEFORE GOING TO "B"

	A.	YES	NO	DON'T KNOW
1)	A semi-private room in a hospital?	01	02	94
2)	Care given by a surgeon to a patient in a hospital?	01	02	94
3)	Care given by other doctors to a patient in a hospital?	01	02	94
4)	Maternity bills?	01	02	94
5)	Eye examinations for glasses?	01	02	94
6)	Dental x-rays, fillings and other routine dental care?	01	02	94
7)	Teeth straightening, braces, or orthodontia?	01	02	94
8)	Oral surgery?	01	02	94
9)	Medicines prescribed by a doctor to a patient outside the hospital?	01	02	94
10)	X-rays and diagnostic tests taken while not a bed patient in a hospital?	01	02	94
11)	Care provided by a doctor in his office or at your home?	01	02	94
12)	Counseling by psychiatrists or mental health professionals?	01	02	94
13)	Care in a semi-private room in a nursing home?	01	02	94
14)	Hospitalization for mental illness?	01	02	94
15)	Emergency room care?	01	02	94

A design for achieving prespecified levels of representation for multiple domains in health record samples

Douglas Drummond, Research Triangle Institute

Judith Lessler, Research Triangle Institute

Donna Watts, Research Triangle Institute

Stephen Williams, Research Triangle Institute

Introduction

New hospital facilities are licensed according to the need for the facilities and the particular services they will offer. Information on the extent of use of existing facilities is needed to assess the need for a particular expansion or new facility, and a survey was designed and conducted to this end for a particular state. Because of the effectiveness of the survey design in this setting and its apparent applicability in myriad other settings, this paper was prepared to describe this sample survey methodology with the feeling that others might benefit from the experience.

This methodology can be used to increase the representation of multiple small domains in a sample. The method was developed for a study of hospital services and is potentially useful in a wide range of applications in which prespecified levels of precision are needed for selected subpopulations, or domains, but a listing of the elements with their domain identifications does not exist, and it is too expensive or too burdensome to construct. The method does require, however, the existence of reasonably good measures of relative domain sizes.

The specific example discussed in this paper describes a retrospective study of hospitals in which use information by hospital services was needed. For example, estimates were to be made by bed-service where the latter is a group of beds specifically designated for use by a particular unit or service of the hospital, such as surgery, pediatrics, psychiatry, or intensive care. The proportion of patients using these services varies greatly. General medical/surgical beds have high proportions (0.5–0.85) and specialty services have low proportions (0.01–0.1). Thus, simply sampling and abstracting records without regard to service use would require very large numbers in order to obtain the required precision for the rarely used services (i.e., the rare domains).

The authors wish to express appreciation to the Florida Department of Health and Rehabilitative Services (FDHRS) for permission to use the Florida Acute Care Facility Need Study as an illustration of the methodology presented in this paper. The study was conducted by the Research Triangle Institute and NTS Research Corporation, under contract with the Florida Association of Health System Agencies, Inc. and FDHRS. Project technical monitor was William W. Alfred, statistician at FDHRS.

In a pretest of the study, hospital administrations required that their own staff select, pull, and abstract the sample information. Furthermore, primary concern for the main study was that many hospitals would not cooperate if a substantial burden was placed on them. Thus, the methodology needed to be simple and efficient. Also, the cost of pulling and reviewing patient records makes it desirable both to ascertain whether the record is in the sample and to proceed immmediately with the abstraction when a sample record is identified.

The option of using abstracting service data was considered. Too many of the hospitals, however, did not subscribe to such a service and, overall, record automation was not sufficiently standard nor widespread enough to answer the study needs. As a result, cost considerations narrowed the options to some form of multistage screening design in order to control domain sample sizes. The first stage units were clusters of hospital discharges (i.e., hospitals); the second stage units were individual discharge records within hospitals. Use in patient-days was to be estimated for different services so as to satisfy prespecified precision levels. The estimates of service use in each hospital came from extant data on bed counts by type of bed and number of discharges by type of service (prior year). Other settings in which the method is useful are summarized in the final section of this paper.

Sample selection methods

This section of the paper describes the method of screening and sampling. The discussion assumes that a two-stage sample of hospitals and patient discharge records within hospitals is used. The goal of the sample design is to guarantee prespecified levels of representation in each of several, possibly overlapping, service subpopulations in a cost-effective manner. The service populations overlap because a patient can use more than one service while in the hospital.

The class of designs to be discussed in this section requires some prior notion as to the prevalence of the various services in the discharge record population at each hospital. This information is used in selecting both the hospital and discharge record sample in an effort to realize and overall self-weighing sample of discharge records, by service, while equalizing the total number of record abstractions at each sample hospital. The mechanism for achieving this will be a two-stage sample design in which hospitals at the first stage are selected proportional to a particular composite size measure and discharge records at the second stage are subjected to a

multiphase sequential screening.

Let

M = total hospitals in the frame, indexed by $i = 1,2,...,M$;
D_i = number of discharge records at hospital i, indexed by $\ell = 1,2,...,D_i$;
w_i = the analysis weight associated with sample hospital i;
$w_{i\ell}$ = the analysis weight associated with sample record ℓ in hospital i.

We wish to control the sampling rate for each of K services which are distributed across the population of hospitals.

The following additional notation will facilitate our discussions. Let

K = total services, indexed by $j = 1, 2,...,K$.
$\hat{D}_i$ = estimated number of discharge records at hospital i.
$\hat{p}_{ij}$ = estimated proportion of discharge records receiving service j at hospital i.
n_j = desired number of sample discharge records from service j.

In a self-weighting design the estimated sampling rate for service j is $\hat{r}_j$, where

$$\hat{r}_j = n_j \left[\sum_{i=1}^{M} \hat{p}_{ij}\hat{D}_i \right]^{-1}$$

Define the composite size measure, $\hat{S}_i$, according to

$$\hat{S}_i = \sum_{j=1}^{K} \hat{r}_j \hat{p}_{ij} \hat{D}_i .$$

Selecting m hospitals proportional to size yields

$$w_i = m_i \left[E(m_i) \right]^{-1}$$

where

$$E(m_i) = m\hat{S}_i/\hat{S}_+$$

m_i = number of times hospital i gets selected into the first stage sample,

and

$$\hat{S}_+ = \sum_{i=1}^{M} \hat{S}_i = n_1 + n_2 + \ldots + n_k .$$

The desired service-specific sample size, n_j, is allocated to sample hospitals in proportion to the weighted number of estimated discharge records of that type at that hospital. That is, sample hospital i would be allocated a sample size of n_{ij} service j discharge records, where

$$n_{ij} = n_j w_i \hat{p}_{ij} D_i \left[\sum_{i\varepsilon sample} w_i \hat{p}_{ij} D_i \right]^{-1}$$

assuming n_{ij} does not exceed $\hat{p}_{ij}D_i$. Equivalently, such an allocation rule would require that discharge records indicating receipt of service j be selected at hosital i at the rate of $r_{j\cdot i}$, where

$$r_{j\cdot i} = n_{ij}/\hat{p}_{ij}D_i = n_j w_i \left[\sum_{i\varepsilon sample} w_{ij} p_{ij} D_i \right]^{-1}$$

If an efficient sampling frame which listed the service use indicated in each discharge record could be constructed once the sample hospitals were visited, the K service-specific samples could be directly selected at these specified rates. In practice this is not feasible and patient records must be screened for receipt of service j. However, it is not necessary to screen every sample record for each service to be studied. Specifically, a large sample of records can be selected and screened for the service with the highest sampling rate. Subsamples can then be screened for the services with lower sampling rates.

Algebraically this is described as follows:

Let

$$v_i = \max_{j=1,2,...,K} r_{j\cdot i}$$

and

$$v_{j\cdot i} = r_{j\cdot i}/v_i$$

Selection of the second-stage sample of discharge records at sample hospital i then proceeds as follows:

Step 1: Select an initial epsem sample at rate v_i.
Step 2: Select K subsamples of the Phase I sample at rates $v_{j\cdot i}$ $(j = 1,2,...,K)$ and screen subsample j for receipt of service j. Abstract service-specific information when present.

Discharge records are selected independently at each sample hospital. Proceeding in such a fashion clearly results in a service-specific self-weighting sample. I.e.,

$$w_{ij} = w_i \left[v_i v_{j\cdot i} \right]^{-1} = w_i (r_{j\cdot i})^{-1} = r_j^{-1}$$

where

$$r_j = n_j \left[\sum_{i\varepsilon sample} w_i \hat{p}_{ij} D_i \right]^{-1}$$

Moreover, the expected number of sample discharge records at hospital i being screened for and indicating receipt of service j would be modeled as $E(\hat{n}_{ij})$, where

$$E(\hat{n}_{ij}) = D_i v_i v_{j\cdot i} \hat{p}_{ij} = D_i r_{j\cdot i} \hat{p}_{ij} = n_{ij}.$$

That is, the *a priori* expected total number of service-specific abstractions at sample hospital i, $E(\hat{n}_i)$, is given by

$$E(\hat{n}_i) = \sum_{j=1}^{K} E(\hat{n}_{ij}) = n_{i+}.$$

But

$$n_{i+} = \sum_{j=1}^{K} \frac{n_j w_i \hat{p}_{ij} D_i}{\sum_{i \in \text{sample}} w_i \hat{p}_{ij} D_i}$$

$$= w_i \sum_{j=1}^{K} r_j \hat{p}_{ij} D_i$$

$$= m_i \frac{(n_1 + \ldots + n_K)}{m} \frac{\sum_{j=1}^{K} r_j \hat{p}_{ij} D_i}{\sum_{j=1}^{K} \hat{r}_j \hat{p}_{ij} \hat{D}_i}$$

That is, in expectation, all sample hospitals for which $m_i = 1$ are projected to require an approximately equal number of service-specific abstractions. It is similarly shown that the expected number of abstractions from service j is the desired number, i.e.,

$$\sum_{i \in \text{sample}} E(\hat{n}_{ij}) = n_j .$$

The true number of abstractions, even in expectation, clearly depends on the accuracy of the service-specific prevalence factors, $\hat{p}_{ij}$.

Several comments are appropriate at this time.

1. Differential rate subsampling is particularly effective in the presence of accurate size measures and substantial variation in the second phase subsampling rate for services within a hospital; i.e., $v_{j\cdot i}$: j = 1,2,...,K. If little variation exists, all services under consideration for Phase I sample members could be abstracted.
1. In some cases (e.g., for very rare services), v_i will be exceedingly large. When this occurs, consideration should be given to employing supplementary frame sampling for this service in order to realize a tolerable screening workload.
3. Service-specific abstractions could be increased by employing the rule "abstract all eligible services whenever a record is successfully screened for a particular service." Clearly, however, such a rule causes unequal weighting among the abstracted records receiving that service unless $v_{j\cdot i} = 1$ (j = 1,1,...,K). Employing nested subsamples at Phase 2 would result in the realization of the smallest number of patient discharge records requiring abstraction under such a rule.
4. The proposed differential rate subsampling plan can be implemented on a flow basis over time. This is not true of double sampling for stratification schemes discussed later.
5. Use of an alternative size measure for selecting hospitals will result in unequal abstraction workloads at sample hospitals and in extreme cases might inhibit our efforts to achieve self-weighting.
6. Stratification at the first stage of the design should attempt to better guarantee the ability of the design to achieve the desired sample sizes—selection mechanism focuses primarily on the rate.
7. Knowledge concerning the prevalence of all $2^K - 1$ patterns of service combinations would allow one to consider screening for same. This would be particularly useful in drug-reaction studies and, in general, for controlling on combinations of risk factors. At a minimum, the total number of required abstractions could be reduced (i.e., the current rate subsampling scheme ignores the fact that some sample members could support multiple service-specific samples). Admittedly, however, combining services having different marginal sampling rates will inflate the unequal weighting effects in the design. Field efforts are of course also rendered more complex under more involved screening rules.
8. Overrepresenting population domains in any sample causes deterioration in the precision level otherwise attainable for estimates of overall population parameters. This paper does not address how one decides on the service-specific sample sizes, only how to attain them in a reasonable manner.

Parameter estimation

Many statistical analyses are primarily concerned with the estimation of population totals or ratios of totals (including means and proportions) specific to a given domain of interest, as well as in the approximate precision of the estimation. Each will be addressed in turn. It is emphasized at the outset that major concern rests with exploiting the underlying multiplicities in the design and not with estimation theory per se. To this end, discussions will involve only the treatment of totals un-

der the design (i.e., both first- and second-order properties). Nonlinear functions of totals would presumably be estimated by the same nonlinear function of the estimated totals. In the absence of independent replicates of the design, the precision of these latter estimates could be approximated using the Taylor-linearized form of the statistic. In the presence of independent replicates, point estimates could be formed as the arithmetic average of the individual replicate estimates and their precision unbiasedly estimated by the simple standard deviation between them divided by the square root of the number of replicates.

Estimation of totals. The target population of interest consists of all discharge records at an eligible hospital indicating receipt of at least one of the K study services. Parameters of interest will generally be domain (G) totals either for a specific service (s_j) (e.g., Y_{G,s_j}) or for the overall target population (e.g., Y_G), where

$$Y_{G,s_j} = \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} Y_{i\ell}\, I_G(i\ell)\, I_{s_j}(i\ell) \qquad (1)$$

and

$$Y_G = \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} y_{i\ell}\, I_G(i\ell) \qquad (2)$$

where

$$I_G(i\ell) = \begin{cases} 1 & \text{if discharge record } \ell \text{ at hospital i belongs to domain G} \\ 0 & \text{otherwise;} \end{cases}$$

$$I_{s_j}(i\ell) = \begin{cases} 1 & \text{if discharge record } \ell \text{ at hospital i indicates receipt of service j} \\ 0 & \text{otherwise.} \end{cases}$$

To estimate these parameters, sample data must be assigned analysis weights which reflect both the underlying randomization mechanism of the sample design and the form of estimate desired. For the purposes of this paper, we will employ unbiased linear expansion estimators. For example, the parameters in equations (1) and (2) could be respectively estimated by

$$\hat{Y}_{G,s_j} = \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} \frac{w_i}{r_{j\cdot i}} y_{i\ell} I_G(i\ell) I_{s_j}(i\ell) t_j(i\ell) \qquad (3)$$

and

$$\hat{Y}_G = \sum_{j=1}^{K} \sum_{i=1}^{M} \sum_{\ell=1}^{D} \frac{w_i}{f_{i\ell} r_{j\cdot i}} y_{i\ell} I_G(i\ell) t_j(i\ell) \qquad (4)$$

where

$$t_j(i\ell) = \begin{cases} 1 & \text{if discharge record } \ell \text{ at hospital i is selected to be screened for service j} \\ 0 & \text{otherwise,} \end{cases}$$

and

$f_{i\ell}$ = number of eligible services on discharge record ℓ at hospital i.

Notice that estimators of non-service-specific parameters (e.g., equation 4) must account for patients potentially receiving multiple services.

For a broad range of designs

$$w_i = m_i/[E(m_i)]$$

where

m_i = observed number of times that that hospital i appears in sample,

and

$E(m_i)$ = expected number of times that hospital i would appear in the sample.

For example, under without-replacement sampling, the proposed analysis weight is the usual inverse of the selection probability for each sample member. The same form of weight applies for with-replacement sampling and for probability minimum replacement sampling under this form of unbiased linear expansion estimator.[1] Proceeding in such a fashion will yield unbiased estimates of the intended linear parameters.

Alternative estimators do exist, however. For example, equation (2) could unbiasedly be estimated by

$$\tilde{Y}_G = \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} \frac{w_i y_{i\ell}}{\Pi_{\ell\cdot i}} \cdot I_G(i\ell) t(i\ell) \qquad (5)$$

where

$$t(i\ell) = \begin{cases} 1 & \text{if discharge record } \ell \text{ at hospital i belongs to at least one of the K samples} \\ 0 & \text{otherwise} \end{cases}$$

and $\Pi_{\ell\cdot i}$ denotes the associated second-stage sample inclusion probability for record (i,ℓ). With the nested subsampling strategy,

$$\Pi_{\ell\cdot i} = \max_{j=1,2,\ldots,K} \left\{ r_{j\cdot i} I_{s_j}(i\ell) \right\} \qquad (6)$$

Under independent subsamples,

$$\Pi_{\ell\cdot i} = \sum_{j=1}^{K} r_{j\cdot i} I_{s_j}(i\ell) - \sum_{j<j'}^{K} r_{j\cdot i} i_{s_j}(i\ell) r_{j'\cdot i} I_{s_{j'}}(i\ell) + \sum_{j<j'<j''}^{K} r_{j\cdot i} I_{s_j}(i\ell) r_{j'\cdot i} I_{s_{j'}}(i\ell) r_{j''\cdot i} I_{s_{j''}}(i\ell) \ldots + (-1)^{K+1} \prod_{j=1}^{K} r_{j\cdot i} I_{s_j}(i\ell). \quad (7)$$

The statistics in equations (3) and (5) can be recognized as the Horvitz-Thompson estimators for the domain G totals (i.e., service-specific and overall) when the first stage hospital selections are without replacement. Similar alternative estimates can also be formed in the presence of the screening rule: "If discharge record ℓ at hospital i was screened for and received treatment j, abstract all eligible services." Specifically

$$\bar{Y}_{G,s_j} = \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} \frac{w_i}{\Pi_{\ell\cdot i}} y_{i\ell} I_G(i\ell) I_{s_j}(i\ell) t(i\ell) \quad (8)$$

or

$$\bar{Y}_{G,s_j} = \sum_{g=1}^{K} \sum_{i=1}^{M} \sum_{\ell=1}^{D_i} \frac{w_i y_{i\ell}}{f_{i\ell} r_{g\cdot i}} I_{s_j}(i\ell)\, I_{s_g}(i\ell) I_G(i\ell) t_g(i\ell) \quad (9)$$

Finally, an alternate class of estimators can be developed for addressing the estimation of domain totals comprising the $2^K - 1$ possible disjoint combinations of patient services. For illustration purposes, consider the estimation of equation (1) under such a representation. To this end, there are 2^{K-1} possible service j subdomains under this representation (i.e., presence/absence of service j′ in addition to service j, j′ = 1, 2,..., j−1, j+1,...,K). Clearly, many options exist for forming the required service specific estimates under any particular Phase 2 subsampling strategy. For example, under nested subsamples, one could use the specific service j subsample to estimate only those subdomains not covered by service-specific subsamples having a higher second-stage sampling rate. As such, all component subdomains are covered by a unique sample chosen to maximize the sample size for estimating the subdomain parameter. In the case of independent subsamples at Phase 2, one might opt for estimators based on subdomain estimators, multiplicities, overall inclusion probabilities, or merely some convex combination of the separately available estimates chosen according to some optimality principle.

The choice between these alternative estimators involves consideration of sample size, unequal weighting effect, ease of analysis, and the extent of any overlap between the domains. When the overlap is "small," reliance on the individual service-specific subsamples is suggested. As the degree of overlap increases, the best form of estimator is not so clear cut and some investigation is needed to provide guidance in its selection.

Approximating the variance of an estimated total. The service-specific rate subsampling design proposed in this paper employs two stages of sample selection. For the most part, attention has focused on the allocation and selection of the second-stage sample of discharge records, as well as on the preferred size measure to be employed in selecting a proportional sample at the first stage of the design. Nevertheless, it has been assumed that any candidate first-stage design would furnish an error structure capable of supporting the approximation of precision for study estimates. Admissible designs in this regard include the independent selection of at least two hospitals with or without replacement in each primary stratum and independent replicates of the entire design. In the latter case, the point estimates discussed in the previous subsection could be averaged across replicates and precision unbiasedly estimated by the simple sample variance between replicates divided by the number of replicates. In the remaining admissible designs, variances could be approximated by squared between-hospital differences within primary strata. For without replacement primary selections, such an approximation is known to be conservative and could be corrected through use of the unbiased two-stage Yates-Grundy variance estimator, provided all joint inclusion probabilities are positive. In instances where maximum stratification is used in the first-stage design (i.e., one hospital selected independently in each stratum), adjacent strata can be paired and the variance approximated using the squared difference between paired sample hospitals (i.e., pseudo-replicates) within each collapsed stratum. This latter approach can also be used for independent zone sampling as well as under designs employing probability minimum replacement (PMR) selection methods such as pps systematic or Chromy's (1981) sequential PMR scheme.

The following additional remarks are germane to our discussions:

1. Phase 2 requires epsem subsampling by service but to date has not specified a particular preference of method (e.g., systematic, simple random sampling, etc.). Operational considerations often encourage the use of systematic sampling under a random start-point. Clearly, this would preclude use of a Yates-Grundy variance estimator (i.e., conditional joint inclusion probabilities within a sample hospital are not all positive under systematic sampling).
2. Use of a Yates-Grundy variance estimator will require that proper account be taken of ineligi-

bles in the second-stage sample (i.e., discharge records selected for screening but found not to indicate receipt of the service(s) being screened for).

3. Consideration might be given to replicating designs employing controlled selection at the first-stage and/or multiphase double sampling for stratification. In the absence of this, it may be difficult or impossible to adequately approximate the precision of location parameter estimates.

Efficiency of technique compared to other approaches

One question that naturally arises when considering the proposed approach for better assuring representation of multiple small domains is how it compares in cost and efficiency to other candidate methods. Among the latter, we have chosen two fairly standard methods of comparison:

1. Sampling of hospitals proportional to overall size and simple random sampling of discharge records within hospitals. This results in equal screening workloads among hospitals and equal probabilities of selection for the sample records.
2. Double sampling for stratification within each sample hospital. Here, a large screening sample of records is selected and service utilization determined. The screening sample is then stratified by service and a sample of records selected within each service for full data collection. To facilitate comparison between the methods, we have assumed that the hospitals are sampled with probabilities proportional to overall size. Second-phase subsample sizes would be chosen to yield service-specific self-weighting samples.

As a first step in comparing the three methods (the third method is the use of composite size measures and domain-specific rate subsampling described in the previous section), the variances of estimates that one might typically wish to make were derived. These are shown in Table 1.

Specifically, estimators for four types of population parameters are given: (1) the overall total of some characteristic abstracted from the record without regard to domain membership or service used (e.g., length of stay); (2) the total number of discharges indicating use of service j; (3) the total number of discharges in service j that are also members of some domain G (e.g., males using service j); and (4) the ratio estimate of the proportion of discharges from service j that are in domain G (e.g., proportion of males among users of service j). For the purposes of Table 1, we have used the linear expansion estimators in equations (3) and (4) of the previous section and have assumed that hospitals are sampled with replacement.

In examining the formulas given in Table 1 several points can be noted concerning the characteristics and the relative advantages and disadvantages of the various techniques. Some of these are summarized in Table 2.

When we examine Table 2 it appears that the double sampling for stratification is the preferred procedure overall because it is the only one that allows maximum control over the realized number of abstractions from each of the K services (i.e., provided that the initial screening samples in aggregate realize at least the desired number of discharge records receiving each service, sample size requirements can be met). It certainly would be the method of choice if lists of discharge records with service use information were available and could be easily tabulated. However, in the absence of such a list the operational aspects of carrying out a double sampling for stratification procedure became very costly in that a two-step procedure must be used. Thus, the choice of methods will be heavily influenced by their relative costs and "do-ability."

The following simple example serves to illustrate the increased cost effectiveness of the screening procedure relative to the other two methods.

Assume that we wish to study four subpopulations whose overall representation in the population is shown in Table 3. Suppose further that we require 500 abstractions from each subpopulation and that the following costs for sampling, screening and abstracting must be incurred.

c_1 = cost of sampling and pulling a record from the file ($0.25).

c_2 = cost of screening a record for all services used and recording results ($0.30).

c_3 = per service cost of screening a record for a single service ($0.10).

c_4 = cost of abstracting a record ($1.00).

c_5 = cost per record of constructing frame for double sampling ($0.10).

We will ignore all cost associated with the sampling and induction of hospitals in this simplified example. Doing so results in the following approximate costs.

No control: To expect to yield 500 abstractions from the rarest domain, we would need to sample and abstract 50,000 discharge records where the total sampling and abstraction cost is

$$\text{Total cost} = 50{,}000\ (1.00 + 0.25)$$
$$= \$62{,}500$$

DSS: To expect to yield 500 abstractions from the rarest domain, the initial sample would have to be 50,000 records. Membership in each of the four domains would need to be determined and a sample of 500 records selected for abstraction from each. Assuming that the records must be repulled from the file for abstraction, we obtain

Table 1
Sample estimators and variances for various sampling procedures

Type of Estimate	Type of Sampling Procedure: Hospitals sampled with probabilities proportional to total size. Equal probability sampling of records. No attempt to control domain size. (NO CONTROL)	Hospitals sampled with probabilities proportional to total size. Large screening sample of records sampled for stratification by service. Service specific subsamples selected. (DSS)	Hospitals sampled with probabilities proportion to composite size measure. Service specific screening samples set for each domain based on domain sampling rates. (DOMAIN RATE CONTROL)
Overall Total without regard to domain membership (such as total days of care)	$\hat{Y} = r \sum_{i=1}^{m} \sum^{n_i}_{\ell=1} Y_{i\ell}$ $Var\{\hat{Y}\} = \frac{D}{m} \sum_{i=1}^{M} D_i [\bar{Y} - \bar{\bar{Y}}]^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{D_i}{n_i} [\frac{D_i - n_i}{D_i}] \sum^{D_i}_{\ell=1} \frac{(Y_{i\ell} - \bar{Y}_i)^2}{D_i - 1}$	$\hat{Y} = \frac{D}{m} \sum_{i=1}^{m} \frac{D_i^2}{n_i} \sum_{j=1}^{k} \frac{n_{ij}}{o_{ij}} \sum^{n_i}_{\ell=1} \frac{u_j(i\ell) I_{sj}(i\ell) Y_{i\ell}}{f_{ij}}$ $Var\{\hat{Y}\} \doteq \frac{D}{m} \sum_{i=1}^{M} D_i [\bar{Y}' - \bar{\bar{Y}}']^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{1}{D_i} \sum_{j=1}^{k} (\frac{D_{ij}^2}{o_{ij}}) (\frac{D_{ij} - o_{ij}}{D_{ij}}) S^2_{y'_{ij}}$ $+ D_i^2 \sum_{j=1}^{k} \frac{D_{ij}}{D_i} (\bar{Y}'_{ij} - \bar{Y}'_i)^2$	$\hat{Y} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\pi_i} \sum_{j=1}^{k} \frac{1}{r_{ij}} \sum^{D_i}_{\ell=1} \frac{t_j(i\ell) I_{sj}(i\ell) Y_{i\ell}}{f_{i\ell}}$ $Var\{\hat{Y}\} = \frac{1}{m} \sum_{i=1}^{M} \pi_i [\frac{Y'_i}{\pi_i} - Y']^2$ $+ \frac{1}{m} \sum_{i=1}^{M} \frac{1}{\pi_i} \sum_{j=1}^{k} \frac{D_i^2}{n'_{ij}} [\frac{D_i - n_{ij}}{D_i}] S^2_{y'_{ij}}$
Total number in service j	$\hat{Y}(sj) = r \sum_{i=1}^{m} \sum^{n_i}_{\ell=1} I_{sj}(i\ell)$ $Var\{\hat{Y}(sj)\} = \frac{D}{m} \sum_{i=1}^{M} D_i (p_{ij} - p_j)^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{D_i}{n_i} [\frac{D_i - n_i}{D_i}] p_{ij}(1 - p_{ij})$	$\hat{Y}(sj) = r \sum_{i=1}^{m} \sum^{n_i}_{\ell=1} I_{sj}(i\ell)$ $Var\{\hat{Y}(sj)\} = \frac{D}{m} \sum_{i=1}^{M} D_i (p_{ij} - p_j)^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{D_i}{n_i} [\frac{D_i - n_i}{D_i}] p_{ij}(1 - p_{ij})$	$\hat{Y}(sj) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\pi_i} \frac{1}{r_{ij}} \sum^{D_i}_{\ell=1} t_j(i\ell) I_{sj}(i\ell)$ $Var\{\hat{Y}(sj)\} = \frac{1}{m} \sum_{i=1}^{M} \pi_i [\frac{D_{ij}}{\pi_i} - D_j]^2$ $+ \frac{1}{m} \sum_{i=1}^{M} \frac{1}{\pi_i} \{\frac{D_i^2}{n'_{ij}} [\frac{D_i - n_{ij}}{D_i}] p_{ij}(1 - p_{ij})\}$
Total number in service j that are also members of some domain G	$\hat{Y}(G,sj) = r \sum_{i=1}^{m} \sum^{n_i}_{\ell=1} I_G(i\ell) I_{sj}(i\ell)$ $Var\{\hat{Y}(G,sj)\} = \frac{D}{m} \sum_{i=1}^{M} D_i [\theta_{ijG} p_{ij} - \theta_{jG} p_j]^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{D_i}{n_i} [\theta_{ijG} p_{ij}(1 - \theta_{ijG} p_{ij})]$	$\hat{Y}(G,sj) = r \sum_{i=1}^{m} \frac{n_{ij}}{o_{ij}} \sum^{n_i}_{j=1} u_j(i\ell) I_{sj}(i\ell) I_G(i\ell)$ $Var\{\hat{Y}(G,sj)\} = \frac{D}{m} \sum_{i=1}^{M} D_i [\theta_{ijG} p_{ij} - \theta_{iG} p_j]^2$ $+ \frac{D}{m} \sum_{i=1}^{M} \frac{1}{D_i} \frac{D_{ij}^2}{o_{ij}} [\frac{D_{ij} - o_{ij}}{D_{ij}}] \theta_{ijG}(1 - \theta_{ijG})$	$\hat{Y}(G,sj) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\pi_i} \frac{1}{r_{ij}} \sum^{D_i}_{\ell=1} t_j(i\ell) I_{sj}(i\ell) I_G(i\ell)$ $Var\{\hat{Y}(G,sj)\} = \frac{1}{n} \sum_{i=1}^{M} \pi_i [\frac{D_{ij} \theta_{ijG}}{\pi_i} - D_j \theta_{jG}]^2$ $+ \frac{1}{n} \sum_{i=1}^{M} \frac{1}{\pi_i} \{\frac{D_i^2}{n'_{ij}} [\frac{D_i - n_{ij}}{D_i}] \theta_{ijG} p_{ij}(1 - \theta_{ijG} p_{ij})\}$

Table 1 (cont'd)

Type of Estimate	Hospitals sampled with probabilities proportional to total size. Equal probability sampling of records. No attempt to control domain size. (NO CONTROL)	Hospitals sampled with probabilities proportional to total size. Large screening sample of records sampled for stratification by service. Service specific subsamples selected. (DSS)	Hospitals sampled with probabilities proportion to composite size measure. Service specific screening samples set for each domain based on domain sampling rates. (DOMAIN RATE CONTROL)
Ratio estimate of proportion of service j that are in domain G	$\hat{\theta}(jG) = \hat{Y}(G,sj)/\hat{Y}(sj)$	$\hat{\theta}(jG) = \hat{Y}(G,sj)/\hat{Y}(sj)$	$\hat{\theta}(jG) = \hat{Y}(G,sj)/Y(sj)$
	$\mathrm{Var}\{\hat{\theta}(jG)\} \doteq \frac{1}{[Dp_j]^2}\frac{D}{m}\sum_{i=1}^{M} D_i\{p^2{}_{ij}(\theta_{ijG}-\theta_{jG})^2$	$\mathrm{Var}\{\hat{\theta}(jG)\} \doteq \frac{1}{[Dp_j]^2}\,m\sum_{i=1}^{M} D_i\{p^2_{ij}(\theta_{ijG}-\theta_{jG})^2\}$	$\mathrm{Var}\{\hat{\theta}(jG)\} \doteq \frac{1}{[Dp_j]^2}\frac{1}{m}\sum_{i=1}^{M}\frac{D_i^2}{\pi_i}\{p^2_{ij}[\theta_{ijG}-\theta_{jG}]^2\}$
	$+\frac{1}{[Dp_j]^2}\frac{D}{m}\sum_{i=1}^{M}\frac{D_i}{n_i}p_{ij}[\theta_{ijG}(1-\theta_{ijG})]$	$+\frac{1}{[Dp_j]^2}\frac{D}{m}\sum_{i=1}^{M}\frac{1}{D_i}\frac{D_{ij}^2}{o_{ij}}\left[\frac{D_{ij}-o_{ij}}{D_{ij}}\right]\theta_{ijG}(1-\theta_{ijG})$	$+\frac{1}{[Dp_j]^2}\frac{1}{m}\sum_{i=1}^{M}\frac{1}{\pi_i}\frac{D_i^2}{n'_{ij}}\left[\frac{D_i-n'_{ij}}{D_i}\right]p_{ij}\theta_{ijG}(1-\theta_{ijG})$

Notation:

D_i = total hospital discharges

m = hospital sample size

r = inverse of record sampling rate

p_j = proportion of overall population using service j

$D = \sum_{i=1}^{M} D_i$

$\pi_i = \frac{S_i}{S+}$

$S^2_{y_{ij}} = \frac{1}{D_i-1}\sum_{\ell=1}^{D_i}[Y'_{ij\ell}-\bar{Y}_{ij}]^2$

$o_{ij} = \frac{I_{sj}(i\ell)Y_{i\ell}}{f_{i\ell}}$

$Y'_{ij\ell} = \frac{I_{sj}(i\ell)Y_{i\ell}}{f_{i\ell}}$

θ_{jG} = proportion of service j users with characteristic

P_{ij}, θ_{ijG} = corresponding within hospital proportions

n_{ij} = number from original sample in sj at hospital i

D_{ij} = number of discharges from sj in hospital i

$\bar{Y}_i = \frac{1}{D_i}\sum_{\ell=1}^{D_i} Y_{i\ell}$

$\bar{Y} = \frac{1}{D}\sum_{i=1}^{M}\sum_{\ell=1}^{D_i} Y_{i\ell}$

$\bar{Y}'_{ij} = \frac{1}{D_{ij}}\sum_{\ell=1}^{D_i} Y'_{ij\ell}$

$\bar{Y}'_i = \frac{1}{D_i}\sum_{j=1}^{K} D_{ij}\bar{Y}'_{ij}$

For typographical reasons,

$sj = sj$

$r_{ij} = r_{j\cdot i}$

} in main text

$u_j(i\ell)$ = indicator for abstraction in sj sample for service j

n_i = size of initial screening sample at hospital i

n'_{ij} = size of screening sample for sj in hospital i

Table 2
Characteristics of various sampling techniques and their relative advantages and disadvantages

	Item	No Control	DSS	Domain rate control
1.	Overall estimates (Abstracted from medical record)	Equal probabilities of selection. No increase in variance due to unequal weights. Increase in variance due to larger clustering effect. Preferred method if subpopulation estimates not needed.	Unequal probabilities of selection. Estimators must account for multiplicity. Variance inflated due to unequal weighting and reduced sample sizes unless gains from stratification achieved.	Unequal probabilities of selection. Estimators must adjust for service multiplicity. Variance inflated due to unequal weighting and reduced sample size. Some gain due to stratification is possible.
2.	Total number in service j (Determined in screening)	Same characteristics all techniques. Estimate of subpopulation size.	Same	Same (Possibly some variance reduction due to composite size measure, i.e., hospital size measure roughly proportional to number in service) Sample sizes generally smaller.
3.	Total number in service j with a particular characteristic, or total characteristic for service.	Domain estimate for subpopulation totals subject to increased variance due to inefficient frame and clustering. Number of sample cases in subpopulation controlled for only in expectation.	Service subpopulation composed of a design strata and hence no second phase loss of efficiency due to domain estimates. Number of sample cases by service directly controlled. Number of sample cases and finite population correction factors allow maximum control of precision. Preferred procedure if advance estimates of relative subpopulation sizes unreliable and subpopulation estimates required.	Same as for no control. Sample sizes will be smaller for services with high prevalence. If estimates of relative subpopulation size reliable, fairly good control on subpopulation sample should be achieved.
4.	Ratio estimate for proportion of service j that has a characteristic or ratio mean for service j.	Precision equal to that of a sample of $p_{ij} n_i$ records from an efficient frame to the usual order of approximation. Equal probabilities of selection for subpopulation members but no ability to control across subpopulations due to constant n_i.	Precision enhanced due to the advance choice of o_{ij}, i. , o_{ij} can be chosen to be greater than $p_{ij}n_i$. Maximum control on o_{ij} achieved. Preferred procedure if advance estimates of relative subpopulation sizes unreliable and subpopulation estimates required.	Precision equal to that of a sample of $p_{ij} n'_{ij}$ from an efficient frame (usual order of approximation). Greater ability to control $p_{ij}n'_{ij}$ by varying n'_{ij} for various subpopulations. Fairly good control on achieved sample size if estimates of subpopulation sizes reliable.
5.	Operational Aspects—Hospital workload	Equal number of screenings (none) and equal number of abstractions.	Number of screening equal, number of abstractions varies. Abstraction requires second pulling of record. Could control abstractions by use of a composite size measure for first stage selections.	Number of screenings varies from hospital to hospital. Number of abstractions equal. Abstractions done on a flow basis.

Initial sampling costs = 50,000 (0.25) = $12,500
Cost of screening for services used = 50,000 (0.30) = $15,000
Cost of constructing frame for double sampling = 50,000 (0.10) = $ 5,000
Cost of sampling and pulling records to be abstracted = 2,000 (0.25) = $ 500
Cost of abstraction = 2,000 (1.00) = $ 2,000
Total expected cost = $35,000

Rate Control: Again 50,000 records would need to be screened to yield the 500 required from the rarest service. However, subsamples of these records could be screened for use of particular services. Records would be abstracted on a flow basis and would not need to be repulled from the file. Hence,

Initial sampling costs = 50,000 (0.25) = $12,500

Costs of screening for service used

Prevalence	
0.01	50,000 (0.10) = $5,000
0.05	10,000 (0.10) = $1,000
0.10	5,000 (0.10) = $ 500
0.60	833 (0.10) = $ 84

Cost of abstracting records = 2,000 (1.00) = $ 2,000

Total expected costs = $21,084

The rate control approach is by far the least expensive method to use. The differences are heavily influenced by the large volume necessary to get the required numbers from the rarest services. If the four services had prevalences 0.6, 0.1, 0.05 and 0.05, the costs of the competing methods would be

No control	$12,500
DSS	$ 9,000
Rate control	$ 7,084

The example given is admittedly simplistic. However the basic comparison would remain true under a more complex analysis—that is, that the double sampling scheme allows for maximum control of the subpopulation sizes but it is likely to cost considerably more than the rate control procedure. Also respondent burden and thus cooperation may be decreased under the double sampling procedure because on-site time by data collectors would be greater for the two-step procedure.

Illustration of methodology

Introduction. The procedure of two-phase sequential sampling, using domain specific subsampling rates for the different classes identified in screening, was developed and used for a specific situation—sampling patient medical records in the Florida Acute Care Facility Need Study. The purpose of the study was to obtain information on the use of selected health care services in Florida short-term hospitals. Data were collected from a probability sample of 3,436 patient record abstractions in 62 Florida hospitals, for the period October 1, 1978, through September 30, 1979. The study was conducted by the Research Triangle Institute and NTS Research Corporation, under contract with the Florida Association of Health System Agencies, Inc., and the Florida Department of Health and Rehabilitative Services (FDHRS). FDHRS is using study results in a model to project the units of care and equipment that would be required to meet future health needs in the state.

Table 3
Relative subpopulation sizes and required number of abstractions—simple hypothetical case

Subpopulation	*Prevalence of subpopulation*	*Abstraction required*
1	0.6	500
2	0.1	500
3	0.05	500
4	0.01	500
		2000

An objective of the study was to estimate use (in terms of total number of bed-days or procedures; and average length of stay, or average number of procedures per discharge record) for the hospital services listed in Table 4. This objective, together with features of hospital medical records systems and cost constraints of the study, motivated the use of the two-phase procedure. To obtain the desired estimation precision for each service, a certain minimum sample size was needed for each service. The two-phase procedure was used to identify services used for selected discharge records and to select records for abstraction, using different sampling rates for different services.

Table 4
Bed service and procedure categories for which use estimates are needed (Florida Acute Care Facility Need Study)

1. General medical/surgical (includes all beds not specified below)
2. Intensive medical/surgical
3. Intensive coronary
4. Burn
5. Psychiatric
6. Obstetrical
7. Neonatal intensive care
8. Pediatric
9. Cardiac catheterization lab
10. Megavoltage radiation therapy equipment (linear accelerators, cobalt 60, betatrons)
11. Computerized axial tomography units

Overview of the sample design. A two-stage design with stratification imposed on the first stage was used. First-stage sampling units were short-term hospitals, and second-stage units were inpatient discharge records. At the second stage of sampling, the two-phase procedure was used to select patient records, screening for service use and applying different sampling rates for different services. The development of the sample design is described in Williams et al. (1978); Harris et al. (1978a); Williams and Weber (1978); Williams (1978); and Harris et al. (1978b).

All short-term hospitals in the state of Florida were geographically stratified into the nine Health System Agencies (HSAs). The total sample size of inpatient discharge records was allocated so that approximately equal estimation precision should result for each HSA. Within each HSA, hospitals were stratified according to an urban/rural factor if HSA characteristics warranted this. Finally, hospitals were stratified according to size, using the composite size measure alluded to previously. Size measure computation is discussed in the following section. Two hospitals were selected from each stratum, without replacement and with probability proportional to size.

The second-stage frame was the conceptual list of all inpatient episodes in the selected hospitals. In each HSA, the sample size of patient episodes was allocated among strata in that HSA in proportion to stratum size. Sample hospitals within a stratum were assigned an equal number of sample episodes which in turn were allocated among the available services of interest based on the contribution by service to the hospital's size measure. A two-phase selection procedure, described below, was used to obtain the desired number of patient episodes, by service, for each hospital.

Using these methods for hospital selection and allocation of the projected number of record abstractions among hospitals and services, it is possible that the total projected number of abstractions for a service within an HSA may not be allocated. When a given service was offered in an HSA but none of the selected HSA hospitals offered that service, a supplementary stratum of HSA hospitals providing the service was created. This occurred only twice—for psychiatric beds in one HSA and for cardiac catheterization lab in another HSA. The supplementary stratum for cardiac catheterization lab was later omitted when the focus of the study was narrowed from the 11 services originally considered to the 6 major bed services.

It also occurred that although some of the selected hospitals in an HSA offered a given service, the total projected number of abstractions for that service was not met. In this situation the allocated numbers of projected abstractions were adjusted towards meeting the desired total, as described in Williams (1978). These adjustments were of course subject to the limitation of the total estimated HSA use for the given service, as well as the limitation of obtaining feasible (in terms of cost and hospital burden) patient record screening rates within hospitals. In this study the requirement of feasible screening rates was an important concern, not easily met when considering the desired estimation precision for some of the rare services. Instead of dealing with insufficient service-specific projected numbers of abstractions by the methods described above, the problem can be prevented by imposing more control on the selection of hospitals, taking into account the services provided beyond formation of the size measures.

Hospital size measures. Hospital size measures were based on recent use information and desired sampling rates by service. Recent service-specific use data were provided by FDHRS. When use data were missing for some hospital and service, the required value was estimated based on appropriate available information such as number of beds, number of patient-days, number of procedures, or data from hospitals of similar size in that hospital's HSA. Note that the availability, prior to sample selection, of measures of use is a requirement for applying the class of designs discussed in this paper.

An example of size measure calculation, as described in section 2, is given for HSA 1 of this study. Table 5 displays the desired service-specific sample size, n_j in the notation of section 2, and sampling rates $\hat{r}_j$. Use data $(\hat{p}_{ij}\hat{D}_i)$ are given in Table 6 for some of the HSA 1 hospitals. For example, consider the use of service 1 in hospital 235. It was estimated prior to sample selection that 5,424 discharge records in hospital 235 would indicate receipt of service 1. Using the information in Tables 5 and 6, the size measure for hospital 235 is calculated as

$$\begin{aligned}\hat{S}_i &= \sum_{j=1}^{11} \hat{r}_j(\hat{p}_{ij}\hat{D}_i) \\ &= .00048576\,(5{,}424) + .00655222(372) + \\ &\quad .00912863(300) + .00908941(528) + \\ &\quad .00260572(1{,}272) + .00424345(612) \\ &= 18.521.\end{aligned}$$

The size measures for all HSA 1 hospitals are shown in Table 7. Table 7 also displays the stratification of hospitals, the selected hospitals, and the allocation of the sample size among the selected hospitals for HSA 1.

Table 5
Sampling rates, by service, for HSA 1
(Florida Acute Care Facility Need Study)

Service code[1] j	*Desired sample size, n_j (number of record abstractions)*	*Sampling rate $\hat{r}_j$*
1	50	.00048576
2	50	.00655222
3	55	.00912863
4	15	.46875000
5	55	.00908941
6	35	.00260572
7	40	.05063291
8	35	.00424345
9	76	.02697906
10	45	.02906977
11	55	.00623936

[1]Key is given in Table 4.

Two-phase selection procedure. The two-phase selection procedure included the following steps: selecting a first-phase sample of medical records for screening; observing the services used during each selected episode;

then selecting, from the first-phase sample of records, the sample of medical records for data abstraction. This procedure permitted the use of different sampling rates for different services, within the same hospital. This was very useful, in light of the fact that use, as well as desired sample sizes, differed for the various services. The procedure is described in Harris et al. (1978b) and Lucas et al. (1979b).

Table 6
Use (annual number of discharges)[1] for selected HSA 1 hospitals (Florida Acute Care Facility Need Study)

Service Code[2]	Hospital Code 031	235	275	141	009	227
1	1,286	5,424	5,743	4,908	8,522	19,602
2	0	372	156	804	773	768
3	0	300	300	250	662	1,330
4	0	0	0	0	32	0
5	0	528	0	794	901	1,128
6	0	1,272	868	804	816	2,728
7	0	0	0	0	0	0
8	0	612	156	1,020	0	2,484
9	0	0	0	89	0	676
10	0	0	0	368	366	454
11	0	0	0	1,153	1,079	2,922

[1]Use data were provided by FDHRS. When data were not available, use was estimated based on appropriate available information such as number of beds, number of patient-days, number of procedures, or data from hospitals of similar size in that hospital's HSA. Use is equivalent to the product $\hat{p}_{ij}\hat{D}_i$ in the notation of the second section in the text.
[2]Key given in Table 4.

To apply this procedure, sampling rates ($r_{j\cdot i}$) were calculated for each service to be observed within a hospital as the ratio of the following: (1) the desired number of patient records to be abstracted for the service within the hospital, n_{ij}; and (2) the estimated number of total discharges for the service within the hospital during the study period, $\hat{p}_{ij}\hat{D}_i$. The highest service sampling rate, v_i, defined the overall screening rate for the hospital. Relative screening rates, $v_{j\cdot i}$, were calculated for the other services with respect to this overall screening rate. A subsample was selected from the screened records for each service by systematic sampling using the appropriate relative screening rate. When a record in a service's subsample showed use of that service during the episode, the record was abstracted.

To assist the medical records abstractor in the sampling and recording task, Sample Selection Forms were prepared to indicate which patient records to screen and which of those to abstract. An example of a Sample Selection Form is shown in Figure 1. These forms were prepared to the extent shown before being sent to hospitals. Forms were prepared for each month of the study period, independently selecting random starts. Note that, in a retrospective study, the Sample Selection Form allows selection of records for screening, screening for services used, selection of records for abstraction, and abstraction to all be performed during one visit to the hospital.

In the example in Figure 1, the abstractor should list 1/25 of the patient episode identification numbers (which can be linked to specific records by hospital staff), beginning with the fourteenth record of the indicated month. After listing an identification number, the abstractor checks the corresponding record to determine if *any* of the services that are indicated (by an X) were used during that episode; if so, information on *all* services of interest to the study is abstracted for that episode. (This may include information on services not marked X for that patient record on the Sample Selection Form.) If none of the indicated services were used during the episode, then no abstract is completed. Abstracting information on all services of interest used during the sample episode, instead of abstracting information on only the indicated service(s), was done to obtain additional information at a relatively small cost. Once a record is pulled and information on patient characteristics and the sample episode is abstracted, it is usually little additional effort to determine and record the length of stay for other services. We discussed the unequal weight-

Table 7
Sampling frame for hospitals in HSA 1 (Florida Acute Care Facility Need Study)

Stratum	Hospital code	Size measure $\hat{S}_i$	Desired number of record abstractions
1. Rural	031	0.625	
	025	0.746	
	104	0.804	
	245**	0.945	33
	151	1.480	
	024	1.626	
	124	1.699	
	250	1.721	
	093	1.911	
	073	3.019	
	100	3.121	
	053	3.376	
	169	4.385	
	251	4.463	
	205	8.995	
	158**	9.783	34
	235	18.521	
2. Urban—small hospitals	071	0.112	
	233	4.256	
	290	4.880	
	275**	9.474	63
	237	15.375	
	074	19.107	
	154	29.733	
	141**	43.867	64
3. Urban—medium hospitals	009**	57.935	59
	256**	60.316	59
4. Urban—large hospitals	192**	94.458	99
	227**	104.264	100
Total			511

**Hospital is selected.

Figure 1
Sample selection form

NOTICE: All information recorded on this document which would permit identification of an individual or an establishment will be held in strict confidence, will be used only by persons engaged in and for the purposes stated, and will not be disclosed or released to other persons or used for any other purpose.

1. Facility Code: 9 9 9 2. Page 0 1 of 0 1 for sample period

(Month) (Day) (Year) (Month) (Day) (Year)

3. Sample Period: 0 1 - 0 1 - 8 0 through 0 1 - 3 1 - 8 0

4. Type of Patients (Check one): [X] Inpatients [2] Outpatients [3] All Patients

5. Random Start Number: 0 1 4 6. Screening Rate: 1 in 0 2 5

	Medical Record Number (for other patient identification number)	Sample Indicators (If one or more services indicated, abstract for all services for episode)										
		Service										
		1	2	3	4	5	6	7	8	9	10	11
1				X			X					
2			X	X								
3		X		X								
4			X	X								
5				X								
6			X	X			X					
7		X		X								
8			X	X								
9				X								
0			X	X								
1		X		X			X					
2			X	X								
3				X								
4			X	X								
5		X		X								
6			X	X			X					
7				X								
8			X	X								
9		X		X								
0			X	X								

(Month) (Day) (Year)

Completed by: _____ Date: - -

ing effect implications of this procedure earlier. Note that columns corresponding to some services are not marked at all in Figure 1, because prior information from FDHRS indicated that those services are not available at the hospital. In some instances, the service definitions used by a hospital were not the same as the FDHRS service definitions used for this study. When such disagreement occurred, the medical record abstractors reclassified service use to be consistent with the FDHRS definitions during screening and abstraction. This prevented a noncoverage problem that could have otherwise occurred.

Results. Table 8 gives the sample sizes obtained for the six major bed services, for nonfederal hospitals. The estimated total number of episodes during the study period is also given for each of these services. This table illustrates the need for using different sampling rates for different services to obtain the desired sample sizes. For example, suppose that patient records are selected

Table 8
Distribution of sample sizes and estimated population sizes for major bed services, nonfederal hospitals (Florida Acute Care Facility Need Study)

Service	Sample size (number of patient records abstracted)	Proportion of total[1] sample size	Estimated[2] Population size (total number of patient discharges)	Estimated proportion of total[1] populetion size	Stenderd error of the estimeted proportion
General medical/surgical	1,633	.508	1,284,104	.833	.015
Intensive medical/surgical	863	.269	112,959	.073	.007
Intensive coronary	933	.290	77,442	.050	.006
Psychiatric	339	.106	27,090	.018	.003
Obstetric	364	.113	118,875	.077	.010
Pediatric	438	.136	65,988	.043	.006
Total[1]	3,211	1.000	1,541,002	1.000	.000

[1]Episodes involving at least one of the six major bed services. An episode, or visit to the hospital, may involve use of more than one service.
[2]Based on Horvitz-Thompson estimator.

within hospitals by simple random sampling, without regard to service use, that is, the same sampling rate is applied to all medical records in a hospital. It is estimated that 5% of major bed-service episodes would involve use of the Intensive Coronary service. Under simple random sampling, it would be expected that approximately 161 (.05×3,211) patient episodes, in a sample including 3,211 major bed-service episodes, would involve use of the Intensive Coronary service. To obtain 933 Intensive Coronary episodes under simple random sampling, a major bed-service sample size of approximately 18,660 (933÷.05) would be required. This number of medical record abstractions (compare with 3,211) would not be feasible because of cost and the burden on hospital staff. Note also that a major bed-service sample size of 18,660 would be expected to contain approximately 15,544 (18,660×.833) General Medical/Surgical episodes—many more than needed to obtain the required estimation precision. It follows that the use of different sampling rates for different services is appropriate in this situation.

The need for using different sampling rates for different services can also be considered at the hospital level, by examining the difference in relative screening rates among services. As discussed previously, to obtain the desired sample sizes by service from a hospital, an overall screening rate and relative screening rates by service were computed. Table 9 gives an example of screening rates for a hospital. Overall, 1/36 of the records are to be screened; all of these records are to be screened for use of service 9 and service 10. Of the screened records, 150 are to be screened for use of service 1, and so on. Services 4, 5, 7, and 8 are not available in the hospital, according to prior information from FDHRS. For the 62 sample hospitals, the minimum overall screening rate was 1/5, the maximum was 1/85, and the median was 1/20. The mean overall screening rate was .0564, or approximately 1/18, and the standard deviation was .0328. The smallest relative screening rate (among services) for a hospital ranged from 1/543 to 1/1, with a median of 1/46. The

Table 9
Example of relative screening rates by service for a hospital

Overall screening rate	1/36
Relative screening rate for service[1]:	
1	1/50
2	1/4
3	1/3
4	—
5	—
6	2/17
7	—
8	—
9	1/1
10	1/1
11	1/4

[1]Key is given in Table 4.

mean was .0662, or approximately 1/15, and the standard deviation was .1755. The fact that the median smallest relative screening rate was 1/46 indicates that there is a substantial difference, at the hospital level, in the sampling rates needed to obtain the desired sample sizes, by service.

Table 10 shows the distribution of sample sizes and estimated population sizes with respect to the combination of major bed-services used during an episode. This description of "service overlap" can be compared with Table 8. For example, it is estimated (Table 7) that .833 of all major bed service episodes during the study period involved the use of service 1, general medical/surgical. However, it is estimated (Table 9) that .748 involved the use of only service 1. The other service 1 episodes also involved other major bed services. Recall that when a record was abstracted, information was collected on all services of interest used during the episode, not just the screened service(s). This was done to obtain additional information at a very small additional cost. Little information about "service overlap" was available prior to the study, and "service overlap" was not considered in developing screening rates, etc. It may be useful to consider this in designing future studies. Note that more overlap is expected when considering specialized procedures as opposed to just the major bed services. The unequal weighting effect implications of uses of "service overlap" were discussed earlier.

Other applications

Two other hospital-based applications of this technique come immediately to mind. First, the technique could be used in the U.S. National Hospital Discharge Survey (NHDS). The NHDS is conducted by NCHS and consists of some 275,000 record abstractions from over 400 sample hospitals. In spite of the massive size of this survey, precise estimates of the characteristics of people with a specific medical diagnosis cannot be obtained, and data tabulations are published only for fairly broad groups of conditions. The precision of estimates for small diagnostic categories could be improved by employing the method presented in this paper. Thus, a large sample of records would be screened for the particular diagnosis of interest, and all records found pertaining to that diagnosis would be abstracted. A subsample of records would then be abstracted and processed according to existing procedures. Since many hospitals maintain discharge summaries which list the discharge diagnosis, the method would be very easy to apply in these hospitals. In fact, if these summaries were available in computer accessible form and/or summary tabulations by diagnosis were available, the double sampling scheme could be used and exact control of domain size (dischage group) could be obtained.

A second hospital-based application for which the method is ideally suited is studies of drug use within

Table 10
Distribution of sample sizes and estimated population sizes by major bed services used, nonfederal hospitals (Florida Acute Care Facility Need Study)

Services[1] used during episode	Sample size (number of petient records abstrected)	Proportion of total[2] semple size	Estimated population size (total number of patient discharges)	Estimated proportion of total[2] population size	Standard error of the estimated proportion
1	532	.1657	1,152,910	.7482	.0180
2	159	.0495	21,700	.0141	.0035
3	222	.0691	20,912	.0136	.0033
5	313	.0975	25,714	.0167	.0029
6	355	.1106	117,981	.0766	.0096
8	422	.1314	64,387	.0418	.0055
1,2	456	.1420	77,358	.0502	.0048
1,3	485	.1510	44,524	.0289	.0034
Other combinations of two services	118	.0367	7,503	.0049	.0010
Three services	147	.0458	7,969	.0052	.0007
Four services	2	.0006	45	.00003	.00002

[1]Key is given in Table 4.
[2]Episodes involving at least one of the six major bed services.

Table 11
Example settings in which the methodology may be advantageous

Type of study	*Cluster unit*	*Extant information about cluster domain sizes*	*Source of screening information*
Hospital service use	Hospital	Number of beds and number of past discharges by type of service	Patient records
Public opinion or other household related information	Census tract, ED, or BG	Historic population counts by demographic groups	Household screening questionnaire (personal interview)
Drug use	Pharmacy	Total sales by drug category	Customer file
Miiitary personnel study	Military installation	Summary reports on medical services or disciplinary actions	Personnel folder
School discipline	School	Summary statistics on disciplinary actions	Student record
Ground water contamination	County	Historic crop production and information and license application rates for pesticides	Well proximity to pesticide application

hospitals. Many hospitals maintain patient drug profile cards in their pharmacies. These systems typically list basic demographic information, sometimes diagnostic information, and always the drugs that were prescribed for the patient and the amounts dispensed. The patient drug profile system provides an almost ideal case for sampling patients by drug use. Separate samples of patients could easily be selected within types of drug use. Also, since interactions between several types of drugs are often of interest in drug studies, specific sampling rates for combinations of drugs could be set.

The technique need not be limited to health records surveys. The basic setting in which the method is useful is when clusters of elements form the higher stages of sampling and when existing information about the relative size of the domains of interest is available for these clusters in order to set within-cluster screening rates. Table 11 summarizes some additional settings in which the method could be used.

Footnote

[1] Probability minimum replacement sampling as defined by Chromy (1981) requires that the random selection frequencies m_i deviate by less than one selection from their expectation.

Discussion: A design for achieving prespecified levels of representation for multiple domains in health records samples *and* Consumer knowledge of health insurance coverage

Mary Grace Kovar, National Center for Health Statistics

Records may be the primary data source or they may be used as validation. The paper by Drummond, Lessler, Watts, and Williams presents a sample design for using hospital records as a primary data source. The paper by Walden, Horgan, and Cafferata presents data using insurers' records as truth against which household respondents' reporting is validated. Because these papers are so different, they will be discussed separately.

Drummond paper

Over the past few years there has been a lot of work done at the Research Triangle Institute on improving the methods of sampling, collecting, and analyzing data from hospital records. The evidence for that work is in the references given in the paper by Drummond, Lessler, Watts, and Williams. The authors have built on that work to develop an elegant cost-effect approach to sampling from such records when the objective is to produce reliable estimates for subdomains of the population represented by the records.

The problem is simple to state. The objective is to provide reliable estimates of the characteristics of patients in different hospital services based on data derived from a sample of records.

It would be possible to draw a simple random sample but the services vary so widely in size—from 27,000 psychiatric patients through 1,284,000 general medical/surgical patients—that a very large sample would be required to obtain reliable estimates for the smallest service. It would also be possible to draw a sample stratified by service. However, doing so would require listing all of the records by service to have the sampling frame and then going back through the records to abstract those selected in the sample.

Drummond et al. have developed an ingenious method of screening, sampling, and abstracting simultaneously and have demonstrated that it is cost effective for their problem. Their method requires only that some knowledge of the relative size of the services be available in advance. The application of their method in other situations is discussed but the limitations are not made explicit. The decrease in cost using their scheme depends heavily on the large difference in the relative size of the services or, to use a general term, domains. In their simplified example, their controlled scheme costs ⅓ as much as an uncontrolled sample when the smallest domain is 1/60 the size of the largest; it costs almost ⅔ as much when the smallest domain is 1/12 the size of the largest. There would be little gain if there were fewer domains and their relative size approached unity. The authors do not address the question of the break-even point.

The suggested applications do not address the question of multiple uses of data from a single survey. The technique is designed to be efficient for producing estimates for one type of domain. Many surveys are designed to produce estimates for a number of different kinds of domains. In the National Hospital Discharge Survey, for example, estimates by age, sex, race, hospital size, and geographic region are as important as the estimates for diagnostic categories. A sampling scheme designed to minimize costs and variances for estimates by diagnostic categories would not necessarily be the most efficient one for those other purposes.

Walden paper

The National Medical Care Expenditure Survey was carefully designed to permit cross-checking of household reporting and records. Such a design is unusual. We have too few opportunities to obtain information about the same people from two or more sources. The designers of the survey should be commended as well as the authors—Walden, Horgan, and Cafferata—who have used that survey design to investigate consumers' knowledge about their health insurance coverage. The authors should also be commended for going beyond the methodological considerations to point out the implications for public policy. In doing so they have enlarged the area open to discussion because there can be differing interpretations of the same data.

There are always problems with using records as "truth," and this paper reveals some of them, including some that I did not expect. There is no indication that the data from either the Health Insurance/Employer Survey or the Uninsured Validation Survey was checked for accuracy; we have to presume that it could not be and was not. Yet people checking records have been known to make errors in reporting and we have no idea of the magnitude in this study. It is a one-way check only. It would be interesting to know the error rate for a group of insurers asked to ascertain coverage when there were no claims to be paid.

We do not know that the complete universe was checked. If a respondent reported coverage but gave the wrong insurer's name, the record check would reveal no coverage when in fact there was. If the respondent was

employed and reported no health insurance and the employer confirmed that there was no coverage through employment, the household members could still have had coverage from another source. If the respondent was not employed and failed to report coverage, no check was made.

The reliability of the records becomes more suspect when one considers the specific benefits. Trained coders employed by the survey coded the benefits from the policies. Yet the coders could not determine from the policies whether 11% of the population had psychiatric coverage, 14% had maternity coverage, or 21% had nursing home coverage.

These methodological problems are pointed out not to criticize the survey or the paper, but as background for the policy implications. While the authors are probably correct in stating that "[A]merican consumers are not in every respect knowledgeable about their health insurance coverage," it is difficult to be certain about the extent of their lack of knowledge from this study.

After reading the questions carefully, I would be uncertain how to answer some of them even though I have carefully read our policy. If I do not have coverage for a specified benefit until after a specified amount has been spent, do I answer yes or no? More critical is that there is little indication of whether the lack of knowledge is important. It is possible that those who don't know whether they are covered for maternity care or orthodontia are people who are old enough that they never expect to use either; those who don't know about nursing home coverage may be very young. Do I answer yes or no? If the amount allowed on the policy is so little it pays only for a small fraction, do I answer yes or no? The problem is not that I don't know what will be reimbursed; I do know, but I don't know how to answer the question.

Comparison of three data sources from the National Medical Care Expenditure Survey: Household questionnaire, household summary, and medical provider survey

Judith A. Kasper, National Center for Health Services Research

The National Medical Care Expenditure Survey (NMCES) collected data on health care use and expenditures from three sources—a household questionnaire, a computerized summary document used for updating and correcting household data, and a medical provider survey (MPS) of physicians and hospitals who treated household respondents. Both the summary and the medical provider survey were undertaken as means to improve data quality—the summary to provide respondents with the opportunity to report previously unknown data or to correct erroneously reported data; the medical provider survey to augment cost and diagnostic data reported by household respondents. This paper examines the relationships among these data sources and the effect of multiple data sources on the quality of expenditure data. Some observations are offered concerning the analytical complexities with regard to multiple data sources and improvements in data quality in light of these complexities.

The data

NMCES was a one-year panel design survey of 40,000 individuals.[1] Five interviews plus a brief clean-up interview were conducted to collect health care use and expenditure data for 1977. At the second through fifth interview, respondents received a computerized summary of use and expenditure data reported in the previous interview. They were to review this information and make any needed additions or corrections. In particular, the summary was to allow respondents a means to provide more complete charge and payment data at a later date if they were unknown at the time of the interview. The medical provider survey was a record check or verification procedure to obtain expenditure and diagnostic data from physicians and hospitals who treated a sample of household respondents during the year. A sample of respondents was selected for this record-check procedure rather than all respondents, primarily for reasons of cost.[2]

This paper benefited from helpful suggestions by Marc Berk, Steve Cohen, Dan Horvitz, Lou Rossiter, Dan Walden, and Renate Wilson. The author thanks John Carrick for his prompt, careful typing and Angelita Manuel and Sandy Smoot of Social and Scientific Systems, Inc., for their excellent programming support. The views expressed in this paper are those of the author, and no official endorsement by the National Center for Health Services Research is intended or should be inferred.

Table I shows the relationships among the household questionnaire, summary, and medical provider survey. This schema applies to the three types of events included in the medical provider survey—hospital stays, ambulatory care physician visits and inpatient physician visits; only hospital stays and ambulatory care physician visits will be examined in this paper. The complexity of relationships demonstrated in Table I (there are 12 types of records based on all possible combinations of data sources) stems both from having three data sources rather than two (most surveys with record checks have only one household data source) and from collecting data at the event level (visit or stay) rather than the person level.[3] While all the events for an individual are either in the MPS sample or not, an individual in the sample may have any combination of records of type A-F-N, B-G, C, H-O, I, or J, and an individual not in the sample may have any combination of records of type C, D-K, E, L, or M. Frequencies of hospital stays and physician visits by record type are given in Table I with a description of each type of record. Other tables reference the record types in Table I.

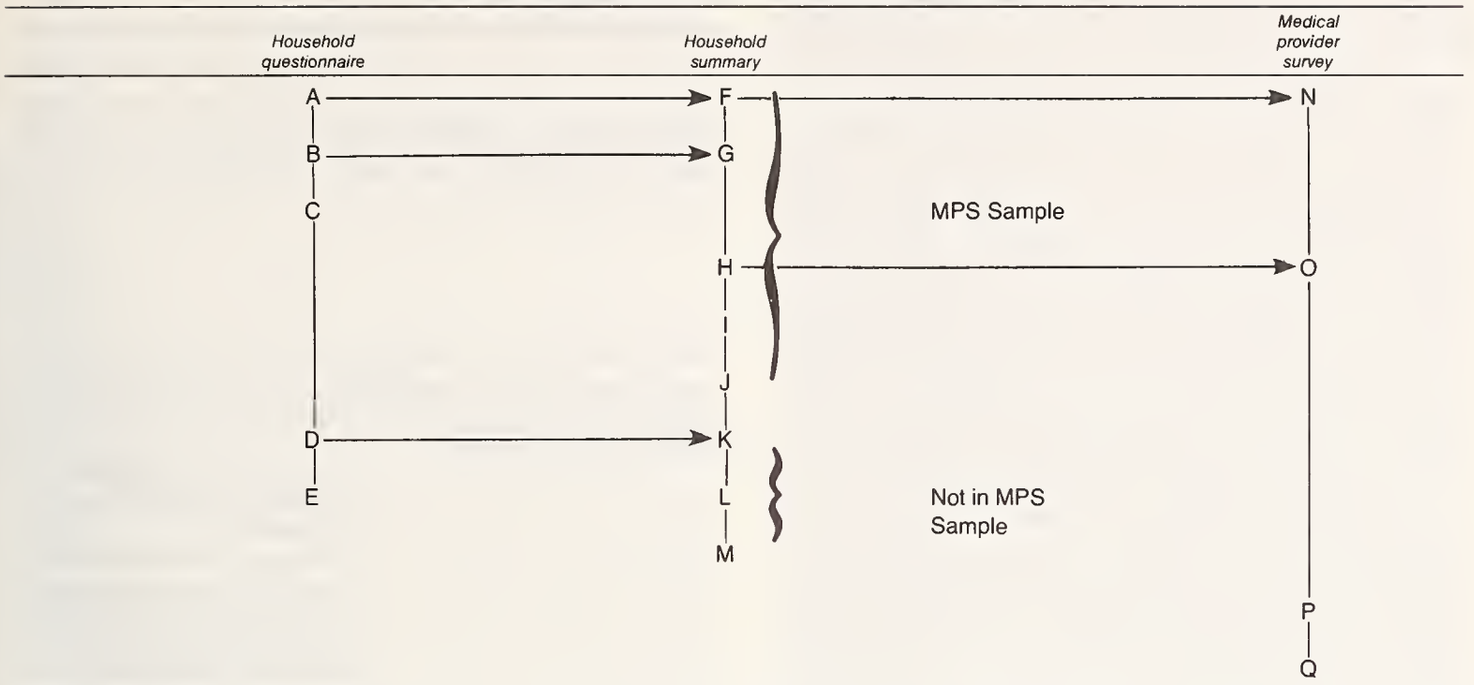
Estimating the number of physician visits and hospital stays

Two issues face the analyst of data from multiple sources. First, is there concurrence among data sources about whether an event occurred? Second, is there agreement about the characteristics of that event, for instance the charge or diagnosis associated with a physician visit?

Multiple sources of data about the same event virtually guarantee some disagreement. The first determination of whether an event occurred involves differences between the household questionnaire and summary. An event reported during an interview with a household was recorded in the questionnaire with information about where the health care was provided, the charge, sources of payment, and waiting time for appointments and treatment. The summary sent out for later review carried a very abbreviated version of this information: provider name and address, charge, and source of payment. When the final summary and questionnaire data were compared, there were 233 hospital stays and 7,268 physician visits that appeared only in the questionnaire data (type C, Table I), representing about 4% of all stays and visits reported there, and 63 stays and 1,167 visits that appeared only in the summary (types H-O, I and L), about 1% of all summary visits and stays.

Most of the changes made in the summary were of the sort anticipated, i.e., changes in charges or source of

Table 1
Types of records across three data sources—household questionnaire, household summary, and medical provider survey (MPS)



# of hospital stays	# of physician visits[a]	Type	Description
2,067	30,737	A-F-N	An event reported in the questionnaire, retained on the summary, and reported by a medical provider
1,055	34,080	B-G	An event reported in both household data sources but not confirmed by a medical provider
233	7,268	C	An event reported on the questionnaire, subsequently removed from the summary and from eligibility for the medical provider survey, viewed as a correction to the summary by the respondent
2,994	102,620	D-K	Events not selected for the medical provider survey
218	3,341	E	Events determined to be duplicate reports
23	186	H-O	Events reported on the summary and by a medical provider, respondents incorrectly introduced new events directly on the summary otherwise these events would be of the A-F-N type
12	323	I	Events reported only on the summary; these cases represent the same kind of field error as type HO; they are equivalent to type B-G
10	319	J	Events determined to be duplicate reports by respondents in the medical provider survey
28	658	L	Events entered directly on the summary, like types H-O and I, but for persons not selected for the medical provider survey
17	641	M	Events determined to be duplicate reports by respondents not selected for the medical provider survey
40	13,777	P	Events reported only by a medical provider
122	2,889	Q	Events determined to be duplicate reports by a medical provider

[a]Excludes 12,814 visits by MPS sample respondents to non MD/DO providers of health care.
Source: National Center for Health Services Research.

payment amounts. However, the removal from and addition to the summary of some visits and stays was probably inevitable. For all NMCES analyses, the summary has been regarded as the "best" household report of events and related expenditure data. Given the small levels of disagreement between questionnaire and summary in the number of stays and visits occurring, this seems a reasonable decision. However, this decision was not based on our ability to determine from the data that the summary was right and the questionnaire wrong, but rather on our belief that the reviewing process improved data rather than the reverse. There probably are some instances in which the questionnaire is more accurate in reporting a visit or stay, either because the respondent

who reviewed summary data was a different person from the one who originally reported it or through field or processing error. One of the major difficulties in arriving at a "best estimate" of number of events from multiple data sources is that the analyst usually has no evidence to determine which is the "right" answer.

Disagreements between households and medical providers as to whether a visit or stay took place are even more difficult to resolve. Some previous studies with record-check data have chosen provider reports or denials of events over household data (for example, Andersen et al., 1979). Marquis (1980), however, makes a convincing argument that in studies with what he calls an AB design (a household survey conducted first and then a record check on persons reporting occurrences of events), record-check data should not be used to arrive at a "best estimate" of how many events occurred. His primary argument is that the AB design is inadequate for this purpose since respondents who report no events are not included in the verification and false negatives cannot be detected. Two other major assumptions would be necessary in combining household and medical provider survey data to arrive at a "best estimate" on how many events occurred: (1) the assumption of no error in the record check data, so that events said not to occur by the verification source are rejected and previously unreported events are accepted; and (2) the assumption of face validity of events for which the verification source did not respond or for these same cases assuming an error rate based on the experience of the medical provider survey sample (in the case of NMCES this includes events reported by respondents not selected for the verification procedure). Apart from the design consideration raised by Marquis, the difficulty in making the two assumptions above is demonstrated by examining events not confirmed by the medical provider survey.

Table 2 lists four reasons why respondent reported events were not confirmed in the medical provider survey.[4] If the physician denied seeing the patient (1 in Table 2) or confirmed some events but not others (and perhaps reported some new events) (2 in Table 2), one may be willing to assume the provider is correct. It is more difficult to resolve cases that fall into categories 3 and 4 of Table 2, especially for NMCES data where the percentage of persons in the MPS sample affected is substantial (55% of MPS sample respondents with physician visits have at least one physician of type 3 or 4, and 32.5% of respondents with stays have at least one hospital stay of type 3 or 4). To form a "best estimate" of number of visits or stays for NMCES by combining these two data sources would require a decision for the provider or respondent in cases 1 and 2 (Table 2) and an adjustment for cases 3 and 4 based on the experience of cases 1 and 2 (the approach used in Andersen et al., 1979). A similar adjustment or some other approach using a weighting or imputation technique would be required for the sizeable number of events not included in the medical provider survey sample. Given the NMCES study design and the substantial adjustments to the data necessary to form a best estimate, Marquis's findings suggest there is no evidence that the estimates resulting from this process would be less biased than those based on the household summary or medical provider survey data alone. This is not to say that a best estimate dataset cannot or even will not be constructed from NMCES data. However, constructing such a

Table 2
Events not confirmed by the medical provider survey (B-G, I)

Reasons for no confirmation:		
	1.	medical provider responded that patient was not treated or was not a patient
	2.	medical provider responded but did not confirm some events (includes both inadequacies of the matching process and respondent or provider error in reporting the occurrence of visits or stays)
	3.	medical provider responded but no events could be matched to respondent reported events (includes both inadequacies of the matching process and respondent or provider error in reporting the occurrence of visits or stays)
	4.	medical provider did not respond
Persons with physician visits		
#	%	
4,727	44.2	Respondents for whom all physicians responded[a]
3,644	34.0	Respondents for whom at least one physician did not respond (4) or for at least one physician no events could be matched (3)
2,334	21.8	Respondents for whom all physicians did not respond (4) or for all physicians no events could be matched (3) or both
Persons with hospital stays		
#	%	
1,614	67.5	Respondents for whom all hospitals responded[a]
87	3.6	Respondents for whom at least one hospital did not respond (4) or for at least one hospital no events could be matched (3)
690	28.9	Respondents for whom all hospitals did not respond (4) or for all hospitals no events could be matched (3) or both

[a]A physician or hospital was considered to have responded if at least one visit or stay to that provider was of type A-F-N or H-O or the provider responded he did not treat the patient. This definition is significantly different from a field response rate for providers.
Source: National Center for Health Services Research.

dataset presents serious theoretical and practical problems.[5]

Reporting expenditure data in the questionnaire and summary

Differences in reporting characteristics of events is the second major issue raised by multiple data sources.[6] The rest of this paper is concerned with comparisons of reported charge and source of payment data between the questionnaire and summary, and the summary and medical provider survey.

Table 3 indicates that for almost 40% of hospital stays respondents were unable to provide charge data either in the questionnaire or in later reviews of the summary, but in about 14% of stays charge data was picked up on the summary. About 23% of physician visits had no charge data in either the summary or the questionnaire, while for 22% a charge was obtained later in the summary review. For almost 90% of visits or stays with charge data in both data sources, the amount reported was the same. This suggests the summary picked up some charge data not initially reported, particularly for physician visits. However, corrections on the summary to previously reported charge data were infrequent since the initial questionnaire amount and the summary amount were the same for 88% to 93% of all stays and visits.

The lower half of Table 3 shows that the family appeared as a source of payment for hospital stays more often in the summary (40.7% of stays) than in the questionnaire (19.5% of stays). For physician visits, the family is represented as a source of payment for about 60% of the visits in both, although the family appears as a source of payment for the same visits only for 47.5% of visits. When family was reported as a source of payment in both the questionnaire and summary, an amount was usually reported in both as well (84.3% for hospital stays, 89.2% for visits). For about three-quarters of physician visits with an amount paid reported in both summary and questionnaire, the summary amount was the same as the initial questionnaire amount. However for 24.6% of visits the summary amount paid by the family was lower. For hospital stays with an amount reported in both the questionnaire and summary, changes in the amount paid by the family were more frequent. For 16.0% of stays the summary amount paid was lower than the questionnaire amount, and for 21.2% it was higher. More changes to the summary were made with regard to family as a source of payment than for the charge. Changes in the amount paid by the family occurred more often for stays than visits. Third-party payers are more often involved in paying for hospital care, which may lead to greater uncertainty about what the family will pay out of pocket in initial questionnaire reports.

Table 3
Comparison of questionnaire and summary charges and family payment for stays and visits (A-F-N, B-G, D-K)

			Hospital stay		*Physician visit*	
			#	%	#	%
Charge for Stay or Visit						
Questionnaire		Summary				
Missing		Missing	2,270	37.1	39,206	23.4
Present		Missing	109	1.8	1,471	0.9
Missing		Present	823	13.5	37,492	22.4
Present		Present	2,914	47.6	89,271	53.3
$	=	$	2,562	87.9	82,858	92.8
$		$	229	7.8	3,798	4.3
$		$	123	4.2	2,615	2.9
Source of Payment						
Questionnaire		Summary				
Not family		Not family	3,379	55.2	43,657	26.1
Family		Not family	244	4.0	24,432	14.6
Not family		Family	1,542	25.2	19,736	11.8
Family		Family	951	15.5	79,615	47.5
Amount not reported		Not reported	33	3.5	1,692	2.1
Amount reported		Not reported	48	5.0	2,800	3.5
Amount not reported		Reported	68	7.2	4,115	5.2
Amount reported		Reported	802	84.3	71,008	89.2
$	=	$	504	62.8	52,156	73.4
$		$	128	16.0	17,492	24.6
$		$	170	21.2	1,360	1.9

Tables 4 and 5 examine the relationship within the summary between reporting charge and source of payment data. A charge was reported for 61.1% of all hospital stays in the summary. There is a clear relationship between sources of payment and ability to report the total charge. The family or private insurance were much more likely to be paying for care if a charge was reported; for instance Table 4 shows private insurance as a payer for 70.6% of stays where a charge was reported. The family paid some of the charge for about half of the stays with a reported charge. A charge for a hospital stay was somewhat less likely to be reported for those under 6 or 65 or older and for the lowest-income group. When no charge was reported, Medicaid was much more likely to be reported as a source of payment for hospital care (34% of stays with no reported charge compared to 4.9% of stays with a reported charge). This suggests that persons who pay some portion of their bill or receive statements from insurance payers are more likely to be able to report charges. When Medicaid pays, billing and payment are between the provider and the government or some intermediary and the patient generally remains ignorant of charge and payment data. Since Medicare pays for almost all persons 65 or older, it was less clearly associated with ability to report a charge.

Table 5 shows a similar pattern for physician visits. Three-quarters of all physician visits in the summary had a reported total charge. The lowest-income group was more likely not to report a charge. The family was named as a source of payment for three-quarters of visits with a reported charge. Medicaid was more likely to be a source of payment for children and the low-income group and within these groups charges were less likely to be reported when Medicaid paid.

With regard to charge and source of payment data, there appeared to be a strong relationship between ability to report information and types of payment for care since the latter affected what information was available to patients. For charge data the summary appeared to have more impact in reducing missing data than in changing data once reported. More changes were made to summary data with regard to whether the family paid for care and the amount paid by the family, particularly for hospital care. Despite the use of the summary, however, from one-quarter to one-third of charge data remained missing and had to be imputed prior to analysis.

Table 4
Summary reports of charges and associated sources of payment for hospital stays (A-F-N, B-G, H-O, I, D-K, L)

	Number of stays	% of all stays	% of steys with a source of peyment by[a]				
			Family	Private insurance	Medicaid	Medicere	All other
	Stays with cherge reported						
Total	3,774	61.1	52.5	70.6	4.9	16.9	4.3
Age							
Under 6 years	416	52.1	55.5	57.7	6.2	0.0[b]	3.6
6–18	333	63.8	49.8	78.4	3.3	0.0	4.5
19–54	1,804	66.0	52.4	76.4	4.5	0.7	4.2
55–64	511	65.4	53.4	78.3	3.1	10.2	7.8
65 or older	710	52.9	51.8	54.1	7.2	80.7	2.7
Income							
Less than $12,000	1,434	50.8	54.0	54.9	9.7	30.6	5.2
$12,000–$19,999	1,067	67.1	54.3	78.9	2.5	10.1	3.2
$20,000 or more	1,273	72.1	49.4	81.2	1.6	7.2	4.3
	Stays with no charge reported						
Total	2,405	38.9	22.7	34.0	29.7	20.6	12.0
Age							
Under 6 years	383	47.9	29.2	15.9	41.0	—[c]	9.9
6–18	189	36.2	14.3	32.8	46.6	3.7	9.0
19–54	930	34.0	16.6	38.9	28.5	2.6	17.4
55–64	270	34.6	18.5	34.8	28.1	14.8	14.8
65 or older	633	47.1	31.9	37.9	20.2	67.1	4.9
Income							
Less than $12,000	1,388	49.2	21.1	24.0	40.3	27.0	11.2
$12,000–$19,999	524	32.9	27.1	44.6	17.7	14.5	13.4
$20,000 or more	493	27.9	22.3	51.1	12.6	9.1	12.8

[a]Sources of payment add up to greater than 100% because a stay may have more than one source of payment.
[b]0.0 indicates quantity greater than 0.0 but less than 0.5.
[c]— = quantity zero.
Source: National Center for Health Services Research.

Table 5
Summary reports of charges and associated sources of payment for physician visits (A-F-N, B-G, H-O, I, D-K, L)

	Number of physician visits	% of all visits	% of stays with a source of payment by[a]				
			Family	Private insurance	Medicaid	Medicare	All other
	Physician visits with charge reported						
Total	127,522	75.6	74.9	24.7	2.0	8.4	14.0
Age							
Under 6 years	10,340	74.8	70.1	14.9	2.3	0.0[b]	22.7
6–18	19,094	74.1	72.3	24.4	2.0	0.2	16.9
19–54	57,303	74.4	73.9	28.6	1.6	0.5	14.2
55–64	16,818	77.9	80.2	27.3	1.6	3.8	11.3
65 or older	23,967	78.8	77.8	17.9	3.1	40.6	9.5
Income							
Less than $12,000	45,712	66.7	73.3	17.2	4.6	15.7	14.4
$12,000–$19,999	34,587	80.7	75.8	26.3	0.8	5.2	14.1
$20,000 or more	47,223	82.5	75.8	30.8	0.4	3.6	13.6
	Physician visits with no charge reported						
Total	41,085	24.4	11.0	26.1	29.7	8.7	25.4
Age							
Under 6 years	3,481	25.2	7.8	14.5	45.8	0.3	29.6
6–18	6,663	25.9	9.5	22.7	37.8	1.7	25.1
19–54	19,746	25.6	10.6	31.8	25.1	3.2	28.6
55–64	4,765	22.1	13.0	28.7	28.4	5.3	22.3
65 or older	6,430	21.2	13.8	16.8	28.1	40.3	16.2
Income							
Less than $12,000	22,838	33.3	8.8	2.2	44.1	11.8	21.3
$12,000–$19,999	8,246	19.2	12.1	36.1	16.2	6.3	30.7
$20,000 or more	10,001	17.5	15.0	42.6	8.1	3.7	30.5

[a]Sources of payment add up to greater than 100% because a stay may have more than one source of payment.
[b]0.0 indicates quantity greater than 0.0 but less than 0.5.
Source: National Center for Health Services Research.

Table 6
Comparison of summary and medical provider survey charges for stays and visits (A-F-N, H-O)

Charge for Stay or Visit			Hospital Stay		Physician Visit	
Summary		Medical provider survey	#	%	#	%
Missing		Missing	110	5.3	1,981	6.4
Present		Missing	82	3.9	1,807	5.8
Missing		Present	695	33.2	4,746	15.3
Present		Present	1,203	57.6	22,389	72.4
Charge reported	=	Charge reported	528	43.9	11,958	53.4
		by 0–5%	98	8.1	180	0.8
		by 6–25%	84	7.0	1,549	6.9
		by 26+%	97	8.1	2,309	10.3
Total			279	23.2	4,038	18.0
		by 0–5%	170	14.1	207	0.9
		by 6–25%	95	7.9	1,688	7.5
		by 26+%	131	10.9	4,498	20.1
Total			396	32.9	6,393	28.6

Reporting expenditure data in the summary and medical provider survey

Leaving aside stays or visits reported only in the summary or medical provider survey, how much agreement exists about the characteristics of events reported by both? Table 6 shows about 72% of physician visits and about 60% of hospital stays have a reported charge in both the summary and medical provider survey. For 33.2% of the hospital stays, providers reported a charge where the patient did not. For those stays where a charge was reported by both data sources, the charges were the same for 43.9% of stays and 53.4% of visits, however, where disagreements occurred the medical provider was more likely to report a higher charge (32.9% for stays, 28.6% for visits).

Table 7 compares summary and medical provider survey reports of sources of payment and amounts paid. An interesting pattern emerges here. There is a high level of agreement about source of payment between patients and hospitals in cases where private insurance, Medicare, and Medicaid paid for care (60.0% to 75.3%). There is less agreement about when the family paid; in 46.2% of the cases, both report the family paid, and for the rest one source reports family payment while the other does not. For physician visits, about 70% of the time there was agreement that the family paid. But physicians are much less likely to report Medicare or private insurance as payers (for 62.6% of visits, the patient reports private insurance and the provider does not, 63.3% for Medicare). One explanation is that patients and physicians see different sides of the same

Table 7
Comparison of summary and medical provider survey reports of sources of payment (A-F-N, H-O)

Source of payment for stay or visit			*Hospital Stay*		*Physician Visit*	
Summary	*MPS*		#	%	#	%
Family	Not family		289	24.4	4,662	20.7
Not family	Family		348	29.4	2,176	9.7
Family	Family		547	46.2	15,699	69.6
Amount reported		Reported	458	83.7	15,097	96.2
Amount not reported		Reported	89	16.3	602	3.8
Private	Not private		241	17.2	6,580	62.6
Not private	Private		130	9.3	909	8.6
Private	Private		1,031[a]	73.5	3,026	28.8
Amount reported		Reported	771	75.4	2,249	74.3
Amount not reported		Reported	252	24.6	777	25.7
Medicaid	Not Medicaid		92	24.6	1,173	34.8
Not Medicaid	Medicaid		58	15.5	432	12.8
Medicaid	Medicaid		224	60.0	1,768	52.4
Amount reported		Reported	41	18.3	351	19.9
Amount not reported		Reported	183	81.7	1,417	80.1
Medicare	Not Medicare		55	9.2	2,756	63.3
Not Medicare	Medicare		92	15.4	548	12.6
Medicare	Medicare		449	75.3	1,051	24.1
Amount reported		Reported	206	45.9	661	62.9
Amount not reported		Reported	243	54.1	390	37.1
All other sources	Not all other sources		91	17.6	1,671	21.4
Not all other sources	All other sources		323	62.4	4,376	56.2
All other sources	All other sources		104	20.1	1,745	22.4
Amount reported		Reported	32	30.8	713	40.8
Amount not reported		Reported	23	22.1	271	15.5
Amount not reported		Not reported[b]	49	47.1	761	43.6

[a]Includes 8 cases not shown where the medical provider survey reported private insurance as the source of payment but gave no amount.
[b]Normally, the provider could not report a source of payment without an amount because sources of payment were listed on the questionnaire and providers either filled in an amount or left a blank space. However, providers were asked to specify other sources of payment not listed and so could report on "other" source of payment without giving an amount paid.
Source: National Center for Health Services Research.

transaction. As fewer physicians accept assignment, they become less able to accurately report who ultimately paid for care since they are unaware of reimbursements by insurers to the patient. For Medicaid, there is more agreement but for 34.8% of visits the patient reported Medicaid as a payer while the physician did not. Whether these are instances of the physician withholding information cannot be determined. It should be pointed out that the questionnaire design did not allow a provider to indicate Medicaid or any other payer as a source of payment if the provider did not know the amount.[7] It is clear that when hospitals and physicians report Medicaid as a payer, they are far more able to provide amounts than are patients (about 80% of stays and visits where both sources reported Medicaid paid were missing dollar amounts in the summary but not the MPS). The "all other" payer category is interesting because providers reported many more stays and visits than patients that were paid by payers other than family, private insurance, Medicaid or Medicare. The large number of cases here suggests the possibility that some may actually belong in one of the other source of payment categories, in particular the Medicare or Medicaid categories, if providers reported "state" or "federal" as other payers.

Turning to the amounts reported, both providers and patients reporting a family payer were likely to report an amount that the family paid (83.7% for hospital care, 96.2% for physician visits). Hospitals and patients were both likely to report private insurance as a payer and to provide an amount paid. Hospitals were better able to report the amount Medicaid paid and were able to report an amount Medicare paid in about half the cases where Medicare was an agreed source of payment but the respondent did not give an amount. As mentioned, the questionnaire did not give providers the opportunity to name a source of payment without giving an amount paid. The exception was for "all other" payers (providers

Table 8
Estimates of mean charges and sources of payment for hospital stays from the original summary, medical provider survey, and imputed summary

	Original summary (OS)		*Medical provider survey (MPS)*		*Imputed summary (IS)*[a]	
	Mean charge[b]	*# stays weighted (in thousands)*	*Mean charge*[b]	*# stays weighted (in thousands)*	*Mean charge*[b]	*# stays weighted (in thousands)*
Total	$1,239	20,504	$1,302	23,944[c]	$1,425	32,770
Age						
Less than 6	595	2,438	510	3,097	748	4,592
6–18	780	1,846	826	1,766	907	2,744
19–54	1,067	10,133	1,136	11,333	1,175	14,996
55–64	1,841	2,669	2,065	2,594	2,055	3,903
65 or older	1,986	3,418	1,921	5,153	2,318	6,535
Family income[d]						
Less than $12,000	1,351	7,146	1,415	10,401	1,649	13,752
$12,000–$19,999	1,108	6,007	1,158	6,479	1,216	8,850
$20,000 or more	1,238	7,280	1,266	6,930	1,307	10,006

	Average percent paid by[e] *Family*			*Private*			*Medicaid*			*Medicare*			*All other*		
	OS	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*
Total	17.8	10.1	16.7	57.5	48.3	47.7	2.9	8.7	10.3	10.7	17.2	14.6	3.2	5.5	9.5
Age															
Less than 6	30.8	19.2	32.4	54.6	47.7	37.4	4.6	15.9	17.3	0.0[f]	0.5	0.0	2.8	5.2	10.4
6–18	17.5	9.7	15.9	71.7	58.2	61.9	1.4	10.7	14.1	0.0	2.5	1.4	2.2	5.9	5.8
19–54	18.4	10.0	15.7	69.1	62.7	61.4	3.2	8.1	10.0	0.0	1.4	1.0	3.6	7.5	10.6
55–64	13.6	7.7	13.6	61.7	54.9	55.3	1.6	10.0	7.5	8.0	9.1	10.6	5.7	6.5	12.1
65 or older	10.2	6.2	10.2	14.1	10.4	12.9	2.8	4.6	6.1	56.6	71.1	64.0	0.9	0.6	6.6
Family income[d]															
Less than $12,000	21.8	10.6	17.5	37.8	28.8	27.5	5.9	14.9	18.8	20.8	30.6	24.8	4.0	3.6	10.7
$12,000–$19,999	16.1	9.9	16.1	66.5	61.9	59.7	1.7	5.9	5.2	7.0	9.0	8.7	2.2	4.8	8.9
$20,000 or more	15.1	9.6	16.2	69.7	64.7	64.7	1.0	2.2	3.2	3.9	4.8	5.6	3.2	8.9	8.7

[a]Information on procedures for imputing missing data for hospital stays will appear in a forthcoming NMCES data preview on use and expenditures for inpatient services. Some stays for newborns and in nursing homes not usually included in hospital stay estimates have been retained for purposes of comparability with the original summary and medical provider survey data.
[b]Excludes stays with zero or missing charge. Missing charges have been removed from the imputed summary by the imputation process but remain in the original summary and MPS. This is the major reason for differences in number of stays on which the estimates are based.
[c]This is not an unbiased national estimate of number of stays because the MPS weight does not account for partial nonresponse of providers for persons with more than one provider of hospital care.
[d]Excludes persons with negative income.
[e]Excludes sources of payment with zero or missing charge. Percentages do not add up to 100 for medical provider survey or original summary because they have not been edited to make all sources of payment cover the total charge as has been done for the imputed summary.
[f]0.0 indicates quantity greater than 0.0 but less than 0.5.
Source: National Center for Health Services Research.

were asked to supply a name which was coded) and here providers did leave the amount missing in about two-fifths of the cases where both sources reported an "other" source of payment.

In the small percentage of visits where providers and patients agreed private insurance or Medicare had paid for physician care, both reported an amount paid in the majority of cases. Again, however, physicians were much more capable of reporting amounts paid by Medicaid.

Tables 8, 9, and 10 focus on estimates of charges and sources of payment from the three datasets.[8] Despite disagreements among data sources concerning the occurrence of specific events or their characteristics, the overall mean estimates from each data source are not strikingly different. The mean charge for a physician visit is the same across data sources (Table 9). The mean charge for a hospital stay in the original household summary, prior to imputation for missing charges, is only slightly lower than the other two estimates (Table 8).

For hospital charges, while the original summary estimates are usually lower, the differences are not substantial (Table 8). The average percentage paid by various sources was arrived at by summing all payment amounts that were not zero or missing for a source of payment type and calculating the percentage of total charges this represented. Missing data remains in the original summary and MPS; for example, in Table 8 the sources of payment shown accounted for only 89.8% of all MPS hospital total charges.

While hospitals and respondents reported the family paid in about the same percentage of stays (Table 7), the MPS generally reported a lower percentage of the total paid by the family than did the summary (10.1% overall versus 17.8% for the original summary). Private insurance was more likely to be reported as a payer for hospital care by the respondent and the percentage of charges covered by private insurance also was higher (57.5% compared to 48.3% for the MPS). The hospitals, on the

Table 9
Estimates of mean charges and sources of payment for physician visits from the original summary, medical provider survey, and imputed summary

	Original summary (OS)		*Medical provider survey (MPS)*		*Imputed summary (IS)*[a]	
	Mean charge[b]	*# visits weighted (in thousands)*	*Mean charge*[b]	*# visits weighted (in thousands)*	*Mean charge*[b]	*# visits weighted (in thousands)*
Total	$26	541,403	$28	489,432[c]	$27	722,582
Age						
Less than 6	18	47,638	18	47,341	18	64,496
6–18	21	82,155	24	71,783	23	113,608
19–54	28	249,934	33	218,496	30	339,325
55–64	27	67,775	25	61,996	29	88,420
65 or older	27	94,001	26	89,816	29	116,734
Family income[d]						
Less than $12,000	25	172,227	33	186,059	27	266,784
$12,000–$19,999	25	152,719	23	127,891	26	190,203
$20,000 or more	27	215,076	25	173,979	28	263,604

	Average percent paid by[d]														
	Family			*Private*			*Madicaid*			*Medicare*			*All other*		
	OS	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*	*OS*	*MPS*	*IS*
Total	67.4	63.0	53.9	20.9	8.6	21.6	1.4	6.7	7.5	5.4	3.2	5.7	1.9	6.3	9.8
Age															
Less than 6	79.4	65.2	62.5	15.6	5.4	14.1	2.3	14.4	12.1	—[e]	0.0[f]	0.0	1.1	4.6	9.4
6–18	70.8	58.9	56.1	23.1	11.2	22.2	1.5	11.1	9.4	0.0	0.6	0.5	1.9	5.7	9.8
19–54	67.5	62.0	52.7	25.4	10.6	26.6	1.4	6.1	6.9	0.0	61.8	0.7	2.6	8.2	11.5
55–64	68.7	69.3	56.9	23.0	9.7	23.4	0.9	2.7	6.1	3.3	1.4	3.5	1.6	5.7	8.9
65 or older	56.8	63.4	48.3	7.9	2.2	9.1	1.3	3.4	6.0	27.8	14.6	30.0	0.5	3.3	6.0
Family income[d]															
Less than $12,000	67.2	51.8	48.3	12.2	5.3	12.5	3.5	14.0	16.2	10.9	6.6	10.5	2.0	6.2	11.4
$12,000–$19,999	69.0	69.1	58.6	22.6	10.0	24.5	0.7	3.0	3.4	3.6	1.2	3.5	1.6	6.6	8.9
$20,000 or more	66.2	70.5	56.1	26.7	10.8	28.6	0.0	1.7	1.6	2.3	1.2	2.3	2.0	6.1	8.9

[a]Information on procedures for imputing missing data for physician visits will appear in a forthcoming NMCES data preview on use and expenditures for inpatient services.
[b]Excludas visits with zaro or missing charge. Missing cherges hava baan removed from tha imputad summary by the imputation process but remain in the original summary and MPS. This is the major reason for diffarences in numbar of stays on which the estimates are based.
[c]This is not an unbiased national astimate of numbar of visits. Excludes about 7,000 visits which wara covared by flat fee charges for purposes of comparability with the original and imputed summary data because tha MPS weight doas not account for partial nonresponse of providers to persons with more than one provider of physician care. Maan charges may increase when these visits ara included since flat faa visits tand to have a higher charge.
[d]Excludas sources of paymant with zaro or missing charga. Percenteges do not add up to 100 for medical provider survey or original summary because they have not been edited to make all sources of paymant cover the totel charga as has baen done for the imputed summary.
[e]— = Quantity zaro.
[f]0.0 indicates quantity graatar than 0.0 but less than 0.5.
Source: Netional Cantar for Health Sarvicas Raseerch.

other hand, attributed a larger share of payment for charges to Medicaid and Medicare payers. These trends held across age and income groups.

Table 9 shows similar data for physician visit charges and sources of payment. Again there was very little variation across charges in the three data sources. Physicians reported a lower percentage of all charges as paid by the family and private insurance and a higher percentage by Medicaid, as did hospital sources in the previous table. But a larger percentage of payment of charges was attributed to "other" sources by physicians (see also in Table 7).

It should be noted that, for physician care, the sources of payment shown only accounted for 87.8% of all charges, and for hospital care 89.8% of all charges. If the remaining percentages were accounted for through editing or imputation of missing data, some of the differences discussed above might disappear. In addition, given the superior ability of MPS respondents to report amounts paid by Medicaid (Table 7), it is not surprising that this represents a higher percentage of total dollars paid.

It is interesting to note the relationship of the imputed summary estimates to the original summary and MPS estimates.[9] For hospital care, the imputed summary most resembles the original summary in the percentage paid by the family, but is closer to the MPS estimate for private insurance and Medicaid. The imputed summary estimate for Medicare falls midway between the others. For physician care, the imputed summary estimate of percentage of the total charge paid by the family is lower than either the original or MPS estimates, closer to the MPS estimate for Medicaid as in Table 8, and it approximates the original summary estimate for private insurance. The Medicare estimates are similar for the total population, but for the most relevant age group, 65 or older, the imputed summary estimate is much closer to the original summary estimate. No standard errors were presented here, so some of the differences pointed out may not be statistically significant. Yet the distribution of charges among various payers is significantly affected by who is perceived to be paying for care, and Table 7 suggests some differences in perception between patients and providers. In addition, the particular mix of respondent reported data and MPS data in the imputed summary is subject to these differences in perception. A different mix of household and MPS data might change estimates and apparently would also alter the representation and share of charges paid of various sources of payment.

Table 10 is simply one more way of looking at the relationship between reporting charge and source of payment data. Looking only at hospital stays where the family was named as a source of payment, MPS sources still reported that families paid a smaller percentage of the charge for the stay (21.4% versus 38.1% for the original summary) but reported a higher mean charge than the summary data. For physician visits, the reverse occurred, with physicians reporting that families paid a higher percentage of the charge for visits where they reported the family paid (94.8% compared to 81.9% in the original summary). For visits where physicians re-

Table 10
Mean charge and percent paid by source of payment reported in the original summary, medical provider survey and imputed summary

	Original summary		*Medical provider survey*		*Imputed summary*	
	Mean charge per stay or visit[a]	*Average percent paid*[b]	*Mean charge per stay or visit*[a]	*Average percent paid*[b]	*Mean charge per stay or visit*[a]	*Average percent paid*[b]
When source of payment for hospital stay was:						
Family	$1,135	38.1	$1,414	21.4	$1,275	41.5
Private	1,235	83.9	1,237	77.8	1,343	80.6
Medicaid	1,337	84.0	1,502	68.2	1,705	86.7
Medicare	2,033	83.1	1,925	80.2	2,211	82.9
Other	1,456	70.4	1,508	72.8	1,912	80.6
When source of payment for physician visit was:						
Family	$23	81.9	$22	94.8	$24	81.7
Private	37	76.6	47	81.4	37	80.8
Medicaid	28	82.6	63	78.7	29	94.2
Medicare	35	72.0	47	74.9	36	75.3
Other	29	79.7	35	93.2	32	90.9

[a]Excludes stays or visits with zero or missing charge.
[b]Excludes sources of payment with zero or missing charge.
Source: National Center for Health Services Research.

ported private insurance, Medicaid or Medicare as payers, they also reported higher mean charges than did household respondents. The difference is particularly striking for Medicaid where the mean visit charge in the original summary is $28 while the MPS charge is $63.

Conclusions

The NMCES design presented many challenges to those involved with data collection and to analysts. This paper focused on some of the complexities which arise from trying to incorporate data from three of the major sources of data in NMCES—the household questionnaire and summary and the medical provider survey.

One of the major goals of the summary was to reduce the level of missing expenditure information by providing a mechanism to update this data. It appears to have served this purpose for a substantial number of cases. However, one-quarter to one-third of hospital and physician charges remained missing at the end of the survey period. One probable explanation for some of this missing data is the association demonstrated here between Medicaid as a source of payment and the inability to report charge data. This suggests that if a summary-type instrument is used, it might be targeted to persons with missing data for whom Medicaid is not a payer. Changes in initially reported charge data and removal of visits or stays from the summary were infrequent and cannot conclusively be demonstrated to have improved data quality.[10]

Several issues come to mind with regard to medical provider surveys for record check purposes. It is unlikely that funding could be obtained today to do a complete provider verification of a large household sample. Doing a verification on a sample of household respondents and using this data to adjust household reported data remains technically possible but theoretically very problematic for the reasons discussed in the early part of this paper. However, comparisons of household and medical provider data provide some interesting insights into the shortcomings of both data sources.

Mean charge data for the total population and across age and income classes are quite similar and surprisingly stable across different data sources. However, Table 10 suggests there may be more variation in charge for stays with certain types of payers. Thus the mean charges for low-income children or elderly blacks may not prove as stable across data sources.

Table 7 suggests that household respondents may be a better source than physicians for information on who ultimately paid for physician care. However, there are some major areas of disagreement with regard to family and Medicaid payers for hospital care, which are not easily resolved for one side or the other. It is clear that providers are far more capable of providing Medicaid payment data than are household respondents. Acquiring Medicaid payment data without provider sources remains problematic. Given that verification studies are less likely to be done, options such as imputation from external data sources (average Medicaid physician payments for broad classes of procedures by state or region for example) may need to be explored. Administrative records from the Medicaid program are another possibility. The National Medical Care Utilization and Expenditures Survey (a panel design survey sponsored by the National Center for Health Statistics and the Health Care Financing Administration) is using both national Medicare claims data and Medicaid claims data in four states to supplement household reporting. Their experience should shed some light on the advantages and limitations of such data.

While the final role of the MPS in the NMCES analyses is yet to be determined, it has already proved useful in missing data imputation procedures. It presents an interesting test of the limits of various data adjustment procedures and has helped define the shortcomings and strengths of household-reported expenditure data.

Footnotes

[1] Funding for NMCES was provided by NCHSR, which cosponsored the survey with the National Center for Health Statistics. Data collection for the survey was done by Research Triangle Institute, N.C., and its subcontractors, National Opinion Research Center of the University of Chicago and Abt Associates, Inc., of Cambridge, MA, under contract HRA 230-76-0268.

[2] Cox (1980) describes the sampling procedures and other aspects of the medical provider survey.

[3] The 1970 CHAS/NORC Survey is an example of a person-level survey with two data sources. Andersen et al. (1979) is a methodological study of this survey.

[4] Among the reasons are instances in which the computerized matching algorithm developed to match household and medical provider survey data was unable to achieve a match. Developing this matching algorithm was a lengthy, complex process. More information is available in Cooley (1981).

[5] Initial efforts in this direction involved attempts to adjust data for persons not in the medical provider survey sample using knowledge about the relationship between sample household and medical provider reports. While adjustments to total population mean expenditures or numbers of visits were acceptable,it proved difficult to maintain distributional properties of the data across age, race, and income groups that were not biased to a statistically significant degree. For information on these strategies, see Cox (1979) and Williams (1979).

[6] NMCES actually had multiple data sources, or multiple reports about the same data, within the household component of the survey because of its panel design. For instance reports of insurance coverage were collected in each interview and employment data was collected at two points. Furthermore certain information, like work status, appeared in both employment and limitations data, and whether or not families reported AFDC income had implications for Medicaid coverage. Discrepancies among these reports became obvious because we followed one population over time. Some discrepancies appeared to be legitimate reflections of changes over time, others did not. But in any case, the editing process was more complex due to multiple reports within the household data about the same events or characteristics.

[7] The following sources of payment were listed on the questionnaire and providers were asked to indicate amounts paid by each when applicable: the patient or his/her family, Medicare, Medicaid, Workers' Compensation, Blue Cross/Blue Shield, other insurance company(ies) (specify), other source(s) (specify), don't know source. Presumably, the

Medicaid space would be left blank both when the provider didn't know the amount or when it wasn't applicable, and the two cases cannot be differentiated.

[8] A comparison of original summary and imputed summary charge and source of payment data for dental care, prescribed medicine, eye care, and other medical equipment and supplies is given in Rossiter and Cohen (1981).

[9] While there are some zero charge events in the imputed summary, all missing charge data has been removed and sources of payment were made to add to 100% of the total charge. The imputed summary is an amalgam of MPS and summary data. For records of type A-F-N and H-O, if charge data was missing in the summary they were attributed from the MPS if available. All other missing data were then inputed from like summary cases using a hot deck procedure.

[10] This assumes a narrow function for the summary. The NMCES Summary also served as a means of clarifying provider names and addresses for the medical provider survey (Holt, 1981). In addition, this paper did not examine the extent to which the summary improved source of payment data beyond examining the amount the family paid in Table 3.

Dual-frame sampling in the Community Hospital Program Access Evaluation

Sara Segal Loevy, Center for Health Administration Studies, University of Chicago

Introduction

Dual-frame sampling uses two overlapping frames, usually list and area frames, to survey a population. Working with both list and area sampling frames raises several methodological questions which must be addressed. These questions derive from (1) the sampling process, such as determining the relevant domains for the respective frames and the relative size of the samples to be drawn from each; (2) field work, the complexity and cost of designing and collecting data from two types of ultimate sampling units, i.e., housing units in the area and particular people from the list; (3) data processing and analysis issues associated with constructing appropriate weights for combined subgroups from the two samples, estimating design effects and resultant standard errors, and assessing biases in the data stemming from differential nonresponse or from the definition of the list frame itself.

Considerable interest in the dual-frame approach has been expressed in recent years, particularly in the context of supplementary list samples of target groups of interest which may introduce relative cost efficiencies in the collection of information over a straight area probability sampling approach (Armstrong, 1979; Casady et al., 1981; White and Massey, 1981). One particular concern addressed in the literature focuses on the problem of constructing unbiased estimates from the two samples taking into account the overlap between the two frames and the joint sampling probability of individuals in this overlap (Bosecker and Tortora, 1978; Hartley, 1962; Ibid, 1974). Some research has dealt with the nonsampling problems associated with clearly delimiting the boundaries of the overlapping domains and the field costs and effort in working with two types of samples (Beller, 1979; Bosecker, 1978; Henning et al., 1978; Vogel, 1975).

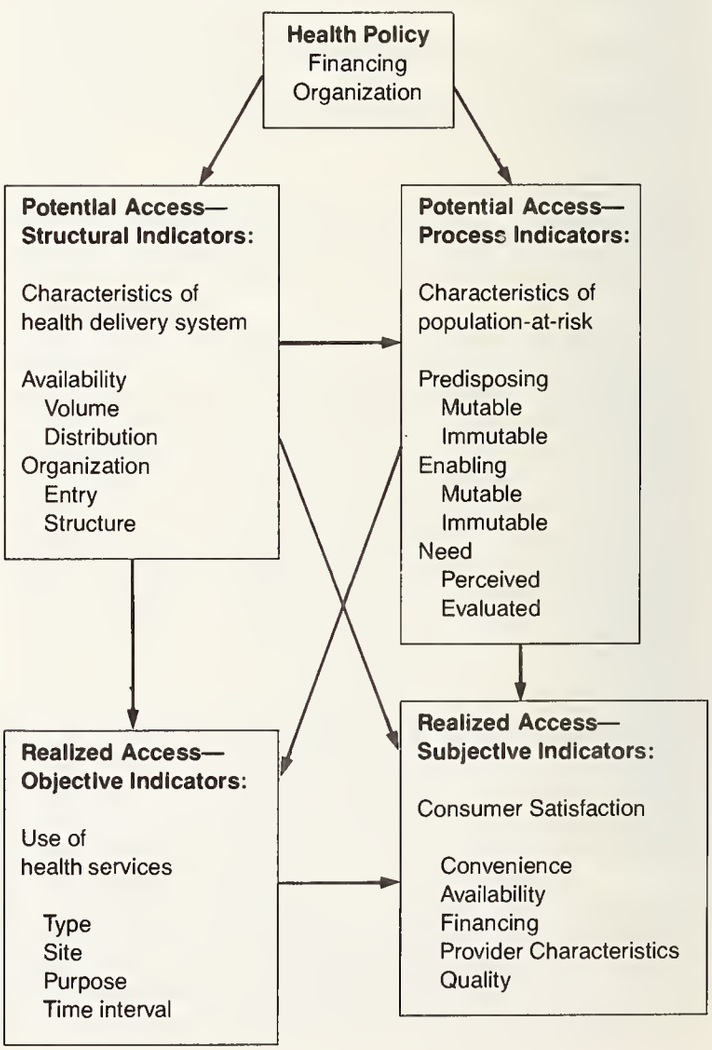
This paper provides information on the problems and solutions associated with carrying out a modified dual-frame sampling approach in a large social survey-based program evaluation. The paper summarizes the sampling, data collection, and processing and analyses issues associated with this approach. This information should prove useful to other people interested in implementing comparable research designs.

The research reported here was supported by a grant from the Robert Wood Johnson Foundation, Princeton, New Jersey. The author appreciates the contributions of the following Community Hospital Program Access Evaluation staff members: Lu Ann Aday, Ronald M. Andersen, Martha J. Banks, and Virginia Martin.

Background

In 1974, the Robert Wood Johnson Foundation initiated the Community Hospital Sponsored Primary Care Group Practice Program in an effort to promote swifter and more equitable access to medical care and to provide primary care on a continuing basis to families (RWJF, 1974; Block et al., 1978). The grant program was designed to assist community hospitals and medical staffs in strengthening the hospitals' role in ambulatory care by developing hospital-sponsored primary care group practices. Fifty-three programs received awards of up to $500,000 each; the money was intended to offset the planning and operating deficits of the hospital-spon-

Figure 1
Framework for the study of access



sored group practices for four to five years (Block et al., 1980).

To evaluate program effectiveness, the Foundation contracted with the Center for Health Administration Studies at the University of Chicago to study the ability of the Community Hospital Program to improve access to medical care in the communities it serves (Aday et al., 1978).[1] The evaluation is based upon the Center's behavioral model of access to medical care (Figure 1). This model "implies that characteristics of the delivery system (the availability and distribution of health care providers and facilities, for example) and characteristics of the population at risk in an area (their age, health status, insurance coverage, and income levels, for example) reflect the probable or *potential* levels of access to medical care whereas utilization and satisfaction measures may be considered indicators of actual or *realized* access to services" (Aday et al., 1981).

Study hypotheses

The central hypothesis of the access evaluation is that the Community Hospital Programs (CHPs) improve access to medical care in the communities they serve. The evaluation design, a modified panel study with two waves of data collection (1978–79 and 1981), uses both area and list sampling frames. The area samples provide estimates of community access measures, with the Phase I and Phase II data permitting estimates of longitudinal changes in access. Given the proportionally small size of the list samples relative to the community size, the list samples were not combined with the area samples to estimate community measures (Table 1).

Other hypotheses focus on the difference in access between site users and nonusers.

The list, or patient, samples provide the primary source for estimating the access experience of site users.

Table 1
Number of patients, service area size, penetration rates, cost ratios, and target number of completed cases, Phase I and Phase II

	Sites											
	(01)	*(02)*	*(03)*	*(04)*	*(05)*	*(06)*	*(07)*	*(08)*	*(09)*	*(10)*	*(11)*	*(12)*
Phase I												
A. Pts. ever seen by CHP	1128[a]	4967	—	7975	287	1640[a]	2024	1055[a]	6666	519	1212[a]	1700
B. Pts. elig. for interview[b]	749	2810	—	3989	165	990	1565	794	3935	390	738	757
C. Service area population	150,000	19,300	45,000	496,000	15,900	60,000	19,000	35,000	222,000	20,000	80,000	600,000
D. Penetration rate = B/C	.005	.146	—	.008	.010	.017	.082	.023	.018	.019	.009	.001
E. List:Area cost ratio	1.0	1.0	—	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F. Target # completed cases[c]	930	915	920	920	925	1035	915	900	920	975	1032	925
1. List	340	290	—	360	55	275	275	225	360	105	370	235
2. Area	590	625	920	560	870	760	640	675	560	870	662	690
Phase II												
A. Pts. ever seen by CHP	7470	7224	4662	22,658	5672	8024[d]	4482	2278[d]	—	2646	5088[d]	1192[e]
B. Pts. elig. for interview[b]	5689	4825	4110	10,990	3932	3866	2946	1485	—	1736	3414	989
C. Service area population	138,000	18,400	53,500	426,000	22,900[f]	72,600[f]	22,300	41,600	—	24,900	78,600	629,000
D. Penetration rate = B/C	.041	.262	.077	.026	.171	.053	.132	.036	—	.070	.043	.001
E. List:Area cost ratio	1.55	1.55	1.55	1.65	1.55	1.55	1.35	1.55	—	1.35	1.55	1.75
F. Target # completed cases	850	875	850	825	850	850	875	850	—	875	850	825
1. List	350	300	350	375	300	350	350	300	—	350	350	275
2. Area	500	575	500	450	550	500	525	550	—	525	500	550

[a] At Phase I, an additional 481 persons who were expected to use Site *01* were sampled from; Site 06, *208* additional persons; Site *08*, 1*070*; and Site 11, 1*600*.
[b] Differs from "A" because those who died, were institutionalized, or lived outside the service area are excluded.
[c] Target number includes an assigned cushion.
[d] At Phase II, an additional 236 persons who were expected to use Site *06* but hadn't yet were sampled from; Site *08*, 391 persons; Site 11, 1327 persons.
[e] Lower than Phase I due to record deactivation by the site.
[f] Service areas enlarged at Phase II.

User to nonuser comparisons, both longitudinal and cross-sectional, are possible by comparing users from both the list and area samples with nonusers from the area samples.

Study design

The access evaluation was conducted with a stratified sample of twelve CHPs. The original sampling frame, constructed in July, 1977, consisted of the 43 sites already funded. Six sites were then eliminated from the frame: one in Hawaii, due to data collection costs, and 5 due to an estimated low probability of surviving the 4-year period. From the 37 remaining eligible sites, a stratified sample of 12 sites was drawn. The strata were created from assessments by the Community Hospital Program Technical Advisory Committee of (a) the sites' potentials for having measurable effects on access indicators, (b) organizational strengths, (c) geographic location, and (d) hospital ownership (Catholic or non-Catholic) (Aday et al., 1981). Since 10 of the 12 practices were already providing services by the first wave (Phase I), the design is not a true pre-post evaluation design.

Shortly after the Phase I field period, one group practice (Site 09) withdrew from the program. Consequently, Phase II data were not collected from this group practice. In addition, the Site 08 originally sampled was involved in a hospital merger, and its physical location was questionable at the beginning of the Phase I field period. This uncertainty led to dropping it from the sample. A new, not randomly selected, Site 08 was chosen; data from this group practice will be presented as a case study.

Sampling decisions

The study originally intended to focus on longitudinal changes in the access experience of Phase I site nonusers who had become site users by Phase II. The original design called for completing 900 interviews at each of the two time periods (Phase I and Phase II) for each group practice site. Respondents would enter into the survey through a simple two-stage cluster sample of households within the site's service area as defined by patient origin studies. Early estimates predicted that 30% of the respondents would be site users at the time of reinterview, yielding enough site users to make statistically valid comparisons between users and nonusers.

The patient origin studies permitted better estimates of the sites' projected penetration rates, the proportion of users in the service areas. The origin studies indicated that area samples alone would yield inadequate numbers of users (Table 1). Penetration rates, a function of both patient load and service area size, ranged from a low of nearly zero to a high of 14.6% at Phase I. Given the low penetration rates, screening households for users would have substantially increased the costs of the study. Furthermore, screening would have revealed the focus of the study, which was politically inadvisable in some sites. To ensure an adequate number of users in the evaluation, supplementary list samples of site patients were incorporated into the study design.

Using the previously determined target of 900 completed cases, the general decision was made to complete approximately 300 cases in the list sample and 600 cases in the area sample at Phase I. The factors relevant to deciding the relative number of list and area cases in each site, generally in order of importance, included (1) the size of the patient load at the time of sampling, which varied from the one site not yet seeing patients at Phase I to the site with a patient load of 8,000, and the concomitant penetration rate, (2) the Phase II projected patient load and penetration rate, (3) relative size of the standard error of the difference of the list sample estimates to the area sample estimates, and (4) the importance of interviewing users at Phase I relative to interviewing nonusers at Phase I who would subsequently become users at Phase II. Interviewing Phase I early users, the last point listed, seemed valuable as a means of having both before and after measures for some individuals over time. Because the cost of completing a list case relative to an area case was unknown at Phase I, relative cost was not factored into the case distribution decision.

In the second phase of interviewing, the target number was increased in the list sample and reduced in the area. The distribution ranged from 325 to 375 in the list sample and from 475 to 525 in the area sample. An additional criterion was folded into the Phase II list-sample size decision, that is, capitalizing on the number of reinterviews while maintaining equal sampling fractions between the Phase I patients and the newly registered Phase II patients. Estimates of relative cost, based on the earlier sampling and field experiences, were also used in determining the relative number of list and area cases (Table 2).

Patient origin studies

To define the service area for a site already seeing patients, a random sample of patient addresses was tabulated and mapped. Service areas, at Phase I, include 74% to 89% of current patients. This range reflects an effort to inscribe a service area that corresponds to a known population base, such as census tracts, and to maximize both the proportion of patients in the sampling frame and the overall concentration (or penetration rate) of patients in the service area. For sites not yet seeing patients (Site 03) or seeing only few patients (Site 10), Phase I service areas were defined from information supplied by group practice and hospital staff.

The Phase I service area definitions were verified at Phase II through new patient origin studies of site users. Based on the results of the second origin study, the Phase II service areas were enlarged for two sites (05 and

Table 2
Example of factors considered in deciding Phase II (list to area cost ratio)

List:Area cost ratio	Penetration rate	# Cases		Standard Error of Estimate, 25% or 75% Items, for Different Analytic Groups				
		List	Area	Users[a] All to Nonusers[c] All	Users[b] 1/3 to Nonusers 2/3	Phase I area to Phase II area	Phase I list to Phase II list % growth in patient load 100%[d]	300%[e]
1.55	1%	250	650	3.6	5.8	2.3	3.8	5.4
		300	572	3.4	5.3	2.4	3.5	5.0
		350	495	3.4	4.9	2.5	3.3	4.6
		400	417	3.4	4.6	2.6	3.0	4.3
		450	340	3.4	4.3	2.8	2.9	4.1
	10%	250	650	3.5	5.4	2.4	3.6	5.1
		300	572	3.4	5.0	2.5	3.4	4.7
		350	495	3.4	4.7	2.6	3.1	4.4
		400	417	3.4	4.4	2.7	3.0	4.2
		450	340	3.5	4.2	2.9	2.8	4.0

[a]Users defined as everyone who has been to site (entire list sample).
[b]Restricted definition of users; assumes 1/3 list sample as users, 2/3 as nonusers.
[c]Community estimates.
[d]Growth rate permits 1/2 of Phase II list to be reinterviews from Phase I.
[e]Growth rate permits 1/4 of Phase II list to be reinterviews from Phase I.

06). For the other sites, the Phase I samples were also used to enlarge the original Phase I patient samples. The Phase II patient samples thus include patients using the sites since the Phase I cut-off dates.

Field work

Using dual-frame sampling created special fieldwork problems, particularly with the list sample. Interviewers had to work with two different listing forms,for example, since the list sample required interviewing specific, named respondents. In the area sample, however, one adult and one child (if present in the family unit) were randomly selected for interviewing according to a random number chart pasted into the listing form.

Three categories of list-sample cases required special consideration and effort during the field period: people who were also selected into the area sample (ASO), people who moved out of the area (MOSA), and people who moved but could not be located (MCL) (Table 3). The ASO cases were dropped from the list sample and retained in the area sample. Retaining the ASO cases in the area sample was one method for maintaining the representativeness of the community estimates. The random selection of respondents within a household in the area sample, however, meant that the original list-sample person, a probable user, was not always interviewed. Determining MOSAs often involved considerable work in establishing new addresses.

The last category, the MCLs, required costly and time-consuming tracking as the field staff attempted to find them by contacting neighbors, the post office, utility companies, and the CHP sites for updated addresses. In anticipation of this problem at Phase II, respondents at Phase I were asked to provide names and phone numbers of friends and family most likely to know the respondents' whereabouts. Despite these efforts, some proportion of the list sample remained unlocated in every site.

Greater losses in both MOSA and MCL categories occurred at Phase II, partially due to the increased age of the list frame. Community out-migration also influenced the Phase II losses in areas experiencing industrial loss (Site 03 and 06) or housing stock change (Site 02). In two communities (Site 01 and 12), highly mobile populations affected the number of MCL cases.

Response rates. The response rate goal was set at 80%, for both time periods and both samples. In most sites, this goal was achieved. The final response rates in the area samples, in which refusals and not-at-homes were encountered at the household as well as individual level, assume that the relative number of eligible respondents in noninterviewed households equaled the relative number in interviewed households.

The response rates in the list sample include an additional category of nonrespondents, the MCL group. Adjusting for this group of nonrespondents requires a slightly more complex assumption than that used for

refusals in the area sample. The assumption was that some MCLs had moved from the service area and were, consequently, ineligible for interviewing. To compute the final response rate, the arithmetic solution was that the proportion of MCLs still living in the service area (and thus eligible for interviewing) equaled the proportion of patients who were located at new addresses within the service area to *all* patients who were located at new addresses within *or* outside the service area:

$$MCL_{Adj.} = MCL_{Unadj.} \left(\frac{Moves_{In}}{Moves_{In} + Moves_{Out}} \right)$$

Analytic issues

Yield of users. One of the more difficult analytic issues in this study is the definition of a CHP site user. In the broadest terms, anyone who has ever been to the site, i.e., has a medical record, qualifies as a user. A narrower definition, which better fits the access framework, restricts users to those who consider the site their regular source of care.

This study used dual-frame sampling to ensure that adequate numbers of CHP site users would enter the study. If the sample had been confined to an area frame,

Table 3
Case disposition and response rates, area and list samples, Phase I and Phase II

	Site											
	(01)	(02)	(03)	(04)	(05)	(06)	(07)	(08)	(09)	(10)	(11)	(12)
Phase I												
Area sample												
Completed cases	647	635	884	500	870	750	671	701	550	872	581	669
Response rate[a]	78%	85%	89%	73%	83%	83%	85%	90%	79%	89%	84%	72%
List sample												
ASO	1	31	—[d]	0	10	11	35	9	1	10	11	0
MOSA	82	82	—	56	5	72	84	30	73	6	78	102
Known	14	57	—	38	3	63	63	25	51	5	55	44
Assumed from MCL	68	25	—	18	2	9	21	5	22	1	23	58
Other ineligible[b]	9	18	—	7	21	4	9	10	3	0	24	25
Eligible noninterview	109	49	—	94	5	47	25	22	73	3	22	75
Located	60	23	—	88	5	41	19	20	48	3	14	63
Assumed elig. from MCL	49	26	—	6	0	6	6	2	25	0	8	12
Completed cases	273	286	—	339	50	255	279	224	340	117	383	117
Response rate[a]	71%	85%	—	78%	91%	84%	92%	91%	82%	97%	94%	54%
Phase II[c]												
Area sample												
Completed cases	503	544	512	379	577	489	541	620	—[e]	516	537	524
Response rate[a]	77%	79%	86%	70%	80%	83%	89%	89%	—	84%	80%	68%
List sample												
ASO												
MOSA	7	25	10	0	25	7	45	12	—	24	9	0
Known	120	104	114	128	38	96	102	62	—	47	106	136
Assumed from MCL	46	51	62	75	32	63	58	38	—	41	58	65
Other ineligible[b]	74	53	52	53	6	33	44	24	—	6	48	71
Eligible noninterview	10	45	19	8	7	16	5	9	—	0	12	10
Located	252	105	80	83	64	72	21	47	—	53	45	135
Assumed elig. from MCL	60	75	49	70	62	47	21	43	—	52	28	106
Completed cases	92	30	31	13	2	25	0	4	—	1	17	29
Response rate[a]	375	283	364	338	324	347	356	314	—	318	389	239
	71%	73%	82%	80%	84%	83%	94%	87%	—	86%	90%	64%

[a]The final response rates for both the community sample and the patient sample were calculated by assuming that the relative number of eligible respondents in noninterviewed households equaled the relative number in interviewed households. In the patient sample, this meent that we assumed that, of the patients we could not locate, the proportion who still lived in the service area was the same as the proportion of located patients who had moved within the service area to all located patients who had moved from the address we hed obtained from their patient records. Phase II patient sample adjusted response rates are preliminary and subject to minor changes.
[b]Died, institutionalized, duplicete case, etc.
[c]Preliminary, subject to minor revisions.
[d]No patient sample, site not yet seeing patients.
[e]Site withdrew from program.

the actual number of users, using either definition, would have been statistically inadequate (Table 4). Ignoring the political implications of screening, it is clear that the costs of household screening would have been impractical even in sites with high penetration rates (Site 02, 07, 10) or high proportions of patients who consider the site their usual source of care (Site 10).

The list sample, by definition, channels a sufficient number of broadly defined users into the study. Applying the restricted definition of user, however, reduces the number of analytically available users. That is, some list-sample respondents known to have visited the site did not identify it as their regular source of care. This reduction is minimal in some sites (Site 10). In other sites, the loss is large enough to limit some of the planned, more complex analyses. Factors assumed to be important in explaining these losses include physician turnover (Site 02 and 11), particularly when physicians leave to open nearby practices (Site 06 and 07), severe staffing problems (Site 11), and local competition for patients (Site 04). These unpredictable factors, nonetheless, do not explain the entire loss of users from either sample.

The yield of narrowly defined users from the dual-frame sample meets most of the study's analytic needs. Clearly, an area frame alone would have been inadequate. Adding the list frame, despite the additional field problems and the extensive reduction in analytic users, proved to be a fairly effective method for drawing a rare population into the study.

It seems that the utility of dual-frame sampling depends in part on the governing characteristic of the list frame. When the frame is organized around an immutable characteristic, such as age or sex, or a mutable characteristic subject to little change, such as income or education, the observed yield should be high. When the frame is organized around a characteristic subject to change or interpretation, as with the user variable, the observed yield may be lower than expected. This decrease in yield is particularly critical when the characteristic serves key analytic objectives.

Weighting. Since the area and list samples were not

Table 4
Expected[a] and observed[b] number of site users from area and list samples, Phase I & Phase II

Number of users[c]	Sites (01)	(02)	(03)	(04)	(05)	(06)	(07)	(08)	(09)	(10)	(11)	(12)
Phase I												
Area sample												
Expected	3	93	—[d]	4	9	13	55	16	10	17	5	1
Observed	2	78	—	9	16	26	24	16	2	7	4	0
Observed-expected	−1	−15	—	+5	+7	+13	−31	0	−8	−10	−1	−1
List sample												
Expected	273	286	—	339	50	255	279	224	340	117	383	117
Observed	114	181	—	249	16	33	88	106	221	40	136	39
Observed-expected	−159	−105	—	−90	−34	−222	−191	−118	−119	−77	−247	−78
Combined sample												
Expected	276	379	—	343	59	268	334	240	350	134	388	118
Observed	116	259	—	258	32	59	112	122	223	47	140	39
Observed-expected	−160	−120	—	−85	−27	−209	−222	−118	−127	−87	−248	−79
Phase II												
Area sample												
Expected	21	143	39	10	99	26	71	22	—[e]	36	22	1
Observed	16	85	23	6	69	10	30	23	—	39	10	1
Observed-expected	−5	−58	−16	−4	−30	−16	−41	+1	—	+3	−12	0
List sample												
Expected	375	283	364	338	324	347	356	314	—	318	389	238
Observed	115	140	223	132	192	58	56	204	—	212	121	62
Observed-expected	−206	−143	−141	−206	−132	−289	−300	−110	—	−106	−268	−176
Combined sample												
Expected	396	426	403	348	423	373	427	336	—	354	411	239
Observed	131	225	246	138	261	68	86	227	—	251	131	63
Observed-expected	−265	−201	−157	−210	−162	−305	−341	−109	—	−103	−280	−176

[a]Expected users based on actual number of completed cases and on broad definition of "been to CHP site," equal to penetration rate in area sample. In list sample, expected does not take into account the portion of the list sample who had not yet been to site in S01, S08, and S11.
[b]Observed users based on restricted definition of those who consider CHP site as regular source of care.
[c]Unweighted n.
[d]Site not open.
[e]Dropped from Phase II.

combined to estimate community measures of access to medical care, creating relative sample weights to combine the two is not an issue. The weights required by each sample are thus determined independently.

Most of the list samples were self-weighting. In a few sites, potential groups of users were undersampled at Phase I and required additional weights to reflect the rate of undersampling. In one site, for example, we knew that the pediatrician was leaving the site before the field start to open a private practice down the road and would presumably be followed by some of the CHP patients. To protect against this loss of users, his patients were sampled at ½ the rate of other patients and thus require weights of 2.

Selection of area households, in general, was also self-weighting. Area sample weights thus reflect only the probability of being selected within a family unit. That is, the adult respondent is weighted up by the number of adults within the family and the child respondent weighted up by the number of children. These individual weights prevent sampling bias. Clusters were sampled equally in all but two sites (Site 06 and 07). In these two sites, peripheral segments expected to yield relatively few users were undersampled. These segments require additional weights to reflect the rate of undersampling.

To provide comparability with the area frame, the list frame was limited to patients residing within the site's service area. The list frame thus represents a subset of the area frame. Every user, using either definition, had the same probability of being sampled, that is, the probability of being sampled through the list or area sample or both [P = p(Area) + p(List) − p(Area + List)].

These equal probabilities resolve the theoretical problem of combining users from the two samples when comparing users with nonusers. An additional weighting factor was added to the combined group of users, however, which permits giving each interview equal weight. The additional weights reflect the ratios of the unweighted list and area cases to the total number of unweighted cases.

Standard errors. In most survey research, one sampling method is used. Standard errors of estimates, consequently, are based on straightforward formulas. The dual-frame process used in this evaluation, however, contains two different sampling methods: a systematic random sample in the list frame and a simple two-stage cluster sample in the area frame. Each sampling method demands a different approach to calculating standard errors. As a result, the two sampling methods complicate calculating the standard errors of estimates for the combined user group.

Standard errors of estimates from complex surveys, such as those using cluster samples, are traditionally calculated by computing simple random-sample standard errors and then multiplying by the square root of the design effect. The design effect takes into account the design complexities; it is the ratio of the variance of the sample to the variance of a simple random sample with the same n. The theoretical loss in precision due to the design effect in a cluster sample is often more than compensated by the increased sample size permitted by the cost efficiency of cluster sampling. In this study, the design effect for each site was obtained by computing standard errors for a sample of estimates and generalizing the results.

To being calculating standard errors of estimates for the combined user group, an estimated proportion (or mean) was calculated for the combined, weighted user group. The standard error of the estimate was then computed, using the unweighted number of users in the denominator to represent the actual number of cases. The standard error was then refined to take into account the generalized design effect (DEFF) present in the area sample. The adjustment involves multiplying the unweighted fraction of area users in the user group by the design effect, so that:

$$SE_{\text{adj for DEFF}} = SE_{SRS}\left[DEFF\left(\frac{n_{area}}{n_{area + list}}\right) + \frac{n_{list}}{n_{area + list}}\right]^{1/2}$$

Adjusting for the design effect requires several stages of custom programming, beginning with computing generalized standard errors and ending with site specific adjustments in the combined user group. These processes significantly increase the time and effort involved in data analysis. The delay, however, appears both unavoidable in light of the two sampling methods and justifiable in meeting the needs of the study.

Measurement biases. The organizing characteristic of the list frame may create measurement biases in some of the estimates of interest. In this study, the list sample domain, or organizing characteristic, is patients ever seen at the CHP site. Some key outcome measurements in the access framework focus on use of health services within the past year, for example, number of visits to regular source of care and number of contacts with all providers. Since most of the sites opened in the 18 months preceding Phase I field work, the majority of list sample respondents had seen a medical provider within the last year. These use measures, consequently, also served as criteria for list frame eligibility. Phase II data also contain these biases as the Phase II list frame included new patients, people who first used the site after Phase I sampling. Given the increased age of the entire Phase II list frame, the degree of bias is somewhat less at Phase II.

One method of controlling these biases involves deleting the newest users, those who first used the site within the year preceding fieldwork, from some of the use analyses. Identifying these newer users may be done in several ways. For sites that use medical record numbers

and maintain record number logs, the more recent users may be deleted based on record numbers assigned after the cut-off date. The study's identification numbers incorporate the medical record numbers. Another method depends on responses given to a question asked at Phase II: "When did you first go to [site name]?" Newer users could then be identified according to this response. Yet another approach would be to limit certain analyses to those seeing a doctor within the year.

Another access process measure, travel time to regular source of care, may also be biased. This bias stems from the geographic definition, the site's service area, of both list and area domains.

If other major health-care providers are located near the perimeter of the CHP service area, the CHP service area may overlap the centers of other provider-service areas. Nonusers, respondents who consider these other providers their regular sources of care, may spend less time traveling for care because the area sample encompasses only the centers of other service areas. Conversely, if other providers are located in a ring outside the perimeter of the CHP service area, the area sample may include a disproportionate number of other providers' more distant patients. This situation biases nonusers' travel time in the other direction.

Controlling the bias in this measure may be costly or difficult. The most obvious solution, geographically plotting other providers and estimating their service areas, is costly and error prone. The geographers' solutions, such as distance functions, may be useful if readily available. The simplest solution may be to acknowledge the bias and cautiously interpret differences in travel time between users and nonusers.

List frames are sometimes used to increase the number of respondents from a rare population. The organizing characteristic of the frame usually relates to some of the measurements of interest, as in this study, and may result in some measurement bias. When using a list frame, it is crucial to identify potential measurement bias and, whenever possible, find a method to control the bias.

Conclusions

Dual-frame sampling created numerous problems in the evaluation of the Community Hospital Program. The complications began with the sampling process, extended into the fieldwork, and persisted throughout data processing and analysis. Some of the problems were anticipated, such as estimating design effects and standard errors, while others were not, such as the yield of analytic users from the list frame. The problems, however, are resoluble.

Despite the difficulties, dual-frame sampling met a critical requirement of the design. That is, it permitted a sufficient number of users to enter the study so that key user and nonuser comparisons could be made. Given that the sampling design met this need, the complications imposed by dual-frame sampling are justifiable.

Footnote

[1] In addition, RWJF funded an organizational evaluation of the group practices. This study, conducted by the University of Washington School of Public Health, documents the organizational and financial development of the groups (Shortell and Dowling, 1978).

Discussion: Comparison of three data sources from the National Medical Care Expenditure Survey *and* Dual-frame sampling in the Community Hospital Program Access Evaluation

Kent Marquis, Bureau of the Census

Both of these papers raise a very difficult methodological issue. My goal will be to make these issues explicit, mention some potential solutions, and point out areas where basic methodological advances are needed.

Record check issues

Kasper's paper addresses nonsampling errors in the National Medical Care Expenditure Survey. In that survey, the designers expected a good deal of missing data and inaccurate reporting on the costs of family medical services. So, for part of the household sample, they included a record check that sought cost information from the doctors and hospitals mentioned by the households. They also used a panel survey design and an expenditure summary form so that respondents could report the costs of their health services in the future, say, after the bills arrived at the household.

Although the paper doesn't explicitly treat this point, my impression is that this combination of design features brought the cost item nonresponse rates down to very manageable levels. If so, this represents a very important technological advance in health cost surveys.

But how good is the provider record check? Kasper has decided it is not good enough to use for some purposes. What are her results? Why question their usefulness? And what needs to be done to get useful data? And, in looking at these issues, we can also address a question that has surfaced often at this conference: Is more reporting a sign of better reporting?

Let me pose a simplified record-check example to illustrate some general points about the procedure and then go on to discuss Kasper's data directly.

On a desert island containing 10 people, we will do a census[1] to estimate the annual per capita doctor visit rate. There are two doctors on the island who give us access to their billing records. If we do a full design record check and match records to survey reports, we get the outcome shown in Table 1. Note that cases e and f are not visit response errors but irrelevant match errors; the respondent reported the number of visits correctly (one); the error was on one of the variables used for matching, causing a match error but not an error in our count of desert island doctor visits.

The cross-classification table and estimates (for the full design) are shown in Table 2.

Table 1
Match of records to survey reports

Case	*Questionnaire*	*Record*	*Comment*
a	visit to X	visit to X	yes match
b	visit to Y	visit to Y	yes match
c	visit to X	none	Telescoping response error
d	none	visit to Y	Forgetting response error
e	visit to Y	none	Match error: name of doctor incorrectly reported
f	none	visit to X	

Table 2
Cross-classifications and estimates

Questionnaire \ Record	Visit	No visit	
Visit	A) 2 (Cases a,b)	B) 2 (Cases c,e)	4 visits
No visit	C) 2 (Cases d,f)		(N = 10 persons)
	4 visits		

Overreport = B/(A+B) = .50
Underreport = C/(A+C) = .50

$\bar{Q}$ = Questionnaire mean = (A+B)/N = .40

$\bar{R}$ = Record mean = (A+C)/N = .40

$\bar{S}$ = Average response bias = $\bar{Q}$-$\bar{R}$ = .00

Relative questionnaire bias = $\bar{S}/\bar{Q}$ = .00

But there are other ways to design the record check that save a lot of money when dealing with a large population of providers. The early record-check designs might be called AC designs. These are the ones that led to the conclusion that more reporting is better reporting. On the desert island, to do an AC design, we would go to the records first, find all the people with recorded visits, interview them to see if they will report their visits, and observe the results shown in Table 3.

We observe the matched cases, the forgetting error, half of the match errors, and conclude that the response bias is negative (predominantly underreporting). And after doing several such studies with different records, subject matters, and similar results, we would conclude: (1) that people usually underreport (forget or deliberately omit) and (2) that a new survey procedure that elicits more reports must be getting better reports. This

Table 3
AC design results

Questionnaire \ Record	Visit	No visit	
Visit	A) 2 (Cases a,b)	B) 0[a]	Underreport = .50 Overreport = .00[a]
No visit	C) 2 (Cases d,f)		
	4 visits		

[a]It is possible to observe cases in the B cell using the AC design such as when the record shows 1 visit for a person and the person reports 2 visits. But the AC design does not necessarily permit us to see all of the B-cell errors.

is one basis for "the more the better" criterion.

If we conducted an AB design on the desert island, we would administer the questionnaire first and ask the doctors to verify that each reported visit took place. Our results would be as shown in Table 4.

Table 4
AB design results

Questionnaire \ Record	Visit	No visit	
Visit	A) 2 (Cases a,b)	B) 2 (Cases c,e)	4 visits Underreport = .00 Overreport = .50[a]
No visit	C) 0[a]		

[a]It is possible to observe some, but not necessarily all, of the C-cell cases such as when the respondent reports only one of multiple visits to a single provider.

We would observe the matches, the telescoping error, and the other half of the match errors. A series of this kind of study would not lead us to conclude that more is better.

We can see that the type of incomplete design used will determine our conclusions about the sign of the response bias (whether underreports or overreports are larger). This happens for two basic reasons: (1) because only one class of true response error is (completely) observed and (2) because half of the irrelevant (e.g., match) errors appear as true directional response errors. In theory, the full design allows the irrelevant errors to "cancel each other out" so we are not misled about the size of the response bias and allows us to observe all of the true response errors so we are not misled about the sign of the average net bias.

To conclude that "more reporting is better reporting," we require that full design record-check studies consistently show either a negative average response bias or a predominance of underreport versus overreport errors. This appears not to be the case either for hospital stays (Marquis, 1978) or physician visits (Marquis et al., 1979).

The NMCES record check is a modified AC design, one that should produce more overreports than underreports and one that will cause irrelevant errors to appear mainly as overreports.[2] From Table 1 in Kasper's paper, the record check cross-classifications for hospital stays and physician visits are shown in Table 5.

On intuitive grounds, the author and the majority of the audience expressed the view that these data, analyzed in this way, are not giving a realistic picture of the survey response biases. It seems unreasonable to believe, for example, that half of the physician visits reported by households were made up or telescoped (true overreport errors). To distinguish between the true errors and the irrelevant or random errors, the full complement of relevant C-cell observations is needed, but the record-check design did not permit this to occur.

To make methodological progress, we need to develop practical record-check designs for health expenditure studies that do not rely on strong assumptions to produce bias estimates. In theory, to get an unbiased estimate of response bias, a record-check design should produce unbiased estimates for the A, B, and C cells. In practice, this is often not feasible unless the population of relevant records is easily located and accessible. We also need ways of distinguishing between the true response errors and the irrelevant errors that inflate the estimates of response error variance in record-check evaluations.

Table 5
Record check cross-classifications for hospital stays and physician visits

Hospital stay records

Questionnaire summary \	Yes	No	
Yes	2,090 (items AFN & HO)	1,067 (items BG & I)	3,157
No	40 (item P)		
	2,130		

Overreport = 1,067/3,157 = .34
Underreport = 40/2,130 = .02

Physician visit records

Questionnaire summary \	Yes	No	
Yes	30,923	34,406	65,329
No	13,777		
	43,700		

Overreport = 34,406/65,329 = .53
Underreport = 13,777/43,700 = .31

Sampling issues

Loevy's task is to say something about a community's use of a hospital-based outpatient facility, and the bulk of her paper is about a series of technical and operational problems that arose by using contemporary data to select a sample of a future target population. The specific estimation objectives aren't mentioned in the paper so it isn't possible to comment on the appropriateness of her solutions or about whether the most important threats were addressed. Instead, I'll make some comments about the general problem of using today's information to sample target populations at other time points.

In health surveys, today's data are used to design samples to represent both the present and the past. The general problem is that we sample from nonstationary populations: people are eligible today but not necessarily yesterday or tomorrow, or people are ineligible today but eligible yesterday or tomorrow. This is because of geographic mobility, births, deaths, transitions to and from the military and other institutions, reformulated family and household groupings, and refusals to continue participating in surveys. As Carl Morris might say, we base our sample designs on the URN model when the reality we seek to represent requires a SIEVE model (Morris et al., 1980).

Loevy has used a dual-frame approach to the nonstationary population problem, sampling both geographic area frames and list frames of past users. She describes a number of problems but it appears to be too early to say whether her approach has succeeded in getting the research closer to its estimation objectives.

Other solutions to the nonstationary population problem include:

1. The rotating panel design of households. The panel feature allows prospective study of each cohort eligible at a beginning time, and the rotation feature provides for more or less continuous resampling of new cohorts of eligible units. In theory, people who move in and out of survey eligibility status can be reselected during future periods of eligibility. The procedure is efficient for making area estimates of change but this is less than satisfactory if each person's complete history is necessary for estimation. In the last case, expensive tracing procedures may be needed.
2. Sample compensation. If one expects losses in certain groups, one can oversample these groups initially and still end up with the sample sizes required to meet estimate precision requirements. It is usually necessary to assume that the future losses occurred at random with respect to the estimates of interest.
3. A possible third alternative is a truncated version of the panel design and tracing method. One panel of dwelling units is sampled and whenever the people composition of the dwelling unit changes, ad hoc decisions are made about who is to remain in the study. The underlying decision criterion is to preserve the relationship of variables (in a people by variable matrix) rather than to preserve a representative sample of people. This is very much like the principles underlying experimental design.

A second problem with the URN design involves measurement error. So far I've discussed only true change issues in eligibility variables. Usually the measures on which we base our sample selection are imperfect; they reflect not only the true characteristic of interest (such as "permanent" income) but also transitory (true) changes, and various kinds of response errors.

The effects of measurement errors depend on how the measures are used in sample selection. If the measures are used to form sample strata and if probability samples are selected from all strata, then measurement error will usually just decrease the precision gains that are theoretically expected from stratification. If measures are used to screen population elements for eligibility (forming some strata excluded from selection), then biases can occur. Misclassifications will exclude unique classes of eligible units and include ineligible units. The latter can be eliminated at the analysis stage but the former cannot be recovered easily.

So, health survey methodology can also benefit from advances that produce realistic alternatives to the URN sample selection model and models that also take measurement errors into consideration.

Footnotes

[1] We could do a sample survey, but, for illustration, we will ignore the random sampling issues.

[2] Some of the modifications that move the design from the pure AC version closer to the full version are asking doctors and hospitals about visits for all household members and for services rendered within a long time interval.

Open Discussion: Session 4

Consumers' knowledge of their health insurance coverage

In responding to the discussant, Walden indicated that they encountered substantial questionnaire design problems in the process of developing the health insurance supplement. They were not as comfortable with the questionnaire as they would have liked when they went into the field. It is hard to get at the notion of deductibles. Wilensky explained further that the design allowed respondents to report deductibles that went across different people in the family, but the coding may not have handled them well. Kovar was not sure that respondents counted as deductibles those that crossed people. The question is in the use of the term.

It is much easier for them to tell how not to ask questions, said Walden, than how to ask them. It is clear that people cannot be asked about coverage plan by plan. People in the pretest thought about all their plans together when they responded about coverage for a particular type of service. For example, people with insurance to supplement Medicare consider Medicare and the supplemental plans together when asked if insurance will pay for things. Wilensky added that the health insurance supplement of the National Medical Care Expenditure Survey (NMCES) took an abstract approach about general coverage without referencing any event. Respondents were first asked if they had coverage for a type of medical care, and, if so, the amount the coverage paid. The Rand study took a different approach and used a hypothetical situation. Respondents were asked if their policies would pay if they had a specific type of illness. Wilensky then asked if anyone knew how the NMCES comparisons of knowledge about health care access compared to other surveys where knowledge was measured in a different way.

Marquis responded that the Rand study used abstract questions about knowledge in the first interview site. The abstract questions were of the nature, "Are you covered for hospitalization (doctor visits)?" They discovered a substantial underreporting of coverage when they compared policies obtained through a record check procedure, especially for coverage of doctor visits outside of hospitals. These early findings agreed with the work of R. Andersen and others. They then designed an actual experiment to compare the results of the abstract questions with a less abstract approach such as, "Well, if you spent $1,000 on a hospital bill and then you had a doctor visit outside the hospital, would the cost for that doctor visit be covered?" The results from the less abstract questions were in much closer agreement with the record check results than those obtained from the abstract questions. This led them to the principal that if they really wanted to know, they had to ask specifically.

Satin suggested that the relevance of the paper was summed up by Walden's comment that the questions asked in the survey may not have been relevant to the problem. A person may not know information because they have no need to know; this says nothing about their ability to get the information. The fact that people do or do not have knowledge on the tip of their tongues concerning the full ramifications of their health insurance says nothing about their ability to make a decision if presented with options. The wrong question is often asked; the question should elicit information necessary for policy decisions.

Kulka wondered why we want to measure knowledge of insurance coverage? A study currently being done for the Health Care Financing Administration was designed for the specific goal of finding out knowledge of coverage of Medicare supplementary plans. They oversampled new policy owners, assuming that they would have a greater knowledge of the policies than those who had purchased policies earlier. The results, however, showed they had even less knowledge than NMCES respondents. It was the recent *users* of insurance who knew their coverage.

Kulka was interested in studies attempting to validate data. Validation studies generally deal with over- and underreporting, and these have a basic well defined model. Comparing dollar amounts moves away from this over- and underreporting model and toward the direction used by the paper. This new type of model is needed to understand the validity of Cannell's interviewing techniques and the validities of telephone versus personal interview modes of data collection.

Sample designs and screening methods for increasing the representation of multiple small domains in health record surveys

Drummond responded to a question on the relevance of his study to the National Hospital Discharge Survey (NHDS) and on what to do about overlapping populations. Suppose a population is split into two groups, one containing 1% of the population and the other group 99%. Nothing else is known about these groups. If 100 records of each type were desired, a straight sample of 10,000 records would be required, producing 9,900 records for the larger group. Blind screening can be very expensive with little return when the groups of interest are rare subpopulations. His design says that if the decision-making needs can be specified beforehand, he can select a more efficient sample with a minimum use of resources. The procedures represent the minimum con-

trol that should ever be considered. Some subpopulations cannot be defined by strata variables, since they are the ultimate sample units. When there is a limitation of resources, we are forced to use a method like this. The reference to NHDS was made as a humorous aside. The large NHDS with around 400 hospitals and 250,000–300,000 records still can't be used to estimate many types of diagnostic categories. If they wanted to, they could achieve much greater control by using a technique such as this. Maybe they could use it every third time.

On the other point, about 5% of the paper was devoted to the problem of overlapping populations, when people can be in more than one subpopulation. This can be easily handled with multiplicity estimators. However, the multiple groups represented by a single record is an advantage of the methodology. For example, a combination of specific drug use can be defined as the screening requirement for a drug study. Here we force multiplicity in. We want to know what happens when drugs A, B, and C are used in combination.

Sudman found the paper interesting, but limited. The general approach to achieving equal sample sizes of two groups, e.g., black and white people, has been to take a general sample based on the population size of the smaller group, and then subsample the larger group. This deals with the same issue addressed by the paper, but is simpler since only the population size is needed. The advantage of the procedure in this paper, if there is one, is the reduction in variability caused by differential reactions of people in various categories. The disadvantage is that it requires knowledge of the fraction of discharges of each type of interest. Is this available without going through an initial investigation? The difference in efficiency is important to consider in relation to the cost. Drummond responded that self-weighting can be achieved without knowing the fraction of discharges of each type of interest, but equality of workload cannot be achieved without knowing it.

Warnecke observed that not all hospitals have the same case-load mix. Are data available beforehand about the differences in service mixes for secondary and tertiary facilities? Would you want to confine the analysis to secondary hospitals? Kovar commented that one hospital can have a very high proportion of a type of service and another hospital have none. In that sense, there is partial control. Drummond responded that the service mix needs to be known, but it is difficult to obtain. He could make the tertiary service its own hospital or make it a special service center with its own sampling probability.

Comparisons of data sources in the National Medical Care Expenditure Survey

Kasper commented that the cell containing stays reported by the household but not by the hospital includes matching problems. A complex matching computer algorithm was used, but it did not match everything it should have. For example, a household-reported stay in January might not have been matched if the hospital reported the stay as being in December (but that might be a poor example).

Dual-frame sampling in the Community Hospital Program Access Evaluation

Loevy agreed with Marquis that there are important problems in panel designs if characteristics are not immutable.

Stimson was not clear on how the area sample was selected. Loevy clarified that it was a two-stage area probability design, where the areas were restricted to the geographic service areas of the program. Patients on the list sample also had to live within the geographical confines of the same area. Stimson then commented that this might explain why there was such a high dropout rate over time. It probably varies greatly with distance from the service center. A gravity-type probability model could be used to give a differential weight that varies with distance from the core of the service area. This would give a better idea of the true dropout in the program.

Drummond asked if the area frame took dominance over the list sample. If so, there is no possible way to remove overlapping estimates when the list sample cases did not turn up the actual area sample. There are ways to estimate variance in mixed frame samples, however. One is to make individual estimates from each of the samples and then combine them. Multiplicity in this type of situation is catastrophic.

SESSION 5:
Hiring, training, and monitoring interviewers

Chair: Robert Fuchsberg, Division of Health Interview Statistics, National Center for Health Statistics

Recorder: Owen Thornberry, Division of Health Interview Statistics, National Center for Health Statistics

U.S. Bureau of the Census random-digit-dialing experiments: An analysis of job requirements for telephone interviews

Barbara H. Lacey, Bureau of the Census

Introduction

The United States Bureau of the Census is conducting a series of experiments primarily to investigate the response rates, sampling bias, and costs of the Census surveys conducted by telephone, using random digit dialing as the basic method of sampling. A major interest in these experiments is the determination of how and to what extent the typically high quality of the Census surveys, generally conducted in person, is affected by the telephone interviewing method of data collection. Because it is believed that interviewer competence has an impact of the quality of surveys and their cost efficiency, an investigation of the requirements of the telephone interviewing job was undertaken.

In order to assess a job applicant's ability to perform the established job requirements, employee selection procedures were developed through a content validation strategy. This strategy establishes and documents the relationship and representativeness of the procedures to the content and context of the job. Further validation of the procedures is planned, first through cross-validation of the content validity of the procedures, and second through criterion-related validation of the procedures to assess their value in predicting job performance. The purpose of this paper is to discuss the content validation of telephone interviewer selection procedures for the Census Bureau's random-digit-dialing (RDD) experiments, with particular focus on the job analysis which is the essence of a content validation strategy.

Methodology

Job analysis—secondary sources. The first phase of the RDD telephone interviewer job analysis involved a review of existing written job data, the so-called secondary sources of job information. For jobs already in existence within an agency, these secondary sources normally include formal position descriptions, statements of job instructions and procedures, organizational charts, or the like. In the case of the RDD job analysis, however, the chief secondary sources included verbal and unpublished written reports, training manuals and applicant and employee evaluation forms provided by some public and private survey research organizations, and work manuals and self-studies for interviewers involved in the Census Bureau's surveys.

Survey research organizations referenced in the literature as having conducted empirical research or having otherwise reflected considerable interest in interviewer selection were contacted by letter. Exhibits 1 and 2 are the contact letter and a list of the firms contacted, respectively.* The letter requested the results of any job analyses underlying the firms' personnel selection programs and information on the nature of the selection procedures themselves. Eleven responses were received to the 27 inquiries.

The literature search revealed a variety of opinions and often contradictory empirical findings as to which criteria were critical to job success for the telephone interviewer and should therefore be applied in employee selection. The books and articles used in the review appear in the appendix of references at the end of this paper.

The primary focus of most of this literature was on the personal characteristics of interviewers who tend to perform well on the job, rather than on the behaviors required for successful job performance. Under the legal guidelines which provide direction for Census validation activites, the Uniform Guidelines on Employee Selection Procedures (USEEOC, 1978), a focus on observable work behaviors and/or on knowledges, skills, and abilities that are operationally defined is preferable to a focus on personal traits. The validation of selection procedures purporting to measure the personal traits of interviewers would require a more extensive, complex, and arduous validation effort (namely, construct validation), for which time, personnel, or funding were unavailable in this survey. (Resource availability was a major consideration affecting the choice of methodology for several aspects of this study.)

Record was made of those work behaviors or operationally defined knowledges, skills, and abilities identified in the literature as necessary for either personal or telephone interviewing. Recorded were 105 reasonably discrete job requirements, with note as to their applicability for personal and/or telephone interviewing. There was a deliberate effort on the part of the researcher to avoid consolidating job requirements derived from different sources where there was evidence to suggest that differences in the requirements, although slight, were intended. Considerable consolidation would have most assuredly reflected the researcher's bias. It was believed that if indeed the requirements identified were redundant, then that would be revealed in the second phase of the job analysis, which entailed a survey of primary sources. Job requirements having applicability for telephone interviewing and mentioned by two or

* Note: Exhibits appear at the conclusion of this paper.

more secondary sources were earmarked for inclusion in this survey.

Job analysis—primary sources (the Job Requirements Survey). Each preliminary job requirement established through the investigation of secondary sources and not already expressed as an observable work behavior was carefully expressed in operational terms with the aim of preserving the meaning intended by the secondary source. The requirements were then grouped into the following categories:

Interviewing Skills
Voice Characteristics
Reading and Other Verbal Skills
Clerical Skills
Survey Content and Mechanics
Equipment Operation
General Work Habits
Other General Job Requirements

All identified work behaviors were subsumed under one of these headings and incorporated into a survey instrument called the Job Requirements Survey(see Exhibit 3). Space was provided for the survey respondents to list additional observable work behaviors not already specified on the survey instrument.

The main objectives of the Job Requirements Survey were to establish the relevance of the listed work behaviors to telephnone survey interviewer work, to determine their relative importance to the overall success of surveys, and to determine the need for an interviewer to perform these behaviors upon entry into the job. Subsidary objectives were to determine the degree to which the behaviors differentiated quality of work among telephone interviewers, to identify behaviors to be covered in training or in job performance measurement techniques such as monitoring, and to determine the relative difficulty of training interviewers in the behavior. The questions and scales used on the survey instrument to derive this information were a modification of those recommended by the Office of Personnel Management for job analysis purposes (USCSC, 1975).

General information was requested about the survey respondents and the organizations which employed them. (See Exhibit 3.) Respondents' names were requested to ensure that only one completed survey questionnaire was tabulated for any one respondent. To find out about the nature of the sample, other information was requested on the position the respondent occupied and the number of years of experience he or she had in work directly related to telephone interviewing. Information requested about the respondents' employer included the size of the telephone interviewing staff normally maintained, the average years of experience of all telephone interviewers employed by the organization, and the average number of hours per day the telephone interviewers worked.

Survey research organizations which responded in the first phase of the job analysis, as well as several other recommended organizations, were contacted by telephone to request their assistance in the Job Requirements Survey. Survey questionnaires were mailed to key contact persons at the consenting organizations. A list of these organizations is provided in Exhibit 4. The contact persons were advised of the types of respondents being sought for participation in the survey: persons who had been working with telephone interviewers in a supervisory or managerial capacity for at least one year, and persons highly knowledgeable about telephone interviewer work through directly related personnel management or research experience. The criteria for respondent eligibility were repeated in the cover letter to each survey questionnaire (see Exhibit 5).

Results of analysis

Of the 16 survey research groups asked to participate in the Job Requirements Survey, employees from 13 of them responded with 58 completed questionnaires. Table 1 provides the number of questionnaires received from each responding organization,[1] as well as some other interesting information on the organization's telephone interviewing staff. One-third of the respondents were employed with a single commercial survey research firm. Excluding one organization, which is a fledgling to telephone interviewing though not to survey research, the normal telephone interviewing staff for the re-

Table 1
Job requirements survey—telephone survey interviewers—responding organizations

Research organization	*No. of respondents*	*Normal staff size*	*Average yrs. experience*	*Average hrs. worked/day*
G-A	1	—[a]	—	—
C-A	3	130–175	.5–2+	5–7.5
A-A	1	50	4	5.5
A-B	1	20	5	5
G-B	9	10–40	≤1	5–6
C-B	6	150–250	1.5–3	6
C-C	1	68	4	6
C-D	20	300	.7	6
C-E	2	13–18	2–4	8
A-D	4	6–10	1.5	6
A-F	3	15–30	2–5	3–5
C-F	2	572	2+	6
C-G	5	50	.75	6
Total	58			

[a]Hiring not yet begun.

sponding organizations ranged from a low of 6 to a high of 572, with a median staff size of about 50 interviewers. The average job tenure of their interviewers varied from 6 months to 5 years, with 8 organizations reporting averages between 1.5 and 5 years. The average working hours per day ranged from 3 to 8 hours with most staffs working a 5- or 6-hour day. These figures suggest that the responding organizations had more than just a transitory involvement in telephone survey operations.

As shown in Table 2, 55 of the 58 respondents served in positions from monitor to assistant director of the survey research firm. Three persons did not provide their job titles. Three of the 55 respondents who did provide titles were involved in the training aspects of telephone survey work. Most of the positions were at least titularly designated supervisory or mangerial. Also shown in Table 2 is the mean number of years of experience the respondents had in work directly related to telephone interviewing. Two respondents had as little as 10 months' experience in directly related work, while one had as much as 20 years' experience. Another two respondents indicated that they had worked 30 years and 10 years in survey research, respectively, but they could not determine what proportion of those years involved telephone research. The mean for the 53 respondents providing meaningful data on experience was 5.6 years of experience. Hence, the average respondent could base her or his responses to the Job Requirements Survey on ample experience in telephone survey interviewing.

Questions A, B and C on the survey were used to determine the most important work behaviors for employee selection purposes. These questions are shown in Table 3. Also shown is the number of responses obtained for each observable work behavior listed, and the percentage of the respondents responding in specific ways to each question for each work behavior.

Most work behaviors listed were shown to be highly relevant for the telephone interviewer job, with 90% or more of the respondents indicating that these are behaviors in which their telephone survey interviewers (TSI's) are involved. Although opportunity was provided for respondents to identify other observable work behaviors not already listed on the survey, only five persons added any behaviors, and only one of these behaviors was named by at least two respondents: works effectively with minimal supervision.

In cases where a particular behavior was not relevant for their interviewers, respondents were instructed to skip the other questions regarding that behavior. Therefore, the sample size for questions B and C is sometimes smaller than that for question A.

At least 60% of the respondents perceived it *critical* to the overall success of their surveys that TSI's perform 82% (32) of the 39 listed behaviors well, in accordance with instructions. Perhaps due to the nature of the survey work in private and university survey research organizations, with which the majority of the respondents were affiliated, only 20% of these respondents viewed the maintenance of interview data confidentiality as a critical behavior. All of the respondents from the two governmental organizations represented in the sample thought this behavior critical. Very few respondents perceived any of the behaviors as unimportant to overall survey success.

Question C addressed the necessity for performing a work behavior when first employed, prior to any job training. Although some work behaviors were critical to overall job success, it was not necessarily important for a TSI to be able to perform these behaviors upon entry into the job. Such behaviors were referred for coverage in training. The behaviors in Table 4 were identified by at least 90% of the respondents as relevant (question A), by at least 60% of the respondents as critical to overall job success (question B), and by at least 70% of the respondents as essential or desirable for TSI's when first employed (question C). The behaviors are stated in abbreviated form in this table and not as they appear on the survey questionnaire.

Table 2
Job requirements survey—telephone survey interviewers—respondents

Position title	*Number*	*Mean years' experience*
Assistant department head	1	10.0
Assistant director, responding organization	1	20.0
Assistant manager, marketing telecentral	1	12.0
Assistant supervisor	2	1.4
Associate director/dept. head of operations	2	6.3
Chief of data collection	1	3.5
Director of interviewer quality control	1	4.5
Director of interviewer training	1	7.0
Division director	1	4.0
Field assistant	1	1.3
Field director	1	15.0[a]
Field manager	2	7.3
Field supervisor/coordinator	8	5.1
Field/telephone survey supervisor	5	5.8
Field training manager	1	6.0
Group manager	1	9.0
Head of survey service facility	1	30.0[a]
Interviewer trainer	1	3.5
Manager/chief of telephone survey center	5	7.6
Monitor	6	2.6
Operations manager	1	6.0
Production coordinator	1	6.5
Project director, CATI	1	6.0
Senior survey specialist	1	5.0
Supervisor of executive interviewing	1	4.0
Supervisory/lead statistical interviewer	2	2.5
Supervisory statistical assistant	1	3.0
Survey statistician	4	6.5
No response re title or years experience	3	—
Total	58	5.6[b]

[a]Not all involving telephone interviewing.
[b]Excluding persons who could not determine number of years in telephone interviewing.

Table 3
Job requirements survey—telephone survey interviewer (TSI)—summary for questions a, b, and c

Observable work behaviors	A: Is this a behevior in which your TSI's are involved et all? N	Yes (%)	No (%)	B: How important is it to the overall success of your surveys that TSI's perform this behavior well, in accordence with instructions? N	Critical (%)	Somewhat important (%)	Not at all (%)	C: How necessary is it that a TSI be able to perform this behavior when first employed (prior to any job training)? N	Essential (%)	Desirable (%)	Not necessary (%)
1.0. Interviewing skills											
1.1. probes neutrally or nondirectively to clarify or expand "don't know," ambiguous or incomplete responses	58	100	—	58	91	9	—	58	9	48	43
1.2. controls the subject matter and pace of the interview by tactfully limiting extraneous talk and assuring that the interview is not rushed	58	100	—	58	74	26	—	58	10	47	43
1.3. establishes and maintains rapport through a pleasant, courteous manner and appropriate, neutral reinforcement, to place respondent at case and to encourage cooperation, truthfulness, and confidence	58	100	—	58	90	10	—	56	17	64	19
1.4. listens attentively to responses, demonstrated by allowing adequate time for respondent to answer before next interviewer behavior and by showing sensitivity to need for probing, repetition, or other clarification of questions, reassurance of confidentiality, encouragement to cooperate, etc.	58	100	—	58	81	9	—	58	14	66	21
1.5. handles respondent statements or questions aptly and smoothly	58	100	—	58	60	40	—	58	3	57	40
1.6. initiates interview contacts and subject matter, household, or demographic questions with confident, positive (non-apologetic) approach	58	100	—	58	71	29	—	58	5	50	45
1.7. communicates with respondent with neutrality and objectivity, avoiding expressions of approval, sympathy, dismay, etc.	58	100	—	58	81	19	—	58	10	53	36
1.8. conducts interviews in businesslike, professional manner, avoiding oversociableness	58	100	—	58	78	22	—	58	14	53	33
1.9. maintains confidentiality of individual interview data	57	98	2	56	20	77	4	56	20	77	4
2.0. Voice characteristics											
2.1. speaks clearly and distinctly (articulates)	58	98	2	57	72	28	—	57	51	46	4
2.2. demonstrates voice quality through moderated pitch, modulating tone, and appropriate inflection	58	98	2	57	46	54	—	57	23	70	7
2.3. speaks at rate which is not too fast to be easily understood by the respondent or so slow that it drags	57	100	—	57	60	40	—	57	25	70	5
2.4. speaks with proper volume, not too loud or too soft for a respondent with normal hearing ability	58	100	—	58	48	52	—	58	19	72	9
3.0. Reading and other verbal skills											
3.1. reads and follows written directions, including exact order of questions and any skip instructions given on questionnaire	57	98	2	56	91	9	—	56	27	61	13
3.2. reads questionnaire orally verbatim with proper location and duration of pauses and timing of phrases, and maintaining an even pace of about two words per second	56	95	5	53	68	32	—	53	13	70	17
3.3. reads in a conversational manner, showing proper inflection in questions and emphasis on key words	55	96	4	53	62	38	—	53	13	70	17
3.4. uses English (or other required) language correctly	54	100	—	54	63	37	—	54	39	54	7
4.0. Clerical skills											
4.1. records responses with any qualifying remarks accurately and completely onto survey questionnaire or other required forms	56	98	2	55	85	15	—	55	22	49	29
4.2. records responses legibly and neatly onto required forms	57	100	—	57	77	23	—	57	32	53	2
4.3. transcribes information accurately from one form to another	56	86	14	48	75	25	—	48	27	50	19

Table 3 continued

Observable work behaviors	A: Is this a behavior in which your TSI's are involved at all? N	Yes (%)	No (%)	B: How important is it to the overall success of your surveys that TSI's perform this behevior well, in accordance with instructions? N	Critical (%)	Somewhat important (%)	Not at all (%)	C: How necessary is it that e TSI be able to perform this behavior when first employed (prior to any job treining)? N	Essential (%)	Desirable (%)	Not necessary (%)
4.4. keeps accurate and complete records to indicate call results	57	100	—	57	67	33	—	57	23	37	40
4.5. uses references (e.g., manuals) effectively to seek out information regarding survey procedures, content, etc.	47	87	13	41	49	51	—	41	12	39	49
4.6. classifies and codes accurately responses, types of living quarters, etc.	47	94	6	44	75	25	—	44	14	43	43
4.7. uses verbal and numeric rating scales appropriately to classify responses	50	90	10	45	76	24	—	45	11	33	56
4.8. edits or reviews completed forms for accuracy and legibility, using edit criteria provided	52	88	12	46	57	43	—	46	15	26	59
5.0. Survey content and mechanics (application of knowledge)											
5.1. introduces oneself and the survey, including sponsorship, authority, confidentiality, and purpose of survey	57	100	—	57	81	19	—	57	16	33	51
5.2. converts refusals or abates reluctance by explaining the importance of the survey or specific survey questions and/or emphasizing the confidentiality of responses	58	100	—	58	76	24	—	58	10	29	60
5.3. explains the need for household, demographic, or socioeconomic information in response to concerns expressed by respondents	58	100	—	58	78	22	—	58	9	41	50
5.4. explains questions to enchance understanding and relieve doubts by making minor neutral modifications in wording which do not alter frame of reference or question objective or otherwise bias responses	54	63	37	34	65	32	3	34	12	47	41
5.5. explains sampling technique or method of respondent selection when requested	58	100	—	58	41	55	3	58	3	22	74
5.6. screens telephone respondents to identify eligible respondent(s) based on survey guidelines	58	100	—	58	81	19	—	58	9	26	66
5.7. helps respondent to understand his/her role by clarifying respondent's task during interview	57	82	18	47	57	43	—	47	4	40	55
6.0. Equipment operation											
6.1. uses applicable telephone systems (WATS, FTS, etc.) with ease, applying knowledge of various telephone signals	57	98	2	56	63	34	4	56	9	36	55
6.2. operates cathode-ray tube (CRT) terminal properly	54	70	30	37	83	17	—	36	0	30	70
7.0. General work habits											
7.1. demonstrates dependability and reliability through attendance and punctuality	58	100	—	58	81	19	—	58	43	50	7
7.2. carries out work assignments with efficiency (productivity)	58	100	—	58	74	26	—	58	24	47	29
7.3. follows conscientiously instructions given in briefings, classroom training, self-study materials, etc.	58	100	—	58	81	9	—	58	33	43	24
7.4. works effectively under close supervision	52	98	2	51	63	37	—	51	18	61	22
8.0. Other general job requirements											
8.1. availability for required (productive) hours	54	94	6	51	67	31	2	46	37	53	—
9.0. Other observable behaviors not covered above (please specify. Avoid personal traits not operationally defined in terms of specific observable behaviors.)											
9.1. _____											
9.2. _____											
9.3. _____											
9.4. _____											

Table 4
Critical work behaviors necessary upon job entry and important for ranking purposes

	Work behaviors	Essential upon entry	Importance in differentiating superior overall work performance (%)[a] Very	Somewhat	Not at all	Ranking
1.3	establishes and maintains rapport		68	30	2	
1.4	listens attentively		78	20	2	R
1.9	maintains data confidentiality[b]	X	52	48	—	
2.1	speaks clearly	X	71	29	—	R
2.3	speaks at acceptable rate	X	56	44	—	
3.1	reads and follows questionnaire directions	X	84	14	2	R
3.2	reads questionnaire verbatim		64	27	9	
3.3	reads in a conversational manner		61	27	11	
3.4	uses English correctly	X	65	25	10	
4.1	records responses accurately	X	79	21	—	R
4.2	records responses legibly	X	75	25	—	R
7.1	demonstrates dependability	X	69	28	4	
7.2	carries out work assignments efficiently	X	73	24	2	R
7.3	follows training instructions conscientiously	X	80	20	—	R
8.1	is available for required working hours	X	73	24	2	R

[a]Percentages based only on those respondents identifying behavior as necessary upon job entry.
[b]Critical for Census Bureau work.

It is important to note that although behavior 1.9 (maintains data confidentiality) had not been identified by the respondents as critical to the overall success of their surveys, it was perceived by more than 70% of the respondents as being desirable upon job entry. As indicated earlier, governmental survey organizations tended to view the importance of this behavior differently from other organizations because of differences in the nature of the surveys. For this reason, and because the maintenance of data confidentiality is critical for census work by law, this behavior was included in the list of important entry-level behaviors. Those behaviors in Table 4 which were identified by at least 20% of the respondents as being *essential* upon entry into a telephone interviewing job are so marked.

Question D of the job requirements survey inquired about the effectiveness of each behavior in differentiating TSIs with superior overall work performance. The responses to questions D were important primarily for those work behaviors which met the job relevance, job importance, and entry-level need criteria, since only these behaviors would be given further consideration for selection purposes. Hence, only these behaviors could conceivably be used for ranking purposes, that is, to distinguish the minimally qualified from the better qualified candidates. Table 4 shows the responses to question D. Behaviors which at least 70% of the respondents identified as very important in differentiating superior performers were considered for use as ranking criteria in selection.

Other information obtained through the survey could be useful for planning training or monitoring activites, rather than selection. Work behaviors identified as critical for job success, but unnecessary when first employed (question C), were nonetheless important for the job. It was very important to cover and assess in training all critical work behaviors which would not be assessed during the selection process. Also for training purposes, it may be important to know how difficult it is for TSIs to grasp and perform a particular work behavior; hence, question E. Therefore, for all respondents saying that a behavior was unnecessary upon job entry, their responses to question E were tabulated. These responses are shown in Table 5 only for those work behaviors which are most important for training coverage since they are not covered in selection. Other behaviors covered in selection may also be important for more extensive and comprehensive treatment during training, but these are not shown here. Note that few of the work behaviors were perceived by many respondents as being very difficult to grasp and perform. Work behavior 5.2 on converting refusals, however, was viewed by 40% of the respondents as being very difficult, while 37% reported it to be moderately difficult.

Behaviors considered important for monitoring purposes were those which met the job relevance and job importance criteria. These behaviors were grouped according to the proportion of the respondents perceiving them as critical to survey success, and referred to other census researchers responsible for designing the monitoring system. The referral document appears as Exhibit 6. Behaviors viewed by the respondents as very important in differentiating TSIs with superior work performance are asterisked.

Discussion of results and their use in selection procedure development

Although the analyst in the Job Requirements Survey had hoped for a larger respondent sample representing more survey research organizations, the size and qualifications of the sample obtained appear to be fully satis-

Table 5
Critical work behaviors important for training coverage and the difficulty in their performance

	Work behaviors	Difficulty in performance (%)		
		Vary difficult	Modarately difficult	Easy
1.1	probes neutrally	16	44	40
1.2	controls the interview	—	68	32
1.5	handles respondent's concerns aptly	9	52	39
1.6	initiates interview contacts and survey areas with confidence	—	32	68
1.7	communicates neutrally and objectively with respondent	—	52	48
1.8	conducts interview in businesslike manner	—	42	58
4.4	keeps accurate and complete call records	—	48	52
4.6	classifies and codes responses accurately	5	53	42
4.7	uses verbal and numeric rating scales appropriately	8	40	52
5.1	introduces self and survey properly	—	17	83
5.2	converts refusals or abates reluctance	40	37	23
5.3	explains need for survey information in response to concerns	3	62	34
5.6	screens respondents for eligibility	—	37	63
6.1	uses applicable telephone systems with ease	—	19	81
7.4	works effectively under close supervision	9	18	73

factory for purposes of this analysis. Most respondents were apparently well-qualified to participate in the analysis, judging from the nature and length of the experience they had in telephone interviewing work. Of considerable value was the variety of perspectives brought to the analysis by respondents whose positions ranged from the executive to the monitor or line supervisor, and whose years of experience in survey research ranged from 10 months to 30 years.

Because the work behaviors included in the survey instrument were based on a fairly extensive literature seach, we can be reasonable confident that it included a rather complete list of behaviors. However, if other behaviors, such as "works effectively with minimal supervision" suggested by two respondents, had appeared on the list, these too may have been assessed by a large proportion of the respondents as relevant and important to the telephone interviewer's job.

As stated in the introduction to this paper, the chief purpose of the job analysis was to develop a content valid procedure or procedures to *select* telephone interviewers. Because the use of the analysis for training or monitoring purposes was not the chief focus of this survey, that use is not treated in this paper.

Table 4 contains the job analysis results which were crucial to the development of the selection process. The work behaviors listed there were examined to determine how they could best be assessed. The closer the content and the context of the selection procedure are to the work behaviors, the stronger the basis for showing content validity (USEEOC, 1978), therefore it was decided that for our purposes a work sample procedure would be preferable to other commonly used selection procedures such as a written test or a personal interview. We believed that a carefully constructed work sample procedure would provide for an objective, reliable, and valid means of assessing the ability of applicants to perform telephone interviewing work.

Research on the work sample approach has been primarily in trade and technical occupations. In these areas, the evidence has supported its use not only because of its validity for predicting job proficiency and training success, but also because it appears to be useful in reducing turnover and discrimination against minorities (Karren, 1980).

Not every behavior listed in Table 4, however, could reasonably by assessed through a work sample procedure or test. If behavior 1.9 (maintaining data confidentiality) can be assessed at all, it must be through on-the-job performance techniques. Behaviors 7.1, 7.2, and 7.3, which call for a demonstration of dependability, productivity, and the ability to follow training instructions, respectively, can be better assessed through a reference-check procedure. And, finally, behavior 8.1 (availability) can be better assessed by the applicant directly through an application self-report procedure. The remaining behaviors designated for assessment through a work-sample procedure are shown in Table 6.

Several of the survey organizations who responded to the initial request for job analysis information reported using various work sample procedures for selection purposes. However, none of these organizations indicated that their procedures were based in any formal job analysis. Nevertheless, information provided did reinforce our intent to pursue a work sample approach to telephone interviewer selection. The reduction of turnover rates and the maintenance and improvement of high performance on the telephone interviewing staff were cited as valuable benefits of a work sample procedure by at least one company. Major facets of the selection process reported by the three companies which gave detailed information on their work sample procedures are shown in Table 7. As shown in the table, all three organizations held a recruiting session during which important job information was shared and applications were filed. These sessions also provided the opportunity for

Table 6
Work behaviors to be assessed in work sample procedure

	Ranking Behaviors
1.4	Listen attentively to responses, demonstrated by allowing adequate time for respondent to answer before next interviewer behavior and by showing sensitivity to need for probing, repetition or other clarification of questions, reassurance of confidentiality, encouragement to cooperate, etc.
2.1	Speaks clearly and distinctly (articulates)
3.1	Reads and follows written directions, including exact order of questions and any skip instructions given on questionnaire
4.1	Records responses with any qualifying remarks accurately and completely onto survey questionnaire or other required forms
4.2	Records responses legibly and neatly onto required forms
	Other Behaviors
1.3	Establishes and maintains rapport through a pleasant, courteous manner and appropriate, neutral reinforcement, to place respondent at ease and to encourage cooperation, truthfulness, and confidence
2.3	Speaks at rate which is not too fast to be easily understood by the respondent or so slow that it drags
3.2	Reads questionnaire orally verbatim with proper location and duration of pauses and timing of phrases, and maintaining an even pace of about two words per second
3.3	Reads in a conversational manner, showing proper inflection in questions and emphasis on key words
3.4	Uses English (or other required) language correctly

self-screening. The recruiting session, self-study, and the try-out or test interview work together to provide a realistic job preview, which serves to screen out applicants who are not likely to perform well on the job and/or to stick with the job for any length of time; hence, the beneficial effects on work force quality and job tenure.

The self-study materials provided to job applicants by the two companies in Table 7 representing the academic community introduce a variation in the simple work sample procedure, a variation which is suggestive of the miniature training and evaluation (MT&E) approach to selection. The MT&E approach is a method by which applicants are trained to perform specific work behaviors which are representative of the job; they are then evaluated on their ability to perform these behaviors through a work sample procedure. The work sample test, then, in effect is measuring the applicant's ability to learn a set of behaviors, rather than his or her ability to perform behaviors acquired through earlier experiences. According to Thomas (1980), the MT&E is "based on the conception that applicants who can demonstrate the ability to learn and perform a sample of tasks which incorporate essential elements of the job should be able to learn and perform successfully on the job given adequate on-the-job training. The primary objective of a miniature training and evaluation test is to assess the extent to which applicants have the potential to reach a satisfactory standard of performance on the job or at the end of training."

Viewing the measurement of learning ability or "trainability" as a primary objective of the telephone interviewer selection process, it was decided that a MT&E variation of a work sample procedure would be appropriate for census purposes. The chief disadvantage of the MT&E procedure, the expense of administration in terms of personnel resources, was carefully weighed against the potential savings to be reaped through the use of such a procedure. If the procedure functions as expected, the realistic job previews which the MT&E procedure provides and the required demonstration of critical job behaviors would be useful in screening out applicants who are either unqualified or not highly motivated for telephone interviewing work. The elimination of such applicants during selection should save the employer costly replacement and retraining activities resulting from excessive turnover and/or poor performance. Therefore, it is hoped that the expense of selection using MT&E will be more than compensated by the gains in the productivity and stability of the Census Bureau's telephone interviewing staff.

There are three major phases to the MT&E procedure as developed for the Census Bureau: (1) training the job applicant on the important entry-level work behaviors in a miniature or mini-training sesssion; (2) administering the work sample test, which consists of a test interview with a rehearsed respondent; and (3) rating the applicant's test performance. Phase 1 of the MT&E

Table 7
Work sample selection procedures as reported by three responding organizations

Selection process components	*C–D*	*A–F*	*A–C*
Preliminary screening on voice quality through phone call in response to ad	X		
Recruiting session providing information on job, pay, hours, etc., and opportunity for application	X	X	X
Application blank as kind of simulation task			X
Self-study materials to prepare for trial interview		X	X
Trial (test) questionnaire	X	X	X
Telephone trial interview with training supervisor, recruiter or telephone manager	X	X	X
Prepared script for respondent in trial interview		X	
Rating of trial interview using standard criteria		X	

procedure was executed during a recruiting session. The recruiting session began with an introduction to the telephone interviewer job including information on working hours, pay, benefits, working conditions, and general responsibilities. Applicants were given opportunity to screen themselves out after this brief job introduction if they so desired. The remaining applicants took part in the first phase of the MT&E procedure, mini-training. The recruiter, using a verbatim script, trained the applicants in the following areas which included all the important work behaviors derived from the job analysis:

1. Asking questions exactly as worded (Behaviors 3.2, 3.4)
2. Asking every appropriate question in the correct order (Behavior 3.1)
3. Asking questions using appropriate style (Behaviors 2.1, 2.3, 3.3)
4. Clarifying questions when necessary (Behaviors 1.4, 3.1)
5. Preparing to address respondents' concerns (Behaviors 1.4, 3.1)
6. Establishing and maintaining rapport (Behavior 1.3)
7. Recording responses correctly and completely (Behaviors 4.1, 4.2)

During the mini-training session, the applicants were free to ask questions for clarification of anything covered during the training, which, of course, included all material pertinent to the subsequent test interview. Questions on other material were not accepted in order to enhance the standarization of the recruiting session, to preclude giving any applicant or group of applicants an unfair advantage, and because answers to such questions were not required for successful performance on the work sample test.

During the mini-training session, applicants were also provided with copies of the questionnaire to be used in the test interview, one which closely simulated the RDD questionnaire they would use on the job, and a self-study manual which summarized and reinforced the classroom training. These materials were to be taken home for further study and practice. Practice interviewing at home was encouraged.

An adjunctive segment of the recruiting session consisted of the administration of a written aptitude test of skills found to be relevant for telephone follow-up jobs during the 1980 Decennial Census. This test was administered for research purposes only; its results were not used in the selection of RDD telephone interviewers. Test scores will be correlated with job performance criteria in order to determine if there is any significant statistical relationship between performance on such a written aptitude test and telephone interviewing performance.

Each applicant was scheduled to return to the Census Bureau to administer a test interview, the work sample test, a few days after the mini-training session. Using a clean, serialized copy of the test questionnaire, each applicant telephoned a rehearsed respondent and conducted the interview, which took an average of 5 to 10 minutes. The rehearsed respondent responded to interview questions from a highly standardized script. Standard responses were provided for the questions that should have been asked if the applicant interviewers followed the correct skip patterns, as well as for questions which should have been skipped. The interviews were taped by the rehearsed respondents for subsequent rating by two independent raters.

Phase 3 of the MT&E procedure involved the evaluation of test performance in terms of errors made. Highly structured rating worksheets were used by trained raters to rate the completed questionnaires and the taped interviews. Careful training of the raters and use of standardized rating instruments worked together to enhance the reliability of the ratings. The inter-rater reliability has been estimated to be 0.98. Also, raters were instructed to listen to each tape twice to ensure the accuracy of her or his ratings. The rating process for each rater took an average of 30 minutes per applicant.

A trial of the MT&E procedure was conducted prior to its use in selection, using a small sample of Census Bureau employees at various grade levels. As a result of the trial, administrative and technical modifications were made to improve the procedure, including clarifying definitions and providing examples for complex questionnaire instructions. Other improvements were made in the use of visual aids, the degree of classroom participation, the rating instructions, and in the statement of the rating criteria themselves.

Although exhibits of the rating worksheets used to rate candidates cannot be provided here, it is nevertheless important at this juncture to demonstrate the linkage between the critical entry-level behaviors derived from the job analysis and the rating criteria, shown in Table 8. Several opportunities were provided in the test interview for assessing each of the work behaviors, and most work behaviors were assessed through several rating criteria. Only one behavior, 3.1 (reading and following written directions), was assessed on both the questionnaire and the taped interview. Raters were provided detailed instructions on how to assess each criterion, followed by an opportunity for practice and evaluation of their comprehension of the rating procedures.

An error rate was computed for each work behavior separately, for all ranking behaviors, and finally for all behaviors. Error rate is defined as the ratio of the observed error frequency to the number of possible errors on the test. The average error rate for the two raters for all behaviors was used to decide which applicants qualified for the job. The average error rate for the ranking behaviors was used to rank qualified applicants into three bands—highly qualified, well-qualified and minimally qualified. Only qualified candidates who re-

Table 8
Linkage between important work behaviors and test interview rating criteria

Work behaviors derived from job analysis	*Rating criteria used to assess the behavior (in terms of errors made in test interview)*	
	Interview	*Questionnaire*
1.3. establishes and maintains rapport through a pleasant, courteous manner and appropriate, neutral reinforcement, to place respondent at ease and to encourage cooperation, truthfulness, and confidence	Unpleasant or discourteous manner in responding to R[a] concern Rude or unpleasant in handling R confusion Unduly long delays between questions without explanation Interview not closed courteously	
1.4. listens attentively to responses, demonstrated by allowing adequate time for respondent to answer before next interviewer behavior and by showing sensitivity to need for repetition or other clarification of questions, reassurance of confidentiality, encouragement to cooperate, etc.	Not attentive to R remarks (asks question already clearly answered) Not responsive to R concern Adequate time not allowed for R answer Not sensitive to need for clarification	
2.1. speaks clearly and distinctly (articulates)	Word not spoken clearly or endings deleted Generally slurred speech	
2.3. speaks at rate which is not too fast to be easily understood by the respondent or so slow that it drags	Speech too fast or too slow	
3.1. reads and follows written directions, including exact order of questions and any skip instructions given on questionnaire	Question not skipped as instructed Failure to address R concerns as instructed in manual Failure to clarify question as instructed Terms not defined as instructed Applicable questions skipped, repeated or asked out of order Other questionnaire instructions not followed	R information not recorded as instructed Boxes not marked as instructed Numerical codes for R information not recorded No response marked Interviewer check items not marked
3.2. reads questionnaire orally verbatim with proper location and duration of pauses and timing of phrases	Question not completed or reworded after interruption Deletions, additions, rewordings and/or improper pauses in reading questions	
3.3. reads in a conversational manner, showing proper inflection in questions and emphasis on key words	Read question in monotone without inflection Key words (bold type or underscored) not expressed loudly and clearly	
3.4. uses English (or other required) language correctly	Improper English usage in answer to R concern Number of subject and predicate not agreed in question	
4.1. records responses with any qualifying remarks accurately and completely onto survey questionnaire or other required forms		Numerals not recorded accurately Appropriate boxes not marked for given response Names, places, and other information not recorded accurately Information recorded for wrong household member
4.2. records responses legibly and neatly onto required forms		Numerals not easily read Words not easily read

[a]R = Respondent.

ceived favorable reference checks could be offered jobs as telephone interviewers.

Conclusion

The aim of this study was to develop and implement a selection procedure to employ telephone interviewers which was objective, reliable, valid, fair, and economically feasible. A job analysis of the telephone interviewer's job was conducted which involved a literature search and a job requirements survey of survey research organizations. Thirteen survey organizations responded to the survey, providing information on the nature of the work performed by their telephone interviewers, the importance of various work behaviors, the need for these behaviors upon the entry into the job, their effectiveness in differentiating superior performers, and the difficulty in their performance. Information was also provided on work sample procedures in use. On the basis of this information, a content valid miniature training and evaluation procedure, which incorporates a work sample test representative of critical job elements, was developed.

There are three major components to the MT&E procedure: training the applicants for important work behaviors, administration of the work sample test, and evaluation of test performance. We hope that by providing a representative job sample, the MT&E procedure will screen out job applicants who are not motivated for telephone interviewing work and who may have difficulty learning critical work behaviors. We also expect that it will screen out many of those who are less likely to perform successfully on the job. The reliability of the MT&E procedure has been shown to be very high. Although the procedure is obviously more costly than other more commonly used selection procedures, we anticipated that its reliability and validity in making inferences about trainability and future job performance and its utility in reducing turnover will more than compensate for the expense of administration. In any case, through current research, several means are being investigated to improve the cost effectiveness of the procedure, chiefly through modifications in the rating process.

In spite of the strong evidence of content validity based on the job analysis of a sample of survey research organizations throughout the nation, further research is planned to crossvalidate the MT&E procedure on a sample of Census Bureau telephone interviewers. There is a need for assurance that the selection procedure is fully representative of telephone interviewing work as performed at the Census Bureau. An analysis of the telephone interviewer's job will be conducted after the interviewers have had the opportunity to gain some experience in their job duties. Intervierwers, monitors, supervisors, and facility managers will be included in the job analysis sample.

Still further validation of the MT&E procedure is being considered to estimate the predictive validity of the procedure, if technically feasible (that is, given adequate sample size, reliable performance criteria, and other necessary conditions). Statistical evidence of the degree to which the MT&E procedure predicts training and job performance is desirable.

There is yet much research to be accomplished for the purpose of evaluating and improving the miniature training and evaluation procedure in current use at the Bureau of the Census, but the prospects of its utility in employing telephone interviewers appear promising. With a sound job analysis at its base, there is reason for optimistic expectations for selecting a reasonably competent telephone interviewing staff through an MT&E procedure. Job climate factors may well have an adverse effect on performance in time and together with the relatively low pay and the attraction of permanent full-time work, they may negatively influence job tenure. Nevertheless, we hope to realize a clear and positive gain from having used an MT&E procedure firmly based in the job.

Footnote

[1] The names of responding organizations are not provided. Organizations are coded such that the first letter represents the type of firm and the second letter the individual company. The types of firms are as follows: C—private commercial firms; A—firms affiliated with the academic community; and G—governmental organizations.

Appendix of references

Andrews, L.
1974 "Interviewers: recruiting, selecting, training and supervising." Pp. 124-132 in R. Feber (ed.), Handbook of Marketing Research. New York: McGraw-Hill.

Barioux, M.
1952 "A method for selection, training and evaluation of interviewers." Public Opinion Quarterly (spring):128-130.

Bingham, W., B.V. Moore, and J.W. Gustad
1959 How to Interview. New York: Harper and Bros.

Blankenship, A.B.
1977 Professional Telephone Surveys. New York: McGraw Hill.

Bliesch, W.
Undated "Interviewer guidance, supervision and control." In Seminar on Fieldwork Sampling and Questionnaire Design, Part 1. Amsterdam: European Society for Opinion and Marketing Research.

Boyd, J.L., Jr. and B. Shimberg
1971 Handbook for Performance Testing: A Practical Guide for Test Makers. Princeton: Educational Testing Service.

Campion, J.E.
1972 "Work sampling for personnel selection." Journal of Applied Psychology 56:40-44.

Cannell, C.F., S.A. Lawson, and D.L. Hausser
1975 A Technique for Evaluating Interviewer Performance. Ann Arbor: University of Michigan, Survey Research Center.

Chilton Research Services
Undated "Standard telephone interviewer recruiting and training program." Internal report.

Collins, W.A.
1970 "Interviewers' verbal idiosyncrasies as a source of bias." Public Opinion Q. 37:416-422.

Guest, L.
1947 "A study of interviewer competence." International J. of Opinion and Attitude Research 1:17-30.

Hanson, R.H., E. Marks, and National Opinion Research Center
1958 "Influence of the interviewer on the accuracy of survey results." J. of the American Statistical Association 53:635-655.

Harris, M. (ed.)
1956 The Social Survey: Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing. London: The Social Survey Division, Central Office of Information.

Horton, R.L., and D.J. Duncan
1978 "A new look at telephone interviewing methodology." Pacific Sociological Review 21:259-273.

Hyman, H.H.
1954 Interviewing in Social Research. Chicago: University of Chicago Press.

Karren, R.
1980a "Alternative Selection Devices: The Work Sample and the Interview." Paper presented at National Council on Measurement in Education Convention. Boston, MA.
1980b The Work Sampling Approach to Personnel Selection (PRR-80-11). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center.

Marks, E.S., and W.P. Mauldin
1950 "Response errors in Census research." J. of the American Statistical Association (Sept.):424-438.

O'Leary, L.R.
1973 "Fair employment, sound psychometric practice, and reality: a dilemma and partial solution." American Psychologist (Feb.):147-150.

Payne, S.L.
1974 "Data collection methods: telephone surveys." Chapter 4 in R. Feder (ed.), Handbook of Marketing Research. New York: McGraw-Hill.

Plumlee, L.B.
1980 A Short Guide to the Development of Work Sample and Performance Tests (2d ed.,pp.80-83). Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center.

Reilly, R.R., and W.R. Manese
1979 "The validation of a minicourse for telephone company switching technicians." Personnel Psychology 32:83-90.

Research Information Center, Inc.
Undated Interviewer Training Manual and Monitor Fact Sheet.

Richardson, S.A., B.S. Dohrenwend, and D. Klein
1965 Interviewing—Its Forms and Functions. New York: Basic Books, Inc.

Schmidt, F.A., A.L. Greenthal, J.E. Hunter, J.G. Berner, and F.W. Seaton
1977 "Job sampling v. paper-and-pencil trads and technical tests: adverse impact and examinee attitudes." Personnel Psychology 30:187-197.

Sheatsley, P.B.
1950 "An analysis of interviewer characteristics and their relationship to performance, part I." International J. of Opinion and Attitude Research 4:473-498.
1951a "An analysis of interviewer characteristics and their relationship to performance, part II." International J. of Opinion and Attitude Research 5:79-94.
1951b "An analysis of interviewer characteristics and their relationship to performance, part III." International J. of Opinion and Attitude Research 5:191-220.

Siegel, A.I., and B.A. Bergman
1975 "A job learning approach to performance prediction." Personnel Psychology 28:325-339.

Smith. J.M.
1972 Interviewing in Market and Social Research. London: Routledge and Kegan Paul, Ltd.

Temple University, Insitute for Social Research
Undated Observed/Taped Interview Evaluation Form.

Thomas, P.H.
1980 Trainability Testing: The Miniature Training and Evaluation Approach to Selection (PRR-80-9). Washington, DC: Office of Personnel Management, Personnel Research and Development Center.

University of Chicago, National Opinion Research Center
1978a Introduction to Interviewing.
1978b Try-Out Questionnaire.

1979 "Manual for Recruiting Interviewers," in Field Management Handbook, Sec.VI.
1980 Reference Check Form (rev. 10/80).

University of Illinois, Survey Research Laboratory
1980 Recruiting Record (7/80). An evaluation for try-out questionnaire.
1981 Try-Out Questionnaire and Question-by-Question Specifications for the Try-Out Questionnaire (10/81).
Undated Application Evaluation. A reference check.

University of Michigan, Survey Research Center
1980 Monitoring Form Code Sheet (5/19/80 rev.).

U.S. Bureau of the Census
1980a National Crime Survey: Interviewer's Manual (NCS-550)
1980b "Michigan RDD experiment: interviewer debriefing." Internal report.
1980c Quarterly Housing Survey: Interviewer's Work Manual (725).
1982a Current Population Survey: Interviewer's Manual (CPS-250).
1982b Current Population Survey: Self-Study Manual (CPS-271).
1982c National Crime Survey: Self-Study Manual(NCS-521).

U.S. Civil Service Commission
1973 Job Analysis—Developing and Documenting Data: A Guide for State and Local Governments (BIPP 152-35). Washington, DC: U.S. Civil Service Commission, Bureau of Intergovernmental Programs.
1975 Job Analysis for Improved Job-Related Selection: A Guide for State and Local Governments(BIPP 152-63). Washington, DC: U.S. Civil Service Commission, Bureau of Intergovernmental Programs.

U.S. Department of Labor, Manpower Administration
1972 Handbook for Analyzing Jobs. Washington, DC: U.S. Government Printing Office.

U.S. Equal Employment Opportunity Commission, U.S. Office of Personnel Management, U.S. Department of Labor, U.S. Department of Justice, and U.S. Department of Treasury
1978 "Uniform guidelines on employee selection procedures (1978)." Federal Register 43 (166):38290-38315.

Womer, S., and H. Boyd
1951 "The use of a voice recorder in the selection and training of field workers." Public Opinion Q. (Summer):358-363.

Exhibit 1
Request for secondary source information

[Letterhead]

[Address]

Dear [Name]

My staff and I are in the process of developing employee selection procedures which are valid for use in hiring personal or telephone interviewers for survey research at the U.S. Bureau of the Census. I understand that you employ interviewers for survey and/or marketing research purposes. I am seeking the results of any job analyses underlying your personnel selection program, and information on the nature of the selection procedures themselves (e.g., tests, interviews, rated applications, biodata instruments, miniature training and experience procedures). Although it may not be possible for you to share with us the actual selection procedures for security reasons, I would still appreciate any information about the procedures which you are at liberty to provide. The documentation of any job analysis you have conducted in order to develop your selection procedure would be most helpful.

A response is requested by October 30, 1981. If you have any questions, you may call me on 301-763-5955. Thank you very much for any assistance you can provide.

Sincerely,

BARBARA H. LACEY
Chief, Personnel Research
Branch

Exhibit 3
Job requirements survey—Telephone survey interviewers
General information on Organization and Survey Respondent

1. Organization
   a. Name of your organization_____
   b. Location (City, State)_____

2. Telephone Interviewing Staff
   a. Size of telephone interviewing staff normally maintained _____interviewers
   b. Average years of experience of your telephone interviewers _____years
   c. Average number of hours per day your telephone interviewers work _____hours

3. Respondent
   a. Your name _____
   b. Your position _____
   c. Years experience in work directly related to telephone interviewing _____years

Exhibit 2
Contacts for secondary source information

A.C. Nielsen Company
Northbrook, IL

American Institute of Research
Washington, DC

Behavioral Sciences Laboratory
University of Cincinnati
Cincinnati, OH

Bureau of Applied Social Science
Columbia University
New York, NY

Burke Marketing Research, Inc.
Darien, CT

Chilton Research Services
Radnor, PA

Creative Marketing Enterprises, Inc.
Toledo, OH

Crossley Surveys, Inc.
New York, NY

Data Group, Inc.
Elkins, Pk, PA

Gallup Organization, Inc.
Princeton, NJ

Institute for Social Science Research
University of California at Los Angeles
Los Angeles, CA

Institute for Survey Research
Temple University
Philadelphia, PA

Market Facts, Inc.
Arlington, VA

Marketing Research Corp. of America
Stamford, CT

Market Opinion Research
Detroit, MI

National Family Opinion
Northwood, OH

National Opinion Research Center
University of Chicago
Chicago, IL

Opinion Research Corporation
Washington, DC

Research Information Center
Phoenix, AZ

Sindlinger & Company, Inc.
Media, PA

Statistical Research, Inc.
Westfield, NJ

Survey Research
General Foods
White Plains, NY

Survey Research Center
University of Michigan
Ann Arbor, MI

Survey Research Laboratory
University of Illinois
Urbana, IL

Trendex, Inc.
Westport, CT

Valley Forge Information Center
King of Prussia, PA

Walker Research, Inc.
Indianapolis, IN

Exhibit 4
Contacts for job requirements survey

Chilton Research Services
Radnor, PA 19089

Computer-Assisted Telephone Interviewing
U.S. Bureau of the Census
Suitland, MD 20233

Institute for Social Science Research
University of California at Los Angeles
Los Angeles, CA 90024

Institute for Survey Research
Temple University
Philadelphia, PA 19122

National Center for Health Statistics
Hyattsville, MD 20782

National Opinion Research Center
Chicago, IL 60637

Opinion Research Corporation
Princeton, NJ 08540

Rand Corporation
Santa Monica, CA 90406

Research Information Center, Inc.
Phoenix, AZ 85020

Research Triangle Institute
Research Triangle Park, NC 27709

Survey Research Center
University of California at Berkeley
Berkeley, CA 94720

Survey Research Center
University of Michigan
Ann Arbor, MI 48109

Survey Research Laboratory
University of Wisconsin
Madison, WI 53706

Survey Research Laboratory
University of Illinois
Urbana, IL 61801

Walker Research, Inc.
Indianapolis, IN 46250

WESTAT
Rockville, MD 20850

Exhibit 5
Job requirements survey—
Telephone survey interviewer
introduction and instructions

[Letterhead]

Dear Respondent:

The U.S. Bureau of the Census is planning an experiment which will investigate the effectiveness and feasibility of telephone interviewing in conducting its surveys. Experimental findings on response rate and in other areas will depend to a large extent on the quality of the telephone interviewing staff. For that reason, the development of valid interviewer selection procedures has been identified as a major project of the experiment. Underlying these procedures must be a thorough job analysis, i.e., a systematic and comprehensive investigation and documentation of the elements and requirements of the telephone interviewer's job.

Your assistance is requested in the present job analytic survey if you have been working with telephone interviewers in a supervisory or managerial capacity for at least one year. Also eligible to respond are persons who are highly knowledgeable of telephone interviewer work through directly related personnel management or research experiences.

Work behaviors involved in the telephone survey interviewer's job have been identified and are presented in the attached survey questionnaire. These behaviors were derived from the following: verbal and written reports on selection procedures, training manuals and evaluation forms provided by public and private survey research organizations such as yours; books and journal articles on survey interviewing; and work manuals and self-studies for personal interviews involved in the Census Bureau's surveys. Observable work behaviors are the focus of the job analysis. Even personal characteristics, so far as possible, have been couched in operational terms.

If you are an eligible respondent, your task is to evaluate the listed work behaviors as indicated in the questionnaire, and to amend the listing as you think necessary. Your evaluation will assist us in determining the relative importance of the behaviors, the need for each upon entry into the job, and the effectiveness of each in distinguishing superior performers. If, after c*arefully* examining the listed behaviors, you discover that other important behaviors performed by members of your telephone interviewing staff have been excluded, please amend the list as indicated. If at all possible, try to fit all behaviors into areas presently listed. Where additions are necessary, please be certain that each behavior is observable and operationally defined.

Please return the questionnaire in the enclosed self-addressed envelope by *December 4, 1981.* If you have any questions regarding this survey, you may call me on 301-763-5955. Although your response is voluntary, your cooperation is urged and will be appreciated very much.

Sincerely,

BARBARA H. LACEY
Chief, Personnel Research Branch
Personnel Division

Exhibit 6
Telephone survey interviewer job analysis—
Important job-related work behaviors
for possible assessment in performance monitoring

The work behaviors below were identified for inclusion in this list of behaviors recommended for assessment in performance monitoring procedures as follows:

1. at least 90% of the respondents perceived the behaviors as relevant for telephone survey interviewer work at their respective facilities;
2. 95% or more of those responding to the survey item perceived it at least "somewhat important" to the overall success of the surveys that these behaviors be performed well in accordance with instructions, while 60% or more of those responding perceived it "critical" that they be performed thus.

Of the work behaviors identified using the criteria given above, some were perceived as "very important" in differentiating telephone survey interviewers with superior overall work performance. Those perceived by at least 75% of those responding to the survey items as "very important" in differentiating are marked with an asterisk(*).

Group I (90% or more of respondents see it as "critical" that these behaviors be performed well)

*1. Probes neutrally or nondirectively to clarify or expand "don't know," ambiguous, or incomplete responses.
2. Establishes and maintains rapport through a pleasant, courteous manner and appropriate, neutral reinforcement, to place respondent at ease and to encourage cooperation, truthfulness, and confidence.
*3. Reads an follows written direction, including exact order of questions and any skip instructions given on questionnaire.

Group II (80-90% of respondents see it as "critical" that these behaviors be performed well)

*1. Listens attentively to responses, demonstrated by allowing adequate time for respondent to answer before next interviewer behavior and by showing sensitivity to need for probing, repetition, or other clarification of questions, reassurance of confidentiality, encouragement to cooperate, etc.
2. Communicates with respondent with neutrality and objectivity, avoiding expressions of approval, sympathy, dismay, etc.
3. Records responses with any qualifying remarks accurately and completely onto survey questionnaire or other required forms.
4. Introduces oneself and the survey, including sponsorship, authority, confidentiality, and purpose of survey.
5. Screens telephone respondents to identify eligible respondent(s) based on survey guidelines.
6. Demonstrates dependability and reliability through attendance and punctuality.
*7. Follows conscientiously instructions given in briefings, classroom training, self-study materials, etc.

Group III (70-79% of respondents see it as "critical" that these behaviors be performed well)

1. Controls the subject matter and pace of the interview by tactfully limiting extraneous talk and assuring that the interview is not rushed.
2. Initiates interview contacts and subject matter, household or demographic questions with confident, positive (non-apologetic) approach.
3. Conducts interviews in businesslike, professional manner, avoiding over-sociableness.
4. Speaks clearly and distinctly (articulates).
5. Records responses legibly and neatly onto required forms.
6. Classifies and codes accurately responses, types of living quarters, etc.
7. Uses verbal and numeric rating scales appropriately to classify responses.
*8. Converts refusals or abates reluctance by explaining the importance of the survey or specific survey questions and/or emphasizing the confidentiality of responses.
9. Explains the need for household, demographic, or socioeconomic information in response to concerns expressed by respondents.
10. Carries out work assignments with efficiency (productivity).

Group IV (60-69% of respondents see it as "critical" that these behaviors to be performed well)

1. Handles respondent statements or questions aptly and smoothly.
2. Speaks at rate which is not too fast to be easily understood by the respondent or so slow that it drags.
3. Reads questionnaire orally verbatim with proper location and duration of pauses and timing of phrases, and maintaining an even pace of about two words per second.
4. Reads in a conversational manner, showing proper inflection in questions and emphasis on key words.
5. Uses English (or other required) language correctly.
6. Keeps accurate and complete records to indicate call results.
7. Explains questions to enhace understanding and relieve doubts by making minor neutral modifications in wording which do not alter frame of reference or question objective, or otherwise bias responses.
8. Uses applicable telephone systems (WATS, FTS, etc.) with ease, applying knowledge of various telephone signals.
9. Works effectively under close supervision.

The effect of training and supervision on common measures of field interviewer performance

Floyd J. Fowler, Jr., Center for Survey Research, University of Massachusetts/Boston

Thomas W. Mangione, Center for Survey Research, University of Massachusetts/Boston

Introduction

There has been considerable research demonstrating that the way the interviewer does his or her job affects the data that are collected in a survey interview. Early studies by Katz (1942) showed the importance of standardized wording. Probing and recording effects on data were identified by Hyman (1954). Cannell and Fowler (1964) reported data which suggested the importance of the interviewer as a motivator. Reinforcement of respondents, intended or unintended (Marquis et al., 1972), the pace of the interview (Marquis and Cannell, 1971) and the way the interviewer structures the interview and explains the study to respondents (Cannell, Oksenberg, and Converse, 1977a) are other aspects of interviewer behavior that have been clearly linked to the quality of survey data. However, there is remarkably little wisdom, much less data, about what to do to affect the way interviewers perform on the job.

In fact, the researcher has three kinds of decisions to make about interviewers that might be expected to affect performance. First, there are decisions about selection. Studies such as those by Schuman and Converse (1971) and Erlich and Reisman (1961) and Robinson (1946) show that interviewer demographic characteristics can affect answers in certain situations. However, for most studies, existing literature provides little guidance about how to select interviewers.

Second, there are decisions about how much and what kind of training to give interviewers. Survey organizations have developed strong collective views about this over the years, but there is great diversity in those views. Many surveys are carried out by interviewers who are never directly trained by the research organization. The training received by professional market interviewers and others who do polling is obviously extremely uneven and, at the moment, unknown. In addition, it is not uncommon for community surveys to use interviewers who receive only a few hours of training. Even among professional academic survey organizations, the training sessions can last from a scant two days through a full five days. Moreover, organizations differ in the extent to which they have additional training or retraining after initial formal training sessions are over. The point is that interviewers doing surveys receive very different degrees of initial training. As early as 1942, Friedman found that a little training helped a lot in reducing gross interviewer effects. However, we have not learned much since then about the utility or disutility of various amounts of interviewer training.

A third decision involves the strategies for supervising interviewers once they are trained. It is not uncommon for an interviewer to receive no feedback whatsoever while he or she is in the field on a particular project. Sometimes one or two early interviews will be reviewed, and an interviewer will hear about any egregious errors that are identified.

A critical issue with respect to supervision is the kind of information that is available to supervisors for monitoring interviewer performance. All organizations have access to information about production, efficiency (hours per interview), and response rates, although the timeliness with which that information is available can differ greatly. In addition, of course, completed interviews can be reviewed. From such reviews, a supervisor can ascertain whether an interviewer is appropriately following skip instructions, obtaining answers of some sort to all the appropriate questions, writing legibly, and meeting question objectives in other ways. In addition, if survey instruments contain open-ended questions, perusal of the answers gives some indication of whether an interviewer is approximating verbatim recording, though, of course, there is no real information of how accurately answers are being recorded.

A critical point to note is that reviewing completed interviews provides no information about how the interviewer is performing the essential role of asking questions and probing inadequate answers; and it provides only scant information about the quality of the interviewer's recording of answers. In fact, unless interviews are tape recorded or observed, interviewers can receive no supervision about how they carry out their data gathering.

It also is important to note that survey organizations usually cannot tell whether interviewers affect the data they collect. When samples are assigned to interviewers on the basis of convenience, any differences between answers that an interviewer obtains and study averages can be attributed either to sample difference or interviewer differences. It is not possible to sort them out. For almost all studies, the quality of data collected by an interviewer is both unknown and unknowable.

This paper is a first report stemming from a large-scale field experiment designed to identify the links

The research reported in this paper was supported by grant number HS04189 from the National Center for Health Services Research.

between the kind of training and supervision that interviewers receive and the way they perform their jobs. The ultimate test of how much difference training and supervision make is how they affect the data interviewers collect, that is, the amount of error and the amount of bias interviewers introduce into the answers they obtain. The full project analysis will address exactly those issues.

This preliminary analysis uses only a limited part of the data, but it is the part most likley to be available to researchers: costs, response rates, and evaluation of completed interviews. The focus of our analysis is the relationship between these common criteria for interviews and the kind of training and supervision interviewers receive. In fact, one would not expect to find strong relationships, because the measures are not likely to be indicative of how inteviewers do the main part of the job for which they are trained. However, if there is no association between general training and our usual measures of performance, we have gained an important perspective on problems of quality control and evaluation in field interview studies.

Methods

Sixty interviewers without previous professional survey research experience were recruited and randomly assigned to one of four training programs. Each training program used the same manual. The messages and techniques communicated in each training program were also identical. What differed was the time trainees spent in formal training and, as a result, the amount and kind of experience they had in working with the various ideas and procedures they were supposed to use.

Level 1 training was the shortest training program we responsibly could devise. Interviewers read a training manual before an approximately five-hour traing session, which included a two-hour briefing on the study purposes and specific question objectives, a one-hour lecture on procedures to be used by standardized, nondirective interviewers, one half-hour on sampling procedures, one hour on pay forms and administrative matters, and one half-hour demonstration of a practice interview.

Level 2 training lasted about two days. It provided more opportunity for discussion and some opportunity for interviewers to practice role playing their interviewing techniques. Level 3 training lasted approximately five days. Level 4 training lasted approximately ten days. Each level added more practice, more supervised role playing, more discussion, and, in the case of the longest training session, some additional readings about the background and reasons why interviewers are trained to perform as they are.

Once interviewers had completed training sessions, they were randomly assigned in a balanced design to one of three levels of supervision. The supervisory programs were structured as described below.

Each interviewer had a weekly telephone conversation with a field supervisor. During that contact, the supervisor provided the interviewer with feedback on his or her work for the preceding week. The content of the feedback was carefully structured.

What we called Level 1 supervision provided interviewers only with feedback about the number of hours they were putting in, their efficiency, and their response rate. Interviewers were rated on each of these facets of their performance each week and were told what their rating was. If their performance was rated as "needing improvement," the supervisor had specific suggestions that she gave to interviewers to improve their performance. The supervisor initiated no feedback on any other topic, though she answered questions brought up by interviewers on any topic.

Level 2 supervision added to the above feedback the results of a review of a sample of completed interviews. Those aspects of interviewer performance which can be evaluated by looking at a completed interview schedule were systematically evaluated: legibility, following skip instructions, meeting question objectives, recording probes, and apparent verbatim recording. Those evaluations were communicated to the interviewer each week in addition to the ratings of productivity and response rates.

Interviewers assigned to Level 3 supervision tape recorded all of their interviews. In addition to the feedback given to Level 2 supervision interviwers, there was systematic review of a sample of taped interviews. Systematic evaluation was made of the way in which interviewers asked questions exactly as worded, used appropriate nondirective probes, handled respondents in getting them to choose response alternatives to closed questions, recorded answers and handled the interpersonal aspects and the pace of the interview.

Interviewers each received an assignment of 40 addresses that constituted a random subsample of the total sample. Thus, each interviewer's assignment was statistically equivalent to every other interviewer's assignment.

The questionnaire for the study took about 45 minutes to administer. It included a carefully structured sample of questions typically used in health services research. Questions covered use of health services, health status and health conditions, everyday practices and lifestyle likely to be relevant to health, health beliefs, mental health, and standard demographic questions. The questionnaire included a high percentage of items taken directly from commonly used health survey instruments, including the National Health Interview Survey. Selection of items was balanced to include adequate samples of sensitive and nonsensitive items, difficult and easy items, attitudinal and factual items, open and closed questions.

The sample was an area probability sample drawn from six communities in suburban Boston. The pro-

cedure was to interview an objectively selected adult in each chosen household. Kish selection table procedures were used to designate an adult (Kish, 1965).

In order to better compare interviewers, two constraints were placed on their efforts to enlist the cooperation of respondents. First, interviewers were restricted to a total of six calls (plus a seventh to keep a definite appointment), with at least three calls having to occur after five o'clock on a weekday or a weekend. In this way, comparisons of costs and rates of finding people at home would reflect interviewer efficiency and not willingness to exert unlimited effort for callbacks.

Second, there was, of course, no transfer of assigned addresses in the event that an interviewer encountered a reluctant respondent. Interviewers were instructed to attempt to leave a respondent who did not want to be interviewed immediately in a frame of mind that would permit a second attempt to convince the respondent to cooperate. However, if an interviewer obtained an "informed refusal," (that is, when he/she was convinced that the respondent was fully informed about the study and had made a conscious decision not to cooperate), the interviewer was credited with a final refusal. Thus, response rates totally reflect the interviewers' effectiveness in presenting the study and how well they used the six calls they had at their disposal.

The specific analysis carried out here looks at three measures of interviewer performance. First, interviewer costs are usually a salient concern to any survey organization. Therefore, one measure of performance was the average number of hours per interview.

Second, response rates are, of course, an important part of the quality of data that an interviewer obtains. For each interviewer, we calculated the fraction of occupied housing units at which he/she succeeded in obtaining a completed interview.

Third, survey organizations have the option of reviewing complete interviews as one measure of interviewer performance. At least five interviews from each interviewer were rated in a standardized way as outlined above. Those ratings of how well the interviewer appeared to complete interview schedules constitute another readily available set of performance measures that are evaluated below.

Results

Looking first at cost per interview, there is not a clear basis for predicting any relationship between the training program to which interviewers were assigned and their efficiency in completing interviews. In all training sessions interviewers were encouraged to make long trips and to plan their trips efficiently. In the longest training session, there also was one exercise in which interviewers planned a hypothetical trip, which was discussed with other interviewers and a supervisor. That exercise provided an opportunity for somewhat more discussion about how to be efficient in trip planning than anything else.

With respect to supervision, all levels of supervision received exactly the same feedback with respect to productivity and cost per interview. However, for Level 1 supervision, response rates, productivity, and efficiency constituted the sole focus of the feedback, while interviewers in the other supervisory programs received feedback on other topics. Thus, it is possible that the importance of efficiency would seem greater to those interviewers who did not receive feedback about the quality of their interviews and their interviewing.

Table 1 shows the calculation of hours per interview. It can be seen clearly that there is no association between the training program to which an interviewer was assigned and interviewer efficiency. There is, however, a statistically significant difference associated with the level of supervision. Those interviewers assigned to Level 3 supervision, who tape recorded their interviews and received feedback on the quality of their interviewing, averaged an hour per interview more time than the interviewers assigned to the other two strategies of supervision.

Table 1
Hours per interview by level of training and supervision

	Hours	*Number of interviews*
Supervision		
1 (Production only)	3.8	511
2 (Plus questionnaire review)	3.7	442
3 (Plus tape review)	4.7*	423
Training (in days)		
1	3.9	355
2	4.2	296
5	4.0	344
10	4.0	381

*Significantly different from 1 and 2, .05 level of confidence.

Response rates obviously are a critical part of an interviewer's performance. There is some reason to think that increased training might be helpful in improving response rates. Even though the majority of training focused on the actual skills of interviewing, longer training sessions provided more opportunity for discussion of the purposes of the survey and of the problems that might be encountered in enlisting respondent cooperation. In addition, one feature of the longest traning program was that interviewers went out in the company of a supervisor and actually knocked on some doors. There also was an exercise in which there was extensive role playing about how to enlist respondent cooperation. Thus, one might expect those who received more training should be somewhat advantaged with respect to response rates.

With respect to supervision, the only prediction would be the same as that with respect to efficiency: those whose supervision was restricted to feedback about response rates and costs might pay more attention to response rates than would interviewers who received feedback on many aspects of the way they did their jobs.

Table 2 presents the data. No association can be seen between the length of training interviewers received and their average response rates. Although the response rate for the group receiving the most training was as good as any, and their rate of obtaining refusals was the best of the four groups (though not to a statistically significant degree), the least trained group did virtually as well. There is no apparent reason for thinking that the small curvilinear trend is meaningful.

Table 2
Response rates by level of training and supervision

	Refusal rate	Overall response rate	Number of occupied HUs
Supervision			
1 (Production only)	21	70	727
2 (Plus questionnaire review)	23	67	683
3 (Plus tape review)	22	65*	722
Training (in days)			
1	21	69	548
2	26	63	490
5	24	67	529
10	19	69	565

*Significantly different from 1, 0.5 level of confidence.

In contrast, there is a clear significant pattern in Table 2 associating response rates with the supervision program to which an interviewer was assigned. Those interviewers who tape recorded interviews had a significantly lower response rate than those who received feedback only on production and response rates, with the middle level of supervision also falling in the middle with respect to response rates.

However, it is important to note that the groups did not differ with respect to the rate at which they obtained refusals. Rather, the main difference in response rate resulted from the rate at which noninterviews occurred for reasons other than refusal. Note, the coverage was the same for all three levels of supervision; that is, each interviewer was required to make no more than six calls (with the possibility of a seventh call only if it was to keep a definite appointment), with at least three of those calls occurring after five on a weekday or a weekend.

A third set of criteria came from evaluation of completed interviews. At least five interviews taken by each interviewer were reviewed on four dimensions: legibility and quality of recording; verbatim recording; following skip instructions properly; and meeting question objectives. Each rated interview was evaluated on these four dimensions using a four-point scale, where 1 was unsatisfactory, 2 was needs improvement, 3 was satisfactory, and 4 was very good. The same two people reviewed interviews from all interviewers, regardless of level of supervision. Table 3 presents the results by the training and supervision program to which an interviewer was assigned.

It is clear from simple visual perusal that there is no association between the amount of training an interviewer received and apparent performance on the four dimensions of interviewing that could be assessed from looking at a completed interview schedule. Similarly, there was not a significant association between the kind of supervision an interviewer received and ratings from review of questionnaires.

The consistent pattern of no effect of training may lead readers to wonder whether there were any real differences in the training experiences. Since the main focus of training is teaching interviewers how to carry out the interview process, and since none of the performance measurers considered thus far had anything to do with how interviewers carry out the question and answer process, it is not surprising that there were no associations. A full examination of the significance of

Table 3
Ratings* of review of completed interviews by level of training and supervision

	Average Supervision Ratings of:				
	Recording procedures	Verbatim recording	Following skip instructions	Meeting q. objectives	Number reviewed
Supervision					
1 (Production only)	3.9	3.9	3.7	3.0	95
2 (Plus questionnaire review)	3.6	3.9	3.9	3.0	90
3 (Plus tape review)	3.4	3.8	3.7	2.9	100
Training (in days)					
1	3.6	3.8	3.7	2.9	75
2	3.7	3.9	3.9	2.9	65
5	3.6	4.0	3.9	3.2	70
10	3.6	3.8	3.7	2.9	75

*Rating scale from 1 to 4, with 4 being best. None of the differences in the table meets usual standards for statistical significance.

different levels of training must await analyses which we have not yet begun. However, from our monitoring of tapes of interviews for one-third of the interviewers, we were able to obtain some reading on whether the training affected the way that interviewers handled tape recorded interviews.

When interviews were taped recorded, some of those interviews were reviewed for six potential problem areas: reading questions exactly as worded, appropriate probing of open questions, appropriate probing of closed questions, recording answers verbatim, appropriate interpersonal behavior, and the pace and tone of the interview. Since only five interviewers from each level of training tape recorded their interviews, and since only five interviews per interviewer were reviewed, these data can only be taken as suggestive. However, from Table 4 it appeared that the amount of training interviewers received was associated with the evaluations of tape recorded interviews. In particular, there seems to be a difference between those who received the briefest training and those who received at least two days of training on five of the six ratings. Although these figures cannot be taken as conclusive, the data in Table 4 provide some evidence that interviewers who received more training behaved differently in some discernible ways.

Table 4
Ratings of taped interviews by training level

	Average Supervisor Ratings of:						
Training (in days)	Reading questions	Probing open questions	Probing closed questions	Verbatim recording	Interpersonal behavior	Pace/ tone	Number reviewed
1	2.8	2.7	3.1	3.0	3.9	3.9	23
2	3.7	3.0	3.4	3.6	3.9	4.0	20
5	3.2	2.9	3.6	3.3	4.0	4.8	21
10	4.0	3.3	3.6	3.5	3.8	3.9	23

Note: None of the differences in the table meets usual standards for statistical significance.

Discussion

The principal reason for undertaking this project was to develop data on which researchers and funders of research could set guidelines for the appropriate level of investment in the training and supervision of interviewers. Heretofore, there has been no basis for setting such standards, and, not suprisingly, procedures and practices with respect to training and supervision of interviewers vary widely.

These preliminary results provide ample evidence as to why such procedures do vary widely. Normally, those supervising interviewers have access to only a very limited body of information for evaluating interviewers. Cost, response rates, and whether interview schedules are completed and filled out adequately are the only aspects of an interviewer's performance that can be readily reviewed. Response rates obviously have some bearing on the quality of data that an interviewer collects, and costs and the correct filling out of questionnaires are obviously among appropriate considerations when evaluating interviewers; however none of these measures has anything to do with the main task for which we train interviewers, that is, carrying out a standardized, nondirective interview.

The fact that none of these measures is related to the amount of training interviewers receive is predictable. This is not to say that additional training focused specifically on response rates or costs would not produce results. Perhaps it could. However, the typical training program focuses primarily on interviewing procedures.

The findings also provide a clear explanation for why training programs vary so greatly. If the amount of training does not clearly relate to a change in observable interview performance, cost-conscious researchers and funders of data collection can easily justify brief training programs. Our data simply support what many researchers have known for years: even with minimal training, interviewers can "get the job done" insofar as "the job" is commonly assessed.

The findings on supervision, particularly the effects of taping, are also perplexing. It is clear that tape recording interviews entails extra cost for equipment and listening to the tapes. It was not expected that taping would adversely affect interviewing time or response rates.

With respect to interviewing time, we thought that taped interviews perhaps took longer to administer. However, there was no difference in average interview length by supervision type. Lower response rates slightly increase the cost per completed interview, but the difference in this respect does not explain the hour per interview difference. We know that some interviewers were spending extra time listening to their tapes to edit their interviews. However, we do not think that practice was prevalent enough to explain the difference observed.

Our main hypothesis is that interviewers did not like tape recording their interviews. Inevitably, they knew that all study interviewers were not taping, and there were many complaints about having to tape. Interviewers who taped were least diligent about keeping their telephone appointments for feedback with supervisors. It was clear that tape recording was anxiety producing for at least some significant number of interviewers. Our guess is that this anxiety accounts for a major part of the increased costs; interviewers who do not like the interview process are likely to go home sooner from the field, producing shorter, less cost-effective trips. It also seems likely to play a role in the slightly higher rate of nonresponse; interviewers who are anxious about doing interviews may be less successful at finding people at home.

The preliminary data presented in Table 4 gives some indication that training does affect interviewer performance, though we will have to await further analysis to assess the value of training for overall error reduction. It seems clear on the surface that tape recording provides a valuable way to effect quality control in the field. However, the results of our experiment highlight some heretofore unanticipated costs which we may or may not be able to reduce.

The fascinating point on which to conclude is the uncertainty that currently exists about standards for interviewers. Unless interviews are tape recorded or observed, the quality of data collection is unmeasured. In that case, interviewers are evaluated on other grounds, which, as we have seen, do not seem to be related to the amount of general interviewer training received. In that context, it makes sense to provide a minimum of training.

One of the great potential benefits of telephone interviewing is that it allows for timely, appropriate evaluation of interviewer performance. At this time, it is not clear how to achieve that same kind of control over the quality of performance in the field, and how much that control is worth. It seems certain that these are important issues, however; it is also important to be aware that we do not yet know how to deal with them. We expect to have much better answers as our own analyses proceed.

Improving the training of survey interviewers*

Stanley Presser, Survey Research Center, University of Michigan

This paper describes some alterations in interviewer training that we are in the process of making at the Survey Research Center. This work has been carried out mainly by Pamela Guenzel, Tracy Berckmans, and Lois Oksenberg; Charles Cannell has provided general direction. We began making changes quite recently and do not as yet have a final product. Thus this is an interim report on work in progress. I will focus on the nature of our dissatisfaction with past training and then outline the kinds of changes now being implemented. The best way to begin is by providing an overview of our most recent interviewer training program.

Training of new SRC interviewers has been carried out by regional field supervisors who recruit and hire trainees. It takes place during a five-day period immediately before the start of an actual field period. On the morning of the first day the supervisor gave a lecture on the character of survey research, the nature of the interviewer's role with special attention to interviewing ethics, and the kinds of techniques used in interviewing. She then conducted a demonstration interview with one of the trainees to illustrate good interviewing practice. This was followed by round robins—the trainees breaking into pairs and taking turns role playing interviewer and respondent. During the role playing, the supervisor listened to each pair of trainees and provided feedback as appropriate. This was then supplemented by general discussion. Both the role playing and the supervisor's demonstration interview made use of the actual questionnaire that was to be employed on the study beginning at the end of training.

As homework for the first evening each trainee was assigned a practice interview—again with the actual study questionnaire—to take with a friend or family member. The next morning as the first order of business, the supervisor evaluated these interviews. This was followed by continued training on interviewing techniques using lectures, demonstrations, and role playing. In addition, administrative matters such as filling out the Interviewer Time and Expense Journal were introduced. On the second evening another practice interview was assigned, though this time the assignment involved knocking on doors to take the interview with a stranger. The homework practice interviews, lectures, demonstrations, and role playing were also used on succeeding days to cover sampling, respondent selection, and other topics.

A number of features of this approach to training troubled us. The first was its lack of standardization. The approach was unstandardized in at least three ways: it varied from supervisor to supervisor, it varied from study to study, and it varied from trainee to trainee even within a given training session. It varied between supervisors because not all trained in the fashion I have outlined. Some supervisors departed from the agenda by, for example, introducing topics in a different order. Instead of waiting until the afternoon of the second day to train on filling out Time and Expense Journals, some supervisors introduced that early on the first day. In addition, supervisors differed in the amount and kind of coverage they gave to various topics.

The training varied from study to study because the questionnaire used in the practice interviews and the role playing changed as the study changed. Interviewers who were trained before our annual Panel Study of Income Dynamics, for example, were trained on a questionnaire that has no open attitude items. By contrast interviewers who were trained before our biennial election study were trained on a questionnaire that has a wealth of open attitude items. Clearly, the opportunity to learn probing skills on such questions varied dramatically between the two sessions.

Finally, the sessions varied from trainee to trainee because the experience of the practice interviews and the role playing depended on who happened to be interviewed. We exerted no control over the kinds of answers and problems the trainee confronted in these exercises. Sometimes these proved to be very difficult, other times quite simple.

In addition to the unstandardized nature of past training, some of the materials and assignments we had been using were not well suited to training. The immediate use of the study questionnaire illustrates this point. Many questionnaires are not appropriate for early stages of training. In part I've already referred to this with the example of the Panel Study of Income Dynamics questionnaire which has no open attitude questions. Obviously, it is not a good vehicle for developing probing of such questions. But there are other problems as well. Following the skip patterns in the Panel Study of Income Dynamics questionnaire requires an intimate familiarity with a number of complicated study concepts; it is necessary to understand the definition of family unit, dwelling unit, and sample membership in order to proceed through the questionnaire. These study-specific concepts prove confusing to many trainees and interfere with learning the fairly simple principles of skip patterns.

* The author is indebted to Tracy R. Berckmans for sharing her store of knowledge about interviewer training.

A similar problem existed in other ways in the kinds of training we had been doing. Early in training one wants to emphasize elementary skills such as reading the question as it is written, verbatim recording, and nondirective probing. Yet by sending the trainees out to knock on doors before they had fully acquired these skills we were having them practice under conditions more difficult than necessary. Moreover, doing this made it hard to evaluate their performance. Supervisors could look at the questionnaire itself to see whether answers appeared to have been recorded verbatim but, as pointed out previously, there are many other essential elements of interviewing that are impossible to evaluate simply by looking at a filled-in questionnaire.

Exhibit 1
Parts III and IV: Clarification and probing for answers

Exercise #1—Listening and Rating

Directions: Listen to each example as you follow along in the Questionnaire. If no probe or clarification is used, mark NO PROBE in the left margin. If a probe is used, identify it by its abbreviations and then indicate whether it was used correctly or incorrectly (e.g., RQ−, RQ+). Identify clarifications in the same way (C+/−). Appropriate skips are already recorded on the Questionnaire. Refer to your Job Aids and to the Q-by-Qs across from the Questions if necessary.

Assume the R has one child.

Exercise #2—Listening, Clarifying, and Probing

Directions: Listen to each example. Decide which clarification or probe to use and write it out verbatim under the question on the Questionnaire. Include neutral prefaces where appropriate.

Assume the R has one child.

So both the immediate use of the study questionnaire and the early introduction to actual survey conditions seemed unwise to us. Imagine trying to train doctors or pilots in this fashion! It seemed more sensible to us to simplify the task and teach it in parts under somewhat more artificial conditions. This then, along with the goal of standardization, was our primary aim in redesigning the interviewer training program.

We have now produced a standard set of training materials to be used by all supervisors. The materials have two distinguishing characteristics. The first is that they are built around a training questionnaire and associated respondent scripts. We devised a questionnaire solely for training purposes to replace the actual study questionnaire during the first few days of training. Various versions of this questionnaire are to be used in the role playing exercises. Each version of the questionnaire has an associated respondent script to be used by the trainee who role-plays the respondent. Both the questionnaires and scripts vary along a difficulty or complexity dimension—the easier ones to be used earlier on, the more difficult ones as the session progresses. Furthermore the first two homework practice interviews are now to be taken with the supervisor over the telephone using versions of the training questionnaire; the supervisor uses a pre-written script in playing the role of the respondent. (This is similar to the procedure Barbara Lacey described that is in use at the Census Bureau to evaluate job applicants.) This will relieve the trainee of the burdens attendant on finding and interviewing a stranger, provide control over the interview experience, and perhaps most importantly let the supervisor directly evaluate the trainee's performance.

Tapes of the scripted interviews have also been prepared; some to demonstrate good performance, others to illustrate errors. These tapes are used as the basis of training exercises. The face sheet for two such exercises—on learning to probe and to ask for clarification—may be found in Exhibit 1. In the first exercise, the trainee listens to a taped interview and is asked to identify and then evaluate the probes used by the interviewer. In the second exercise, a similar tape is used and the trainee's task is to decide which clarification or probe ought to be used in each circumstance.

Finally, the trainees themselves are taped doing one of the role playing interviews. They then not only listen to their interview but also rate themselves in the same way that they rate the exercise tapes. This particular feature, having the trainee listen to and grade his or her own interviewing, strikes us as especially promising.

The first central element in this new approach to training then is the set of questionnaire scripts and tapes. The second key feature of the materials is that they are self-instructional. In the past, trainees were asked to read various chapters in our Interviewer's Manual, but our guess was they retained only a limited amount from doing so. On reflection, we felt it was probably unreasonable to expect them to retain a great deal. If trainees are to retain complex information, it seems sensible to have them make use of it as it is presented. That is what we have attempted to do in our new materials, an example of which is provided in Exhibit 2. Once again this is from the section on probing. About three-quarters of the way down the page, a new probe is introduced ("Which would be closer to the way your feel?"). The sentence below it provides information about the probe similar to the material contained in the Interviewer's Manual. Then the very next sentence has the trainee do something—it says "Write this description in your Job Aid now." More information is then presented about when or why the probe would be used. Finally, the trainee is given examples of the use of the probe and asked to rate them.

This programmed learning approach is used for almost all the topics covered in general training. And we

Exhibit 2
Sample on probing from new interviewer training materials

In the situations below, first decide which probes have been used (WT or TM or WM) and put their abbreviations in the blanks. Then rate the use of each probe by adding + or − to each probe abbreviation.

Exact Q#1: "Now looking ahead—do you think that a *year from now* you will be *better off* financially, or *worse off,* or just about the same as now?"

R: "I don't know. It's so hard to predict the future."

22. ▶_____Iwer #1: "Yes, but what do you expect will happen?"
23. ▶_____Iwer #2: "Could you tell me what you mean by that?"

Exact Q#2: "Looking ahead, which would you say is more likely—that in the country as a whole we'll have continuous good times during the next few years or so, or that we'll have periods of widespread unemployment or depression, or what?"

R: "I certainly hope times will be good."

24. ▶_____Iwer #1: "What do you mean by that?"
25. ▶_____Iwer #2: "What do you think?"

7. Which would be closer to the way you feel? (WC)

(WC) is used when the R has narrowed his choices to a particular range. Write this description in your Job Aid *now*. IF the R has not eliminated any choices, you should pause and/or repeat the response options (RQ), rather than use (WC).

A correctly used (WC) is rated WC+.

In the situations below rate each probe as WC+ or WC−.

Exact Q#1: "As to the economic policy of the government—I mean steps taken to fight inflation or *un*employment—would you say the government is doing a good job, only fair, or a poor job?"

R: "Somewhere between good and fair."

26. ▶_____Iwer: "Which would be closer to the way you feel?"

have also tried it in the context of the at-home study-specific training that all interviewers (new and experienced) undergo before working on a new survey. The survey was the Panel Study of Income Dynamics. The instruction book for the study has always been exceedingly complicated. An example of its complexities is given in Exhibit 3, which is a copy of pages 43 and 44 from the 230-page instruction book used in 1981. These pages provide guidelines for whom to include in the family unit, an important part of doing the panel interview. This is the fifteenth year we have done the survey, and the definition of family unit has always proved troublesome. Many interviewers have difficulty with the concept and make mistakes in applying the definition. Thus this seemed a good test of the programmed learning approach, and the instruction book was rewritten using it. Exhibit 4 presents the 1982 instruction book's treatment of the family unit problem. In addition to trying to simplify the description, we added exercises that required the interviewer to use the information presented.

I introduce this example from the Panel Study partly to demonstrate how we are using the new training techniques, but also because it provides the only quantifiable bit of evidence that we have so far on the impact of the

Exhibit 3
Sample from Panel Study of Income Dynamics instruction book

G. General Guidelines for Whom to Include in the Family Unit (FU)

1. Reinterview (Main Family) Situations

Sometimes it is not clear whether someone who lives in a household is actually an FU member. The brief points listed below, in addition to the general guidelines of relation by blood, adoption, or marriage, should help you cope with some of the more unusual family arrangements you may encounter:

a) *Permanence of the living arrangement over time* is an important consideration when making decisions about whether or not a person should be listed as an FU member. Roomers and boarders are *generally* not included, even though they may rent a room from our Head for years. However, *sometimes* we do include them, especially in cases where we suspect "roomer" or "boarder" may be a euphemism. (This happens occasionally with older male/female pairs; see b(2) below.)

b) *Same-sex roommates are almost never moved into the FU.* The only exceptions to this rule are:

1) homosexual couples, where we treat the sample person as Head, ask an Other FU Member section for the partner/friend, and determine total amounts for both for living expenses. *A homosexual relationship will never be treated as a "Head and Wife" situation, no matter how long it lasts.* For purposes of the question sequences, we can think of the partner as an Other Relative or Child of the FU. (Remember to make a note on the cover sheet when you move in same-sex roommates as to whether this is a homosexual relationship.)

2.) when "just friends" move in together and plan on being together for a long time. Older people occasionally do this. (If, however, the respondent joins a large group, such as a religious organization, it is not practical to include everyone in the interview!) We have come across a few cases where a friend of one of the children in the FU moves in and appears to be supported mainly by the Head and Wife. The Nonrelative helps with housework, eats meals with the family, and generally is taken into the bosom thereof. We should include this Nonrelative in the FU.

Whenever you move a same-sex roommate into the FU, explain your reason(s) for doing so, either on the cover sheet, on an Immediate Action form attached to the cover sheet, or in the thumbnail sketch (keeping in mind that the respondent has a right to review his/her own questionnaire, including the thumbnail).

c) *Opposite-sex roommates should be moved into the FU* if it is apparent—from information or observation—that they are "sharing bed and board" with one of our respondents. We call opposite-sex movers-in "friends."

For help in handling specific question sequences for unusual FU situations, see Appendix 2, pp. 69-71.

Exhibit 4
Sample on the family unit from 1982 instruction book

B. The Family Unit

The Family Unit (FU) is the major unit of analysis for this study. Since families change when family members move in and out over the years, you will often need to decide which individuals living within a *housing unit* (HU) (a physical boundary) are actually members of the *family unit* in which you are interested. (See definition of HU on page 39 of your Interviewer's Manual.)

NOTE: *This year* you will be listing the names of all individuals living within the physical household boundaries, then deciding which of those belong in the FU. *HU members are not necessarily FU members.*

There are two types of family unit situations you will encounter: (1) reinterview (main family) situations, and (2) splitoff situations (when a sample member moves out of the main FU).

1. *Main Family Situations*

WHO IS *NOT* ELIGIBLE TO BE AN FU MEMBER?

(a) Same-sex roommates
Exceptions:
(1) Homosexual couples
(2) Close friends who move in together and plan on being together for a long time
(3) An acquaintance of one of the children in the FU who moves in and is supported by the family Head, helps with housework, eats meals with the family, and is generally treated like a member of the family.

(b) Opposite-sex roommates who are *only* roommates.

(c) Boarders
Exception: If you suspect the "roomer" to be living with a member of the family as if they were married.

(d) Stray relatives who are only *temporarily* living with an FU.

(e) A child, originally in the FU and a sample member, who split off (moved out to form his own FU) *and* still has own coversheet and has now moved back into the Household Unit. This individual is still treated as a member of his/her own FU *separate* from the main FU.

WHO *IS* ELIGIBLE TO BE AN FU MEMBER?

Everyone else living there at the time of the interview.

EXERCISE: Indicate whether each individual is an FU member by circling YES or NO.

YES NO 6. Mary is a widow who lives with her three children. She rents a room in the basement to Susan, a nonrelative, to help make ends meet. Is Susan an FU member?

YES NO 7. Jim, a single-member FU, has lost his job. He temporarily moves in with his co-worker, Fred, and Fred's family. Are Fred's family members also members of Jim's FU?

YES NO 8. The Smiths are raising two teenage children of their own and are also the sole support of one of the good friends of their son, who lives with them and functions as a family member. Could this "extra" child be an FU member?

YES NO 9. Harriet's sister, not a sample member, is living with Harriet temporarily until her new apartment becomes available. Is sister Jane an FU member?

YES NO 10. The eldest Jones boy has been interviewed for his own FU since he married and moved to his own apartment several years ago. After losing his job this year, he and his wife moved back in with his parents, who are still interviewed on a separate cover sheet. Are they now members of his father's FU?

procedures. The complexity of the Panel Study and the study staff's demand for near-perfect data have always meant that a large number of interviews have to be returned to the interviewer for correction of problems. Since in many of these cases the interviewer must recontact the respondent, this is an expensive undertaking. In 1981, using the old instruction book, the send-back rate was about 8%, or roughly 550 interviews. This year, using the new instruction book, the send-back rate is running below 5% (the study is 90% finished).

With respect to the training of *new* interviewers, we have no systematic evidence as yet on the worth of the changes I have described. Our hope is that they will increase the probability of turning out well-qualified interviewers. But Floyd Fowler's results demand that I end on a note of caution. To an extent, our training ideas are premised on the belief that is it important to give new interviewers a set of general interviewing skills, not just those required by the particular survey for which they are hired. Whether this is a cost-effective strategy remains an open question.

Open discussion: Session 5

An analysis of job requirements for telephone survey interviewers

Banks began the discussion of Lacey's presentation by raising a question regarding the Job Requirements Survey. Noting that there is considerable disagreement among survey organizations as to critical or essential criteria for interviewer selection, she asked how the particular persons from each organization were chosen to respond to the questionnaire. Lacey explained that the agencies or research organizations themselves chose which persons within the organization should respond.

Banks noted that the ability of the interviewer to explain how the respondent's telephone number was obtained was rated less critical than the modulation of the interviewer's voice. She suggested that the results may have been a function of who was chosen to fill out the job requirements questionnaire and that perhaps there should have been some criteria for a designation of respondents rather than allowing the survey organizations themselves to determine who would respond. This would have allowed a wider range of respondents in terms of job skills, position, and experience. Lacey responded that criteria for selecting respondents were given to survey organizations. These were (1) persons who had been working with telephone interviewers in a supervisory or managerial capacity for at least one year, and (2) persons highly knowledgeable about telephone interviewer work through directly related personnel management or research experiences. The application of these criteria resulted in the respondent types shown in Table 2 of the paper.

Some surprise was expressed at the fact that only about one-half of the respondents to the Job Requirements Survey rated "maintains data confidentiality" as "very important" in differentiating superior overall work performance. A respondent's concern about confidentiality is one important reason for nonparticipation in surveys, particularly in telephone surveys, in which the respondent can't see the interviewer and confidentiality statements and affidavits cannot be handed to the respondent. Lacey pointed out that while only 20% of respondents identified confidentiality as "critical" to the overall success of their surveys, it was perceived by over 70% as being desirable upon job entry. She also suggested that governmental survey organizations tended to view the importance of that behavior differently from other organizations because of differences in the nature of the surveys.

Rouse asked whether the evaluation of performance on the test interview was based on observation or was evaluated only from the listener's point of view. Lacey described the procedure for this evaluation. The applicant came to the Bureau of the Census, sat in a private room with the questionnaire, and telephoned another office where the rehearsed respondent was situated with a tape recorder. This person responded to interview questions from a highly standardized script and recorded the interviews. The raters were not present during the interview; rather they listened to the tapes and reviewed the completed questionnaires. The tape recordings were considered essential for reliable rating by standarized procedure.

Noting that the error rate was the ratio of errors made to possible number of errors, Axelrod asked for clarification on the number and kinds of errors which could be made or were rated in the test interview. Lacey explained that the rating criteria used and the types of errors rated were standardized and were provided in Table 8 of the paper. Each type of work behavior had specific types of errors associated with it. For example, a relevant work behavior was "establishing and maintaining rapport." The specific errors associated with that work behavior were: (1) unpleasant or discourteous manner in responding to respondent concern; (2) rude or unpleasant behavior in handling respondent confusion; (3) unduly long delays between questions without explanations; and (4) not closing the interview courteously. There was a similar fixed number of errors for each of the other work behaviors. The raters recorded each time a specific error was made.

Kovar commented that the Bureau of the Census is engaged in a considerable amount of research and developmental work about the telephone interview and appears to have made a major commitment toward the development of a telephone interview system. Lacey responded that the purpose of the work she described in her paper and of the overall random-digit-dialing research effort was to obtain data for a comprehensive evaluation of the potential of the use of the telephone approach by the Bureau of the Census. The personal interview is costly, and there is a need to reduce the cost of federal surveys. However, any switch from the personal to the telephone approach may involve sacrifices in the quality of data and in response rates. The RDD research is a first step in an evaluation of the nature and extent of these sacrifices, if any, before deciding on the implementation of a telephone system.

The effect of training and supervision procedures on field interview job performance

Cannell began the discussion of the paper by Fowler by emphasizing that while the optimal amount of training for interviewers is still an open question, it is essential

that a minimal amount of training be given. He related an anecdote about one of his graduate students who conducted a mini-experiment with interviewers with no training of any kind. The experiment was quickly terminated after an interviewer insisted that a respondent's family income could not possibly be as high as reported.

Axelrod asked whether Fowler "really accepted his findings," which implied that if you give interviewers some training, it doesn't matter how much. Fowler pointed out that the importance of these findings is to demonstrate that we are not assessing what we are training interviewers for. The main focus of training is to teach interviewers how to carry out the interview process. While the measures used (cost, response rates, and whether interview schedules are completed and filled out adequately) are aspects of an interviewer's performance that can be readily reviewed, they do not represent measures of the quality of the interview process. He acknowledged that he was a little surprised that longer training did not result in higher response rates. This may be because interviewers learn very quickly how to obtain cooperation from respondents. Since samples were randomly assigned to interviewers and the first half of the assignment set was an independent sample from the second half, it will be possible in this study to look at experience effects.

In answering a question on response rates, Fowler noted that the 69% response rate was somewhat lower than that in previous studies at the Center for Survey Research. Normally, with a designated respondent, a response rate of around 75% would be expected. He attributed the lower rate in this study to elements of the study design: (a) interviewers were restricted to a total of six calls, (b) cover sheets could not be transferred among interviewers, and (c) none of the interviewers had any previous interviewing experience. Each of these was required by the methodological design of this study.

Andrews noted that Fowler had stated several times that the lack of significant differences in his study was because he was not assessing what the interviewers were trained for, and he asked specifically what we are training interviewers for. Fowler responded that the main task for which we train interviewers is to carry out a standardized nondirective interview. Andrews asked how one would measure that. Fowler said that a perfectly standardized interview should not have any effect on the answers respondents give. In this study, there are three measures of how much interviewers influence answers. The main measure is the extent to which variance can be associated with the interviewer, following the approach used by Kish (1962) and elaborated by Groves and Kahn (1979). In addition, some questions were repeated in a reinterview of respondents, thereby allowing test-retest reliability to be used as an indicator of interviewer quality. Finally, bias can be assessed by comparing means obtained by different interviewers.

Frey asked how Fowler would measure training effectiveness. Fowler stated that one has to tape-record. Unless there is systematic taping, there is no supervision and monitoring of the quality of interviewer performance. The reason all interviews in an interviewer's assignment, rather than a sample, were taped in this study was to minimize interviewer discretion as to which respondents were taped. Taping a sample, however, is feasible as a routine monitoring procedure.

Cannell noted that you cannot always use interviewer variance as a dependent variable and asked what else could be used. Fowler responded that that was one of the things he would be able to find out from this study. The way in which interviewers performed on the taped interviews and the measures of quality, such as the interviewer variance, can be related to the outcome variables that can be routinely assessed, such as evaluation of the quality of completed interviews.

Cannell drew attention to the importance of the sensitivity of the outcome variable. Fowler agreed and pointed out that when you have interviewers nested with their samples, it is hard to disassociate the distinctive role of the interviewers from the real characteristics of their samples. This study can produce a reasonable case that what you hear on a tape and some of the criteria you can observe are indicators of the interviewer quality.

Andrews concluded the discussion with the suggestion that a multitrait design could be used to generate some estimates of construct validity to serve as another outcome measure.

A program for interviewer training

Axelrod opened the discussion of the Presser presentation by emphasizing that this was a most useful and profitable session. Taken together, the three presentations represent a solid program for recruiting, training, and maintaining interviewers.

Sudman commented that if you ask the typical interviewer to name the hardest part of the job, most will say obtaining cooperation. He expressed concern about what the Survey Research Center (SRC) is giving up by the elimination of practice field work. Presser explained that they are not eliminating experience in training where the interviewer goes out into the field and knocks on doors. Rather, they are eliminating this during the first half of training; in the past the interviewers went out on the second night of training, they now go out on the fourth and fifth nights.

Sudman inquired as to whether there were any training materials specific to gaining respondent cooperation. Presser responded that they were still very much in the developmental phase of this work and currently only have the revised materials on principles for interviewing techniques which he had described in his presentation. Their hope is to have the time and resources to do the same kind of work in some of the other areas. Sudman suggested that obtaining respondent cooperation is one

of the areas where we get the greatest variance and thus should be a high priority for revising training materials. Cannell agreed but expressed uncertainty as to how to do that. Groves pointed out that this is being done on the telephone through reluctant-respondent role playing. There was general agreement that this was mcuh easier to do on the telephone.

A contact training procedure that some organizations use apparently successfully was described. This involves having some of the most experienced interviewers act as devil's advocates at training sessions for new interviewers, to pose the kinds of contact problems they have encountered in the field. This approach is considered much more effective than a lecture on the topic by a supervisor.

Velez asked on what basis the decision was made that giving a generalized set of skills in interviewing is better than giving very specialized skills relevant to the specific study. Presser agreed that that is a dilemma we face. It is very expensive to recruit, hire, and train interviewers; ideally we would like to think of these as investments we can amortize over time rather than as one-time expenditures. Unfortunately, we find it difficult to retain interviewers because we don't have enough work for them. As Presser noted in his presentation, extensive generalized training in interviewing techiques may not be cost effective. Simply on the basis of cost it may make more sense just to train interviewers for the specific project they will be working on.

Cannell suggested that even if training is just for a specific study, it is best to break training into two sessions. First, indoctrinate the interviewers on general procedures without getting involved in the meaning of specific questions. After they get the concepts down and are somewhat comfortable in handling techniques, then introduce the content of the questionnaire.

Morton-Williams described related aspects of interviewer training in England. A training handbook with exercises is used in conjunction with the interviewer's manual. It starts off with some examples with the answers given and is followed by questions which the interviewers answer and which are marked by the training officer prior to the initial training session. While this approach has not been evaluated in terms of its effectiveness as a training tool, the training officers feel it is effective and that it saves them a lot of time in the actual training session. Not only does this exercise provide the new interviewer with concrete examples prior to initial training, it also helps them during training to understand the types of errors which can be made.

Another procedure relates to actual field experience during training. A short training questionnaire is used for this. The supervisors will take out several interviewers, and several interviewers will go out on their own and administer the short training questionnaire. Then all of them will meet and exchange information on their experiences. This is done to help the interviewers become more comfortable about knocking on doors and to actually train them in making an initial approach even though it may be a somewhat different approach from that of the actual study they will be working on.

Walden referred to the previous comment by Cannell (breaking training into two sessions) and pointed out that that was exactly the kind of trade-off decision faced at the time of training on the National Medical Care Expenditure Survey (NMCES). There was a sense among the staff at NCHSR and NCHS and at RTI and NORC that breaking interviewer training into two sessions with some time between sessions was a good idea. However, the opposite decision was made because of cost considerations. The NMCES had a large area probability sample with interviewers located throughout the United States, and the expense of having them travel to the training site twice was considered prohibitive.

Cannell noted that he wasn't necessarily suggesting that there had to be two sessions separated by a period of time. Rather, he felt that the essential point was to separate training in techniques from training on the specific questionnaire. For example, with a four-day training session, rather than introducing the complexities of the questionnaire on the first or second day, devote the first two days to training in interviewing techniques and the last two to the questionnaire. Get the techniques down first so the interviewers will have an idea of what their role is and then try to apply it to a specific questionnaire.

Fuchsberg suggested that the procedure Cannell described was appropriate only for new interviewers. With experienced interviewers the techniques are reviewed rather briefly prior to specific training for the study questionnaire. Presser noted that SRC provides new interviewers with five days of training. Experienced interviewers receive only study-specific training unless there is evidence of a need for retraining on general techniques.

Sudman commented that the trade-off between training in interviewing techniques and study-specific training is an important one when training time is limited—which more often than not it is. One may conclude that study-specific errors could be avoided by spending less time on basic techniques (for example, on open-ended questions if there aren't any in the specific study) and more time on study-specific aspects. On the other hand, particularly if there is reasonable attrition between studies, there is the danger of going in the other direction and having higher error rates on the problematic items because of limited training on techniques.

SESSION 6:
Survey methods for rare populations

Chair: Seymour Sudman, University of Illinois

Recorder: Ronald Czaja, Survey Research Laboratory, University of Illinois

Locating patients with rare diseases using network sampling: Frequency and quality of reporting*

Ronald Czaja, University of Illinois

Richard B. Warnecke, University of Illinois

Elizabeth Eastman, University of Illinois

Patricia Royston, National Center for Health Statistics

Monroe Sirken, National Center for Health Statistics

Diane Tuteur, Illinois Cancer Council

Introduction and discussion of research

The rapidly increasing costs of health care have created a need to relieve those faced with financial devastation in the wake of serious, chronic illness. Cost-effective programs must be developed, and these depend on national estimates of costs associated with such illness. Surveys employing traditional sampling frames and interviewing methods have not provided these estimates because identifying a large national probability sample of patients is difficult and because relatives and health care providers often limit the access to patients which is needed to obtain accurate and verifiable reports of direct and indirect costs.

In response to these difficulties, the National Cancer Institute (NCI) contracted with the National Center for Health Statistics (NCHS) for a series of survey experiments to develop and test methods that might be incorporated into the National Health Interview Survey (NHIS) to obtain data on the cost of cancer care. The NCHS staff and staff of the Survey Research Laboratory (SRL) at the University of Illinois designed and implemented these experiments. This paper concerns two experiments designed to evaluate the feasibility of using network sampling techniques to identify cancer patients in a general population survey. The following questions were asked to assess the feasibility of network sampling: (1) Will a known cancer case be reported either in the patient's household or in the household of a relative? (2) How accurately will the cancer site be reported? (3) How accurately will the date of diagnosis be reported? (4) How accurately will the names and addresses of patients in the network be reported?

Sampling concerns. Conducting national surveys to estimate the cost of an illness requires identifying a national probability sample of recently diagnosed cases of the disease, confirming the diagnoses with the health care sources, obtaining cost information from patients, and then verifying those data with the providers. Because cancer is very rare and has a highly variable survival rate, identifying a national sample of cancer patients requires special sampling procedures. One of the two strategies usually employed is based on a sampling frame of medical care providers, and the other is based on a sampling frame of households.

When a sampling frame of medical care providers is employed, patients are contacted through their health care source. A provider sampling frame requires an unbiased procedure for selecting hospitals, clinics, and other facilities that maintain records of patients. Once the facilities are identified and their cooperation obtained, the records are screened to obtain a list of eligible patients. Access to patients then requires permission from the physician and often from the health care institution and finally depends on contacting the patient.

The household survey approach typically employs an area probability sample. As part of the interview, the household members are screened to identify patients, who are then requested to participate in a later interview. During that interview, the health care data of interest are obtained along with written consent to verify this information with the health care provider.

The medical care provider frame may seem more efficient, since the patients are identified from formal records and so their diagnoses are not in question. Using this ready and accurate source also avoids the high costs of screening for a rare disease like cancer in the general population. Nevertheless, several recent efforts to estimate cancer costs using a provider frame have been unsuccessful (Kalsbeek et al., 1977; Eldred et al., 1977; Robins et al., 1978). Although more than 80% of the relevant treatment facilities permitted access to records for screening to locate patients, they would not release the patient data or permit any patient contact without the physicians' permission. The physicians' and patients' refusals to allow the interviews limited the respondents to less than 30% of the presumably eligible patients.

On the other hand, traditional household surveys are subject to large sampling errors because most serious illnesses have low prevalence rates. Even with extensive screening, identifying enough cases to provide an adequate sample is difficult and costly. When cancer is studied, this problem is exacerbated, since cancer refers to a number of distinct disease entities, each of which has a different etiology. The stage at diagnosis is also important in many studies, and the need for detailed staging data further complicates the development of a sampling frame that will yield sufficient numbers of cases to avoid large sampling errors.

* This research was supported by the National Cancer Institute under contract no. 233-79-2081. The authors would like to thank Susan Albert and Nancy Lipse for their editorial assistance.

Response bias also has been a problem with traditional survey approaches. Coverage errors result from patient reluctance to mention diseases like cancer in an interview. They also result from omitting certain sources of eligible respondents from the sampling frame. The sources may be long-term care institutions that house many elderly people likely to have a disease such as cancer. Reporting errors result from the patients' inability to accurately identify their illness.

Using network sampling for household surveys may avoid some of these drawbacks. The basic difference between network sampling and traditional sampling is the counting rule applied to define case eligibility (Sirken, 1972a). Traditional surveys employ a counting rule that considers persons eligible for the study if they are identified in their own households. Multiplicity counting rules, employed in network sampling, include individuals identified by the households of specified relatives as well as by their own as eligible respondents. For example, in a household network survey to identify cancer patients, a respondent in a given household would be asked to identify cancer patients in that household and in households of close relatives. Relatives so identified are then recruited for the study. As part of the interview, the number of all relatives eligible to name the patient is obtained so that the probability of selection for each patient can be computed.

Since the late 1950s, Monroe Sirken and colleagues at the NCHS have been using network sampling to estimate prevalence of various kinds of illness (Sirken et al., 1959; Kramm et al., 1962). Particularly since 1970, they have published a series of articles working out most of the major theoretical problems associated with the technique (Sirken, 1970a; Ibid, 1970b; Ibid, 1972b; Sirken and Levy, 1974; Sirken, 1975; Sirken et al., 1975; Nathan, 1976; Levy, 1977a; Ibid, 1977b). One problem that remains unsolved is the cost of conducting network surveys. Network sampling may be more expensive than regular sampling because obtaining a complete list of eligible network sample members adds to the interviewing time. If fewer interviews ultimately are required to locate an adequately large sample, however, this extra time is justified. Network sampling may be particularly appropriate for studies involving people who are hard to identify through conventional sampling methods because their disease is rare, they are institutionalized, or they are reluctant to report their own condition.

In order to avoid overestimating cases due to relatives' positive misreports, the patients' permission to examine medical records is necessary, and thus the ultimate issue remains whether patients will acknowledge their cancers.

Design of experiments 1 and 2. The first experiment of our study was to assess whether cancer patients would be reported in their own households as part of a general health interview. Experiment 2 was designed to test network sampling and reporting procedures. In effect, the two experiments reversed the proposed strategy for the national survey. In Experiment 1 patients were interviewed, and names and addresses of relatives eligible for their network were obtained. In Experiment 2 these relatives were interviewed to ascertain whether they would be accurate reporters of the patients' cancers, of the names and addresses of the patients, and of the numbers of other eligible network relatives.

The sample for Experiment 1 was obtained from two regional tumor registries. They provided us with a sample of 325 patients whose cancers according to their records had been diagnosed within the three years preceding the experiment. From the information provided by the patients in Experiment 1, a sample of 205 relatives living in Illinois was assembled. These 530 respondents were combined with a larger sample of decoy households.

The registry patients were selected from a stratified sample. The strata were defined by geographic region, disease site, and diagnostic period. The geographic regions were the Chicago Standard Metropolitan Statistical Area (SMSA), other Illinois SMSAs, and non-SMSA areas in Illinois. The disease sites were grouped as follows: colon and rectum; breast; cervix uteri, corpus uteri, and ovary; prostate; kidney and bladder; leukemia, lymphoma, and Hodgkin's disease; oral and larynx; melanoma; and miscellaneous (all other sites except skin). The third stratum included three diagnostic periods: August through July for 1977–78, 1978–79, and 1979–80. All miscellaneous cases had to be diagnosed in the year prior to the interview, since their expected survival term was less than one year. As far as possible, cases were to be equally distributed among the cells in the sample design.

A particularly sensitive issue in Experiments 1 and 2 was protection of the sample members' right to privacy. Although it was important to know the identity of all cases in the sample to assess the quality of reporting, disclosing their illnesses was left as their prerogative. Thus, interviewers and staff at the SRL did not know the identity of any case unless the cancer was reported in an interview, and patients and their relatives were unaware of the basis of their selection for the study. Because this confidentiality was so important, a third party managed the sample.

The Illinois Cancer Council (ICC), one of the 21 comprehensive cancer centers established under the 1971 National Cancer Act, was this third party. The ICC selected patients from the registry samples and decoys from reverse telephone and street directories by address. The decoys and patients were combined in one list, which did not distinguish between them and then forwarded to the SRL. As the completed interviews returned to the SRL, staff abstracted certain information the ICC needed to identify relatives of registry patients and new patients, and then the relative households were integrated with new decoy households for the Experiment 2 sample. This second sample was continuously

Table 1
Response rates for registry patient, relative, decoy, and new patient households by sample disposition

Sample disposition	Registry patient N	(%)	Relative N	(%)	Decoy N	(%)	New patient N	(%)	Total N	(%)
In-scope cases										
Completed interview	264	(89)	162	(84)	363	(91)	26	(87)	815	(89)
Refused interview	26	(9)	27	(14)	35	(9)	2	(7)	90	(10)
Other	5	(2)	4	(2)	2	(0)	2	(7)	13	(1)
Total in-scope	295	(100)	193	(100)	400	(100)	30	(101)	918	(100)
Out-of-scope cases										
Completed interview	15		8		3		—		26	
Moved out of Illinois or moved and not locatable	6		0		32		—		38	
Dead, lived alone	8		1		5		—		14	
Other	1		3		2		—		6	
Total out-of-scope	30		12		42		—		84	
Total sample	325		205		442		30		1,002	

integrated with the Experiment 1 sample, as both were worked as if they were one. After completion of field work, all information identifying the patient was destroyed.

Response rates. As shown in Table 1, cases were defined as in-scope and out-of-scope. Households of patients who were not enumerated as a member of the household or deceased with no survivors, who had moved out-of-state, or who could not otherwise be located were out-of-scope.

Overall, 8% of the households in the initial sample were out-of-scope. Thirty of the 325 registry patient households were so classified, and 15 of these 30 did not contain the designated respondent. Seventy-five percent of the out-of-scope relative households also did not contain the designated respondent. Patients who had died more than one year ago and relatives who had died were also designated as out-of-scope. In contrast, the decoy cases were mostly so defined because the selected surname had moved out of state or had moved and left no forwarding address.

The response rate for the entire sample was 89%, 10% refused to be interviewed, and 1% was not interviewed for other reasons. Patients and decoys participated more than 89% of the time, while relatives cooperated in only 84% of the interviews, yielding a refusal rate 5% higher than the rate for patients and decoys. As relatives were interviewed after their related patients, they could have been forewarned by the patients about the content and length of the interview and demurred. Eighty-seven percent of the new patients cooperated, and half of them were contacted following an interview with a relative. Decoy and new patient households are not included in this analysis.

Results

Completeness of reporting

Registry patient households. Rates of reporting cancer for registry patient and relative households are shown in Table 2. These rates are analyzed by the following patient characteristics: age, sex, race, marital status, education, household income, geographic area of residence, vital status, disease site or type, diagnostic period, and whether the patient was present at the interview.

Eighty-nine percent of the patients selected from the participating registries were reported. Of those reported, 48% were reported in the core of the NHIS questionnaire, and the remainder were identified in the supplement designed for this study. Whether or not the registry patient households reported the cancers seems to depend on race and the disease site. In white patient households, 93% of the cancers were reported; but in nonwhite patient households, the rate dropped to 63%. Since the number of nonwhite respondents was very small (26 Black and 4 other races), this 30% difference, while substantial, must be considered provisional.

All (100%) of the patients with cancers of the rectum, breast, corpus uteri, larynx, melanoma, and lung were reported by their households. Two additional sites reported with very high frequency were colon (95%) and oral (92%). Patients with cancers of the prostate, bladder, and cervix uteri were reported with frequencies ranging from 73% to 77%. Since most of our interviewers were

Table 2
Registry patient reporting rates for registry patient households and relative households by characteristics of patients

Patient characteristics	Reporting rates			
	Registry patient		Relative	
	%	Total N	%	Total N
Total reported	89	264	80	162
Age				
<65	90	135	82	83
≥65	88	129	77	79
Sex				
Male	86	133	72	83
Female	92	131	87	79
Race				
White	93	234	84	147
Nonwhite	63	30	40	15
Marital status				
Married	91	181	—	—
Widowed, divorced, separated	86	64	—	—
Never married	83	18	—	—
Education				
Grades 1–11 completed	84	109	—	—
High school completed	93	88	—	—
Some college completed	92	67	—	—
Household income				
<$15,000	88	130	—	—
≥$15,000	92	110	—	—
Refused, didn't know	88	24	—	—
Place of residence				
Chicago SMSA	83	82	—	—
Other SMSA	90	81	—	—
Non-SMSA	94	101	—	—
Vital status				
Living	89	242	80	150
Deceased	96	22	75	12
Cancer site				
Colon	95	19	86	14
Rectum	100	11	100	8
Breast	100	32	95	19
Cervix uteri	77	13	44	9
Corpus uteri	100	13	100	6
Prostate	73	33	65	23
Bladder	77	22	69	16
Kidney	88	8	83	6
Leukemia, lymphoma, Hodgkin's disease	84	25	91	11
Oral	92	24	82	17
Larynx	100	10	100	5
Melanoma	100	28	100	9
Lung	100	11	67	12
Miscellaneous	80	15	57	7
Diagnostic period				
≤12 months	92	92	84	58
13–24 months	85	86	67	48
>24 months	91	86	86	56
Patient present at interview				
Yes	89	187	—	—
No	91	66	—	—
Not recorded	82	11	—	—

female and all prostate and most bladder patients were male, a possible explanation for these lower rates may have been sex difference between interviewers and patients. When investigated, however, male patients were more likely to be reported when present at the interview than when they were absent or their presence was unrecorded.

The relationship between race and patient household reporting is further examined in Table 3. As the table indicates, the nonwhites tended to be younger and less well educated than the whites and earned less than $15,000 annually more often than did the whites. Since these variables differentiate the nonwhites from the whites, race may thus confound the effects of these variables on the reporting rates. When nonwhites are excluded from the analyses, the variables that characterized the nonwhites do seem to have less of an effect on the reporting rates than they do when the entire sample is considered (Table 2). For example, differences in income are not associated with differences in registry patient reporting rates for whites but are associated with differences in these reporting rates for the total sample. Similarly, three more cancer sites are reported 100% of the time when nonwhites are excluded than when they are included in the analysis. Only reporting of bladder, cervix, and prostate cancers is less than 85% when nonwhites are not considered. In general, the nonwhite patient households reported cancers with considerably less frequency than did the white patient households.

In summary, households with cancer patients tend to report them. Certain variables seem to affect this reporting, particularly the race of the patient and the kind or site of cancer. Demographic factors have minimal effect.

Relatives. As shown in Table 1, the number of in-scope interviews conducted in the households of relatives is about 100 less than the number of interviews in registry patient households. Fifty-four of the 264 registry patient households did not have an eligible network relative in Illinois, and 5 refused to provide address information for an eligible relative. Hence, the initial sample of relative households was 205. During the interviewing process, 12 cases were found to be out-of-scope, and 31 cases could not be interviewed, which yielded 162 interviews with in-scope relative households. All tables in this paper dealing with relatives are based on these 162 completed interviews.

Table 3
Registry patient reporting rates for registry patient households by characteristics of patients controlling for race of patient

Patient characteristics	Reporting rates			
	White		Nonwhite	
	%	Total N	%	Total N
Total sample	93	234	63	30
Age				
<65	96	115	60	20
≥65	90	119	70	10
Sex				
Male	88	122	64	11
Female	97	112	63	19
Marital status				
Married	93	167	64	14
Widowed, divorced, separated	92	52	58	12
Never married	86	14	75	4
Education				
Grades 1–11 completed	90	88	62	21
High school completed	94	83	80	5
Some college completed	95	63	50	4
Household income				
<$15,000	93	104	65	26
≥$15,000	93	108	50	2
Refused, didn't know	91	22	50	2
Place of residence				
Chicago SMSA	91	56	65	26
Other SMSA	92	78	33	3
Non-SMSA	94	100	100	1
Vital status				
Living	92	214	64	28
Deceased	100	20	50	2
Cancer site				
Colon	100	17	50	2
Rectum	100	11	0	0
Breast	100	24	100	8
Cervix uteri	83	12	0	1
Corpus uteri	100	13	0	0
Prostate	82	28	20	5
Bladder	76	21	100	1
Kidney	100	7	0	1
Leukemia, lymphoma, Hodgkin's disease	86	22	67	3
Oral	91	22	100	2
Larynx	100	9	100	1
Melanoma	100	27	100	1
Lung	100	11	0	0
Miscellaneous	100	10	40	5
Diagnostic date to patient interview date				
≤12 months	96	82	60	10
13–24 months	88	78	50	8
>24 months	93	74	75	12
Patient present at interview				
Yes	93	162	64	25
No	92	62	75	4
Not recorded	90	10	0	1

As can be seen in Table 2, 80% of the registry cancer patients were reported in the households of randomly selected relatives. This rate is 9% *less* than the rate at which these same cancer patients were reported in their own households. Relatives were more likely to report the cancer when the patient was white, female, or when the cancer had been diagnosed in the year immediately preceding the interview or at least two years preceding it. When compared with households of patients, the relative households tended to report all sites but leukemia, lymphoma, and Hodgkin's disease less frequently than did patient households.

Table 4 presents reporting rates by the characteristics of the relative households. The characteristics that appear most important are whether the relative resided outside the Chicago SMSA or whether the head of the household was educated beyond high school. Other less significant variables also add to a consistent picture. First, when the relative was the patient's child or was present at the interview, reporting was better than when the relative was a sibling or not present at the interview. Second, when the head of the relative household was

Table 4
Registry patient reporting rates for relative households by characteristics of relatives

Relative characteristics	Reporting rates	
	%	Total N
Types of relative selected		
Sibling	75	84
Child	85	78
Selected relative present at interview		
Yes	83	109
No	72	46
Not recorded	86	7
Sex of respondent		
Male	82	74
Female	76	81
Not recorded	86	7
Age of head of household		
18–39	87	61
40–64	76	70
≥65	74	31
Education of head of household		
Grades 1–11 completed	64	45
High school completed	83	52
Some college completed	88	65
Household income		
<$15,000	70	50
≥$15,000	84	103
Refused, didn't know	78	9
Place of residence		
Chicago SMSA	67	48
Other SMSA	84	43
Non-SMSA	86	71

Table 5
Registry patient reporting rates for relative households by patient and relative characteristics controlling for patient sex

Patient and relative characteristics	Patient sex: Male %	Male Total N	Female %	Female Total N
Total sample	72	83	87	79
Patient characteristics				
Age				
<65	78	32	84	51
≥65	69	51	93	28
Race				
White	77	77	91	70
Nonwhite	17	6	56	9
Vital status				
Living	73	74	87	76
Deceased	67	9	100	3
Cancer site				
Colon, rectum	89	9	92	13
Breast	0	0	95	19
Cervix and corpus uteri	0	0	67	15
Prostate	65	23	0	0
Bladder, kidney	71	17	80	5
Leukemia, lymphoma, Hodgkin's disease	83	6	100	4
Oral, larynx	82	11	90	10
Melanoma	43	7	100	5
Miscellaneous	80	10	88	8
Diagnosis date to patient interview date				
≤12 months	80	30	89	28
13–24 months	56	27	81	21
>2 months	81	26	90	30
Relative characteristics				
Type of relative selected				
Sibling	63	38	85	46
Child	80	45	91	33
Selected relative present at interview				
Yes	78	54	87	55
No	63	27	84	19
Not recorded	50	2	100	5
Sex of respondent				
Male	84	37	81	37
Female	64	44	92	37
Not recorded	50	2	100	5
Age of head of household				
18–39	85	34	89	27
40–64	66	35	86	35
≥65	57	14	88	17
Education of head of household				
Grades 1–11 completed	45	20	80	25
High school completed	79	24	86	28
Some college completed	82	39	96	26
Household income				
<$15,000	63	27	78	23
≥$15,000	79	52	90	51
Refused, didn't know	50	4	100	5
Place of residence				
Chicago SMSA	61	23	72	25
Other SMSA	74	23	95	20
Non-SMSA	78	37	94	34

relatively young (less than forty) or relatively affluent (household income in excess of $15,000), reporting was more likely than when these attributes were not present.

Table 5 shows reporting rates when the patient's sex is controlled. The relative households reported 87% of female patients but only 72% of the male patients. Reporting female patients was most likely when relatives were relatively well educated, in households with an income in excess of $15,000, or in households located outside the Chicago SMSA.

More differences appear in reporting male patients. Children were more likely than siblings and younger relatives were more likely than older relatives to report male patients. Reporting was also more likely when the relative was present at the interview. Finally, as with female patients, male patients were most likely to be reported by relatives who were relatively well educated, were affluent, or lived outside the Chicago SMSA.

Some analysis of relative household reporting was also conducted controlling for race (data not presented). As with the patient household reporting, relatives of white patients generally reported more frequently than relatives of nonwhite patients. These patterns were most evident when overall the reporting was poorest. For example, white relatives with little education reported 77% in comparison with the overall rate of 64%; when income was under $15,000, whites reported relatives with cancer 79% of the time compared with the combined rate of 70%.

In summary, the patient and relative characteristics influenced reporting by relative households, but these characteristics differ from those that influenced patient household reporting. With the exceptions of the sex and race of the patient, the characteristics of the relative and the relative's household appear to have been more important than characteristics of the patients in influencing the relative household reporting rate. Children are more likely than siblings to report the cancer. The age and education of the head of the relative's household, the respondent's sex, and the geographic location

of the household all affect the frequency with which a relative household reports a patient. As noted, the patient's race also appears significant, but the small numbers in these cells makes interpretation of this variable difficult.

Completeness of reporting by patient and relative pairs. The discusion thus far has addressed differences between patient and relative households in reporting cancer patients. The next issue of interest is the extent to which information not obtained from the patient household is obtained from the selected relative households. This section will consider (1) the extent to which the patient and relative households reported the same patient; (2) the number of patients not reported in their own households but reported by relatives; and (3) the extent to which reporting bias was consistent in pairs of households, that is, the number of patient and relative pairs that did not disclose the cancer diagnosis of the patient.

To address these issues, the 162 pairs for which interviews were conducted in the patient household and its relative household were examined. In 77% of these pairs, both households reported the patient, but in 6%, neither household reported the patients. The remaining pairs were split. The patient household reported the cancer but the relative household did not in 14% of the pairs, and the relative household reported the cancer but the patient household did not in 3% of the pairs.

As a means of locating patients, network sampling is very efficient. Since 77% of the cancers were reported twice and 94% by at least one member of the network, the chances of locating cancers in the general population would be improved considerably by including both relative and patient households in a survey. Finally, the data suggest that siblings are less informative than are children, and so a network sample might be limited to children and members of the patient household.

Accuracy of reporting cancer site

Patient households. While network sampling has proved to be an efficient means of locating patients, whether the data obtained from these respondents are of high quality must also be investigated. To determine this quality, respondent descriptions of the cancer site or type were compared with the registry information. These comparisons were coded into three categories. A response was considered an exact match when the two descriptions were identical. A response was a close match when the reported site was anatomically close to that listed in the registry. For example, a close match occurred when the respondent mentioned cervix and the registry site was corpus uteri. When a respondent identified a site totally dissimilar to the one listed in the registry, the response was classified a "no match." For example, if back of the neck was mentioned when the tumor was a lymphoma, it was considered a no match. A response was also coded as a no match when the registry site was listed as melanoma and the respondent mentioned only a part of the body, such as ear or back, and did not use the term melanoma. The 31 responses that mentioned multiple sites were coded as exact matches if one of the mentioned sites matched the registry site and as close or no matches based on the rules described above.

The quality of cancer site reporting for patient households is shown in Table 6. Of those patient households that reported a cancer, 78% accurately reported the site or type of cancer, 8% were close in describing it, and 14% did not agree with the registry records. The accuracy of reporting varied considerably by site. Households in which patients had cancers of the breast, bladder, larynx, and lung and leukemia, lymphoma, and Hodgkin's disease, were accurate in describing the cancers more than 90% of the time. When patients had cancers of the cervix, corpus uteri and rectum, their households reported sites that frequently matched the registry information to a least some degree. Households of patients with the remaining cancers except miscellaneous ones identified the cancers completely incorrectly less than 20% of the time. The low accuracy rate for the miscellaneous category is deceptive. Of the 12 responses classified as miscellaneous, 2 gave an exact description of the cancer site, 1 was close, and 9 were coded as no matches. Of these 9, 5 had an unknown primary and, therefore, an exact or close match was not possible. A sixth case had a cancer of the connective tissue, which is very difficult to describe correctly. Thus, overall, the matching was quite good.

Patient vital status was the most significant patient characteristic that affected accuracy of site reporting. The households of living patients exactly reported the cancer type or site 79% of the time and incorrectly reported the cancer only 12% of the time. Only 62% of the households of deceased patients were accurate and 38% were inaccurate in their descriptions of the cancer. These findings are tentative, as there were only 21 deceased cases.

Moderate differences in accuracy were related to the patient's race and education. Households of white patients had a higher proportion of exact matches and a lower proportion of no matches than did the nonwhite households. Eighty-seven percent of patients with some college correctly reported the site, but only 77% of those with a high school degree and 72% of those with 11 years of education or less correctly identified the site. Education and vital status were associated with accuracy of reporting but were not associated with rate of reporting. As with rates of patient household reporting, the race of the patient was related to accuracy of reporting.

Relative households. Data bearing on the accuracy of reporting the cancer site by a respondent in the relative household are presented in Table 7. Seventy percent of the relative households reported the site accurately, 11% were close in describing the site or type of cancer, and 19% did not agree with the registry. In general, relative

Table 6
Accuracy of reporting cancer site for cancer patient households by characteristics of patients

Patient cherecteristics	Accurecy of reporting: Exact Match	Close Metch	No match	Total %	Total N
Total sample	78%	8%	14%	100%	236
Age					
<65	81	10	9	100	122
⩾65	74	7	19	100	114
Sex					
Male	82	8	10	100	115
Female	74	9	17	100	121
Race					
White	78	9	12	99	217
Nonwhite	68	0	32	100	19
Marital status					
Married	79	9	12	100	165
Widowed, divorced, separated	71	9	20	100	55
Never married	87	0	13	100	15
Education					
Grades 1–11 completed	72	10	18	100	92
High school completed	77	8	15	100	82
Some college completed	87	6	6	99	62
Household income					
<$15,000	75	7	18	100	114
⩾$15,000	81	10	9	100	101
Refused, didn't know	71	10	19	100	21
Place of residence					
Chicago SMSA	75	6	19	100	68
Other SMSA	80	11	10	101	73
Non-SMSA	78	8	14	100	95
Vital status					
Living	79	9	12	100	215
Deceased	62	0	38	100	21
Cancer site					
Colon	78%	6%	17%	101%	18
Rectum	36	64	0	100	11
Breast	94	0	6	100	32
Cervix uteri	50	40	10	100	10
Corpus uteri	69	15	15	.99	13
Prostate	83	0	17	100	24
Bladder	100	0	0	100	17
Kidney	86	0	14	100	7
Leukemia, lymphoma, Hodgkin's disease	95	0	5	100	21
Oral	59	23	18	100	22
Larynx	100	0	0	100	10
Melanoma	82	0	18	100	28
Lung	91	0	9	100	11
Miscellaneous	17	8	75	100	12
Diagnostic date to patient interview date					
⩽12 months	74	5	21	100	85
13–24 months	78	10	12	100	73
>24 months	81	12	8	101	78
Patient present at interview					
Yes	78	8	14	100	167
No	72	12	17	101	60
Not recorded	100	0	0	100	9

Table 7
Accuracy of reporting cancer site for relative households by characteristics of patients and relatives

Patient and reletive characteristics	Accurecy of reporting: Exact match	Close match	No match	Total %	Total N
Total sample	70%	11%	19%	100%	129
Patient characteristics					
Age					
<65	72	10	18	100	68
⩾65	67	13	20	100	61
Sex					
Male	72	8	20	100	60
Female	68	14	17	99	69
Race					
White	69	12	19	100	123
Nonwhite	83	0	17	100	6
Vital status					
Living	70	13	18	101	120
Deceased	67	0	33	100	9
Cancer site					
Colon	83	8	8	99	12
Rectum	13	75	13	101	8
Breast	89	11	0	100	18
Cervix uteri	25	25	50	100	4
Corpus uteri	83	0	17	100	6
Prostate	80	0	20	100	15
Bladder	91	0	9	100	11
Kidney	80	0	20	100	5
Leukemia, lymphoma, Hodgkin's disease	80	0	20	100	10
Oral	43	14	43	100	14
Larynx	80	20	0	100	5
Lung	89	0	11	100	9
Melanoma	38	25	38	101	8
Miscellaneous	50	0	50	100	4
Diagnosis date to patient interview to date					
⩽12 months	65	10	24	99	49
13–24 months	69	9	22	100	32
>24 months	75	15	10	100	48
Relative Characteristics					
Type of relative selected					
Sibling	73%	10%	17%	100%	63
Child	67	13	20	100	66
Selected relative present at interview					
Yes	66	12	22	100	90
No	79	9	12	100	33
Not recorded	83	17	0	100	6
Sex of respondent					
Male	67	13	20	100	61
Female	71	10	19	100	62
Not recorded	83	17	0	100	6
Age of head of household					
18–39	68	13	19	100	53
40–64	77	8	15	100	53
⩾65	56	17	26	99	23

Table 7 continued

Patient and relative characteristics	Accuracy of reporting				
	Exact match	Close match	No match	Total %	Total N
Education of head of household					
Grades 1–11 completed	62	14	24	100	29
High school completed	70	12	19	101	43
Some college completed	74	10	16	100	57
Household income					
<$15,000	66	11	23	100	35
≥$15,000	72	12	16	100	87
Refused, didn't know	57	14	29	100	7
Place of residence					
Chicago SMSA	66	9	25	100	32
Other SMSA	78	11	11	100	36
Non-SMSA	67	13	20	100	61

households did not have difficulty in describing the cancer site. Breast, bladder, and lung cancers were accurately reported approximately 90% of the time. Cancers of the colon, corpus uteri, prostate, kidney, and larynx, as well as leukemia, lymphoma, and Hodgkin's disease were correctly described between 80% and 85% of the time. Relatives had most difficulty accurately reporting oral cancer; only 57% of the respondents could provide an exact or close description of that site. No or very small differences were associated with other patient or relative characteristics.

Accuracy of site reporting by patient and relative pairs. When compared with the registry data, the pairs of patient and relative households were accurate in their description of the cancer site 63% of the time and were close 4% of the time. In 19% of the pairs, patient households provided an exact or close description when the relatives were close or incorrect. The relative households were more accurate than patient households 6% of the time. Both households in 8% of the pairs were incorrect. The relative households contributed only slightly to the accuracy of site reporting, but the quality of their information was as high as the patient households' four of every five occasions.

Accuracy of reporting date of diagnosis. Both types of houselhds were asked to provide information about the date of the cancer diagnosis. All hospitals do not use the same event to mark the official date of diagnosis, but use one of the following dates: (1) the date of admission to the hospital for diagnostic tests, (2) the date of admission for surgery, or (3) the date clinical or histological diagnosis was made. The following tables will compare how well the date of diagnosis reported by the patient agrees with that contained in the registry. Accuracy will be rated according to the number of months between the reported and registry dates of diagnosis.

Patient households. Table 8 presents the intervals between the patient household reported date of diagnosis and the registry date of diagnosis. Seventy percent of these households reported a date within one month of the registry date, and 4% reported a date more than one month *after* the registry date of diagnosis. Approximately 25% of the patient households reported a date of diagnosis more than one month *before* the registry date of diagnosis. Of these, 9% reported a date six months or less before and 16% reported a date at least seven months before the registry date. These data indicate that patients or members of their households report the date of diagnosis quite accurately.

An interesting feature of these reports is the pronounced tendency to mention a date earlier than the registry date. Three explanations are plausible. First, cancer is not a discrete event; there are symptoms or warning signals associated with its onset. The patients and other household members may have recalled signs that occurred many months before the patient actually saw a doctor. Although we asked when the patient was first told by a doctor that he/she had cancer, the respondent quite possibly reinterpreted the question to be, "When did you first realize you might have cancer?" Our data are based on only one question, and so we are unable to evaluate this possibility.

A second and more likely explanation attributes the differences to telescoping of responses. Telescoping occurs when a respondent remembers the event as occurring earlier (as in our situation) or later than it actually occurred. One means of determining telescoping is to compare the variation between the dates of diagnosis with the length of time since diagnosis. Those who were diagnosed most recently should most accurately report their date of diagnosis. Data in Table 8 show that of those respondents diagnosed within 24 months of the interview, approximately 75% of their households reported a date of diagnosis that was within 1 month of the registry date; however, only 59% of the households in which patients were diagnosed more than 2 years before the interview reported the date with such accuracy. If 6 months from the registry date is used as the limit for accurate reporting, the pattern is even clearer. For those diagnosed not more than one year before the date of the interview, 92% of their households reported a date within 6 months of the registry date; for those diagnosed 13–24 months from the date of the interview, 85% of the reports were correct within 6 months of the date of diagnosis; and for those diagnosed more than 24 months before the date of interview, 63% of their households reported a date of diagnosis that was within 6 months of the registry date.

A third possible explanation for the discrepancies between the reported and registry dates of diagnosis concerns the selection of cases for the sample. Our instructions to the registries were to include only those cases that received an initial diagnosis of a primary cancer at the reporting institution, but primary cancer in our study could have been a second primary. Some of our cases thus may have had an earlier diagnosis for another cancer.

This possibility was pursued by examining the ac-

Table 8
Accuracy of reporting date of diagnosis for patient households by characteristics of patients

Patient characteristics	Interval between reported and registry dates of diagnosis							
	>1 month after DDX	±1 month from DDX	>1–6 months before DDX	≥7 months before DDX	Cannot classify	Total %	±6 months from DDX	Total N
Total sample	4%	70%	9%	16%	1%	100%	80%	236
Age								
<65	2	77	9	11	2	101	87	122
≥65	6	62	9	22	1	100	73	114
Sex								
Male	7	67	8	17	2	101	77	115
Female	1	73	10	16	1	101	83	121
Race								
White	4	71	9	15	1	100	81	217
Nonwhite	5	58	10	26	0	99	68	19
Marital status								
Married	5	68	9	17	1	100	79	165
Widowed, divorced, separated	2	74	7	15	2	100	82	55
Never married	0	80	7	14	0	101	87	15
Education								
Grades 1–11 completed	4	66	10	18	1	99	76	92
High school completed	2	77	6	13	1	99	84	82
Some college completed	5	66	11	16	2	100	81	62
Household income								
<$15,000	4%	69%	11%	15%	1%	100%	81%	114
≥$15,000	4	74	9	12	1	100	85	101
Refused, didn't know	0	52	0	43	5	100	52	21
Place of residence								
Chicago SMSA	9	65	13	10	3	100	81	68
Other SMSA	3	73	4	21	0	101	77	73
Non-SMSA	1	72	10	17	1	101	82	95
Vital status								
Living	4	69	8	17	1	99	79	215
Deceased	5	76	14	5	0	100	90	21
Cancer site								
Colon	0	79	11	11	0	101	89	18
Rectum	9	64	9	18	0	100	73	11
Breast	0	66	12	19	3	100	78	32
Cervix uteri	0	60	20	20	0	100	80	10
Corpus uteri	0	85	0	15	0	100	85	13
Prostate	8	63	8	21	0	100	75	24
Bladder	6	47	6	35	6	100	59	17
Kidney	29	71	0	0	0	100	71	7
Leukemia, lymphoma, Hodgkin's disease	9	86	5	0	0	100	90	21
Oral	5	73	9	14	0	101	86	22
Larynx	0%	80%	0%	20%	0%	100%	80%	10
Melanoma	0	79	7	11	4	101	86	28
Lung	0	73	18	9	0	100	91	11
Miscellaneous	0	50	17	33	0	100	67	12
Diagnosis date to patient interview date								
≤12 months	1	76	14	8	0	99	92	85
13–24 months	3	74	10	12	1	100	85	73
>24 months	8	59	3	28	3	101	63	78

Table 8 continued

Patient characteristics	Interval between reported and registry dates of diagnosis							
	>1 month after DDX	±1 month from DDX	>1–6 months before DDX	≥7 months before DDX	Cannot classify	Total %	±6 months from DDX	Total N
Cancer patient present for interview								
Yes	4	70	8	18	1	101	78	167
No	5	70	12	11	3	101	83	60
Not recorded	0	78	11	11	0	100	89	9

curacy of site reporting for the 19 cases that reported the date of diagnosis prior to 1977. Of these cases, 12 (63%) of their households were exact in their description of the cancer type or site, 2 (11%) were close in their description, and 5 (26%) were inaccurate. These results compare favorably with the results that indicated 86% of the patient households gave an exact or close description of their cancer and 14% were in error (Table 6). Therefore, the telescopic reporting by this proportion of the sample is not due to their having another cancer but to the length of time between date of diagnosis and the interview.

Accuracy of reporting the date of diagnosis by other patient characteristics is also examined in Table 8. Typically, accuracy within one month of the registry date falls between 65% and 75% for these variables. Differences in accuracy of more than 10% are associated with age, race, marital status, and education; the best reporters were those in households with patients who were under 65, white, never married, and moderately well educated.

The most variation in the table is associated with cancer site. Reporting a date of diagnosis that was within one month of the registry date varied from a low of 47% of households with bladder cancer patients to a high of 86% for those patients with leukemia, lymphoma, and Hodgkin's disease. Miscellaneous cases were also associated with very poor information about date of diagnosis. Households of patients with bladder and miscellaneous cancers also reported the date of diagnosis as being more than seven months after the registry date. For the remaining sites, the patient's reported dates were within one month of the registry date at least 60% of the time.

Also shown in Table 8 is the proportion of patient households where the reported date of diagnosis was within six months of the registry date. Using this broader interval, the accuracy of the patient household responses increased by about 10% across all independent variables. As noted earlier, the telescoping relationship between date of diagnosis and date of the interview is most clearly evident when the six-month interval is used.

Relative households. The relative household's accuracy of reporting the patient date of diagnosis is shown in Table 9. The most obvious conclusion from these data is that relative households could not provide accurate data. Only 40% reported a date within one month of the registry date, 13% reported a date more than one month after the registry date and 33% reported a date more than one month before the registry date. Fifty-seven percent of the relative households reported a date of diagnosis that was within six months before or after the registry date, and 29% (not presented in table) reported a date that was within seven months before of after the registry date. We were unable to classify 14% of the cases.

The telescoping by patient households was also apparent in the reporting by relative households. The proportion of relative households in which a respondent gave a date more than one month before the registry date was two and one-half times the proportion that reported a date more than one month after the registry date.

Little more is learned when we examine the relative's accuracy in reporting by various patient and relative characteristics. The accuracy rates for most variable categories cluster aroung the total sample results for reporting a date that was within one or six months of the registry date. When the interval between the diagnosis date and the patient interview date was less than or equal to 12 months, however, the accuracy was higher; 59% reported a date that was within 1 month of the registry date, and 76% reported one that was within 6 months of the registry date. When the one- and six-month boundaries are used, children were more accurate reporters than were siblings, and when the relative's head of household was under age 40 or had completed high school, accuracy improved considerably.

Quality of name and address reporting by patient and relative households

Network sampling is efficient if it can identify persons with specific characteristics at more than one household. This depends on the willingness and accuracy with which respondents report information about their relatives who live outside their households. If these "source" respondents report relatives with the specific characteristics, the names and addresses of the relatives must then be obtained so that an interview can be conducted to confirm the source report and obtain additional information.

Completion of Experiment 2 depended on cooperation from the registry patient households in providing names and addresses of siblings and children. Only 2% of the registry patient households refused to provide this

Table 9
Accuracy of reporting date of diagnosis for relative households by characteristics of patients and relatives

Patient and relative characteristics	Interval between reported and registry dates of diagnosis: >1 month after DDX	±1 month from DDX	>1–6 months before DDX	≥7 months before DDX	Cannot classify	Same year	Total %	±6 months from DDX	Total N
Total sample	13%	40%	10%	23%	10%	4%	100%	57%	129
Patient characteristics									
Age									
<65	13	43	12	18	10	4	100	62	68
≥65	13	38	8	28	10	3	100	52	61
Sex									
Male	12	40	8	23	13	3	99	55	60
Female	14	41	12	22	7	4	100	59	69
Race									
White	13	42	10	22	10	4	101	45	123
Nonwhite	17	17	17	33	17	0	101	67	6
Vital status									
Living	14	38	11	22	11	3	99	57	120
Deceased	0	67	0	22	0	11	100	67	9
Cancer site									
Colon	8%	50%	17%	25%	0%	0%	100%	67%	12
Rectum	0	50	0	50	0	0	100	50	8
Breast	22	18	11	28	22	0	101	39	18
Cervix uteri	25	25	0	50	0	0	100	25	4
Corpus uteri	0	83	0	0	0	17	100	83	6
Prostate	20	40	0	20	20	0	100	53	15
Bladder	18	36	0	18	18	9	99	54	11
Kidney	0	20	0	20	60	0	100	20	5
Leukemia, lymphoma, Hodgkin's disease	20	50	10	10	0	10	100	70	10
Oral	14	43	7	29	0	7	100	57	14
Larynx	0	20	20	20	20	20	100	40	5
Lung	0	68	22	11	0	0	101	89	9
Melanoma	25	25	38	12	0	0	100	75	8
Miscellaneous	0	50	25	25	0	0	100	75	4
Diagnosis date to patient interview date									
≤12 months	4	59	12	20	2	2	99	76	49
13–24 months	22	38	6	28	0	6	100	56	32
<24 months	17	23	10	21	25	4	100	40	48
Relative characteristics									
Type of relative selected									
Sibling	17	30	11	20	16	6	100	45	64
Child	9	51	9	25	5	2	101	67	65
Selected relative present at interview									
Yes	13%	36%	13%	26%	7%	6%	101%	56%	90
No	15	46	3	18	18	0	100	58	33
Not recorded	0	83	0	0	17	0	100	83	6
Sex of respondent									
Male	15	39	10	25	8	3	100	52	61
Female	13	37	11	23	11	5	100	60	62
Not recorded	0	83	0	0	0	17	100	83	6

Table 9 continued

Patient and relative characteristics	Interval between reported and registry dates of diagnosis: >1 month after DDX	±1 month from DDX	>1–6 months before DDX	≥7 months before DDX	Cannot classify	Same year	Total %	±6 months from DDX	Total N
Age of head of household									
18–39	11	47	15	19	8	0	100	70	53
40–64	15	43	6	23	9	4	100	55	53
≥65	13	17	9	31	17	13	100	35	23
Education of head of household									
Grades 1–11 completed	14	28	3	24	24	7	100	38	29
High school completed	14	54	7	23	0	2	100	68	43
Some college completed	12	37	16	21	10	4	100	60	57
Household income									
<$15,000	11%	17%	20%	28%	17%	6%	99%	46%	35
≥$15,000	14	52	6	19	7	3	101	63	87
Refused, didn't know	14	14	14	43	14	0	99	43	7
Place of residence									
Chicago SMSA	25	28	9	22	9	6	99	47	32
Other SMSA	6	50	6	28	8	3	101	61	36
Non-SMSA	12	41	13	20	12	3	101	61	61

network information, and we were unable to find the relative at home to confirm the address information in only 1% of the cases (data not presented). Approximately 20% of the registry patients did not have any siblings or children or did not have and siblings or children residing in Illinois. Of the 197 addresses where a confirmed relative resided, the patient household provided complete and correct information in 140 (70%) of the cases. The problem responses (30%) included incorrect information (17%) and incomplete information (13%). We were able to locate these relatives in every situation, even when the town name was incorrect.

The quality of information about names and addresses provided by the patient households was examined by various patient and relative characteristics (data not presented). In general, patient households reported more accurate information about the addresses of children (76%) than about the addresses of siblings (66%). The poorest quality of information came from households with patients who had only a high school education (61%) and from patient households where the relative lived in a non-SMSA county (63%).

When the relative household reported the registry cancer patient, the name and adress of the patient was requested. Patient name and address information was not asked in 33% of the relative households, because 20% of the households did not report the patient; 6% of the households reported the patient but gave a date of diagnosis greater than five years prior to the interview; and 7% of the households reported the patient as dead. The relative household provided complete and correct information about the patient's name and address 67% of the time, incomplete information 14% of the time, and incorrect information 14% of the time. Six percent of the relative households refused to provide address information for the patient.

The quality of reporting bore no relationship to patient age and sex, but differences were found related to other variables (data not presented). In general, relative households provided most complete and correct information if it was a child's household, if the relative was present at the interview, if the respondent was a male, or if the relative lived in the Chicago SMSA. The quality of information was relatively poor when the head of the relative household had less than a high school education and when the patient resided in an SMSA other than the Chicago SMSA.

Accuracy of name and address reporting by the patient and relative household pairs. Finally, the success of network sampling depends on the accuracy with which the respondent reports the size of the family network. When persons can be identified by more than one household, the probabilities of being selected into the sample are not equal but depend on the size of each person's network. For the survey estimators to be unbiased, the data provided by the cancer patients must be weighted by the number of households that may identify the patients. It is important, therefore, that the "source" household and/or the household with the desired characteristics be accurate in their reporting of the network size.

Table 10 presents the percentage of agreement between the patient and relative household pairs about the size of the family networks. There was a high degree of agreement between sibling and patient households regarding the total number of siblings 17 years of age or older (89%) and the total number of siblings 17 years of

Table 10
Rates of agreement about network size by patient and relative household pairs and presence of patient and relative at interview

	Number of siblings 17 or older		*Number of siblings 17 or older in Illinois*		*Number of children 17 or older*		*Number of children 17 or older in Illinois*	
	%	N	%	N	%	N	%	N
Type of pair								
Sibling and patient households	89	56	91	56	93	55	93	55
Child and patient households	88	49	94	48	94	52	83	52
Presence at interview								
Patient and relative present	92	53	94	53	91	54	83	54
Patient present, relative not present	81	21	90	20	91	22	91	22
Relative present, patient not present	86	22	91	22	100	22	95	22
Neither present	89	9	89	9	100	9	89	9

age or older who resided in Illinois (91%). The child and patient households reached an even stronger consensus concerning the size of these networks (88% and 94% respectively).

Consensus regarding the number of the patient's children 17 or older reached 93% agreement between sibling and patient households and 94% agreement between child and patient households. When we asked the number of children in Illinois, the proportion of agreement was again 93% for the sibling and patient households, but it declined by 10% for the child and patient households. This decrease reflects disagreement between nine pairs. In seven of the nine pairs, there was overreporting or underreporting by one child. In six of the nine instances, the patient household reported more children than did the child's household.

Whether the patient and/or the selected relative was present at the interview had a small impact on the proportion of agreement. Agreement about siblings was highest when both were present, but agreement about the number of children was highest when only the relative was present. In fact, when only the patient or selected relative was present, there was more agreement on all four questions when the relative was present, rather than the patient. Surprisingly, there was a low proportion of agreement about the number of children in Illinois when both the patient and relative were present.

Summary and conclusions

Concerns about how reliable patient and relative households are when responding to questions about cancer seem to be unfounded. Households with cancer patients tend to report them, although this seems to be affected by the patient's race and, to a lesser degree, by the type of cancer. The relative households selected for this experiment reported patients outside their household with considerable accuracy, if not as completely as the household in which the patient lived. The relative household reporting is influenced by characteristics of both the patient and relative, but the most important seemed to be the relative characteristics. Children reported their parents' cancers more often than siblings reported their siblings' cancers. Given this finding, the extent to which the information about the cancer is managed and censored within the kinship network would be an interesting issue for further research.

Seventy-seven percent of the cancers were reported by both patient and relative households, and 94% were reported by at least one. Patient households quite accurately reported cancer sites and dates of diagnosis. Relative households reported site data with considerable accuracy, but they reported the date of diagnosis with less accuracy. In general, both patient and relative households telescoped the diagnosis date and tended to report it as being earlier than it actually was. This is of mild concern, even though the patient would be contacted in the national survey to confirm site and date of diagnosis. Allowances must be made for deviations, especially as the length of time since diagnosis increases.

A more central question is the amount of information that can be obtained about the network. Very few patient households asked about the names and addresses of relatives refused this information. Moreover, even when we received inaccurate and incomplete information we were able to locate all relative households. When the relative household reported the cancer, the patient name and address data they provided were comparable to the data the patient households provided. Network sampling is thus an effective means of identifying cancer patients in a community sample.

Ascertaining suitable methodological approaches for identifying rare medical populations*

Beth B. Rothschild, National Analysts Division, Booz-Allen & Hamilton, Inc.

Lucy B. Wilson, National Analysts Division, Booz-Allen & Hamilton, Inc.

Introduction

The research project reported here was an investigation into the incidence, prevalence, and costs of the disease entity, multiple sclerosis (MS). In keeping with the other papers in this session, we were confronted with the task of developing an innovative research design suitable to the study of this rare medical event. However rarity was only one of many factors influencing the design choice. It is our contention that scarcity or rarity in and of itself is not the overriding design concern, nor does it necessarily mandate innovative methodological approaches. For example, consider Mercedes-Benz owners. These individuals are relatively scarce in the general population. However, they would be adequately captured with either a straightforward intercept or list survey technique, since they are easily recognized and ownership status can be readily verified.

Thus, medical rarity itself does not necessarily pose significant methodological challenges. Rather, there are certain event-related and data-related needs which emerge, in conjunction with frequency of occurrence, as key factors in choosing among alternative data collection strategies. By event-related factors, we are referring to variables associated with the health or medical condition, whereas data-related factors are those involving the research questions.

This paper will focus on the identification of several of these event- and data-related factors which have an impact on the selection of a research design. We will illustrate these factors with a description of some of the unique event- and data-related problems associated with our study of multiple sclerosis. Finally, we will describe the research approach we selected and the results of the investigation.

Event-related issues. We have classified event-related factors into the following groups: (1) rarity or frequency of occurrence, (2) ease of detection, (3) diagnostic consensus, (4) patient accessibility. Let's address each of these.

In brief, the *actual or presumed frequency of the occurrence* under investigation is a primary event-related element. Many important health conditions are rare events by any objective measure (brain tumors, epilepsy, multiple births, etc.).

In addition, as demonstrated in our Mercedes-Benz example, ease of detection is another important event-related factor. At issue is whether the condition could be detected with an area probability or list sample design. Aspects of the ease of detection issue are:

1. Is the nature of the "beast" apt to produce *undetected* or *undiagnosed* cases? Will the condition be ever so slight that victims do not seek medical help or consultation? How large is this pool expected to be? What impact will these cases have if they go unreported?
2. Are there patients with this condition who will be *uninformed* about their illness; that is, will health care providers refrain from telling patients of their suspected or tentative diagnosis? How large is this pool expected to be? What impact will these cases have if they are disregarded?
3. What is patient acceptance like? Are there informed patients who, when asked about their condition, will *deny or reject* the fact that they have the disease or that they are part of the "rare" population? Is this because they disregard the diagnosis or because of the social stigma or trauma which is presumed to accompany the illness? How large is this pool expected to be?
4. Lastly, how frequently, if at all, will individuals be *misclassified*? Will false positives and/or false negatives distort the data collected?

Another event-related issue is *diagnostic consensus.* For many medical conditions, such as cancer or pregnancy, specific clinical tests will confirm or establish the diagnosis. Other conditions, however, present significant challenges because no specific diagnostic tests actually confirm the disease's existence; multiple sclerosis and psychological disorders are examples.

Patient accessibility is another area of concern, particularly if survival rates associated with the condition are low or if significant limitations or impairments result form the condition. Both physical and emotional disabilities may impede one's ability to identify and subsequently interview the rare population in question.

Data-related issues. While the event-related elements must occupy a prime position in deciding on a suitable methodological approach, the study's objectives and the information desired also play important roles. Are population *projection* and *enumeration* more important than identifying *intergroup differences* or trends? Is the nature

* The research reported here was supported under contract #NIH-N07-NS-4-2335 from the National Institute of Neurological and Communicable Diseases and Stroke.

of the information desired—nose counting, behavioral, psychological—best captured *prospectively* or *retrospectively*? What, if any, time and cost constraints must be taken into account?

Once the event-related issues and data-related needs are fully outlined and understood, the feasibility of alternative methodological approaches can be assessed.

Selecting a methodology to study MS

To illustrate the importance of these factors and different, creative ways of addressing them, we turn to the National Study of the Incidence, Prevalence, and Costs of Multiple Sclerosis. This study was conducted by the National Analysts Division of Booz-Allen & Hamilton under the sponsorship of the Office of Biometry and Field Studies, National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) of the National Institutes of Health.

Let us begin by describing the nature of this disease. A working medical definition of multiple sclerosis is that it is a demyelinating disease of the central nervous system characterized by periods of exacerbations and remissions. In lay terms, the myelin or protective fatty coating of the nerve sheath breaks down. This breakdown or lesion may cause a variety of symptoms (numbness, loss of coordination, vision problems, and the like), depending on where in the nervous system demyelination occurs. Periods in which the symptoms occur are called exacerbations. As the myelin is rebuilt or scars over, the symptoms will partially or completely disappear and the patient enters a symptom-free period called remission.

The rarity of MS. At the outset of our research, no precise national estimates of the frequency of MS had been compiled except those from site-specific or geographically restricted U.S.-based studies. It was estimated that prevalence rates per 100,000 population varied from 10 in New Orleans to 92 in Rochester, Minnesota. Annual incidence rates per 100,000 population ranged from .4 in New Orleans to 3.6 in Rochester, Minnesota. (Percy et al., 1971; Westlund and Kurland, 1953). Regardless of these geographic differences, the consensus was that multiple sclerosis is a relatively rare disease among Americans. In fact, an average of the New Orleans and Rochester prevalence estimates suggested that about 50 persons per 100,000 or 1 in 2,000 Americans would have MS. Thus, the first criterion—scarcity or rarity—was established.

The difficulty of detecting MS. One of the most challenging dilemmas confronting us was the disease and patient detection issue. MS is a disease which cannot be established with certainty on the basis of specific clinical tests. A diagnosis is often made through exclusion of other diseases as well as through extended observation of the clinical course exhibited by the patient. Furthermore, sometimes the symptoms are so mild that they go virtually unnoticed or unattended by the patient or physician. Thus, some cases remain undetected or undiagnosed because medical attention is never solicited or, if sought, the symptoms never present themselves to the physician.

Uninformed MS cases. Complicating this already difficult detection problem is the fact that physicians may withhold a tentative MS diagnosis from patients and/or family members until sufficient time has passed—often several months or years—to confirm their original hypothesis. Even after the diagnosis is confirmed, physicians may choose to withhold this information based on their knowledge of their patients. Thus, some cases remain uninformed about their MS for several years.

Patient denial of MS. No known pattern of remission and exacerbations occur with MS. Hence patients are uncertain as to when and in what form the next episode will occur. Such unpredictability, particularly as it might affect one's job, education, or family plans, can take a severe toll on patients. Thus, even informed patients may deny their MS diagnosis for fear their employment will be jeopardized or their family will reject them.

Misclassification of MS cases. As if these detection problems were not enough, this research was further complicated by the similarity between MS and other central nervous system disorders. "The great imitator," as MS has been called, can be mistakenly diagnosed as one of several other demyelinating diseases. Thus, patients may be misclassified as either false positives or false negatives. It has been verified on autopsy that as many as 25% of all patients may go undiagnosed or be incorrectly classified.

Diagnostic accuracy. Needless to say, diagnostic consensus is not readily established for MS. Not all physicians are equally able by training to make distinctions between MS and other similar disorders. It is generally accepted that neurologists and physicians in related specialties are the most accurate diagnosticians. However, patients may not always be diagnosed by such physicians. Thus, a question may arise as to the accuracy of an MS diagnosis.

Accessibility of MS patients. If MS cases are properly identified, the one area in which relatively few problems arise for the researcher is patient accessibility. Most cases of the disease begin in mid-life and peak between 30 and 35 years of age. The average duration of the illness has been estimated at just over 20 years (Leibowitz and Alter, 1973). Therefore, survival rates are good and the opportunity to make patient contact is not limited by duration of the disease. However in the advanced stages of MS, patients may be institutionalized and their communications skills severely impaired.

Requirements for national projectability. The ultimate

goals of this research were to ascertain the national prevalence and incidence of this disease. In addition, our mandate was to estimate the toll which this disease inflicts on patients and their families in terms of the direct and indirect costs, measured in dollars.

Because of the patterned periodicity of MS and its infrequent but sizable expense a prospective diary approach was considered.

Faced with these issues, we determined a household—area probability or networking—sample would not do. Therefore, a survey approach was designed that would facilitate efficient and inclusive case finding, although at the outset it was recognized that problems with undiagnosed and misclassified cases might not be entirely resolved.

The national MS study: Research design

A double sample approach using secondary source data—health care providers' records and not households as the sample frame—was adopted. Use of association membership lists was ruled out since it was determined that many members were friends or relatives of MS patients, rather than those with the disease. More importantly, unaware or denial MS cases would be noticeably absent from these lists.

First, a probability sample of approximately 9,500 health care providers was drawn, and contacts were made to identify MS patients seen by these providers between 1970 and 1975, inclusive. Since MS patients do not ordinarily receive routine care for their MS, a time frame of sufficient length was required to allow patients time to come in contact with the medical community and therefore be reported by surveyed health care providers during the case-finding effort. The second step was to draw, from the list of all MS patients seen, a probability sample of names and to trace individual patients for interview.

Based on the results from a sizable pilot study, those types of physicians and hospitals known to have contact with MS patients were sampled. Those providers recognized as being more skilled in diagnosing MS were sampled with certainty, that is, a census was taken, while a noncertainty sample of other providers was drawn. Neurologists, neurosurgeons and hospitals with approved neurology residency programs, including military and VA hospitals, comprised the certainty stratum. Internists, general practitioners, family practitioners, ophthalmologists, physiatrists, psychiatrists, and all other short-term general hospitals were included in the noncertainty stratum.

In order to preserve the confidential nature of the provider-patient relationship, those who had seen MS cases were asked to supply minimal identification data: first name, last initial, date of birth, sex, and race. Moreover, physicians were supplied with a uniform set of six diagnostic codes ranging from probable to possible MS to minimize variability across diagnoses. Hospital personnel were asked to use the MS codes from the 8th Revision of the International Classification of Diseases, adapted. All were encouraged in writing and verbally to report both definitive and suspected cases as well as informed and uninformed cases. (Among physicians—after adjustment for death, retirement, etc.—there was a 99% response rate about whether they had treated MS patients during the survey period. Approximately 70% of these physicians provided data on the number of cases seen. Comparable data from hospitals ws 93% and 89%).

Using the health care provider information, multiple mentions of the same person from different sources were matched and an unduplicated count of MS patients seen by the sample of providers during the survey period was determined as well as an estimate of duplication in the universe of health care providers. This became the basis from which estimates of the prevalence and incidence were computed. In addition, the "unduplicated" listing of MS patients identified in the sample became the sampling frame for the second stage sample of patients.

The patient sample was a list probability sample of reported persons residing throughout the conterminous U.S. reported to have MS. The data collection effort represented three discrete tasks: initial in-person and closing telephone interviews, separated in time by a 90-day diary effort for recording MS-related costs. These data contributed to the analysis of the costs of the disease and provided insights into the medical, social, and psychological environment of MS victims.

A total of 1,240 initial patient interviews were completed (for a response rate of 76%). Of these, 14% were with MS patients who had not been explicitly informed of or did not admit to their diagnosis. (In these situations, an adapted questionnaire was used which did not explicitly refer to MS.) In cases where the patient was deceased or physically incapacitated, a next of kin or near relative served as a proxy respondent.

The MS study design had two additional elements, both of which were aimed at minimizing potential errors in the disease estimates. First, it was recognized that the potential existed for a health care provider to save work by denying the care of MS patients. To verify "nontreater status" it was decided that a subsample of self-reported nontreaters would be recontacted. It should be noted that, despite complete confirmation of the initial status, this procedure was discontinued after several follow-ups, as physicians considered it a challenge to their integrity. We decided abandoning these efforts would be in the best interests of the study.

Second, to address at least partially the issue of misclassification, a subsample of patients were to be asked to take a medical examination conducted by a Board-certified neurological specialist at no charge. The objective was to arrive at an estimate of the number of false positives. However, these examinations were not performed

because of the infeasibility of conducting them concurrent with this research.

Based on the results of this survey, we concluded that the number of MS cases on prevalence day (January 1, 1976) was 123,000 cases in the conterminous United Sates. That is a rate of 58 patients per 100,000 population. The annual incidence figures for the 1970–1975 period were 8,800 cases for a rate of 4.2 per 100,000. The annual direct and indirect costs of the disease were estimated to be $800 million, which translates into an average of $1,672 direct medical and $4,855 indirect costs per patient.

Thus, the MS research design and data collection strategy took account of as many of the unique event- and data-related aspects of the disease as feasible. Some limitations are still present, however. First, the disease estimates may be underreported if a sizable group of undiagnosed cases exists, since the design did not account for them. Furthermore, if the survey's six-year interval was not large enough to capture all MS persons, the disease prevalence may be understated. It should be noted that estimates of the frequency with which MS patients seek medical attention from the types of health care providers sampled, as derived from actual patient-reported data, reveal an interval of one to two years, suggesting that noncontact during a six-year reporting period would be negligible, if at all. Autopsies of thousands of individuals would reveal the size of the undiagnosed pool. Such procedures, however, were outside the scope of this investigation.

A second limitation is misclassification. Our original intent was to ascertain the number of false positives, as noted earlier. False negatives, however, would not be determined, as this would have required hundreds of examinations of patients diagnosed as having similar neurological disorders. It is important to recognize, however, that although patients with a false positive diagnosis do not have multiple sclerosis, they are currently labeled by the medical community as MS patients; they are treated as if they have the disease and may adopt a life-style reflecting the diagnosis. Therefore, we believe that they should not be discounted or ignored.

Conclusion

In conclusion, the key factors of ease of detection and diagnostic consensus, coupled with the need for precise national data, argued for a secondary source identification approach to determine the incidence and prevalence of MS. The need for up-to-date economic cost data necessitated follow-up of a subsample of MS patients reported. Thus, household-based or networking approaches were dismissed as inappropriate.

The approach we took to the design of this national study of multiple sclerosis placed medical rarity in the context of several other important event- and data-related issues. By explicitly recognizing problems of detection, patient denial, diagnostic accuracy, etc., we were able to create an approach which both captured the rare medical event and enhanced the reliability and validity of the results. Moreover, we were able, *from the onset* to identify the limitations of the design and its impact on the survey results. We encourage other researchers who choose to investigate rare medical or health phenomena to put the scarcity problem in its proper prospective.

Pilot study for a national survey of epilepsy*

F.A. Bryan, Jr., Research Triangle Institute

J.T. Lesser, Research Triangle Institute

M.F. Weeks, Research Triangle Institute

N.N. Woodbury, Research Triangle Institute

Introduction

The Design and Pilot Study of a Methodology for a National Survey of Epilepsy was begun in 1978 and culminated with data collection and analysis in 1981. The objectives of the Pilot Study were to:

1. formulate the Pilot Study Protocol;
2. design and produce field administrative forms;
3. design and produce sampling forms, data abstraction forms, interview schedules, consent forms;
4. select and visit study areas to recruit consultants in local schools of pharmacy;
5. seek endorsement and cooperation from the various professional and voluntary associations in the study areas;
6. seek endorsement and cooperation from national professional and voluntary associations;
7. select a sample of pharmacies in each study area and recruit selected pharmacies into the study;
8. design and implement techniques for sampling and abstraction of prescription data;
9. design and implement survey procedures for physician data collection;
10. design and implement methodology for obtaining multiplicity data from patients;
11. perform evaluation of the results of the study to measure pharmacy, physician, and patient cooperation levels; and
12. provide recommendations on the feasibility of a National Survey of Epilepsy.

Epilepsy is not a disease; it is a characterization of a type of seizure disorder of various etiologies. Nevertheless, persons classified as epileptics have suffered from varying levels of social stigma in our civilization and have experienced the economic burdens of restrictions on employment opportunities and direct costs of therapy. For these reasons, epilepsy has been largely a hidden condition in our society, and any attempt to survey the epileptic population must necessarily deal with problems of identification.

According to recent reports, between two and four million persons in the United States suffer from some form of epilepsy (Office of Scientific and Health Reports, 1976; Commission for the Control of Epilepsy and Its Consequences, 1978). The Commission for the Control of Epilepsy has estimated that about 100,000 people develop epilepsy each year. Thus, the incidence and prevalence of epileptic disorders are relatively low, and difficulties encountered in screening for a hidden population are exacerbated by the rarity of the disorder.

Survey design

A key consideration in the development of any survey design is to identify unique characteristics of the target population which will allow efficient identification. In the case of epileptics, a unique characteristic is that almost all attempted seizure control is through use of medication. Further, the number of drugs used for sei-

Table 1
Case-finding drugs

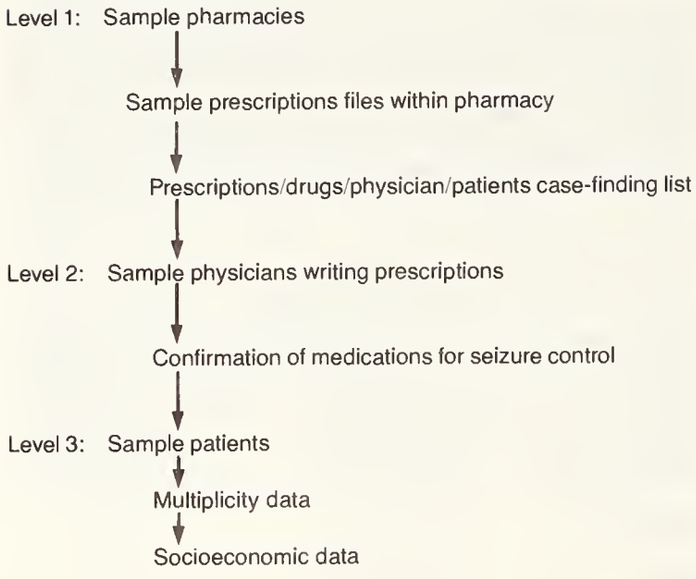
Brand name	*Generic name*	*Manufacturer*
Celontin	Methsuximide	Parke-Davis
Clonopin	Clonazepam	Roche
Depakene	Valporic Acid	Abbott
Diamox	Acetazolamide	Lederle
Dilantin	Phenytoin Sodium	Parke-Davis
Gemonil	Metharbital	Abbott
Mebaral	Mephobarbital	Winthrop
Mesantoin	Mephenytoin	Sandoz
Milontin	Phensuximide	Parke-Davis
Mysoline	Primidone	Ayerst
Paradione	Paramethadione	Abbott
	Phenobarbital	various
Peganone	Ethotoin	Abbott
Phenurone	Phenacemide	Abbott
Tegretol	Carbamazepine	Geigy
Tridione	Trimethadione	Abbott
Zarontin	Ethosuximide	Parke-Davis

* This work was performed under contract number NO1-NS-8-2383 with the National Institute of Neurological Communicative Disorders and Stroke of the U.S. Department of Health and Human Services.

zure control is small. The only difficulty with this list of drugs is that some are frequently used for treatment of problems other than epileptic seizures. Therefore, if these medications were employed for case finding, it would be necessary to verify diagnosis through some other mechanism. A list of case-finding drugs is given in Table 1.

A survey was designed which involved screening prescription files in pharmacies to identify potential seizure victims. Data abstracted from the prescription files was used at a second-level survey of physicians to verify the patient as an epileptic. Finally, confirmed epileptics were to be interviewed (in a national study) to obtain socioeconomic data. Figure 1 provides a diagram of the survey design.

Figure 1
Survey design



A difficulty involved with a survey design of this type is the compounding of nonresponse through the various levels of the survey. Therefore, a principal component of the design effort lay in development of methodology to cope with potential nonresponse and in determination of techniques that could be used to provide estimates of nonresponse bias.

One of the principal concerns in surveying for data on the epileptic population lay in areas of privacy and confidentiality of patient records. In most states, a prescription belongs to three persons: the patient, the physician writing the prescription, and the pharmacy. In order to obtain a release of these data, it was necessary to assure the anonymity of the physician and the patient until each, in sequence, agreed to cooperate with the study. In order to assure this, it was necessary that persons involved as respondents at each level of the survey assist in contacting the appropriate individual at the subsequent level to obtain cooperation. However, relying on respondents to put forth the effort necessary to persuade other individuals to participate in a study of this type, without additional mechanisms for nonresponse conversion, was sure to guarantee failure of the overall effort. Therefore, at each level, the field staff had to stand ready to work as a representative of the current level respondent to recruit respondents at the next level.

In order to provide acceptance of the effort by the professional community, endorsements by the various concerned professional organizations were sought. Also, in order to promote participation of local pharmacies, consultants from the faculties of local schools of pharmacy were recruited to introduce the study to sample pharmacies in their respective areas and to convince these pharmacies of the social value of the study, as well as the professional nature of the work and its ethical character. These local pharmacy faculty members were also to lend assistance to pharmacists in second-level recruitment of physicians to the program. Professional supervisory personnel of the survey staff employed in the project, as well as physician consultants to the project, were used as backup to the pharmacy consultants in obtaining participation of physicians in the program.

Once the physicians were recruited in the program and the epileptic status of the sample patients ascertained, the physicians were asked to request participation of individual epileptic patients. At this point, either staff in the physician's office were to make initial contact or the field data collectors working on behalf of the physician could make the contact directly. In the event that the patient had some question concerning the study which needed response from the survey staff, the patient was provided a toll-free number at which a senior member of the study team could be called. If the patient refused cooperation after contact by the physician's office, no further attempt was made to include the individual in the survey.

During the Pilot Study, only multiplicity data (those necessary to estimate the number of opportunities an individual had to enter the study) were to be collected from patients. These data included the number of prescriptions for case-finding medications filled for the patient during a defined time period and the number of stores or other drug outlets delivering medications in response to these prescriptions.

The Pilot Study interview protocol required personal interviews with pharmacists and physicians and telephone interviews with patients (a mail interview with patients was used as a fall-back if telephone contact was not possible).

Data collection

The Pilot Study data collection was implemented in six judgmentally chosen sites throughout the country. These sites included pharmacies within a a 50-mile radius of schools of pharmacies in the areas. The 50-mile radius was chosen to limit travel and logistical problems

that might otherwise occur. The survey sites were Stockton, California; Storrs, Connecticut; Iowa City, Iowa; New Orleans, Louisiana; Brooklyn, New York; and Chapel Hill, North Carolina.

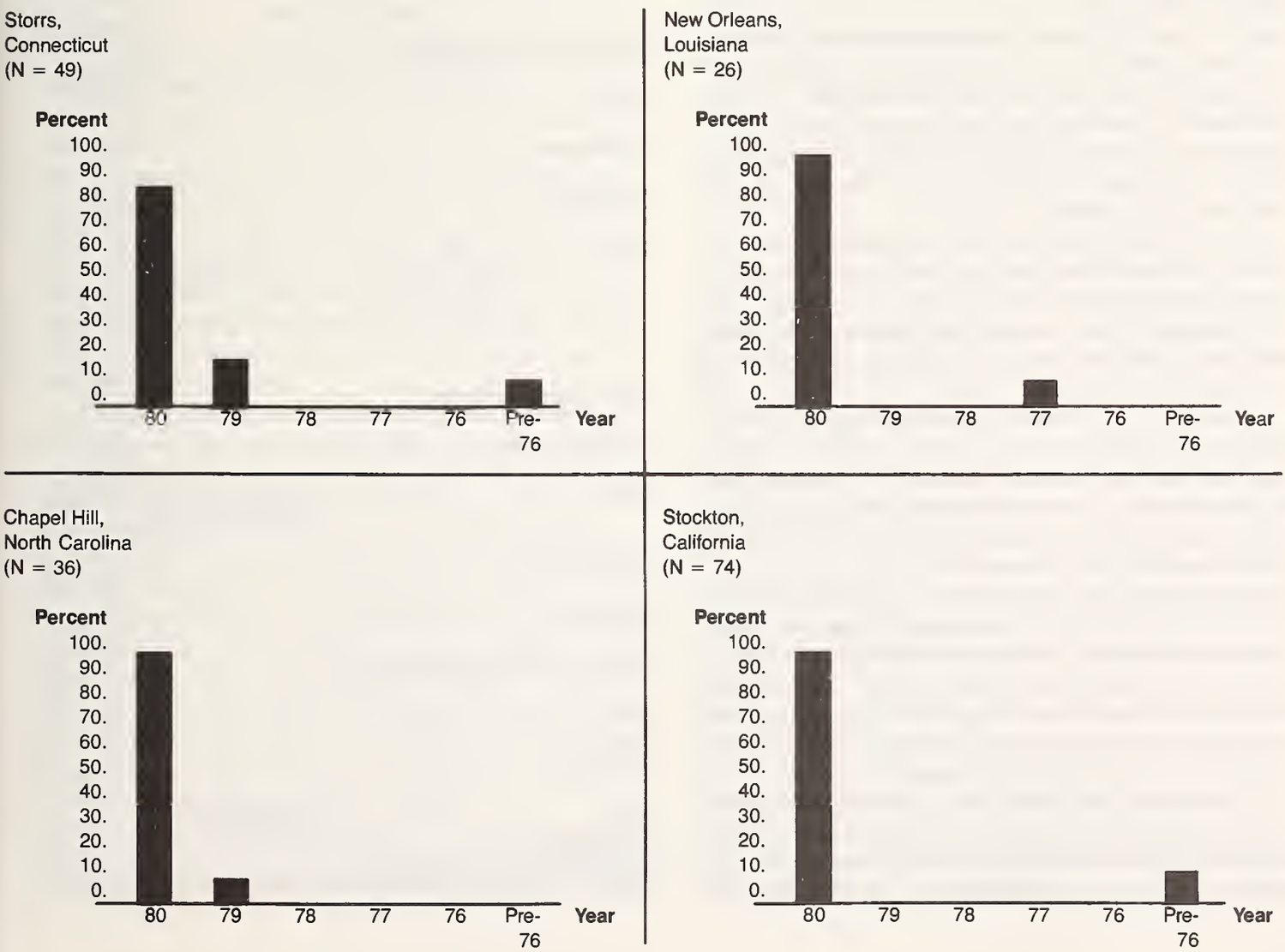
The first stage of the survey consisted of eight pharmacies randomly chosen in each of the six sites. Within each pharmacy, the prescription files were randomly sampled to provide 2,000 prescriptions per store; 880 of these were from the most recent calendar year and 280 from each of the previous four years. The aim of the sampling procedures was to produce data that would represent all prescriptions filled or refilled in 1980, the most recent calendar year. The sampling technique, however, gathered data on refills only as far back as 1976 (a prescription that was written in a previous year had to have been refilled in the current year in order to be admitted to the study). This technique could result in underrepresentation if, in fact, pharmacies were filling a substantial number of prescriptions that were written prior to the five-year study period.

Upon institution of data collection, two of the survey sites were found to be in states (New York and Iowa) that restrict prescription refills to a one-year period. In these cases, 880 prescriptions were drawn from both the current year and the previous year. Refill logs were checked to ensure that the one-year restriction on refills was followed.

In order to estimate completeness of the frame in sites that had no restriction on the refill period, refill logs were examined in all selected pharmacies that maintained these items for selected time periods in the current year. (The time periods were of two weeks' duration. For each pharmacy, four periods were systematically selected with a random start to provide one sampling period in each calendar quarter.) When refill logs were found to contain a prescription predating the prescription file sampling period, the prescription in question was pulled to see if it was written for one of the case-finding drugs. It was determined that the five-year sampling frame was approximately 98% complete in the sites permitting unrestricted refill periods. Results for these sites are shown in Figure 2.

The pharmacy-level data collection was completed in all six sites; however, due to logistical problems, one site

Figure 2
Percent completeness by year



was dropped prior to the physician survey. Table 2 shows the expected and actual yields that resulted from the survey work. Table 3 shows participation rates at each level of the survey.

Table 2
Yields

	Expected	*@5/6**	*Actual*
Persons	336	NA	303
MDs	210	175	168 (63 out-of-scope ◊ 105)
MD interviews	147	122	71 (62 completed interviews)
SDVs**	71	58	57 (44 chronic epileptics)
MD-approved patient contacts	49	41	33

*Six sites were included in the survey design; one of these sites was dropped for logistical reasons prior to the physician level data collection.
**Seizure disorder victims.

Problem areas

The field data collectors used in this survey were pharmacy students from local schools of pharmacy recruited especially for the program by pharmacy administration consultants. These students were trained in interviewing pharmacists, prescription abstraction and reporting, assistance in recruitment of physicians, interviewing physicians, and assistance in recruitment of and interviewing of patients.

The pharmacy administration consultants had the primary responsibility for recruitment of pharmacists to the study and assistance in conversion of nonparticipating physicians. The universe of pharmacies included both chain and community operations, in urban and rural areas.

A number of problem areas developed over the course of the program having to do with the use of pharmacy students as data collectors in primary data acquisition and in the employment of pharmacy administration faculty for recruitment of pharmacists and physician nonrespondent conversion.

The primary difficulty with the students lay in conflicts with their class schedules due in large part to the considerable amount of time required for data collection. Also, the lack of experience in data collection activities exacted a toll from the project. The students did extremely well in abstracting pharmacy data, but much less well in other data collection activities.

The pharmacy administration consultants provided valuable assistance in enlisting the endorsement of local professional societies and in providing access to some of the pharmacies, especially those stores closely allied with the local university. or concerned for the professional status of pharmacists. The recruitment results of the pharmacy administration consultants were much better in community stores than in chain operations.

Use of the pharmacy adminstration consultants in conversion of physician nonrespondents was only marginally effective. This may have been due to a lack of rapport with the medical community in their areas or a lack of the persistence which is so necessary to achieve adequate response levels in any surveyed population.

Participation of chain store operations was most efficiently addressed by contacts with their central offices. While this, at times, could be accomplished by the local consultant, the office was frequently remote from the consultant's location, and the more effective approach was contact by the core staff. This effort was aided to a considerable degree by Dr. J. P. Gagnon, Professor of Pharmacy Administration at the University of North Carolina. Dr. Gagnon was a principal consultant to the project and acted as the primary contact between the program and the pharmacy community.

One unanticipated problem was encountered in a chain operation; namely, that of working in stores with unionized pharmacists. It was impossible to deal with

Table 3
Participation rates

Pharmacy Participation Rates	*Total*	*CT*	*NY*	*NY**	*NC*	*IA*	*LA*	*CA*
Original	27	5	2	7	5	6	5	4
1st conversion	4	—	1	—	2	—	1	—
2nd conversion	8	1	1	3	—	2	1	3
Total	39	6	4	10	7	8	7	7
	39/48 = 81.3%**			10/12 = 83.3%				

Patient Participation Rates		
Patients identified	71	
Eligible	57	(44 chronic epileptics—62% of identified patients)
Contact allowed	33	
Interviewed	20	
Physician Participation Rates (5 sites)		
Original	58	
Conversion	13†	
Total	71	(62 interviewed)

*Speciel substudy.
**Excludes results from the NY special substudy.
†None tried in NY. Averege perticipetion rete in Celifornie, Iowa, Louisiene, end North Cerolina: 73%

this during the Pilot Study. A lot of time, well beyond that available, would have been required to negotiate with union officials to permit data collection activity.

Another special condition was encountered in California. In that state, by regulation, no one except an employee of the pharmacy may have access to the pharmacy files. In California, therefore, it was necessary to persuade pharmacies to actually employ the data collectors for this effort.

Acquisition of store participation was the single greatest problem encountered in the program. An incentive of $50 in professional books (or cash if the store requested) was offered to participants in the study. As shown in Table 3 this resulted in a total of 27 out of the 48 selected pharmacies participating in the study, a participation rate of approximately 56%. With additional work on the part of the pharmacy admnistration consultants, this number was increased to 31, a participation rate of about 65%. This was totally unacceptable, and a special effort was launched to increase the participation of the outstanding stores. To do this, a senior staff member of the research team visited each one of the sites and each of the nonparticipating stores in the sites. During these visits, an additional incentive of $200 to $300 (depending on store size) was offered for participation. This resulted in an additional eight stores agreeing to participate, with a total participation rate of approximately 81%.

. A special situation was encountered in New York (see Table 3). There, the participation rate in the original sample of eight stores was only 50% after all attempts at conversion. In this site, a special substudy was launched in a new sample of 12 stores. The pharmacy administration consultant in this substudy sent a letter of endorsement to the sample pharmacies. Then direct contacts with the pharmacies for recruitment were handled by an experienced field supervisor. Using this approach a participation rate of approximately 83% was obtained with the New York pharmacies, a substantial improvement over the original effort.

Physician participation rates were also a problem. At the physician interview stage, it was recognized that use of the pharmacy students and pharmacy consultants in nonresponse conversion was producing marginal results. Therefore, field supervisors were brought in to assist in this role. Because of time constraints, no physician conversions were attempted in New York. If one includes only the four remaining sites (New Orleans had been dropped at the end of the pharmacy data collection), the total number of physicians eligible for interview was 83. Forty-eight physicians agreed to participate on initial contact and conversion yielded an additional 13 physicians in these four sites. Therefore, a total of 61 agreed to participate, or approximately 73%.

Seventy-one patients were identified during the prescription screening in pharmacies as having obtained case-finding medications. Fifty-seven of these were confirmed to be taking the medications for control of seizure disorders. Of these, 44 were confirmed as chronic epileptics. Physicians allowed contact with 33 of the eligible patients (a contact rate of 75%). Because of the time constraints, only 21 contacts were attempted with patients to recruit them to the study. Of the 21 contacts, 20 resulted in interviews and 1 in refusal.

Discussion of results

From a statistical standpoint the principal concern of the results of this type of survey is production of estimates of disorder prevalence and of socioeconomic impacts. There is no reason to assume *a priori* that participation status of a pharmacy would materially affect estimates produced from survey data of physicians and patients. However, nonresponse at either the physician or the patient level would have to be addressed. The most direct way of doing this would be to embed a number of known epileptics in the sample of patients sent to physicians for review. This would require acquisition of an exogenous frame of such individuals and permission from the individuals involved to use their prescriptions in a seeded sample to be sent to physicians. Physicians would then be contacted and asked to report on the epileptic status of all patients in the sample and to provide permission to contact the patient. The results of physician nonresponse for the known epileptics could then be used to estimate the bias associated with such nonresponse.

Other areas of concern noted above address the use of pharmacy students as interviewers as well as in-pharmacy data collectors. It is apparent from consideration of the Pilot Study experience that such students should be used for abstraction of data from pharmacy records; however, this should be their sole task.

Use of pharmacy administration personnel in local universities for introduction of the study is probably worthwhile from the standpoint of acquiring endorsement of local pharmacy boards and professional organizations. However, the absolute value of such endorsements is not clear.

In one of the survey sites in the Pilot Study, the state pharmaceutical association was positively antagonistic toward any federally supported work. Not only did this association refuse to endorse the study, but they actually advertised against it in their newsletter. In this particular site, the pharmacy participation was 75%, somewhat less than the average for the overall study. But, it is not clear whether the condemnation of the professional association was responsible for the poorer-than-normal performance or whether this was merely due to the political climate. Nevertheless, it appears that even the antagonistic attitude of a professional society would not necessarily harm the program.

Other results of the Pilot Study indicate that the program can be improved by streamlining it and by using individuals who are well versed in their activities. The

results also indicate that professional senior field staff should be used in recruitment of pharmacies and physicians. In addition, any activity which can minimize the load on the pharmacies and the physicians would amplify the participation in the program. In particular, it is recommended that such studies should, if at all possible, employ patient names in the pharmacies for follow-up telephone interviews with the physician in order to confirm the patient as a seizure disorder victim. Furthermore, it appeared from excuses provided for nonparticipation that physician compliance with the study would be enhanced by reducing the amount of information (and thus time commitment) asked of these individuals to an absolute minimum, that is, to the amount required to confirm the diagnosis of epilepsy and to obtain permission to contact the patient.

Since most physicians have only a limited number of patients in such a study (three to five), physician data collection should probably be done by telephone following an introductory lead letter. This would reduce the time required for physician interview (which would enhance participation) and the cost of physician data collection.

Contacts with the patient for multiplicity data only should be done directly from the pharmacy on behalf of the physician if the physician approves; otherwise, it should be from the physician's office. Contacts for personal interview should follow a comparable format for arranging an appointment. Personal interviews would be done on a face-to-face basis with patients.

Table 4 shows sample specifications for a national study of epilepsy based on production of estimates for four regions. Only two years are proposed for screening of prescription files. This is because, in some states (two in the Pilot Study), refills of prescriptions are not permitted for prescriptions more than one year old, and in states where longer refill periods are permitted, the Pilot Study indicated that the two-year time period would provide a frame that was at least 94% complete: (Measures of completeness could, of course, be taken during a national study to determine required adjustment to study results.) Detection rates for patients are estimated on the basis of one unique individual per in-scope prescription in the sample of the prescription files. This was essentially the experience in the Pilot Study and can be expected to be the case in a national investigation.

The other estimates of participation rates indicated in Table 4 are based on the best judgment of the Pilot Study team predicated on suggested modifications to the protocol for implementation of a national survey.

Table 4
Sample specifications—national study of epilepsy

Pharmacy data collection			Total
1. Regions			4
2. Sites (selected using 3-digit ZIP codes)			36
3. Sample stores			432
4. Expected cooperating stores (83%)			360
5. Scripts screened—study year			324,000
6. Scripts screened—previous year			324,000
7. Total scripts screened			648,000
8. Expected in-scope drugs			2,444
	Telephone interview substudy	*Personal interview substudy*	
Physician data collection			
1. Physicians	713	631	1,344
2. Cooperating physicians (80%, 75%)*	570	473	1,043
3. Patients	1,036	860	1,896
4. Eligible patients (62%)	642	533	1,175
Patient data collection			
1. Cooperating patients (80%,75%)*	514	400	914

*Estimated participation rates for the telephone and the personal interview substudies, respectively.

Conducting surveys with mentally retarded youth*

Susan A. Stephens, Mathematica Policy Research, Inc.

Introduction

Usual survey practice is to obtain data on activities, experiences, attitudes, and feelings directly from the individual of interest rather than from some other source. In some instances, records are used as a source for specific administrative data such as participation in transfer programs. But for the broad range of data, self-report is generally regarded as the best single source.

Data completeness and data quality appear to be problematic with certain populations, however, making self-reporting a less well-accepted strategy; the mentally retarded are one such group. The ability to interact with strangers, to respond verbally and to articulate sufficiently to be understood, to conceptualize personal experiences within specific response categories, and to provide complete and accurate answers are issues to be addressed in designing any self-report survey; they are especially critical with the mentally retarded.

The methodological analysis reported here was undertaken as part of an evaluation of a national demonstration program for the mentally retarded. The basic question addressed in the analysis was, Can the mentally retarded be successfully interviewed as part of a research effort? This paper will briefly describe the demonstration program and the research objectives, review past data collection strategies used with the mentally retarded, discuss design issues faced in the development of self-report instruments for this population, and present preliminary findings from the pilot study associated with the evaluation project.

The SW/STETS evaluation and pilot study

Evaluation design. The Supported Work/Structured Training and Employment Transitional Services (SW/STETS) demonstration program is intended to provide structured training and employment services to young adults who have been classified as "mentally retarded." SW/STETS programs provide: (1) work-readiness assessment and training, (2) on-the-job training in the private sector leading to regular employment, and (3) support services to help participants in the transition period once they enter unsubsidized employment. The SW/STETS program is intended to provide these training and employment preparation services to young adults between the ages of 18 and 24 with IQ scores in the 40 to 80 range, concentrating on the moderately to mildly retarded.

The SW/STETS demonstration program is being conducted by the Manpower Demonstration Research Corporation (MDRC). The program is scheduled to operate in five cities: New York, Cincinnati, Tucson, Minneapolis-St. Paul, and Los Angeles.

Mathematica Policy Research, Inc. (MPR), has been contracted by MDRC to conduct two major aspects of the evaluation of the program—an impact analysis to examine the program's long-term effects on participants and a benefit-cost analysis.

The SW/STETS research will address six basic questions:

1. Does SW/STETS improve the labor-market performance of participants (compared to what it would be in the absence of SW/STETS)?
2. Does SW/STETS affect participants' use of other education and training programs?
3. Does SW/STETS reduce participants' use of government transfer programs?
4. Does SW/STETS help participants lead life-styles that are similar to the population norm?
5. How do the characteristics and experiences of participants or of the SW/STETS program influence the effectiveness of SW/STETS?
6. How do the effects of SW/STETS compare with its costs, and how do they compare with the costs and effects of alternative programs?

To answer these questions the research will study two groups of mentally retarded young adults: participants—persons who apply to and are enrolled in the local SW/STETS programs, receiving their services—and controls—applicants who are not enrolled in the programs. Selection to the groups is based on a random assignment procedure. Data on sample members, both participants and controls, will be collected in the following major life areas at intervals over a period of several years:

1. Labor Market Performance
   - Employment
   - Earnings
   - Labor-force participation
   - Job search
   - Job tenure
2. Education and Training Participation
   - School

* This revised paper draws upon a study conducted by MPR for the Manpower Demonstration Research Corporation, described more fully in Bloomenthal et al. (1982). I wish to thank the other authors of that paper, especially Stuart Kerachsky, for their assistance in preparing this manuscript.

Work-experience program
Vocational training
Living-skills classes
Work/study programs

3. Transfer-Program Use
Supplemental Social Security (SSI)
Old Age, Survivors, and Disability Insurance (OASDI)
Medicare/Medicaid
Other Welfare (General Assistance and Aid to Families with Dependent Children)

4. Social and Living Skills
Benefactor relationships
Living arrangements
Family formation
Antisocial behavior
Transportation skills

5. Explanatory Variables
Demographics
IQ
Socio-economic background
Social-support network
Prebaseline employment
Prebaseline school and training

Typically, in a project of this type, the majority of the data for the analysis would best be obtained directly from the sample member. Some data potentially could be obtained from other sources. However, use of "proxies" (reports by "significant others") or records data raise problems of inconsistencies in the availability or quality of data across sample members. For example, program records may be available only for a subset of the sample. Furthermore, the availability and quality of other sources of data such as parents, employers, or counselors may vary as a results of some critical factor, such as program participation or level of independent functioning.

The types of data to be collected and whether respondents would be capable of self-reporting the data were important considerations in designing the data-collection approach in this study. These considerations necessitated a thorough review of previous work with the mentally retarded.

The mentally retarded as a research population

Definition of mental retardation. Mental retardation refers to significantly below average general intellectual functioning combined with impaired adaptive behavior. A mentally retarded person scores at 80 or below on an IQ test, matures and learns more slowly than normal persons, and has some difficulty learning adaptive behavior such as social, vocational, and everyday living skills. Mental retardation occurs before age 22, a criterion which distinguishes it from disabling conditions of later life. About 3% of the population, or more than 6½ million children and adults, are mentally retarded. Most of these (over 90%) are classified as moderately, mildly, or borderline retarded.

Past research efforts with the mentally retarded. Past research on interviewing mentally retarded persons has been limited in volume and varied in success. It indicates that careful consideration must be given to the sample members' expected levels of functioning and communication skills before assuming self-report data collection will prove successful. In addition to their mental deficiencies, mentally retarded youth have: (1) below average emotional and intellectual maturity and relatively little experience in structured interaction with strangers; (2) experiences or expectations of low levels of success at complex verbal and cognitive tasks; (3) an above average incidence of complicating physical and/or emotional impairments; and (4) an above average incidence of limitations arising from relative economic deprivation. Any or all of these factors may play a role in an individual's ability to respond adequately to questions in an interview.

Previous research on this target population has relied on data from various sources and data collection techniques. Probably the most common source of data has been "significant others" (counselors, job supervisors, and parents and guardians, for example) as informants or proxies (see Rusch and Schutz, 1980; Hunt and Zimmerman, 1969; Bogen and Aanes, 1975; Lambert and Nicoll, 1976; Eyman et al., 1979; and Abramowitz, 1980). In many instances, the parents or caretakers are expected to articulate the experiences or capabilities of the retarded persons, who are themselves never interviewed. To date, self-report interviews with the mentally retarded have been used primarily to provide anecdotal details rather than information on major variables for statistical analysis (e.g., Wyngaarden, 1981).[1] However, there is evidence that individuals who are mildly or moderately retarded are willing and able to provide some portion of the data necessary for evaluation research through in-person interviews (for example, see Weinglass, 1980; Richardson, 1979; Gollay et al., 1978; Sigelman et al., 1981; Birenbaum and Re, 1979; and Brolin, 1972).

Expected communication difficulties with the mentally retarded. The IQ range of the respondents in the present study (40–80) and the absence of other major handicaps that would significantly impede interviewing (such as serious visual, hearing, or verbal handicaps or emotional disturbances) suggest that the most important constraints on interviewing are the sample members' (1) accessibility and willingness to be interviewed; (2) comprehension and communication skills; (3) knowledge of the information of interest; and (4) ability to respond reliably.

1. Accessibility and willingness to cooperate. Previous interviewing experience and knowledge about this target population suggests that the sample members are likely

to be both accessible and willing to cooperate with the interview effort (for example, Gollay et al., 1978; Richardson, 1979; Sigelman et al., 1980; Ibid, 1981; and Weinglass, 1980). Furthermore, this population tends to be less mobile than the general population in the same age group and is more often residing with parents or other adults, which makes sample location much easier than usual. However, protective "gatekeepers" (parents, guardians, residential counselors, etc.) may restrict access to the sample members and increase the level of effort necessary to schedule interviews.

2. Comprehension and communication skills. The ability to understand and answer questions is related to intelligence. These abilities can be expected and have been found to be more limited among people with lower IQs (Spradlin, 1964). Nonetheless, recent research findings suggest that some mentally retarded persons do respond to properly constructed interview questions with information found to be consistent with that provided by other sources. Again, an important factor appears to be the IQ of the sample member. The higher IQ groups (above 55) have been found to be substantially better at comprehending questions and communicating responses (Sigelman et al., 1980).

3. Knowledge of the information of interest. Although experience is limited, researchers have found that persons in the IQ range included in this study were knowledgeable about many of the variables of interest (Richardson, 1979; and Weinglass, 1980). However, respondents' knowledge about one particular area of interest, their financial situation (sources and amounts of income, expenses, and assets), varies greatly depending on their level of independence. There may be fairly large gaps in the knowledge of some sample members regarding the details of their financial status, depending on whether they handle their own finances. Also, it is quite common, even among normal IQ youth populations, to find considerable lack of knowledge about total household financial information, especially concerning parents and siblings.

Questions involving recall are also likely to present problems for this group. It may be possible to obtain a level of detail that permits the sample member to state whether an event occurred within some limited period (six months to one year) in the past. However, detailed reporting of dates and other aspects of past experiences is likely to be especially unreliable.

4. Ability to respond reliably and truthfully. Although there is substantial evidence that mentally retarded sample members will participate in interviews willingly and cooperatively, researchers have identified a consistent pattern of acquiescence among retarded respondents in interview settings (Gerjony and Winters, 1966; Rosen et al., 1974; and Sigelman et al., forthcoming b). This is not surprising given general population survey findings which have suggested that acquiescence (yea-saying) is more common among less educated adults and children (Lenski et al., 1960; Wells, 1963; and Rothenberg, 1969).

This problem likely to affect the reliability and validity of self-reported data has been found to be most serious among lower IQ samples and is somewhat less problematic among IQ levels characteristic of this sample (Sigelman et al., 1980). Moreover, various question structures have been found to minimize the opportunity for yea-saying in interviewing (for example, see Sigelman et al., 1981). The solution is not a matter of simply rephrasing questions to reflect a distribution of positive and negative approaches that is skewed negatively, since it is often difficult to reword statements without using double negatives, which create their own response problems. Yea-saying may be result of other response patterns such as a desire to please the interviewer and a tendency to report what is perceived as socially approved or desirable behaviors and attitudes.[2] One approach to minimizing this problem is to use "either-or" questions (forced choice) rather than "yes-no" questions.

General survey development procedures point to the desirability of having more than two response categories, except for the simplest questions.[3] However, long lists of possible responses result in disproportionate choice of the first or last category and are likely to be particularly frustrating for populations such as the mentally retarded. In addition, closed questions may interfere with respondent-interviewer rapport, increase concern over giving the "right" answer, and be the cause of "test anxiety" resulting in poorer quality data and more nonresponse. On the other hand, open-ended questions raise problems associated with specifying appropriate probes, ensuring codable responses, and developing meaningful code categories. On the whole, it is believed that this population is likely to require a somewhat structured question format. Experience suggests that if a uniform set of responses (which can be easily memorized) are maintained and concrete visual cues like color and size of print are provided on response cards, multiple-category reponse formats can be used successfully (Sigelman et al., 1981).

Interviewer behavior as well as question wording and response formats can affect the quality of survey data. One major concern is the number and direction of probes. The mentally retarded are likely to be unsure of their answers and initially respond "don't know" to many questions, both factual and attitudinal. Good interviewing practice requires probing in these cases, as well as when a vague or contradictory answer is given. Simply repeating the question has been found to increase responsiveness. However, excessive probing may provoke biased responses due to "test anxiety" and heightened desire to please the interviewer (Sigelman et al., forthcoming b).

Development of the pilot study

Data to evaluate alternative data-collection strategies for

the SW/STETS evaluation research were derived from a pilot study. Information on baseline data items was collected from interviews with the mentally retarded participants or controls (primary-respondent interviews), proxy-respondent interviews, and the application/enrollment forms completed by the referral counselors before sending applicants to the SW/STETS program.

The pilot study involved conducting all data-collection activities with the research sample in Cincinnati, New York, and Tucson. The pilot-phase design called for interview attempts with all research sample members and an identified proxy for each respondent. Application/enrollment forms for each sample member were received and the data entered for data comparisons.

Criteria for assessing alternative data collection plans. Several criteria were used to compare the data provided by the alternative data sources—namely, accessibility, completeness, and accuracy.

1. Accessibility. A data source must be accessible—that is, the source must be available and/or willing to provide the data in a form that is usable for research.

2. Completeness. A crucial element in comparing data-collection strategies is whether a plan can provide data for all persons in the research sample (sample completeness) and for all data items (data completeness). While no strategy routinely provides usable data on every person and every item, some are more deficient than others.

3. Accuracy. The accuracy of the research findings is directly related to the accuracy of the data. In particular, possible incentives for respondents to misreport their activities (for example, because of acquiescence or a desire to hide activities that are considered unacceptable) are a concern for interview strategies. For records data, the major concern lies with the data-collection processes of the agencies.

Analysis approach. There are two basic limitations to a strictly quantitative assessment of the pilot results: (1) the absence of a known valid data source as a benchmark, and (2) the lack of formal rules or decision criteria. Without a known valid data source, the accuracy of the data cannot be fully evaluated. All three primary data sources considered in this study contain some inaccuracies. The main hypothesis about errors in self-reporting concerns the intellectual limitations of the sample. The resulting errors might be neutral (e.g., from random lack of knowledge) or non-neutral (e.g., from acquiescence). For proxies, errors might stem from incomplete knowledge about the activities of the sample member as well as from possible attempts to minimize dependents' deficiencies. For application/enrollment data, errors could be due to incomplete agency records or to information derived from interviews with clients or proxies rather than from an independent source. These potential problems suggest that none of these sources can serve as benchmark.[4]

The second limitation—the lack of formal rules or decision criteria—has broader implications. This suggests, for example, that, even if a benchmark were available, no definitive standards exist to judge whether the data are of acceptable quality. Of course, minimum error and bias are general goals, but some error and even bias is unavoidable in any method of collecting such a broad set of data.

The absence of a benchmark data source and the lack of formal decision rules suggest that examination of the criteria reviewed above—accessibility, completeness, and accuracy—must rest on subjective assessments as well as on quantitative analysis.

Results of pilot study

General impressions. The pilot data collection included all applicants assigned to participant and control status in three sites from November 1, 1981, through January 31, 1982. In order to be eligible to apply to the SW/STETS demonstration programs, an individual had to be between the ages of 18 and 24, have an IQ in the range of 40 to 80, have no physical or emotional impairments that would prohibit holding a job, and have only minimal previous work experience. All of these conditions were met by the pilot sample.

The median age was 20 years, although the full range was included in the pilot sample. Fifty percent of the sample were classified as mildly retarded, 13% were moderately retarded, and 37% borderline. Slightly more males (53%) than females were in the research sample. The majority (57%) were white, 28% were Black, and 14% Hispanic. Most (93%) of the sample had never been institutionalized and most (82%) lived with their parents at the time of application to the program.

Clearly both the eligibility requirements of the program and other factors resulting in the mix of people applying to the program created a sample different from the total mentally retarded population. It is a young group at the high end of the IQ range without severe secondary problems and with a history of family rather than institutional living. These sample members represent a group of people with potential for employment but who need special support services to reach that goal. Generalizations from the analysis presented here must be made cautiously. There is evidence that the more severely retarded, those with a history of institutionalization, and those with emotional as well as cognitive impairments may be less likely to give reliable, accurate information about themselves (see especially Sigelman et al., 1980; Ibid, 1981).

The baseline pilot interviewers were asked to describe their general impressions of the interviewing process. The following are several key points taken from their responses:

1. Some primary respondents felt that the interview was a test and became nervous. Probed too closely, they lost confidence, jeopardizing the rest of the interview.
2. Parents were sometimes overprotective and tended to

speak for the respondent, or they were hostile to both the respondent and the interviewer. These reactions had a major impact on the quality of the self-report responses. Respondents whose parents were overreacting easily slipped into saying, "I don't know."
3. Many times, both parents and respondents regarded the interviewer as a caseworker. For example, interviewers were asked questions about possible employment opportunities.
4. In general, respondents needed encouragement, praise, and reassurance that they were providing the type of information asked for and were performing adequately as a respondent.

These reactions confirmed the fact that special interviewing problems do exist with this sample, problems that have been found in other studies with a broader range of the mentally retarded. Overall, while this sample represents a particular segment of the mentally retarded, the interviewing process confronted issues that could be expected to arise with most studies of the mentally retarded and other cognitively impaired and partially dependent groups.

The remainder of this section more closely examines the results of the pilot study, focusing on three issues: sample completeness, data completeness, and data consistency.

Sample completeness. Two important factors in evaluating a data-collection strategy based on interviews are (1) the success with which interviews can be completed with respondents and (2) the effort required to obtain those completions. Completion rates and number of contact attempts for both primary sample members and their designated proxy respondents are used in this methodological study to assess the success of the pilot-phase interviewing strategy. The ability to identify a proxy respondent from the primary-respondent interview is also discussed as a measure of the success of the interviewing process.

Table 1
Final completion statuses by experimental status

	By experimental status	
	Participant	*Control*
Assigned sample	58	52
Completed interviews	55	49
Final noncompletions	3	3
Refusal (R)	0	1
Refusal (parent)	0	1
Unable to locate	1	1
Other	2	0
Response rate (Percent of assigned)	94.8	94.2

The excellent overall response rates by both primary and proxy respondents is probably due to a combination of factors: the introduction of the research by the SW/STETS intake couselors, the advance letters sent by the interviewers to both sample members and their parents or guardians, and the efforts that interviewers made to explain the study during their contact with respondents. The $5 respondent payments were well received and also seemed to contribute to respondents' willingness to be interviewed.

The high response rates with both primary respondents (95%) and proxy respondents (99%) make detailed subgroup analysis of completion rates and final-status outcomes inappropriate. Therefore, most of the following discussion reports results only for the total sample and by experimental status (participant or control).

1. Primary respondent final statuses. Table 1 presents the primary respondent final statuses and completion rates for the total sample by experimental status. Most primary sample members were successfully interviewed. The rate of completion varied slightly by experimental status. Control-group members were less likely to have been interviewed; several of these noncompletion final statuses were refusals by control-group members or their parents.

2. Identifying proxy respondents. During the primary-respondent interview, the interviewers noted on a grid all persons who gave the sample member help on critical activities. At the end of the interview, this grid was used to designate the proxy respondent. Proxies were chosen among the helpers in the following order:

1. A live-in parent or relative who gave help with financial management,
2. Any other person who gave help with financial management,
3. A live-in parent or relative, when no help with financial management was received,
4. A social worker or caseworker,
5. Someone who the primary respondent indicated was knowledgeable, when no other criteria were met.

In all cases, proxy respondents were identified from information provided by the primary respondent.

3. Proxy respondent final statuses. Once the proxy respondent was identified through this process, the primary respondent was asked to sign a release form that gave the MPR interviewer permission to contact and interview the proxy. In all cases of completed primary respondent interviews during the baseline pilot period, consent was obtained from the primary respondent. Only in one case was it impossible to complete a successful interview with a proxy—with a non-English-speaking parent in New York City.

The proxies identified and interviewed during the

baseline pilot period were most often parents of the primary sample members (82%); the next most frequent type of proxy respondent was an agency staff member or counselor (8.4%). Foster parents (4.2%) and residential counselors or houseparents (3.2%) were less frequently chosen as proxies. The remaining two proxies had some other relationship with the primary respondent. Over 85% of the designated proxies lived at the same place as the primary sample member; most of these (98%) were parents or foster parents.

4. Primary respondent contacts to reach final status. The number of contacts necessary to reach a final status (whether a completed interview or a final noncompletion status) is a measure of the ease with which an interviewing strategy can be pursued. In the SW/STETS design, all primary sample members and their parents or other persons designated on the application/enrollment form were sent an advance letter. This preliminary contact preceded all other attempts by the interviewer, whether by telephone or in person. Most appointments for the in-person interviews were scheduled by telephone; where there was no home telephone, personal visits were made to the home.

Table 2 shows the number of contacts required to reach final status on the 103 primary sample member contact sheets. The table shows both the overall distribution and the distribution by experimental status. Most final statuses with the primary respondent were reached by the third contact, which was true for each experimental status (69% for participants and 74% for control-group members).

Table 2
Number of contacts to reach final status for primary respondents, total and by experimental status

Number of contacts to final status	Total	By experimental status	
		Participant	Control
1	7 (6.8%)	3 (5.6%)	4 (8.2%)
2	38 (36.9%)	16 (29.6%)	22 (44.9%)
3	28 (27.2%)	18 (33.3%)	10 (20.4%)
4	9 (8.7%)	6 (11.1%)	3 (6.1%)
5	7 (6.8%)	3 (2.9%)	4 (3.9%)
6	6 (5.8%)	2 (3.7%)	4 (8.2%)
7	8 (7.8%)	6 (11.1%)	2 (4.1%)
Total	103	54	49

A regression model was used to test the independent effects of site and experimental status, as well as other characteristics of the primary sample member (race, gender, IQ range, and living arrangements at the time of application to the SW/STETS program), on the number of contacts necessary to reach a final status. No factors appear to significantly affect the number of contacts necessary to reach a final status of primary respondents.

Data completeness. This section discusses the problems of "missing data" for the key study variables, as found in the primary- and proxy-respondent interviews and on the application/enrollment forms. Not all key variables are measured in all three sources, and even when there are similar data the measures are often not identical in their construction. Differences between the data sources and the particular data elements provided by each are summarized in Table 3.

Table 3
Key data items and their sources

Data item	Primary interview	Proxy interview	Application/ enrollment form
Labor market performance			
Employment	X	X	
Job type	X	X	
Hours	X	X	
Earnings	X	X	
Labor-force participation	X	X	
Education and training			
Attendance at school	X	X	
Attendance at training	X	X	
School curriculum	X	X	
Training curriculum	X	X	
Transfer program use			
Receipt of SSI	X	X	X
Receipt of OASDI	X	X	X
Receipt of welfare	X	X	X
Receipt of Medicare/Medicaid	X	X	X
Receipt of food stamps	X	X	X
Amount of SSI	X	X	
Amount of OASDI	X	X	
Amount of welfare	X	X	
Living and social skills			
Living arrangement	X	X	X
Family composition	X	X	X
Money handling	X	X	
Transportation	X	X	

Little or no missing data occur for many of the variables. Examples include the education and training variables (from the interviews), living arrangements and family composition (from all three sources), and other living-skills activities (e.g., money handling, from the interviews). Other types of variables have greater levels of missing data, regardless of source; transfer-program use is the most striking case. Some aspects of labor-market

performance, particularly earnings, also suffer from substantial missing data from both the primary- and proxy-respondent interviews.

The missing interview data found during the pilot-phase period follow the patterns that were expected from the review of the available literature, the experience of MPR's consultants, and MPR's own pretest experience. Key areas are those that involve money—particularly the amounts of earnings and the receipt and amount of transfer payments. In the areas of transfers (both cash and in-kind), patterns to the nonresponse by primary respondents indicate that a "don't know" might indicate a reluctance to say "no" when the question seems ambiguous. From these patterns, an appropriate decision rule was designed for using proxy interviews, which significantly decreased the amount of missing data.

Missing data on the application/enrollment form is a more complex problem. This form was completed by different agencies and staff (varying in the amount of previous contact with the applicant, the quality and quantity of records on file, and whether the referral counselor or the SW/STETS intake officer completed the form) using a variety of sources, but relying in large part on interviews with the applicants and/or their parents. Therefore, the missing data encountered with this form is due to a variety of problems (e.g., lack of records, misunderstanding about how to complete the form, using applicant self-reports in an unsystematic interview format) and may be less tractable overall than the missing data on the interview.

Table 4 provides a guide for the variable-by-variable review of data completeness. Each column (Primary Interviews, Proxy Interviews, and Application/Enrollment Forms) refers to a different total number of cases—96, 95 and 111, respectively.[5] Within each column, individual variables also differ in the number of applicable cases. For example, while each interview and each application/enrollment form contains either an answer or a missing response on the receipt of each type of benefit, the variables on dollar amounts of benefits have different numbers of cases that are logically skipped as inappropriate, depending on whether the receipt variables are coded "yes" or "no." The SSI benefit amount variable

Table 4
Missing data by source

Category of variable	*Variable*	*Primary interviews*	*Proxy interviews*	*Application/ enrollment forms*
Labor market performance	Employment			
	Any job	0	0	NA
	Any paid job	0	0	NA
	Job type	0	2 (3.4%)	NA
	Hours worked per week	2 (3.3%)	1 (1.9%)	NA
	Earnings ($ per week)	7 (12.7%)	16 (34.8%)	NA
	Labor force participation	0	0	NA
Education and training	Any school	0	0	NA
	Any training	0	3 (3.2%)	NA
	School curriculum	0	2 (5.6%)	NA
	Training curriculum	0	0	NA
Transfer program use	Receipt of SSI	12 (12.5%)	1 (1.1%)	7 (6.3%)
	Receipt of OASDI	16 (16.7%)	0	11 (10.0%)
	Receipt of welfare	8 (8.3%)	0	12 (10.8%)
	Receipt of Medicaid or Medicare	8 (8.3%)	2 (2.1%)	21 (18.9%)
	Receipt of food stamps	5 (5.2%)	3 (3.2%)	14 (12.6%)
	Amount of SSI	12 (40.0%)	4 (4.4%)	27 (75.0%)
	Amount of OASDI	6 (50.0%)	0	8 (61.5%)
	Amount of welfare	6 (75.0%)	0	NA
Living and social skills	Living arrangement	0	0	0
	Household composition			
	# parents living with	0	0	1 (0.9%)
	Living with spouse	0	0	1 (0.9%)
	# own children living with	0	0	1 (0.9%)
	Money handling			
	Banks by self	3 (3.1%)	0	NA
	Pays bills by self	0	0	NA
	Pays for purchases by self	0	0	NA
	Transportation			
	Use of subsidized transportation	0	0	NA

has 66 cases logically skipped in the primary-respondent interview dataset, 62 cases in the proxy-interview dataset, and 75 in the form dataset. The numbers of logical skips are not shown in Table 4, but the percentage of missing data for each entry is computed on the number of applicable (i.e., not skipped) cases. Except where noted, all missing interview data are the result of "don't know" responses by the respondent. Missing data on the application/enrollment form are usually the result of blanks on the form.

1. Labor market performance. Of the five labor-market performance variables, only one has a significant amount of missing data—earnings. Of the 55 applicable primary respondent cases, 7 (13%) are missing; of the 53 applicable proxy cases, 16 (30%) are missing. This is the strongest reversal of the expectation that proxy respondents would be fully knowledgeable about the activities of primary respondents. However, when primary respondents have missing data, proxies generally are able to provide an answer, and vice versa; thus, a combined dataset would have almost no missing data on this variable.

2. Education and training. On these variables, proxies again show more missing data than primary respondents. However, in either dataset, the amount of missing information is small—6% at most.

3. Transfer program use. These variables show some of the largest amounts of missing data from all three sources, especially on the amounts of cash transfers. Data completeness from the application/enrollment form seems uniformly poor. However, the pattern of primary- versus proxy-interview missing data is more complex. Primary respondents are less knowledgeable or less confident than proxy respondents about whether they receive benefits and the amounts of these benefits if they do report receiving them. However, a "don't know" response by primary respondents seems to imply that they actually do not receive the benefit, but are reluctant to say so. For example, Table 5 shows that most primary respondents who "don't know" whether they receive a type of transfer-program benefit do not receive that benefit according to the proxy. Primary respondents are perhaps hesitant to say "no" when they are unsure—the reverse of the usual acquiescence problem. The strength and consistency of this pattern supports the assumption that "don't know" responses to the receipt of benefits questions should be equivalent to "no" responses.

The pattern with the amount of cash transfer income is less clearcut. Table 6 compares primary- and proxy-respondent reports of the receipt and amount of the three cash transfers—SSI, OASDI, and welfare. According to the proxies, primary respondents who said they receive SSI but did not know the amount actually receive the benefit. However, when asked about OASDI, only half of the primary respondents who said "yes" to the receipt question but "don't know" to the amount were confirmed by their proxies as receiving the benefit. And the pattern is completely reversed for welfare—according to the proxy respondents, all primary respondents who did not know the amount of welfare do not receive any.

The most likely explanations for these results concern the respondents' understanding of the interview questions. SSI is a common benefit for this population, and

Table 5
Cross-tabulations of missing data for receipt of transfers

SSI		Proxy respondent		
		Don't know	No	Yes
Primary respondent	Don't know	1	10	1
	No	0	49	5
	Yes	0	2	27

OASDI		Proxy respondent			
		Not asked	Don't know	No	Yes
Primary respondent	Not asked	15	0	1	1
	Don't know	1	0	15	0
	No	1	0	48	1
	Yes	1	0	4	7

SSI or OASDI		Proxy respondent		
		Don't know	No	Yes
Primary respondent	Don't know	1	15	1
	No	0	38	3
	Yes	0	1	36

Welfare		Proxy respondent		
		Don't know	No	Yes
Primary respondent	Don't know	0	7	1
	No	0	79	0
	Yes	0	6	2

Food stamps		Proxy respondent			
		Refused	Don't know	No	Yes
Primary respondent	Refused	0	1	0	0
	Don't know	0	0	4	0
	No	0	1	59	3
	Yes	0	0	15	11

Medicaid or Medicare		Proxy respondent		
		Don't know	No	Yes
Primary respondent	Don't know	0	6	2
	No	1	55	3
	Yes	1	7	20

those who receive it recognize the name and can accurately report its receipt. However, since the check is often not handled directly by the primary respondent, its amount might not be known. A "don't know" response on the amount of SSI could almost always be retained as "yes" on the receipt of the SSI variable.

OASDI is less frequently received and its name (it is referred to as "Disability Insurance Benefits from Social Security" in the interview) sounds similar to SSI. When asked about OASDI immediately after SSI, some respondents might confuse the two or feel that they should be receiving one or the other, thus reporting "yes" when in fact they do not receive it. Others, in fact, do receive OASDI but, for money-handling inabilities or other reasons, did not know the amount. No assumption can be made about how to treat "don't know" responses on this question.

Table 6
Cross-tabulations of missing data for amount of cash transfers

SSI receipt and amount		Proxy respondent		
		Not received[a]	Received, doesn't know amount	Received, knows amount
Primary respondent	Not received[a]	60	1	5
	Received, doesn't know amount	1	0	11
	Received, knows amount	1	3	13

OASDI receipt and amount		Proxy respondent		
		Not asked or not received[a]	Received, doesn't know amount	Received, knows amount
Primary respondent	Not asked or not received[a]	81	0	2
	Received, doesn't know amount	3	0	3
	Received, knows amount	2	0	4

Welfare receipt and amount		Proxy respondent		
		Not received	Received, doesn't know amount	Received, knows amount
Primary respondent	Not received[a]	86	0	1
	Received, doesn't know amount	6	0	0
	Received, knows amount	0	0	2

[a]Also includes a few respondents who did not know whether benefits were received.

Primary respondents might also report welfare receipt because of acquiescence (since the welfare questions follow questions about SSI and OASDI). In other cases, however, it might be caused by a misunderstanding on the part of the respondent—reporting family or household receipt of welfare as benefits they personally receive, when in fact the proxy respondent has a clearer idea about how welfare income is received within the family or household. Here, a "don't know" response from the primary respondent on the amount of welfare can almost always be assumed to mean not receiving welfare, at least from proxy reports. Because of this variable pattern, "don't know" responses to questions on the amount of OASDI and welfare, at least, seem to suggest the need for a proxy interview.

4. Living and social skills. These variables, from whatever source, have virtually no missing data.

Data consistency across sources. As described above, there are no benchmark data or widely accepted standards by which the accuracy of alternative data sources can be evaluated in this pilot study. However, an analysis of data consistency across sources, together with the previous discussion of completeness, helps provide inferences about quality. The analysis of consistency is based on cross-tabulation of data from one source versus another.[6]

Generally, the consistency between primary and proxy data is quite high. Where reporting differences do appear, there are indications that the blame rests with proxy respondents as often as it does with primary respondents. Consistency with application/enrollment form data could be examined only on a small subset of the data. Consistency between application/enrollment data and an interview source is high, but not as high as that found between interview sources.

1. Labor market performance. As judged by holding either any job or a paid job, employment sets the pattern of strong correlations between primary and proxy interview data. There are a small number of cases in which the primary reports no job and the proxy reports a job (2 to 3% of all pairs). There are somewhat more cases (13%) in which the opposite occurs—the primary reports a job but the proxy reports no job. While the magnitude of this discrepancy might seem large for such straightforward data items, it is largely explainable and easily understood in the context of the job type.

Reports of job type are generally consistent between primaries and proxies. This pattern is most evident by the large proportion (72.7%) of matched pairs (i.e., matched primary-proxy responses that fall on the main diagonal of the crosstabulation). Equally interesting is the pattern of mismatched responses. Where both primary and proxy respondents report a job, most of the reporting differences appear to be caused by different categorizations of sheltered workshops and day activity centers.

There are also situations in which one respondent

reports a job but the other does not. For the two cases in which primary respondents report no job, the jobs according to the proxies are a sheltered workshop and an "other" type. The former mismatch might reflect different views of the workshop experience, or it might reflect a changed status, since a proxy was sometimes interviewed sometime after the primary respondent was interviewed. The same explanations might apply to the situations in which the proxies report no jobs, but the primary respondents report sheltered workshop or day activity-center jobs. Two cases in which primary respondents report jobs with training and proxies report no jobs are associated with SW/STETS program enrollment. The differential reporting may stem from experimentals who have not yet begun an in-job portion of the program but anticipate having a training job, or from proxies who were not yet fully aware of the very recent change in the primary respondents' job status.

In summary, the differences in reporting employment and job type that do appear are relatively few in number and in a pattern that suggests no serious biases. It is noteworthy that both types of respondents seem equally likely to misclassify jobs.

Data consistency for weekly hours and earnings is also high. Variables were constructed that indicate "matches" for cases in which both respondents report a job and have no missing data. A "match" for weekly hours is recorded when the primary-proxy difference is less than six hours. The match for weekly earnings is recorded when the difference is less than $26.

The matches for the hours and earnings are shown in Table 7. Cases are excluded in which one or both report no job or have missing data. When both primary and proxy respondents report hours, 87% match. When both report earnings, 78% match. Both variables show a high degree of consistency: the smaller number for earnings might be more the fault of proxies than of primary respondents, since the former also have much more missing data for this variable. Therefore, these patterns suggest that primary respondents are a good source of data about details of their employment experiences.

Table 7
Cross-tabulations of hours and earnings data[a]

	Hours per week	*Earnings per week*
Match (with job)[b]	41	21
No match	6	6

[a]Omits cases in which one or both respondents report no job or have missing data. There is more missing deta for proxy than for primery respondents.
[b]"Match with job" for weekly *hours* indicetes a primary-proxy difference of less than six hours. "Match with job" for weekly *earnings* indicates a difference of less than $26.

2. Education and training. A very high degree of consistency is evident between primary and proxy respondents' reports of school attendance (only three reporting differences). Training-program attendance shows thirteen reporting differences. Some proxy respondents report no training-program participation when primary respondents report participation. This problem was discussed above for job type, since a job with training (or vice versa) is considered both a job and training in this analysis.

Consistency is also very good for training curriculum (whether the program involved actual work experiences, classroom job training, or both): only four of the forty-two sets that agree on training-program attendance disagree on curriculum. The source of the modest disagreement stems from either the primary or proxy respondent reporting "other" types of training when the other respondent reports training in a job setting. There is much more disagreement on school curriculum. In one-third of the pairs, one respondent reports participation in an "other" curriculum (which might include learning "about jobs and work") when the other respondent in the pair reports learning "how to do a job" through work experience or through classroom instruction. The source of some of this disagreement might be a slight wording problem in the proxy interview, which biased proxy respondents toward the "other" category.

3. Transfer program use. The consistency of primary- and proxy-respondent reporting of transfer receipt was shown earlier in Table 5. There are two types of transfers, and each type has a different consistency pattern. The first includes transfers that are typically individual and specific—SSI, OASDI, and Medicaid/Medicare. For these transfers there is a very high correlation between the primary and proxy respondents' reports. Furthermore, primary respondents' "don't know" responses are overwhelmingly (91%) associated with no receipt according to proxy respondents.

The second set of transfers includes those that are typically family- or household-based—welfare and food stamps. While the correlations are still quite high, there is a pattern of primary respondents' reporting receipt when proxy respondents and the application/enrollment form do not. This might be the result of an acquiescence problem with primary respondents or a misunderstanding about whether receipt should be reported when the primary respondent is a member of the qualifying unit (i.e., family or household), but not the one to whom the transfer is sent.

High correlation also exists between primary and proxy respondents in their reports of the amount of transfer payments. In the 19 cases in which both report an amount for SSI, OASDI, or welfare, 9 are within $2, 13 are within $35, and 16 are within $65. There are only three large discrepancies.

The correlations between application/enrollment form data and each of the interview sources are also high, but weaker than those between primary and proxy respondents. The higher number of mismatches seems likely to be caused by shortcomings in the application/enrollment form data: (1) much of the data were collected through informal interviews with program appli-

cants or their parents and (2) records data, when used, were outdated.

4. Living and social skills. All three sources provide very similar reports of living arrangement and family composition. The only mismatches for living arrangement are attributable to a few differences in classifying institutions, group homes, and semi-independent facilities. These are not totally distinct concepts, and some confusion in classification is understandable. The only noteworthy mismatches in family composition are the reports of one or two parents. Interviewer observations indicate that this may be attributable to a few mothers "hiding" males in the household.

The correlation between primary and proxy respondent reports of money handling tends to be lower than other relationships. For two of these three living-skills measures (banks by self and purchases clothes by self), the two respondents disagree more than 20% of the time. The match is much better (only 8% disagreement) for the third measure—pays bills by self. Problems of agreement were anticipated for these variables, since they require some level of judgment or opinion. In fact, the correlation is quite high for judgment questions, and these measures of living skills seem very promising, particularly since the proxy responses cannot be presumed to be "correct."

Few transportation questions exhibit enough response variation to permit analysis. For the one variable considered (use of subsidized transportation), only 10 primary respondents and 8 proxy respondents reported use. They agreed in 6 of the cases.

Summary and conclusions

The evaluation project for which this pilot study was conducted will focus on the degree to which the program achieves its objectives, which are to increase participants' employment and earnings and to allow participants to lead more independent lifestyles. As a result, the research must examine data that pertain to differences in the labor-market activites, program use, transfer-payment receipt, residential arrangements, and social and living skills of participant and control group members. Three sources of data were examined as part of this pilot study—primary-respondent interviews, proxy-respondent interviews, and applicant/enrollment forms. This final section will review the results of the pilot study and assess the relative merits of the three sources, particularly the success of the respondent self-report approach. This assessment will return to the three criteria discussed earlier: accessibility, completeness, and accuracy.

Accessibility. Primary- and proxy-interview data are at least as accessible as projected, as indicated by completion rates, and field efforts necessary to obtain completions. Review of the use of the application/enrollment forms indicates that referral-agency records data are less accessible than expected because, in some cases, the records data simply do not exist or are not available to the agencies that refer individuals to the program.

Completeness. The interview sources and the application/enrollment form provided fairly complete sample coverage (i.e., they provided data for all or almost all sample members). There were incomplete interviews but these numbers are well within the acceptable limits based on other interviewing efforts.

The interviews met or exceeded the expectations for data completeness—the ability to provide all the data required for the evaluation. The only important pattern of missing data are associated with amounts of money; even so, these data are rarely missing for both responsdents. Further, there was no case in the pilot study in which the sample member could not provide most of the requested data. In comparison, the application/enrollment form is more limited in data-item coverage.

Accuracy. It was not possible to evaluate data accuracy directly, since there was no true "benchmark" data source. The two interview sources provide highly consistent data, with no single source obviously superior. Application/enrollment form data are also generally consistent with each of the two interview sources for the subset of overlapping variables.

Based on the assessment of the pilot study results, the MPR staff developed a data collection strategy for the evaluation research which relies on baseline data collected from (1) a brief application/enrollment form, (2) in-person interviews conducted with participants and controls, and (3) interviews with proxy respondents only when the primary respondent cannot provide accurate data on key items. This mixed-source strategy was modeled using the pilot sample data with excellent overall results, both in terms of completeness and accuracy.

Mentally retarded young adults can provide factual data about themselves, their activities, and their life circumstances in sufficient detail to undertake a rigorous program evaluation. As the emphasis on deinstitutionalization and mainstreaming grows and as funds become harder to find, there will be considerable interest in evaluating programs and determining which are successful in meeting their goals for the mentally retarded. Self-report by the target group is an appropriate means of gathering at least.some of the information needed for these evaluations.

The application and extension of standard survey practice to samples of the cognitively impaired forces researchers to question many assumptions about how respondents deal with interview tasks. This study has shown that a structured interview instrument, developed with the particular cognitive and interactional limitations of the mentally retarded in mind, can be successfully used by trained interviewers in research studies with this population. Other populations, such as the impaired elderly or the emotionally disturbed, certainly present different sets of issues in research design. However, this study has indicated that a self-report data

collection approach can appropriately be considered with groups usually only asked *about*, rather than asked directly.

Footnotes

[1] Sigelman et al. (1980); Halpern et al. (1977); Richardson (1979); and Weinglass (1980) all used additional collection of data from proxies to supplement or verify information reported by sample members.

[2] There is evidence to suggest that acquiescence and desire to please the interviewer have more influence on the mentally retarded than does concern over giving the normative answer. Studies have found this population willing to discuss and even volunteer information about generally unacceptable behavior. For example, see Weinglass (1980).

[3] A related strategy is to use a multistage approach in which subsequent questions are used to differentiate within groups of people responding to an initial simple dichotomy. Example: "Do you do this?" followed by: "How often?"

[4] One conventional way to evaluate data quality is to obtain access to records on key data items for at least part of the sample. An analysis of the key data for this subsample could then generate information on data quality that could be used to qualify the general study findings. However, this approach is appropriate for only a few data items (e.g., receipt and amounts of some types of transfers) and could not be implemented in sufficient time for this study.

[5] One proxy-respondent interview was not completed due to language problems. The application/enrollment form dataset also contains data corresponding with seven noncompleted primary interviews, seven completed primary interviews that were not available for this analysis, and one February referral not interviewed for this analysis.

[6] The relationship between consistency and characteristics of the respondents were considered but did not show very informative patterns.

Discussion: Survey methods for rare populations

Monroe G. Sirken, National Center for Health Statistics

Introduction

The unique attribute that makes a population rare often makes it difficult to survey. This problem was addressed in the paper on the survey of mental retarded youth. The rareness of the population, especially if a separate sampling frame does not exist, invariably leads to a number of difficult and sometimes insurmountable problems in designing rare population surveys. These kinds of problems were discussed in the papers on the cancer, epilepsy, and multiple sclerosis surveys. In these remarks, I will discuss some of the advantages and disadvantages of the design strategies for rare disease surveys that were reported at this session. In particular, I will focus on various design strategies that are based on alternative sampling frames and counting rules.

Alternate design strategies

Sampling frames and counting rules are very important design features of rare disease surveys. Sampling frames define the enumeration units, and counting rules define the networks of enumeration units that are eligible to report the persons in the rare population. Each of these design features has essentially two options. The sampling frame may be either a frame of households or a frame of establishments, such as hospitals, pharmacies, physicians' offices, etc. The counting rule may be a unitary rule, such as a *de jure* residence rule that uniquely links each person in the rare population to one enumeration unit that is eligible to report him or her in the survey, or it may be a multiplicity rule that links each person in the rare population to a network of one or more enumeration units that are eligible to report him or her in the survey.

The process of designing a rare disease survey involves selecting a sampling frame and a counting rule. Since each of these design features has two options, there are a total of four design option sets to choose from. Each of the four option sets contains a different combination of sampling frame and counting rule as shown in Table 1.

Table 1
Option sets

Option set	*Description*
1. Household frame/ unitary rule	A household survey in which each person with the rare attribute is eligible to be enumerated at only one household.
2. Household frame/ multiplicity rule	A household survey in which each person with the rare attribute is eligible to be enumerated at one or more households.
3. Establishment frame/ unitary rule	An establishment survey in which each person with the rare attribute is eligible to be enumerated at only one establishment.
4. Establishment frame/ multiplicity rule	An establishment survey in which each person with the rare attribute is eligible to be enumerated at one or more establishments.

The cancer survey involved a series of experiments that tested design option sets 1 and 2. A *de jure* residence rule was adopted in option set 1. It had the effect of making persons with cancer eligible to be enumerated only at their own households. A kinship rule was adopted in option set 2. It had the effect of making persons with cancer eligible to be enumerated at the households of their close relatives (children, parents, siblings) as well as at their own households.

The epilepsy survey was based on a frame of pharmacies and in effect its design made it possible to test option sets 3 and 4. In option set 3, epileptics were eligible to be enumerated at the first pharmacy that filled their prescriptions for epilepsy drugs during the reference year. In option set 4, epileptics were eligible to be enumerated at all pharmacies that filled their prescriptions for epilepsy drugs during the reference year.

The multiple sclerosis survey was based on medical provider frames of hospitals and physicians and tested option 4. The survey adopted a counting rule that made these patients eligible to be reported by all hospitals and physicians that had ever treated the patient for the disease.

Comparison of design strategies

As will become apparent, the advantages and disadvantages of the four option sets depend in part on the data requirements of the survey. Particularly important is whether the objective of the survey is merely to count the number of persons with the rare attribute, in order to produce incidence or prevalence estimates, or to collect detailed information on medical care use, direct and indirect costs due to the disease, etc. The advantages and disadvantages of the various option sets also depend on the data requirements implied by the counting rules themselves. For example, the survey collects counting rule related information to determine the number of persons, if any, with the rare disease that are eligible to be reported by each sample enumeration unit, and the number of other enumeration units that are eligible to report each person reported by a sample enumeration unit.

Household surveys. The principal design problems as-

sociated with household surveys of rare populations are that the screening costs of locating households of persons with the rare attributes are expensive and that the rare persons are incompletely enumerated. Let me briefly summarize the pros and cons of option sets 1 and 2 in dealing with these design problems.

The screening costs are substantially smaller for option set 2 than for option set 1 because in option set 2 two or three times as many households are eligible to report persons with the rare attribute. It seems likely, however, that better quality information would be reported by option set 1 than by option set 2 because the patient and the persons living with him would be better informed than his relatives living elsewhere. In the cancer survey, for example, the underreporting rate was about 10% for patients enumerated at their own households and about 20% for the patients enumerated at the households of relatives. Furthermore, some types of information, such as the cost of treating the disease, were collected in the cancer survey only from the cancer patients' own households.

Establishment surveys. Many of the problems in designing establishment surveys are due to the incompleteness of the patient information in the establishments' files. The records in these files rarely contain the kind of detailed information about a patient's current status and activities that is wanted for substantive analysis, and they infrequently contain the information about the patient's transactions with other establishments that would be needed to apply the counting rule. Lacking this information, follow-up surveys with the patient are usually required. However, the patient follow-up surveys are often deferred by a survey with the patient's personal physician which must be conducted solely to obtain the physician's permission to contact his or her patient. Hence, a rare disease survey based on an establishment frame usually involves a survey of the patient's physician, and a patient follow-up survey, as well as the establishment survey. The proliferation of survey adds to the survey costs and to the nonresponse rate.

In the pharmacy survey of epileptics, for example, individuals who had taken specified drugs were identified by the pharmacies that filled their prescriptions. Subsequently, a survey of the physicians who prescribed the drugs was undertaken to determine if the drugs were prescribed for epilepsy and to obtain the physicians' permission to survey the patients. In fact, the pharmacists would not release the names and addresses of the prescription users for the follow-up patient survey unless approval was obtained from the physicians who signed the prescriptions. Since the response rate was about 80% in each of the three surveys, the response rate of the combined surveys was $(.8)^3$ or barely 50%.

On the average, establishments are eligible to report more patients by the multiplicity rule than by the unitary rule. Therefore, a somewhat smaller establishment sample is needed and the survey costs are smaller for option set 4 than for option set 3. On the other hand, the multiplicity rule generally requires more information than the unitary rule, and hence, all of the information needed is less likely to be available in the patients' records. On this basis, option set 3 would be preferable to option set 4. For example, the unitary rule, which links patients to their first transactions with an establishment, requires that the patients' records contain information about their prior medical transactions, but the multiplicity rule, which links patients to all their transactions, requires information about later, as well as prior, transactions.

Concluding remarks

The order of preference of the four option sets for designing rare population surveys is far from settled at this time. The preference is bound to vary somewhat depending on the specific survey objectives and circumstances. In general, the quality of the diagnostic information favors the establishment survey, and the availability of nondiagnostic information favors the household survey.

Unless the problem due to attrition in the response rates in establishment surveys can be resolved, the establishment frame is unlikely to qualify as an acceptable design option for rare disease surveys. One possible strategy for dealing with this problem is the data-broker system that was used with great success in the previously reported cancer survey. It was used there to preserve the anonymity of people who had been selected from cancer registries. If the data-broker system were used in a similar manner to preserve the anonymity of patients reported in establishment surveys, it might persuade the participating establishments that obtaining authorizations from the patients' own physicians was not a prerequisite for conducting patient follow-up surveys.

The data broker preserves the patients' anonymity by serving as a third-party intermediary between the agency that collects the survey data and the establishments and the patients who provide the survey data. The data-broker system might work as follows:

Step 1. The survey agency selects the sample establishments and provides them with protocols for selecting the patients whom they are eligible to report and for abstracting information about them from the establishments' records.

Step 2. The establishments transmit the abstracted information, including personal identifiers, for their patients to the data broker.

Step 3. The data broker merges the data file for patients who were reported by the establishments with a data file of decoy samples that it selected from public directories. It transmits the combined but indistinguishable data files to the survey agency.

Step 4. The survey agency conducts the follow-up patient survey with persons in the combined data files and then returns the merged data files to the data broker.

Step 5. The data broker strips the personal identities from the data files, identifies the records in the file for patients who were originally reported by the establishments, and returns the file to the survey agency.

Step 6. The survey agency processes, tabulates, and analyzes the data that it collected in the establishment survey and in the patient follow-up survey.

Many interesting statistical and ethical questions remain to be answered about the data-broker system including the relative size of the decoy sample selected in Step 3 compared to the number of the patients reported by the establishments in Step 2.

As I noted earlier, the optimum designs for rare population surveys are far from settled, and much more methodological work will be required before the outstanding issues are resolved.

Open Discussion: Session 6

Satin said that there was no discussion of untreated cases in the study of multiple sclerosis. This disease has a preclinical lead time of 10 to 15 years, so ignoring undiagnosed cases could result in substantial underestimates. Wilson said that they had recognized the problem and tried to reduce underestimates by using an expanded case-finding period. In this study, the period between patient recognition of symptoms and physician's diagnosis was usually less than six years. Bryan admitted that the study of epilepsy was limited to only clinical diagnosis because there was no known way of obtaining subclinical cases. Suspicious cases are treated with drugs used for epilepsy so that the problem of false positives may be greater than missed cases.

S. Rice asked about the political problems that were encountered in the epilepsy study. Bryan responded that some pharmacists resisted because the study was perceived as involving government intervention. The Connecticut State Pharmaceutical Association urged its members not to participate because of anger over previous federal intervention.

S. Rice then asked about the value of payment to pharmacists. Bryan responded that this would be essential in a national study and that, while the amounts were large for each pharmacist, they were small when allocated across all interviews. Kasper asked about the implications of rewarding only the uncooperative respondents. Bryan reported that it was acceptable in this pilot study because of the limited number of locations, but would not be appropriate procedure for a national survey.

In response to a question from the chair asking about other special populations, de la Puente reported on a survey which used the American College of Pathology as a data broker. A cooperation rate of 95% was achieved from pathologists.

Drummond asked what mixing fraction is necessary on seeded surveys. Sirken replied it was necessary to weigh the cost of additional decoys versus the danger of loss of confidentiality if the decoy rate is too low. It depends on the prevalence of the condition in the population, but the most important issue is an ethical one. Warnecke reported that on the cancer study, the initial ratio of general population to cancer patients was 2:1, but that this was reduced to 8:1 for cost reasons. This was an effective ratio because the relatives and general public also reported substantial cases of illness which, in effect, further masked the cancer patients. Drummond pointed out that there was a cost to processing and analyzing the data from decoys. Sirken responded that, in the cancer study, the decoys were not selected at random, but were chosen in the same geographic areas as cancer patients in order to reduce travel costs and interviewer concerns, therefore decoy data were not processed. If a random sample of decoys was used, this sample could be used to improve the estimates.

Elinson asked how interviewers felt about not being told the real purpose of the study and the source of the sample. Warnecke reported that interviewers were told that the sample contained a large number of chronically ill persons, but were not given the detailed source of the sample. Cannell reported that in a seeded sample to measure accuracy of number of hospitalizations, interviewers were told only that it was a special sample and that the study was conducted for methodological reasons. During debriefing, interviewers were asked if they had guessed the reason for the study, but none had guessed correctly. Elinson pointed out that this method is similar to the use of double-blind experiments in testing of drugs. Sharp said that interviewers do not like to collect data that will not be used in the analysis unless they are told it is for a screening experiment. Sirken said that the last several remarks pointed out the ethical problems in using decoys relate to both respondents and interviewers. Kulka said that it would be useful to know how brokers and constituencies, as well as respondents, feel about decoy rates.

Fowler pointed out that an alternative to use of institutional frames was to use random-digit-dialing telephone samples of thousands of cases. He reported that he had been asked to screen for a sample of workers who had been fired in the previous four weeks. Fowler refused to do this study because of concerns not only about the very large cost, but also about the effects on interviewer morale of conducting thousands of unsuccessful screening calls. Bryan reported that RTI had considered using RDD procedures for the epilepsy study, but had rejected the idea because of concerns that gatekeepers would underreport epilepsy of other family members in a telephone interview. Warnecke said that all studies of illness are faced with serious gatekeeper problems. This problem is even worse in telephone surveys, if the "wrong" person answers the telephone. In the cancer study, the patients themselves were very willing to talk about their illness and the costs involved.

Wright said that there are special problems with some populations, not because of the rarity of the illness, but because the disease is hidden or because there are problems in obtaining the information even when the respondent agrees to the interview. Warnecke reported that there were no such problems in the cancer study. People were willing to sit for three hours discussing their cancer costs and to go to their files to pull out medical records. Bryan also reported that response was very

good, once the person was contacted. Wilson, however, reported some contrary evidence on MS. Reports on the screener were lower than anticipated, but many people who did not admit having MS on the screener reported symptoms on the general interview that followed.

After a question from the chair, Stephens reported that the feelings of both interviewers and respondents on the survey of the mentally retarded were similar to those of interviewers and respondents on general population samples. Respondents were initially nervous that they were being tested, but by the end of the interview they felt comfortable with the process. Interviewers also felt that the interview was a positive experience. A major problem was the parents of the mentally handicapped respondents who were overprotective and at times volunteered information.

Czaja asked whether the interviewers of the mentally handicapped youth felt that the training they received was adequate for the situations they encountered. Stephens reported, in general, interviewers were satisfied. The experiences of three interviewers on the pilot study were incorporated into the training session in the form of a list of all the behaviors that they encountered and how they handled the situations. Interviewers were also pleased with the very specific probes. The major problem with the interviewer training was that it did not prepare the interviewers to deal with gatekeepers.

Satin said that the mentally ill are an especially difficult population since there is a reluctance to admit mental illness. Nevertheless, the gatekeeper problem in a mail sample of the general population was the same as the gatekeeper problem in the samples of mentally ill patients.

Axelrod asked whether any of these studies of special populations had faced difficulties with institutional review boards. Wilson reported that National Analysts had prepared documents and received approval from several hospital review boards. Warnecke reported that cooperation from hospitals in the cancer study was excellent, in part due to the reputation of the Illinois Cancer Council. Sudman reported that the Institutional Review Board of the University of Illinois was enthusiastic about the use of decoy samples as a procedure for protecting privacy while permitting the research to be done.

References

Abramowitz, M.
1980 Training and Employment of the Mentally Retarded in Private Sector Jobs: An Evaluation of the First 16 Months of the WORC Project. Boston: Transitional Employment Enterprises.

Aday, L.A., and R. Andersen
1974 "A framework for the study of access to medical care." Health Services Research 9:208-220.

Aday, L.A., R. Andersen, and G.V. Fleming
1980 Health Care in the U.S.: Equitable for Whom? Beverly Hills, CA: Sage.

Aday, L.A., R. Andersen, G.V. Fleming, G. Chiu, V. Daughety, and M.J. Banks
1978 "Overview of a design to evaluate the impact of community hospital-sponsored primary care group practices." Medical Group Management 25(Sept.-Oct.):42-46.

Aday, L.A., R. Andersen, S.S. Loevy, B. Kremer, and M.J. Banks
1981 Access to Medical Care: Measurement in a Community Context. Unpublished manuscript. Chicago: Center for Health Administration Studies, University of Chicago.

Aday, L.A., C. Sellers, and R.M. Andersen
1981 "Potentials of local health surveys: A state-of-the-art summary." American J. of Public Health 71 (August):835-840.

Aitkin, D., M. Kahan, and D.E. Stokes
1967 Australian National Political Attitudes. International Consortium for Social and Political Research (ICPSR #7282).Ann Arbor, MI:Institute for Social Research, The University of Michigan.

Alwin, D.
1974 "Approaches to the interpretation of relationships in the multitrait-multimethod matrix." In H.L. Costner (ed.), Sociological Methodology 1973-74. San Francisco: Jossey-Bass.

Alwin, D.F., and D.J. Jackson
1979 "Measurement models for response errors in surveys: Issues and applications." In K.F. Schuessler (ed.), Sociological Methodology 1980. San Francisco: Jossey-Bass.

American Psychological Association
1974 Standards for Educational and Psychological Tests. Washington, DC: American Psychological Association.

Andersen, R., and V. Daughety
1979 "Health insurance coverage and payments." In R. Andersen *et al.* (eds.), Total Survey Error. San Francisco: Jossey-Bass.

Andersen, R., J. Lion, and O.W. Anderson (eds.)
1976 Two Decades of Health Services: Social Survey Trends in Use and Expenditure. Cambridge, MA: Ballinger.

Andersen, R., J. Kasper, M. Frankel and Assoc.
1979 Total Survey Error: Applications to Improve Health Surveys. San Francisco: Jossey-Bass.

Anderson, J.P., J.W. Bush, and C.C. Berry
1977 "Performance vs. capacity: a conflict in classifying function for health status measures. Paper presented at annual meeting of American Public Health Association, Washington, DC.

Anderson, O.W., and R.M. Andersen
1972 "Patterns of use of health services." Pp. 386-406 in H.E. Freeman, S. Levine, and L.G. Reeder (eds.), Handbook of Medical Sociology (2d ed.). Englewood Cliffs, NJ: Prentice-Hall.

Andrews, F.M.
1979 "Estimating the construct validity and correlated error components of the rated effectiveness measures." In F.M. Andrews (ed.), Scientific Productivity. Cambridge: Cambridge University Press/UNESCO.

Andrews, F.M., and R. Crandall
1976 "The validity of measures of self-reported well-being." Social Indicators Research 3:1-19.

Andrews, F.M., and A. R. Herzog
1981 The quality of survey data as related to age of respondent. Unpublished paper. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Andrews, F.M., J.N. Morgan, J.A. Sonquist, and L. Klem
1973 Multiple Classification Analysis. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Andrews, F.M., and S.B. Withey
1974 "Assessing the quality of life as people perceive it." Presentation to the 1974 annual meeting of the American Psychological Association.
1976 Social Indicators of Well-Being: Americans' Perceptions of Life Quality. New York: Plenum.

Andrews, L.
1974 "Interviewers: recruiting, selecting, training and supervising." Pp. 124-132 in R. Feber (ed.), Handbook of Marketing Research. New York: McGraw-Hill.

Aneshensel, C.S., R.R. Frerichs, V.A. Clark, and P.A. Yokopenic
1982a "Measuring depression in the community: A comparison of telephone and personal interviews." Public Opinion Q. 46:110-121.
1982b "Telephone versus in-person surveys of community health status." American J. of Public Health 72:1017-1021.

Armstrong, B.
1979 "Test of multiple frame sampling techniques for agricultural surveys: New Brunswick, 1978." Pp. 295-300 in Proceedings, Survey Research Section, American Statistical Association.

Arthur D. Little, Inc.
1980 "A Study of Consumer Health Insurance Education." Contract No. 200-79-0909. Final Report submitted to Bureau of Health Education, Center for Disease Control, Department of Health and Human Services.

Asher, H.B.
1974 "Some consequences of measurement error in survey design." American J. of Political Science 18:469-485.

Australian Bureau of Statistics
1977-78 Outline of Concepts, Methodology and Procedures Used. Australian Health Survey, 1977-78. Belconnen: Australian Bureau of Statistics.

Australian Bureau of Statistics and Health Commission of NSW
1976 Health Care Surveys in Gosford-Wyong and Illawarra Areas of NSW, 1975. Australian Bureau of Statistics and Health Commission of NSW. Sydney.

Bailar, B.A.
1976 "Some sources of error and their effect on census statistics." Demography 13:273-286.

Barioux, M.
1952 "A method for selection, training and evaluation of interviewers." Public Opinion Q. (Spring):128-30.

Beed, T.W.
1981 Williams Committee Surveys: National Educational Survey, 1977; National Survey of Post-Secondary Teaching Staff, 1977. SSRC Survey Archive, Data Set No. 1477 2 Class 11 (SP).

Beed, T.W., M. Goot, S. Hodgson, and P. Ridley
1978 Australian Opinion Polls, 1941-1977. Hale and Iremonger and the University of Sydney Sample Survey Centre.

Beller, N.D.
1979 "Error profile-multiple frame designs." Pp. 221-222 in Proceedings, Survey Research Section, American Statistical Association.

Bentler, P.M.
1980 "Multivariate analysis with latent variables: Causal modeling." Annual Review of Psychology 31:419-456.

Bentler, P.M., and D.G. Bonett
1980 "Significance tests and goodness of fit in the analysis of covariance structures." Psychological Bulletin 88:588-606.

Bernadt, M.W., *et al.*
1982 "Comparison of questionnaire and laboratory tests in the detection of excessive drinking and alcoholism." The Lancet, No. 8267, February 6.

Berry, C.C., and J.W. Bush
1978 "Estimating prognoses for a dynamic health index, the weighted life expectancy, using the multiple logistic with survey and mortality data." Pp. 716-721 in Proceedings, Social Statistics Section, American Statistical Association.

Bielby, W.T., and R.M. Hauser
1977 "Response error in earnings functions for nonblack males." Sociological Methods and Research 6:241-280.

Bielby, W.T., R.M. Hauser, and D.L. Fetherman
1977a "Response errors of black and nonblack males in models of the intergenerational transmission of socioeconomic status." American J. of Sociology 82:1242-1288.
1977b "Response errors of nonblack males in models of the stratification process." J. of the American Statistical Association 72:723-735.

Bingham, W., B.V. Moore, and J.W. Gustad
1959 How to Interview. New York: Harper and Bros.

Birenbaum, A. and M. Re
1979 "Resettling mentally rtarded adults in the community—almost four years later." American J. of Mental Deficiency 83:323-329.

Blankenship, A.B.
1977 Professional Telephone Surveys. New York: McGraw-Hill.

Bliesch, W.
Undated "Interviewer guidance, supervision and control." In Seminar on Fieldwork Sampling and Quesionnaire Design, Part 1. Amsterdam: European Society for Opinion and Marketing Research.

Block, J.A., D. Bourque, R. Froh, and J.G. Lear
1978 "Physicians and hospitals: providing primary care." Medical Group Management 25(Mar.-Apr.):35-38.

Block, J.A., L. Brideau, A.K. Burns, and J.G. Lear
1980 "Hospital-sponsored primary care: The Community Hospital Program." J. of Ambulatory Care Management 3(Feb.):1-13.

Bloomenthal, A.M., R. Jackson, S. Kerachsky, S. Stephens, C. Thornton, and K. Zeldis

1982 SW/STETS Evaluation: Analysis of Alternative Data-Collection Strategies. Princeton, NJ: Mathematica Policy Research.

Bogen, D., and D. Aanes
1975 "The ABS as a tool in comprehensive MR programming." Mental Retardation 13:38-41.

Bohrnstedt, G.W., and E.F. Borgatta
1981 Social Measurement: Current Issues. Beverly Hills, CA: Sage.

Bohrnstedt, G.W., and T.M. Carter
1971 "Robustness in regression analysis." In H.L.Costner (ed.), Sociological Methodology 1971. San Francisco: Jossey-Bass.

Bollen, K.A., and K.H. Barb
1981 "Pearson's r and coarsely categorized measures." American Sociological Review 46:232-239.

Bonham, G.S., and L.S. Corder
1981 National Health Care Expenditure Study, Instruments and Procedures 1: NMCES Household Interview Instruments. DHHS Publication No. (PHS) 81-3280. U.S. Department of Health and Human Services, National Center for Health Services Research.

Boruch, R.F., J.D. Larkin, L. Wolins, and A.C. MacKinney
1970 "Alternative methods of analysis: Multitrait-multimethod data." Educational and Psychological Measurement 30:833-853.

Bosecker, R.R.
1978 "Evaluating alternative methods for determining overlap domain in multiple frame surveys." Pp. 325-330 in Proceedings, Survey Research Section, American Statistical Association.

Bosecker, R.R., and B.L. Ford
1976 "Multiple frame estimation with stratified overlap domain." Report, Sample Survey Research Branch, Statistical Reporting Service, USDA.

Bowman, C.E.
1981 "Blood Pressure errors with aneroid sphygmomanometer." Lancet 1(8227):1005.

Boyd, J.L., Jr., and B. Shimberg
1971 Handbook for Performance Testing:A Practical Guide for Test Makers. Princeton, NJ: Educational Testing Service.

Bradburn, N.
1979 "Respondent burden." Pp. 49-53 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79-3207. Hyattsville, MD: NCHSR.

Bright, M.
1967 "A follow-up study of the Commission on Chronic Illness morbidity survey in Baltimore. I. Tracing a large population sample over time." J. of Chronic Diseases 20: 707-716.
1969 "A follow-up study of the Commission on Chronic Illness morbidity survey in Baltimore. III. Residential mobility and prospective studies." J. of Chronic Diseases 21: 749-759.

Brolin, D.
1972 "Value of rehabilitation services and correlates of vocational success with mentally retarded." American J. of Mental Deficiency 77:644-651.

Brorsson, B.
1980 Mätfel i hälsointervjuer: Litteraturöversikt och studier av intervjuareffkter. Urval, Skriftserie utgiven av Statistiska centralbyrån, Nr 12, Stockholm.

Brownlea, A.A., and C.L.Ward
1976 "Health care access problems in relatively isolated communities in Northern Queensland and the Darling Downs." Pp. 174-192 in Papers of the First Australian and New Zealand Regional Science Association and Department of Geography, University of Queensland, Brisbane.

Bucher, R., C.E. Fritz, an E.L. Quarantelli
1956 "Tape recorded interviews in social research." American Sociological Review 21: 354-364.

Burnham, C.E., and J.T. Massey
1980 "Redesign of the National Health Interview Survey." Pp. 115-118 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Bush, J.W., J.P. Anderson, R.M. Kaplan, and W.R. Blischke
1982 "'Counterintuitive' preferences in health-related quality-of-life measurement." Medical Care XX (5).

Bush, J.W., W.R. Blischke, and C.C. Berry
1975 "Health indexes, outcomes, and quality of medical care." Pp. 313-339 in R. Yaffee and D. Zalkind (eds.), Evaluation in Health Services Delivery. New York: Engineering Foundation.

Bush, J.W., M.M. Chen, and D.L. Patrick
1973 "Cost effectiveness using a health status index: analysis of the New York State PKU screening program." Pp. 172-208 in R. Berg (ed.), Health Status Indexes. Chicago: Hospital Research and Educational Trust.

Bush, J.W., M. Chen, and J. Zaremba
1971 "Estimating health program outcomes using a Markov equilibrium analysis of disease development." American J. of Public Health 61:2362-2375.

Bush, J.S., and S. Fanshel
1970 "Measuring health system output using a health status index." P. 2 in C. Hopkins (ed.), Outcomes Conference I-II. USD/HEW, Public Health Service, Health Services and Mental Health Administration.

Bush, J.W., S. Fanshel, and M. Chen
1972 "Analysis of a tuberculin testing program using a health status index." J. of Socio-Economic Planning Sciences 6:49-69.

Bush, J.W., A.M. Schneider, T.L. Wachtel, and J.E. Brimm
1983 "A simulation analysis of plasma water dynamics and treatment in acute burn resuscitation." Paper presented at the American Burn Association Meeting, New Orleans, LA.

Bushery, J.M., C.D. Cowan, and L.R. Murphy
1978 "Experiments in telephone-personal visit surveys." Pp. 564-569 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Busse, E.W., and E. Pfeiffer (eds.)
1977 Behavior and Adaptation in Late Life (2d ed.). Boston: Little, Brown and Company.

Camburn, D.
1980 Performance in the Daily Health Records. Technical Report No. 5. Health in Detroit Study. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Campbell, A., P.E. Converse, and W.L. Rodgers
1976 The Quality of American Life: Perceptions, Evaluations, and Satisfaction. New York: Russell-Sage Foundation.

Campbell, D.T., and D.W. Fiske
1959 "Convergent and discriminant validation by the multimethod-multitrait matrix." Psychological Bulletin 56:81-105.

Campion, J.E.
1972 "Work sampling for personnel selection." J. of Applied Psychology 55:40-44.

Cannell, C.F., G. Fisher, and T. Bakker
1965 Reporting of Hospitalization in the Health Interview Survey. Vital and Health Statistics, Series 2, No. 6. Washington, DC: U.S. Government Printing Office.

Cannell, C.F., and F.J. Fowler
1964 "A note on interviewer effect in self-enumerative procedures." American Sociological Review 29:276.
1965 Comparison of Hospital Reporting in Three Survey Procedures. Vital and Health Statistics, Series 2, No. 8. Washington, DC: U.S. Government Printing Office.
1977 "Interviewers and interviewing techniques." Pp. 13-23 in U.S. National Center for Health Services Research, Advances in Health Survey Research Methods, Proceedings of a National Conference, 1975. DHEW Pub. No. (HRA) 77-3154. Rockville, MD: NCHSR.

Cannell, C.F., R.M. Groves, and P.V. Miller
1981 "The effects of mode of data collection on health survey data." Pp. 1-6 in Proceedings, Section on Social Statistics, American Statistical Association.

Cannell, C.F., S.A. Lawson, and D.L. Hausser
1975 A Technique for Evaluating Interviewer Performance. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Cannell, C., P. Miller, and L. Oksenberg
1981 "Research on inverviewing techniques." Pp. 389-437 in S. Leinhardt (ed.), Sociological Methodology. San Francisco: Jossey-Bass.

Cannell. C.F., L. Oksenberg, and J.M. Converse
1977a Experiments in Inverviewing Techniques: Field Experiments in Health Reporting: 1971-1977. Hyattsville, MD: National Center for Health Services Research (HRA) 78-3204. Reprinted 1979, ISR Research Report. Ann Arbor, MI: Survey Research Center, Institute for Social Research, The University of Michigan.
1977b "Striving for response accuracy: Experiments in new interviewing techiques." J. of Marketing Research 14(August):306-315.

Cannell, C.F., O. Thornberry, Jr., and R. Fuchsberg
1981 "Research on the reduction of response error: The National Health Interview Survey." Pp. 206-210 in Proceedings, Social Statistics Section, American Statistical Association.

Caplan, R.D., S. Cobb, J.R.P. French, Jr., R.V. Harrison, and S.R. Pinneau, Jr.
1975 Job Demands and Worker Health. Washington, DC: National Institute for Occupational Safety and Health.
1980 Job Demands and Worker Health. ISR Research Report. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Cartwright, A.
1967 "The families and individual who did not cooperate on a sample survey." The Milbank Memorial Fund Quarterly:347-368.

Casady, R.J., and M.G. Sirken
1980 "A multiplicity estimator for multiple frame sampling." Pp. 601-605 in Proceedings, Survey Research Section, American Statistical Association.

Casady, R.J., C.B. Snowden, and M.G. Sirken
1981 "A study of dual frame estimators for the National Health Interview Survey." Pp. 444-447 in Proceedings, Survey Research Methods Section, American Statistical Association.

Center for Disease Control
1980 Weighing and Measuring Children: A Training Manual for Supervisory Personnel.

DHHS (PHS) publication. Washington, DC: Government Printing Office.

Chen, M., and J.W. Bush
1976 "Maximizing health system output with political and administrative constraints using mathematical programming." Inquiry 13:215-227.

Chen, M., J.W. Bush, and J. Zaremba
1975 "Effectiveness measures." Pp. 276-301 in L.J. Shuman, R.D. Speas, and J.P. Young (eds.), Operations Research in Health Care—A Critical Analysis. Baltimore: The Johns Hopkins University Press.

Chromy, J.R.
1981 "Variance estimators for a sequential sample selection procedure." In D. Krewski *et al.* (eds.), Current Topics in Survey Sampling. New York: Academic Press.

Cleland, E.A., N. Kennedy, and R.J. Stimson
1973 "A quasi-experimental investigation of the accessibility of various forms of medical services: the adequacy of children's medical services in Adelaide." Paper presented to the annual conference of the Sociological Association of Australia and New Zealand. Perth.

Cleland, E.A., R.J. Stimson, and A.J. Goldsworthy
1977a "Access costs of households in isolated areas in using city-based high order health care services: a case study in Port Lincoln." Pp. 112-122 in ANZSERCH Proceedings, the proceedings of annual conference of Australian and New Zealand Society for Epidemiology and Research in Community Health. Adelaide.
1977b Suburban Health Care Behaviour in Adelaide. Research Monograph Series 2. Adelaide: Centre for Applied Social & Survey Research, Flinders University.

Cochran, W.G.
1967, 1977 Sampling Techniques. New York: Wiley and Sons.
1968 "The effectiveness of adjustment by subclassifications in removing bias in observational studies." Biometrics 24:295-313.
1970 "Some effects of errors of measurement on multiple correlation." J. of the American Statistical Association 65:22-34.

Cohen, S.B.
1981 An Analysis of Alternative Imputation Strategies for Individual with Partial Data in the National Medical Care Expenditure Survey. Paper presented at the annual meeting of the American Public Health Association Statistics Section. Los Angeles.

Cohen, S.B., and W.D. Kalsbeek
1981 National Health Care Expenditures Study, Instruments and Procedures 2: NMCES Estimation and Sampling Variance in the Household Survey, DHHS publication No. (PHS) 81-3281. U.S. Department of Health and Human Services, National Center for Health Services Research.

Collins, W.A.
1970 "Interviewers' verbal idiosyncrasies as a source of bias." Public Opinion Q. 37:416-422.

Commission for the Control of Epilepsy and Its Consequences
1978 Plan for Nationwide Action on Epilepsy. DHEW Pub. No. (NIH) 78-276.

Conger, A.J.
1971 "An evaluation of multimethod factor analysis." Psychological Bulletin 75:416-420.

Conner, R.J.
1972 "Grouping for testing trends in categorical data." J. of the American Statistical Association 67:601-604.

Cooley, P.C.
1981 "NMCES matching of MPS and household summary data methodology report." Report prepared by Research Triangle Institute for NCHSR under Contract No. HRA 230-76-0268. Unpublished.

Coombs, L., and R. Freedman
1964 "Use of telephone interviews in a longitudinal fertility study." Public Opinion Q. 28:112-117.

Corcoran, M.
1980 "Sex differences in measurement error in status attainment models." Sociological Methods and Research 9:199-217.

Corson, J.S.
1979 Trends in public attitudes toward survey research and what we can do about them. Paper presented at the 34th annual conference of the American Association for Public Opinion Research, Buck Hill Falls, PA.

Cox, B.
1979 "Medical provider survey imputation strategy: utilization variables." Working paper No. 1 prepared by Research Triangle Institute for NCHSR under Contract No. HRA 230-76-0268.
1980 "Construction of sample weights for the medical provider survey." Report prepared by Research Triangle Institute for NCHSR under Contract No. HRA 230-76-0268. Unpublished.

Cox, E.P., III
1980 "The optimal number of response alternatives for a scale: A review." J. of Marketing Research 17:407-422.

Crider, D.M., F.K. Willits, and R.C. Bealer
1971 "Tracking respondents in longitudinal surveys." Public Opinion Q. 35:613-620.

Cronbach, L.J., and P.E. Meehl
1955 "Construct validity in psychological tests." Psychological Bulletin 52:281-302.
Crowne, D.P., and D. Marlowe
1964 The Approval Motive: Studies in Evaluative Dependence. New York: Wiley.
Cutler, T.A., and K.F. Sharp
1981 "Telephone interviewing—some aspects of using the telephone network for interviewing." The Theory and Techniques of Interviewing, Advanced Training Workshop in Survey Methods. Sample Survey Centre and Centre for Applied Social and Survey Research.
Dalkey, N.C., D.L. Rourke, R. Lewis, and D. Snyder
1972 Studies in the Quality of Life. Lexington, MA: Lexington Books, DC Heath & Co.
Davidson, S.M.
1978 "Variations in state Medicaid programs." J. of Health Politics, Policy, and Law 3:54-70.
Davidson, S.M., J.D. Perloff, P.R. Kletke, D.W. Schiff and J.P. Connelly
1982 Variations by State in Physician Participation in Medicaid. Final report of HCFA grant No. 18-P-97159/5.
Deming, W.E.
1960 Sample Design in Business Research. New York: Wiley and Sons.
Dillman, D.A.
1978 Mail and Telephone Surveys; The Total Design Method. New York: John Wiley and Sons.
Dillman, D.A., J.G. Gallegos, and J.H. Frey
1976 "Reducing refusal rates for telephone interviews." Public Opinion Q. 40:66-78
Eckland, B.K.
1968 "Retrieving mobile cases in longitudinal surveys." Public Opinion Q. 32:51-64.
Eilertsen, E., and S. Hummerfelt
1968 "The observer variation in the measurement of arterial blood pressure." Acta Medica Scandinavica 184:145-157.
Eldred, C.A., T.D. Woolsey, W.R. Simmons, G.K. White, and C.M. Haines
1977 "A Report of the Survey of Intracranial Neoplasms." Final Contract Report prepared for the National Institute of Neurological and Communicative Disorders and Stroke.
Elison, J., L. Jimenez, S. Fisher, and J.U. Davis
1977 "Changes in drug use among high school students: some antecedents." In Center for Socio-Cultural Research on Drug Use, Final Report: A Study of Teen-age Drug Behavior. New York: Columbia University.
Environmental Protection Administration, Office of Research and Monitoring
1973 The Quality of Life Concept: A Potential New Tool for Decision-Makers. Washington, DC: U.S. Government Printing Office.
Erlich, J., and D. Reisman
1961 "Age and authority in the interview." Public Opinion Q. 25:39-56.
Evans, J.G., and G. Rose
1971 "Hypertension." British Medical Bulletin 27:37-42.
Eyman, R., G. Demaine, and T. Lei
1979 "Relationship between community environments and resident changes in adaptive behavior: a path model." American J. of Mental Deficiency 83:330-338.
Fanshel, S., and J.W. Bush
1970 "A health-status index and its application to health-service outcomes." Operations Research 18:1021-1066.
Farley, P., and G. Wilensky
1982 "Options, incentives, and employment-related health insurance coverage." In Advances in Health Economics and Health Services Research, Vol. 4. Greenwich, CT: JAI Press, in press.
Farquhar, J.W., N. Maccoby, P.D. Wood, *et al.*
1977 "Community education for cardiovascular health." Lancet (June 4):1192-1195.
Fellegi, I.P.
1974 "An improved method of estimating the correlated response variance." J. of the American Statistical Association 69:496-501.
Finney, D.J.
1960 An Introduction to the Theory of Experimental Design. Chicago: University of Chicago Press.
Fisher, R.S., and Yates, F.
1953 Statistical Tables for Biological, Agricultural and Medical Research (6th ed.). Edinburgh: Oliver and Boyd.
Fitti, J.E.
1979 "Some results from the Telephone Health Interview System." Pp. 244-49 in Proceedings, Section on Survey Research Methods, American Statistical Association.
Frankel, J.
1980 Measurement of Respondent Burden: Study Design and Early Findings. BSSR Report No. 0529-8. Washington, DC: Bureau of Social Science Research.
Frankel, J., and L. Sharp
1981 "Measurement of respondent burden: summary of study design and early findings." Statistical Reporter 81-4:105-111.
Friedman, P.A.
1942 "A second experiment on interviewer bias." Sociometry 15:378-381.
Fuller, W.A., and L.F. Burmeister
1972 "Estimators for samples selected from two overlapping frames." Pp. 245-249 in Pro-

ceedings, Social Statistics Section, American Statistical Association.

Galen, R.S., and S.R. Gambino
1975 Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses. New York: Wiley-Biomedical

Gerjony, I., and J. Winters
1966 "Lateral preference for identical geometric forms: 11. Retardates." Perception and Psychophysics 1:104-106.

Gleason, C.P., and R.D. Tortora
1978 "Successive sampling of two overlapping frames." Pp. 320-324 in Proceedings, Survey Research Section, American Statistical Association.

Golding, S.L., and E. Seidman
1974 "Analysis of multitrait-multimethod matrices: A two step principal components procedure." Multivariate Behavioral Research 9:479-496.

Gollay, E., *et al.*
1978 Coming Back: The Community Experiences of Deinstitutionalized Mentally Retarded People. Cambridge, MA: Abt Books.

Gordon, T., F.E. Moore, D. Shurtleff, and T.F. Dawber
1959 "Some methodologic problems in the long-term study of cardiovascular disease—Observations in the Framingham Study." J. of Chronic Diseases 10:186.

Gordon, T., P. Sorlie, and W.B. Kannel
1976 "Problems in the assessment of blood pressure: The Framingham study." International J. of Epidemiology 5(4):327-334.

Grasso, K.
1980 Termination Interview. Technical Report No. 4. Health in Detroit Study. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Green, V.L., and E.G. Carmines
1979 "Assessing the reliability of composites." In K. F. Schuessler (ed.), Sociological Methodology 1980. San Francisco: Jossey-Bass.

Greenberg, B.G., and Members of the Panel
1981 "The National Health Interview Survey—recommendations by a Technical Consultant Panel." Pp. 18-23 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS) 81:3268. Hyattsville, MD: NCHSR.

Groves, R.M.
1979a "A researcher's view of the SRC computer-based interviewing system: Measurement of some sources of error in telephone survey data." Pp. 88-98 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS)81-3268. Hyattsville, MD: NCHSR.
1979b "Actors and questions in telephone and personal interview surveys." Public Opinion Q. 43:233-244.

Groves, R.M., and R.L. Kahn
1979 Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Guest, L.
1947 "A study of interviewer competence." International J. of Opinion and Attitude Research 1:17-30.

Guilford, J.P.
1954 Psychometric Methods (2d ed.). New York: McGraw-Hill.

Gutgesell, M., G. Verrell, and D. Labarthe
1981 "Pediatric blood pressure: Ethnic comparison in a primary care center." Hypertension 3:39-47.

Halpern, A., *et al.*
1977 "Assessing social and prevocational awareness in mildly and moderately retarded individuals." American J. of Mental Deficiency 82:266-272.

Hansen, M.H., W.N. Hurwitz, and M.A. Bershad
1961 "Measurement errors in censuses and surveys." Bulletin of the International Statistical Institute 38:359-374.

Hanson, R.H., E. Marks, and National Opinion Research Center
1958 "Influence of the interviewer on the accuracy of survey results." J. of the American Statistical Association 53:635-655.

Harris, B.S.H., *et al.*
1978a Development of a Feasible Data Collection Plan for Hospital Data in Florida: Work Plan for Data Collection. Research Triangle Institute Report No. RTI/1662/00-051. Research Triangle Park, NC.
1978b Development of a Feasible Data Collection Plan for Hospital Data in Florida: Structuring of Medical Records in Hospitals in Florida. Research Triangle Institute Report No. RTI/1662/00-021. Research Triangle Park, NC.

Harris, M. (ed.)
1956 The Social Survey: Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing. London: The Social Survey Division, Central Office of Information.

Hartley, H.O.
1962 "Multiple frame surveys." Pp. 203-206 in Proceedings, Social Statistics Section, American Statistical Association.
1974 "Multiple frame methodology and selected

applications." Sankya: The Indian J. of Statistics 36, Series C, Part 3:99-118.

Hauck, M., and M. Cox
1974 "Locating a sample by random digit dialing." Public Opinion Q. 38:253-56.

Hawthorne, V.M., and M. Smalls
1980 "Blood pressure and ambient temperature." British Medical J. 280(6213):567-568.

Health Services Research and Development Center, Johns Hopkins University
1977 Medical Economics Survey-Methods Study: Final report. Submitted to Division of Health Interview Statistics, National Center for Health Statistics. Baltimore, MD: HSRDC, Johns Hopkins University.

Health and Welfare Canada/Statistics Canada
1981 Health of Canadians: Report of the Canada Health Survey. Ottawa: Government of Canada.

Heise, D.R., and G.W. Bohrnstedt
1970 "Validity, invalidity, and reliability." In E. F. Borgatta and G. W. Borhnstedt (eds.), Sociological Methodology 1970. San Francisco: Jossey-Bass.

Held, D., L. Manheim, and J. Wooldridge
1978 "Physician acceptance of Medicaid patients." Staff Paper SP-78B-02. Princeton, NJ: Mathematica Policy Research.

Heller, R.F., G. Rose, H.D.T. Pedoe, *et al.*
1978 "Blood pressure measurement in the United Kingdom Heart Disease Prevention Project." J. of Epidemiology and Comm. Health 32:235-238.

Henning, C.R., H.F. Woltman, and C.T. Isaki
1978 "An application of multi-frame methodology and measurement error research for the 1976 registration and voting survey." Pp. 542-547 in Proceedings, Survey Research Section, American Statistical Association.

Hill, A.D., E.A. Chaples, M.T. Downey, L.D. Singell, D.M. Solzman, and G.M. Schwartz
1973 The Quality of Life in America: Pollution, Poverty, Power & Fear. Chicago: Holt, Rinehart and Winston, Inc.

Hill, D.H.
1978 A Methodological Note on Obtaining MCA Coefficients and Standard Errors. Working Paper #8008. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Hochstim, J.R.
1967 "A critical comparison of three strategies of collecting data from households." J. of the American Statistical Association 62 (September):976-989.

Holt, M.
1981 "The use of summaries of previously reported interview data in the National Medical Care Expenditure Survey: A comparison of questionnaire and summary data for medical provider visits." Pp. 228-246 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS) 81-3268. Hyattsville, MD: NCHSR.

Hopkins, C.E.
1970 Methodology of Identifying, Measuring and Evaluating Outcomes of Health Service Programs, Systems and Subsystems. Outcomes Conference 1-11. Washington, DC: U.S. Government Printing Office.

Horton, R.L., and D.J. Duncan
1978 "A new look at telephone interviewing methodology." Pacific Sociological Review 21:259-73.

Horvitz, D.G.
1952 "Sampling and field procedures of the Pittsburgh morbidity survey." Public Health Reports 67:1003-1012.

Hunt, J., and J. Zimmerman
1969 "Stimulating productivity in a simulated sheltered workshop setting." American J. of Mental Deficiency 74:43-49.

Hyman, H.H.
1954 Interviewing in Social Research. Chicago: University of Chicago Press.
1972 Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities. New York: Wiley.

Hypertension Detection and Follow-up Project (HDFP)
1978 "Variability of blood pressure and the results of screening in the hypertension detection and follow-up program." J. of Chronic Diseases 31:651-667.

Irvin, L., A. Halpern, and W. Reynolds
1977 "Reliability and validity of the social and prevocational battery for mildly retarded individuals." American J. of Mental Deficiency 81:603-605.

Jackson, D.N.
1969 "Multimethod factor analysis in the evaluation of convergent and discriminant validity." Psychological Bulletin 72:30-49.

Johnson, Robert Wood, Foundation. See Robert Wood Johnson Foundation.

Jones, C., P. Sheatsley, and A. L. Stinchcombe
1979 Dakota Farmers and Ranchers Evaluate Crop and Livestock Surveys. NORC Report No. 128. Chicago: National Opinion Research Center.

Jones, R.G.
1981 "Variations in household telephone access: implications for telephone surveys." The Theory and Techniques of Interviewing, Advanced Training Workshop in Survey

Methods. Sample Survey Centre and Centre for Applied Social and Survey Research.

Jordan, L.A., A.C. Marcus, and L.G. Reeder
1979 "Response styles in telephone and household interviewing: a field experiment from the Los Angeles Health Survey." Pp. 116-123 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS) 81-3268. Hyattsville, MD:NCHSR.
1980 "Response styles in telephone and household interviewing: A field experiment." Public Opinion Q. 44:210-222.

Jöreskog, K.G.
1978 "Structural analysis of covariance and correlation matrices." Psychometrika 43:443-477.

Jöreskog, K.G., and D. Sörbom
1978 LISREL IV Users Guide. Chicago: National Educational Resources.

Kahn, R.L., and C.F. Cannell
1957 The Dynamics of Interviewing. New York: John Wiley and Sons.

Kalsbeek, W.D., C.F. Powers, H.J. Harwood, B.S.H. Harris, III, J.R. Batts, T.D. Hartwell, and T.H. Rice
1977 "Final Report: Survey of the Incidence, Prevalence, and Costs of Head and Spinal Cord Injury." Final Contract Report prepared for the National Institute of Neurological and Communicative Disorders and Stroke.

Kalsbeek, W., and J. Lessler
1979 "Total survey design: Effect of nonresponse bias and procedures for controlling measurement errors." Pp. 19-41 in National Center for Health Services Research, Health Survey Research Methods, Second Biennial Conference, 1977. DHEW Publication No. (PHS)79-3207. Hyattsville, MD:NCHSR.

Kane, R.L., and R.L. Kane
1981 Assessing the Elderly. Lexington, MA: Lexington Books.

Kaplan, R.M., and J.W. Bush
1982 "Health-related quality of life measurement for evaluation research and policy analysis." Health Psychology 1:61-80.

Kaplan, R.M., J.W. Bush, and C.C. Berry
1976 "Health status: types of validity and the index of well-being." Health Services Research 11:478-507.
1978 "The reliability, stability, and generalizability of a health status index." Pp. 704-709 in Proceedings, Social Statistics Section, American Statistical Association.
1979 "Health status index: category rating versus magnitude estimation for measuring levels of well-being." Medical Care 17:501-523.

Karren, R.
1980a "Alternative Selection Devices: The Work Sample and the Interview." Paper presented at the National Council on Measurement in Education Convention. Boston
1980b The Work Sampling Approach to Personnel Selection (PRR-80-11). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center.

Katz, D.
1942 "Do interviewers bias poll results?" Public Opinion Q. 6:248-268.

Kempthorne, O.
1952 The Design and Analysis of Experiments. New York: Wiley and Sons.

Keogh, E.
1980 Selectivity of Interviewed Respondents. Technical Report No. 3. Health in Detroit Study. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Keogh, E., and D. Camburn
1982 Selectivity of Diary Respondents. Technical Report No. 6. Health in Detroit Study. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Kipp, D.M., *et al.*
1981 Pawtucket Heart Health Program Health Survey Field Manual. Unpublished monograph.

Kirkendall, W., M. Feinleib, E.D. Freis, *et al.*
1980 "Recommendations for human blood pressure determined by sphygmomanometer." Circulation 62(5):1146A-1155A.

Kish, L.
1965 Survey Sampling. New York: Wiley and Sons.

Klecka, W.R., and A.J. Tuchfarber
1978 "Random digit dialing: a comparison to personal surveys." Public Opinion Q. 42 (Spring):105-114.

Kluegel, J.R., R. Singleton, Jr., and C.E. Starnes
1977 "Subjective class identification: A multiple indicator approach." American Sociological Review 42:599-611.

Kovar, M.G.
1981 A Statistical Profile. Better Health for Our Children: The Report of the Select Panel for the Promotion of Child Health, Vol. 3. DHHS Pub. No. (PHS) 79-55071. Washington, DC.

Kovar, M.G., and R.A. Wright
1973 "An experiment with alternative respondent rules in the National Health Survey." Pp. 311-316 in Proceedings, Social Statistics Section, American Statistical Association.

Kramm, E.R., M.M. Crane, M.G. Sirken, and M.L. Brown
1962 "A cystic fibrosis pilot survey in three New

England states." American J. of Public Health 52:2041-2057.

Krotki, K.P.
1978 "Estimation of correlated response variance." Paper presented to the 1978 annual meeting of the American Statistical Association.

Krotki, K.P., and A. MacLeod
1979 "Two methods of measuring correlated response variance." Paper presented to the 1979 annual meeting of the American Statistical Association.

Krupinski, J., and A. Stoller (eds.)
1971 The Health of a Metropolis. Melbourne: Heinemann Educational.

Lalonde, M.
1974 A New Perspective on the Health of Canadians. Ottawa: Government of Canada.

Lambert, N., and R. Nicoll
1976 "Dimensions of adaptive behavior of retarded and nonretarded public school children." American J. of Mental Deficiency 81:135-46.

Langwell, K., and S. Moore
In Press "A Synthesis of the Research on Competition in the Financing and Delivery of Health Services." National Center for Health Services Research.

Lansing, J.B., and A.C. Wolfe
1971 Working Papers on Survey Research in Poverty Areas. Ann Arbor, MI: Survey Research Center, Institute for Social Research, The University of Michigan.

Leibowitz, U., and M. Alter
1973 Multiple Sclerosis: Clues to Its Causes. New York: American Elsevier Publishing Company.

Lenski, G., J. Leggett, and J. Caste
1969 "Class and deference in the research interview." American J. of Sociology 65:463-467.

Levin, J.
1974 "A rotational procedure for separation of trait, method, and interaction factors in multitrait-multimethod matrices." Multivariate Behavioral Research 9:231-240.

Levy, P.S.
1977a "Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations." J. of the American Statistical Association 72:758-763.

1977b "Estimation of rare events by simple cluster sampling with multiplicity." Pp. 963-966 in Proceedings, Social Statistics Section, American Statistical Association.

Liebow, E.
1967 Tally's Corner. Boston: Little, Brown and Co.

Linn, R.L., and C.E. Werts.
1973 "Errors of inference due to errors of measurement." Educational and Psychological Measurement 33:531-544.

Lissitz, R.W., and S.B. Greene.
1975 "Effect of the number of scale points on reliability: A Monte Carlo approach." J. of Applied Psychology 60:10-13.

Little, Arthur D., Inc. See Arthur D. Little, Inc.

Locander, W.S., S. Sudman, and N. Bradburn
1976 "An investigation of interviewer method, threat and response distortion." J. of the American Statistical Association 71 (June):269-275.

Lucas, D., *et al.*
1979a Implementation of Sample Data Collection Plan: Patient Record Selection. Research Triangle Institute Report No. RTI/1935/00-02I. Research Triangle Park, NC.

1979b Implementation of Sample Data Collection Plan: Hospital Selection. Research Triangle Institute Report No. RTI/1935/00/01I. Research Triangle Park, NC.

Lucas, D., and C. Leininger
1980 Implementation of Sample Data Collection Plan: Patient Sampling Methodology. Research Triangle Institute Report No. RTI/1935/00-03I. Research Triangle Park, NC.

Lucas, W.A., and W.C. Adams
1977 An Assessment of Telephone Survey Methods. R-2135-NSF. Santa Monica, CA: Rand.

McGlashan, N.D.
1978 "The Tasman Bridge collapse and its effects in metropolitan Hobart: medical facilities in a divided city." OP-3, Department of Geography, University of Tasmania, Hobart.

Marcus, A.C.
1982 "Memory aids in longitudinal health surveys: Results from a field experiment." American J. of Public Health 72:567-573.

Mare, R.D., and W.M. Mason
1980 "A multiple group measurement model of children's reports of parental socioeconomic status." Sociological Methods and Research 9:178-198.

Marks, E.S., and W.P. Mauldin
1950 "Response errors in Census research." J. of the American Statistical Association (Sept.):424-38.

Marquis, K.H.
1978 "Inferring health interview response bias from imperfect record checks." Pp. 265-270 in Proceedings, Survey Research Section,

American Statistical Association.

1979 "Survey response rates: Some trends, causes, and correlates." Pp. 3-12 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79-3207. Hyattsville, MD: NCHSR.

1980 Hospital Stay Response Error Estimates for the Health Insurance Study's Dayton Baseline Survey. Rand/R-2555-HEW, May.

Marquis, K.H., *et al.*

1979 "An Evaluation of Published Measures of Diabetes Self-Care Variables." N-1152-HEW. Santa Monica, CA:Rand.

Marquis, K.H., and C.F. Cannell

1971 Effect of Some Experimental Interviewing Techniques on Reporting in the Health Interview Study. Vital and Health Statistics, Series 2, No. 41. Washington, DC: U.S. Government Printing Office.

Marquis, K.H., C.F. Cannell, and A. Laurent

1972 Reporting on Health Events in Household Interviews: Effects of Reinforcement, Question Length, and Reinterviews. Vital and Health Statistics, Series 2, No. 45. Washington, DC: U.S. Government Printing Office.

Marquis, M.S.

1981 Consumers' Knowledge about Their Health Insurance Coverage. Santa Monica, CA: Rand.

Martin, W.S.

1978 "Effects of scaling on the correlation coefficient." J. of Marketing Research 15:304-308.

Martindale, D.

1968 "Verstehen." Pp. 308-312 in International Encyclopedia of the Social Sciences, Vol. 16. New York: The Macmillan Company and The Free Press.

Mason, W.M., R.M. Hauser, A.C. Kerckhoff, S.S. Poss, and K. Manton

1976 "Models of response error in student reports of parental socioeconomic characteristics." In W. H. Sewell, R. M. Hauser, and D. L. Fetherman (eds.), Schooling and Achievement in American Society. New York: Academic Press.

Massey, J.T.

1978 "New NCHS initiatives involving the Health Interview Survey." Pp. 589-593 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Massey, J.T., P.R. Barker, and S. Hsiung

1981 "An investigation of response in a telephone survey." Pp. 426-431 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Massey, J.T., P.R. Barker, and A.J. Moss

1979 "Comparative results of face-to-face and telephone interviews in a survey of cigarette smoking." Paper presented at the meetings of the American Public Health Association.

Miller, P.

1979 "Applying health interview techniques to mass media research." Pp. 101-113 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS)81-3268. Hyattsville, MD: NCHSR.

Miller, P., and C. Cannell

1977 "Communicating response objectives in the survey interview." Pp. 127-152 in P. M. Hirsch, P. V. Miller, and F. G. Kline (eds.), Strategies for Communication Research. Beverly Hills, CA: Sage.

Moffat, R.J., S.P. Sady, and G.M. Owen

1980 "Height, weight and skinfold thickness of Michigan adults." American J. of Public Health 70:1290-1292.

Monsees, M.L., and J.T. Massey

1979a "Application of personal interview survey definitions to a telephone survey." Pp. 7-14 in Proceedings, Social Statistics Section, American Statistical Association.

1979b "Adapting a procedure for collecting demographic data in a personal interview to a telephone interview." Pp. 130-133 in Proceedings, Social Statistics Section, American Statistical Association.

Mooney, H.W.

1962 Methodology in Two California Health Surveys. Public Health Monograph No. 70. PHS Pub. No. 942. Washington DC: Public Health Service.

Morgan, J.N., and J.A. Sonquist

1963 "Problems in the analysis of survey data and a proposal." J. of the American Statistical Association 58:415-435.

Morris, C.N., J.P. Newhouse, and R.W. Archibald

1980 "On the Theory and Practice of Obtaining Unbiased and Efficient Samples in Social Surveys." R-2173-HEW. Santa Monica, CA: Rand.

Moser, C.A., and G. Kalton

1971 Survey Methods in Social Investigation (2d ed.). London: Heinemann Educational Books.

Moustafa, A.T., C. Hopkins, and B. Klein

1971 "Determinants of choice and change of health insurance plan." Medical Care 9:32-41.

Mugford, J. (ed.)
1979 Australian Social Surveys: Journal Extracts 1974-1978. Survey Research Centre, Australian National University. ANU Press.

Namboodiri, N.K., L.F. Carter, and H.M. Blalock, Jr.
1975 Applied Multivariate Analysis and Experimental Design. New York: McGraw-Hill.

Nathan, G.
1976 "An empirical study of response and sampling error for multiplicity estimates with different counting rules." J. of the American Statistical Association 71:808-815.

Nathanson, C.A.
1978 "Sex roles as variables in the interpretation of morbidity data: a methodological critique." International J. of Epidemiology 7:253-262.

Newhouse, J.P., W. Manning, G. Willard, C.N. Morris, *et al.*
1981 "Some interim results from a controlled trial of cost sharing in health insurance." Pub. No. R-2847HHS. Santa Monica, CA: Rand.

Nihira, K.
1976 "Dimensions of adaptive behavior in institutionalized mentally retarded children and adults." American J. of Mental Deficiency 81:215-226.

O'Brien, E.T., and K. O'Malley
1979 "ABC of blood pressure measurement. Reconciling the controversied: A comment on 'the literature'". British Medical J. 2:1201-1202.

Office of Scientific and Health Reports, National Institute of Neurological and Communicative Diseases and Stroke.
1976 "Neurological and communicative disorders: Estimated number and cost." Washington, DC: U.S. Government Printing Office.

O'Leary, L.R.
1973 "Fair employment, sound psychometric practice, and reality: a dilemma and a partial solution." American Psychologist (Feb.):147-150.

Patrick, D.L., J.W. Bush, and M. Chen
1972 "Toward an operational definition of health." J. of Health and Social Behavior 14:6-23.
1973 "Methods for measuring levels of well-being for a health status index." Health Services Research 8(30):228-245.

Paul, O., M.J. Lepper, W.H. Phelan, G.W. Dupertius, A. MacMillan, H. McKean, and H. Park
1963 "A longitudinal study of coronary heart disease." Circulation 28:20.

Payne, S.L.
1974 "Data collection methods: telephone surveys." Chapter 4 in R. Feber (ed.), Handbook of Marketing Research. New York: McGraw-Hill.

Payne, S., C. Bain, M. Gibbings, G. Lupton, J. Najman, J. Ryan, M. Sheehan, P. Sheehan, and J. Western
1977 "Accessibility factors in health care utilization." Pp. 18-28 in ANZSERCH Proceedings, Proceedings of the annual conference of the Australian and New Zealand Society for Epidemiology and Research in Community Health, Adelaide.

Pearson, K.
1913 "On the measurement of the influence of 'broad categories' on correlation." Biometrika 9:116-139.

Percy, A.K., *et al.*
1971 "Multiple sclerosis in Rochester, Minnesota: A 60-year appraisal." Archives of Neurology 25:105-111.

Plumlee, L.B.
1980 A Short Guide to the Development of Work Sample and Performance Tests (2d ed., pp. 80-83). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development Center.

Puska, P., J. Tuomilehto, J. Salonen, *et al.*
1979 "Changes in coronary risk factors during comprehensive five-year community programme to control cardiovascular disease (North Karelia Project)." British Medical J. 2:1173-1178.

Quinn, R.P., B.A. Gutek, and J.T. Walsh
1980 "Telephone interviewing: A reappraisal and a field experiment." Basic and Applied Social Psychology 1:127-153.

Ramsey, J.O.
1973 "The effect of number of categories in rating scales on precision of estimation of scale values." Psychometrika 38(4):513-532.

Reeder, L.G.
1977 "Summary and conclusions." Pp. 1-3 in U.S. National Center for Health Services Research, Advances in Health Survey Research Methods, Proceedings of a National Conference, 1975. DHEW Publication No. (HRA)77-3154. Rockville, MD:NCHSR.

Reilly, R.R., and W.R. Manese
1979 "The validation of a minicourse for telephone company switching technicians." Personnel Psychology 32:83-90.

Research and Training Center on Mental Retardation
1980 A Manual for Training Volunteers. Lubbock, TX: Texas Technical University.

Richardson, S.
1979 "Careers of mentally retarded young persons: Services, jobs and interpersonal relations." American J. of Mental Deficiency 82:349-358.

Richardson, S.A., B.S. Dohrenwend, and D. Klein

1965 Interviewing—Its Forms and Functions. New York: Basic Books, Inc.

Riedel, D.C., D.C. Walden, S.M. Meyers, and R. Wilson (eds.)
In Press Use of Health Care Resources: A Comparative Study of Two Plans. Ann Arbor, MI: Health Administration Press, The University of Michigan.

Robert Wood Johnson Foundation
1974 Community Hospital-Medical Staff Sponsored Primary Care Group Practice Program of the Robert Wood Johnson Foundation. Princeton, NJ: RWJF.

Robings, M., J.I. Cahill, C.M. Haines, W.R. Simmons, and T.D. Woolsey
1978 "Nationwide Study of Stroke: Final Report." Final Contract Report prepared for the National Institute of Neurological and Communicative Disorders and Stroke.

Robins, P.K., and R.W. West
1977 "Measurement errors in the estimation of home value." J. of the American Statistical Association 72:290-294.

Robinson, D., and S. Rohde
1946 "Two experiments with an anti-Semitism poll." J. of Abnormal and Social Psychology 41:136-144.

Robles, R.
1974 Education Transition and Drug Use Progress Report, No. 7. Rio Piedras, Puerto Rico: Medical Sciences Campus of the University of Puerto Rico.

Rogers, T.F.
1976 "Interviews by telephone and in person: Quality of responses and field performance." Public Opinion Q. 40 (Spring):51-65.

Rose, G.A.
1965 "Standardization of observers in blood pressure measurement." Lancet 1:673-674.

Rosen, M., G.R. Clark, and M.S. Kivitz
1977 Habilitation of the Handicapped: New Dimensions in Programs for the Developmentally Disabled. Baltimore, MD: University Park Press.

Rosner, B., and B.F. Polk
1979 "The implications of blood pressure variability for clinical and screening purposes." J. of Chronic Diseases 32:451-461.

Rossiter, L.F., and S.B. Cohen.
1981 "Alternative measures of expenditures for the analysis of health services." In Proceedings of the 43rd conference of the International Statistical Association, Buenos Aires, Argentina.

Rothenberg, G.
1969 "Conservation of number among four- and five-year-old children: Some methodological considerations." Child Development 40:382-406.

Rothwell, N., and G. Bridge
1979 "Discussion of respondent burden." Pp. 55-61 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79-3207. Hyattsville, MD: NCHSR.

Rusch, F., and R.P. Schutz
1980 "Vocational and social work behavior research: An evaluative review." In J.L. Matson and J.R. McCartney (eds.), Handbook of Behavior Modification with the Mentally Retarded. Illinois: Plenum Press.

Sady, S.P., R.J. Moffatt, and G.M. Owen
1981 "Height, weight and triceps skinfold thickness of Michigan children, 1978." American J. of Public Health 71:855-858.

Sagen, O.K., R.E. Dunham, and W.R. Simmons
1961 Health Statistics from Record Sources and Household Interviews Compared. US DHEW, Public Health Service, Series D, No. 5.

Scheffe, T.H.
1959 The Analysis of Variance. New York: Wiley and Sons.

Schmidt, F.L., A.L. Greenthal, J.E. Hunter, J.G. Berner, and F.W. Seaton
1977 "Job sample vs. paper-and-pencil trades and technical tests: adverse impact and examinee attitudes." Personnel Psychology 30:187-197.

Schmitt, N., B.W. Coyle, and B.B. Saari
1977 "A review and critique of analyses of multitrait-multimethod matrices." Multivariate Behavioral Research 12:447-478.

Schuman, H., and J.M. Converse
1971 "Effects of black and white interviewers on black responses in 1968." Public Opinion Q. 35(Spring):44-68.

Sharp, L., and J. Frankel
1981 Correlates of Self-Perceived Respondent Burden: Findings from an Experimental Study. Proceedings of the annual meeting of the American Statistical Association, Detroit, MI.

Sheatsley, P.B.
1950 "An analysis of interviewer characteristics and their relationship to performance, part I." International J. of Opinion and Attitude Research 4:473-498.
1951a "An analysis of interviewer characteristics and their relationship to performance, part II." International J. of Opinion and Attitude Research 5:79-94.
1951b "An analysis of interviewer characteristics and their relationship to performance, part III." International J. of Opinion and Attitude Research 5:191-220.

Shepard, D.S.
1981 "Reliability of blood pressure measurements: Implications for designing and evaluating programs to control hypertension." J. of Chronic Diseases 34:191-201.

Shortell, S., and W. Dowling
1978 "Hospital-sponsored primary care: organizational and financial issues." Medical Group Management (May-June):16-21.

Siconolfi, S.F., E.M. Cullinane, R.A. Carleton, and P.D. Thompson
1982 "Assessing VO_{2MAX} in epidemiologic studies: modification of the Astrand-Ryhming test." Medicine and Science in Sports and Exercise 14(5): 335-338.

Siegel, A.I., and B.A. Bergman
1975 "A job learning approach to performance prediction." Personnel Psychology 28:325-339.

Siemiatycki, J.
1979 "A comparison of mail, telephone, and home interview strategies for household health surveys." American J. of Public Health 69 (March):238-245.

Sigelman, C.K., E.C. Budd, C.L. Stanhel, and C.J. Schoenrock
1981 "When in doubt, say yes: Acquiescence in interviews with mentally retarded persons." Mental Retardation 19(2):53-58.

Sigelman, C.K., *et al.*
1980 "Surveying mentally retarded persons: Responsivesness and response validity in three samples." American J. of Mental Deficiency 84:479-486.
1981 "Issues in interviewing mentally retarded persons: An empirical study." In R. Bruininks *et al.* (ed.), Deinstitutionalization and Community Adjustment of Mentally Retarded People. Washington, DC: American Association on Mental Deficiency.
Forthcoming Communicating with Mentally Retarded Persons: Asking Questions and Getting Answers. Lubbock, TX: Texas Technical University Research and Training Center in Mental Retardation.

Singer, E.
1978 "Informed consent: consequences for response rate and response quality in social surveys." American Sociological Review 43:144-162.
1979 "Telephone interviewing as a black box—Discussion: Response styles in telephone and household interviewing." Pp. 124-127 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS) 81-3268. Hyattsville, MD: NCHSR.

Singer, E., and M.R. Frankel
1982 "Informed consent procedures in telephone interviews." American Sociological Review 47:416-426.

Singer, E., M.R. Frankel, and M.B. Glassman
1983 "The effect of interviewer characteristics and expectations on response." Public Opinion Q. 47 (Spring):68-83.

Sirken, M.G.
1970a "Household surveys with multiplicity." J. of the American Statistical Association 65:257-266.
1970b "Survey strategies for estimating rare health attributes." Pp. 133-144 in Proceedings, Sixth Berkeley Symposium on Statistics and Probability.
1972a "Stratified sample surveys with multiplicity." J. of the American Statistical Association 67:224-227.
1972b "Variance components of multiplicity estimators." Biometrics 28:869-873.
1975 "Network survey." Pp. 332-342 in Proceedings, Bulletin of the International Statistical Institute, 40th Session, Warsaw.

Sirken, M.G., M.M. Crane, M.L. Brown, and E.R. Kramm
1959 "A national hospital survey of cystic fibrosis." Public Health Report 74:764-770.

Sirken, M.G., G.P. Inderfurth, C.E. Burnham, K.M. Danchik
1975 "Household sample surveys of diabetes: Design effects of counting rules." Pp. 659-663 in Proceedings, Social Statistics Section, American Statistical Association.

Sirken, M.G., and P.S. Levy
1974 "Multiplicity estimation of proportions based on ratios of random variables." Journal of the American Statistical Association 69:68-74.

Sloan, F., J. Cromwell, and J.B. Mitchell
1978 Private Physicians and Public Programs. Lexington, MA: D.C. Heath.

Smith, J.M.
1972 Interviewing in Market and Social Research. London: Routledge and Kegan Paul, Ltd.

Snedecor, G.W., and W.G. Cochran
1967 Statistical Methods (6th ed.). Iowa City, IA: The Iowa State University Press.

Sobol, M.G.
1959 "Panel mortality and panel bias." American Statistical Association J. 54(285):52-68.

Sonquist, J.A., E.L. Baker, and J.N. Morgan.
1974 Searching for Structure (rev ed.). Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Sorbom, D.

1975 "Detection of correlated errors in longitudinal data." British J. of Mathematical and Statistical Psychology 28:138-151.

Soucheh, J., J. Stamler, A.J. Dyer, *et al.*
1979 "The value of two or three vs. a single reading of blood pressure at a first visit." J. of Chronic Diseases 32:197-210.

Spradlin, J.
1964 "Language and communication of mental defectives." In N.R. Ellis (ed.), Handbook of Mental Deficiency. New York: McGraw-Hill.

Steeh, C.G.
1981 "Trends in nonresponse rates, 1952-1979." Public Opinion Q. 45 (Spring):40-57.

Stine, O.C., R. Hepner, and R. Greenstreet
1975 "Correlation of blood pressure with skinfold thickness and protein levels." American J. of Diseases of Children 129:905.

Sudman, S.
1976a Applied Sampling. New York: Academic Press.
1976b "Sample surveys." Annual Review of Sociology 2:107-120.

Sudman, S., and N.M. Bradburn
1974 Response Effects in Surveys. Chicago: Aldine.

Sudman, S., N.M. Bradburn, E. Blair, and C.B. Stocking
1977 "Modest expectations: The effects of interviewers' prior expectations on response." Sociological Methods and Research 7:177-182.

Sudman, S., and R. Ferber
1974 "A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products." J. of Marketing Research 11 (May):128-135.

Sudman, S., and L.B. Lannom
1980 Health Care Surveys Using Diaries. Research Report Series. DHHS Publication No. PHS 80-3279. Hyattsville, MD: National Center for Health Services Research.

Sudman, S., W. Wilson, and R. Ferber
1974 The Cost-Effectiveness of Using the Diary as an Instrument for Collecting Health Data in Household Surveys. Report to the Bureau of Health Services Research and Evaluation. Urbana, IL: Survey Research Laboratory, University of Illinois.

Sullivan, J.L., and S. Feldman
1979 Multiple Indicators: An Inroduction. Beverly Hills, CA: Sage.

Sumner, J.
1976 The 1976 LAMAS Frame and Master Sample: Technical Description. Los Angeles: Institute for Social Science Research.

Survey Research Center, Computer Support Group
1981 OSIRIS IV: Statistical Analysis and Data Management Software System. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Taylor, G.
1979 "Observations on the behavior of automated telephone interviewing." Pp. 99-100 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS)81-3268. Hyattsville, MD: NCHSR.

Thomas, P.H.
1980 Trainability Testing: The Miniature Training and Evaluation Approach to Selection (PRR-80-9). Washington, DC: Office of Personnel Management, Personnel Research and Development Center.

Thornberry, O.T., Jr., and J.T. Massey
1978 "Correcting for undercoverage bias in random digit dialed national health surveys." Pp. 224-229 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Thulein, T., G. Anderson, and B. Schersten
1975 "Measurement of blood pressure—a routine test in need of standardization." Postgraduate Medicine 51(596):390-395.

Tietjen, G.L.
1974 "Exact and approximate tests for unbalanced random effects designs." Biometrics 30:573-581.

Tuchfarber, A.J., and W.R. Klecka
1976 Random Digit Dialing: Lowering the Cost of Victimization Surveys. Washington, DC: Police Foundation.

Tucker, C.
1983 "Interviewer effects in telephone surveys." Public Opinion Q. 47 (Spring):89-95.

U.S. Civil Service Commission
1973 Job Analysis—Developing and Documenting Data: A Guide for State and Local Governments (BIPP 152-35). Washington, DC: U.S. Civil Service Commission, Bureau of Intergovernmental Programs.
1975 Job Analysis for Improved Job-Related Selection: A Guide for State and Local Governments (BIPP 152-63). Washington, DC: U.S. Civil Service Commission, Bureau of Intergovernmental Programs.

U.S. Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration.
1974 Data Evaluation and Methods Research, Series 2, No. 61. National Ambulatory Medical Care Survey: Background and Methodology. Rockville, MD: DHEW Publications No. (HRA) 74-1335.

U.S. Department of Health and Human Services,

Health Care Financing Administration
1980 "Medicare program; schedule of limits on hospital inpatient general routine operating costs reporting periods beginning on or after July 1, 1980." Federal Register, Part VIII, Vol. 45, No. 121, pp. 41868-41880.

U.S. Department of Labor, Manpower Administration
1972 Handbook for Analyzing Jobs. Washington, DC: U.S. Government Printing Office.

U.S. Equal Employment Opportunity Commission, U.S. Office of Personnel Management, U.S. Department of Labor, U.S. Department of Justice and U.S. Department of Treasury
1978 "Uniform guidelines on employee selection procedures (1978)." Federal Register 43 (166):38290-38315.

U.S. National Center for Health Services Research
1977 Advances in Health Survey Research Methods: Proceedings of a National Invitational Conference, 1975. DHEW Pub. No. (HRA) 77-3154. Rockville, MD: NCHSR.
1979 Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS)79-3207. Hyattsville, MD: NCHSR.
1981 Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS)81-3268. Hyattsville, MD: NCHSR.
1981-83 National Health Care Expenditures Study, Data Preview Series, Nos. 8-11 and 14-16. Washington, DC.

U.S. National Center for Health Statistics
1958 The Statistical Design of the Health Household Interview Survey. Health Statistics. PHS Pub. No. 584-A2. Public Health Service. Washington, DC.
1963 Measurement of Personal Health Expenditures. Vital and Health Statistics, Series 2, No. 2. Washington, DC: U.S. Government Printing Office.
1966 "Interview response on health insurance compared with insurance records—United States—1960." U.S. Department of Health, Education and Welfare.
1970 "Estimation and sampling variance in the health interview survey." In Vital and Health Statistics, Series 2, No. 38. PHS Pub. No. 1000. Public Health Service. Washington, DC: U.S. Government Printing Office.
1973 Plan and operation of the Health and Nutrition Examination Survey (United States—1971–1973). Vital and Health Statistics, Series 1, No. 10a. DHEW Pub. No. (PHS) 79-1310. Washington, DC: U.S. Government Printing Office.
1977 Plan and operation of the Health and Nutrition Examination Survey (United States—1971–1973). Vital and Health Statistics. Series 1, No. 10b. Washington, DC: U.S. Government Printing Office.
1979 Personal Out-of-Pocket Health Expenses, 1975. Series 10, No. 122. PHS Pub. No. 79-1550. Public Health Service. Washington, DC: U.S. Government Printing Office.
1981 Plan and operation of the Health and Nutrition Examination Survey: 1976–1980. Vital and Health Statistics. Series 1, No. 15. Washington, DC: U.S. Government Printing Office. DHHS Pub. No. (PHS) 81-1317.

Velez, C.N.
1981 Drug Use among Puerto Rican Youth: An Exploration of Generational Status Differences. Unpublished Ph.D. Dissertation. New York: Columbia University.

Verbrugge, L.M.
1979 "Female illness rates and illness behavior: Testing hypotheses about sex differences in health." Women and Health 4:61-79.
1980a "Health diaries." Medical Care 18:73-95.
1980b "Sensitization and fatigue in health diaries." Pp. 666-671 in Proceedings, Section on Survey Research Methods, American Statistical Association.

Verbrugge, L.M., and C.E. Depner
1981 "Methodological analyses of Detroit health diaries." Pp. 144-158 in U.S. National Center for Health Services Research, Health Survey Research Methods: Third Biennial Conference, 1979. DHHS Pub. No. (PHS) 81-3268. Hyattsville, MD: NCHSR.

Viol, G.W., M. Goebel, G.J. Lrenz, *et al.*
1979 "Seating as a variable in clinical blood pressure measurement." American Heart J. 98 (6):813-814.

Vogel, F.A.
1975 "Surveys with overlapping frames—Problems in application." Pp. 694-699 in Proceedings, Social Statistics Section, American Statistical Association.

Walker Research, Inc.
1978 "1978 industry image survey results." The Marketing Researcher (November). 800 Kune Road, Indianapolis, IN.

Walls, R., and T. Werner
1979 "Vocational behavior checklists." Mental Retardation (August):30-35.

Walmsley, D.J.
1978 "The influence of distance on hospital usage in rural New South Wales." Australian J. of Social Issues 13:71-81.

Watts, D.L., and D. Melroy
1980 Implementation of Sample Data Collection Plan: Data Analysis. Research Triangle Institute Report No. RTI/1935/00-01F. Research Triangle Park, NC.

Weinglass, J.
1980 Draft Memorandum on Job Path Research: Preliminary Findings. New York: Vera Institute of Justice.

Wells, W.
1963 "How chronic overclaimers distort survey findings." J. of Advertising Research 3:8-18.

Westlund, K.B., and L.T. Kurland
1953 "Studies on multiple sclerosis in Winnipeg, Manitoba and New Orleans, Louisiana." American J. of Hygiene 57:380-407.

Wheaton, B., B. Muthen, D.F. Alwin, and G.F. Summers
1977 "Assessing reliability and stability in panel models." In D.R. Heise (ed.), Sociological Methodology 1977. San Francisco: Jossey-Bass.

White, A.A., and J.T. Massey
1981 "An investigation of a dual-frame approach for sampling Hispanics." Paper presented at annual meeting of the American Public Health Association.

Whyte, W.F.
1943 Street Corner Society. Chicago: University of Chicago Press.

Wilcox, J.
1961 "Observer factors in measurement of blood pressure." Nursing Research 10:4-20.

Williams, R.L.
1979 "Medical provider survey imputation strategy: expenditure variables." Working paper No. 2 prepared by Research Triangle Institute for NCHSR under Contract No. HRA 230-76-0268.

Williams, S.R.
1978 Development of a Feasible Data Collection Plan for Hospital Data in Florida: The Sampling Design. Research Triangle Institute Report No. RTI/1662/00-04I. Research Triangle Park, NC.
1979 Development of a Feasible Data Collection Plan for Hospital Data in Florida: Analytical Methodology. Research Triangle Institute Report. No. RTI/1662/00-01F. Research Triangle Park, NC.

Williams, S.R., *et al.*
1978 Development of a Feasible Data Collection Plan for Hospital Data in Florida: Preliminary Considerations Relating to Sampling and the Utilization of Extant Data Sources. Research Triangle Institute Report No. RTI/1662/00-01I. Research Triangle Park, NC.

Williams, S.R., and J.H Weber
1978 Development of a Feasible Data Collection Plan for Hospital Data in Florida: Alternative Sampling Designs. Research Triangle Institute Report No. RTI/1662/00-03I. Research Triangle Park, NC.

Wilson, R.W., and E.L. White
1977 "Changes in morbidity, disability, and utilization differentials between the poor and nonpoor: Data from the Health Interview Survey: 1964 and 1973." Medical Care 15:636-646.

Wingo, L., and A. Evans
1978 Public Economics and the Quality of Life. Baltimore, MD: The Johns Hopkins University Press.

Woltman, H.F., A.G. Turner, and J.M. Bushery
1980 "A comparison of three mixed-mode interviewing procedures in the National Crime Survey." J. of the American Statistical Association 75 (September):534-543.

Womer, S., and H. Boyd
1951 "The use of a voice recorder in the selection and training of field workers." Public Opinion Q. (Summer):358-63.

Wright, B.F., and C.F. Dore
1970 "A random zero sphygmomanometer." Lancet 1:337-338.

Wyngaarden, M.
1981 "Interviewing mentally retarded persons: Issues and strategies." In R. Bruininks *et al.* (eds.), Deinstitutionalization and Community Adjustment of Mentally Retarded People. Washington, DC: American Association on Mental Deficiency.

Yaffe, R., S. Shapiro, R.R. Fuchsberg, C.A. Rohde, and H.C. Corpeño
1978 "Medical economics survey-methods study: Cost-effectiveness of alternative survey strategies." Medical Care 16 (August):641-659.

Zeller, R.A., and E.G. Carmines
1980 Measurement in the Social Sciences. New York: Cambridge University Press.

Subject index

Conference participants

Murray Aborn
National Science Foundation
1800 G Street, NW
Washington, DC 20050

Lu Ann Aday
Center for Health Administration Studies
University of Chicago
5720 South Woodlawn Avenue
Chicago, IL 60637

Ronald M. Andersen
Center for Health Administration Studies
University of Chicago
Chicago, IL 60637

John P. Anderson
Department of Community and Family Medicine
University of California, M-022
La Jolla, CA 92093

Frank M. Andrews
Institute for Social Research
The University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Morris Axelrod
Director, Survey Research Laboratory
Department of Sociology
Arizona State University
Tempe, AZ 85281

Martha J. Banks
Center for Health Administration Studies
University of Chicago
5720 South Woodlawn
Chicago, IL 60637

Terrence Beed
Director, Sample Survey Centre
The University of Sydney
Sydney, 2006
New South Wales, Australia

Debbie Bercini
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Charles C. Berry
Dept. of Community and Family Medicine
University of California, M-002
La Jolla, CA 92093

Gordon S. Bonham
Div. of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Steven Botman
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Norman Bradburn
Director, National Opinion Research Center
6030 South Ellis Avenue
Chicago, IL 60637

Bengt Brorsson
Socialmedicinska Institutionen
Uppsala Universitet
S-750 14 Uppsala
Sweden

Fred A. Bryan
Statistical Methodology and Analysis Center
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

James W. Bush
Dept. of Community and Family Medicine
University of California, M-002
La Jolla, CA 92093

Gail Lee Cafferata
National Center for Health Services Research
Stop 3-50, Park Bldg.
5600 Fishers Lane
Rockville, MD 20857

Charles F. Cannell
Institute for Social Research
The University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

R.A. Carleton
Chief of Cardiology
The Memorial Hospital
Pawtucket, RI 02860

Robert Casady
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Steven B. Cohen
Senior Biostatistician
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

John P. Connelly
American Academy of Pediatrics
Dept. of Health Systems Res. and Development
1801 Hinman Avenue
Evanston, IL 60204

Larry S. Corder
Health Care Financing Administration
Oak Meadows Building, Room 1B-10
6340 Security Boulevard
Baltimore, MD 21207

Ronald Czaja
Survey Research Laboratory
University of Illinois, Room 230
400 So. Peoria, Formfit Bldg.
Chicago, IL 60607

Stephen M. Davidson
Center for Health Services & Policy Research
Northwestern University
Evanston, IL 60201

Joseph de la Puente
Director, Pub/Comm Health Personnel Project
American Public Health Association
1015 Fifteenth St., NW
Washington, DC 20005

Carole D. Dillard
Mathematical Statistician
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

K. Downey
Assistant Field Supervisor
Pawtucket Heart Health Program
The Memorial Hospital
Pawtucket, RI 02860

Douglas Drummond
Department Manager
Sampling Research/Design Center
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Elizabeth Eastman
Survey Research Laboratory
University of Illinois
2300 Formfit, P.O. Box 6905
Chicago, IL 60680

Jack Elinson
Professor, School of Public Health
Columbia University
60 Haven Avenue
New York, NY 10032

Jacob J. Feldman
Assoc. Dir. for Analysis/Epidemiology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Joseph Fitti
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Floyd J. Fowler, Jr.
Center for Survey Research
University of Massachusetts
100 Arlington Street
Boston, MA 02116

Joanne Frankel
Bureau of Social Science Research, Inc.
1990 M Street, NW
Washington, DC 20036

Karen Frey
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Robert Fuchsberg
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Jimmie D. Givens
Office of Research and Methodology
National Center for Health Statistics

3700 East-West Highway
Hyattsville, MD 20782

Donald Goldstone
Acting Director
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Fred Gonzalez
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Bernard G. Greenberg
Dean, School of Public Health
Rosenau Hall, 201H
University of North Carolina
Chapel Hill, NC 27514

Robert M. Groves
Senior Study Director
Institute for Social Research
The University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Robert S. Hartford
Dep. Asst. Dir. for International Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Constance M. Horgan
Researcher/Economist
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Daniel G. Horvitz
Vice President, Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Patricia Johnson
150 Hazelton Avenue
Cranston, RI 02902

Judith A. Kasper
Room 850-B
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Sonya Kennedy
Institute for Social Research
The University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Diane Kipp
Pawtucket Heart Health Program
The Memorial Hospital
Prospect Street
Pawtucket, RI 02860

William M. Kitching
Division of Health Examination Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Phillip A. Kletke
American Academy of Pediatrics
1801 Hinman Avenue
Evanston, IL 60204

Mary Grace Kovar
Interview & Exam. Stat. Program
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Richard A. Kulka
Survey Operations Center
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Barbara Lacey
Chief, Personnel Research
Bureau of the Census
FOB 3, Room 2286
Suitland, MD 20233

Judith T. Lessler
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Sara Segal Loevy
Center for Health Administration Studies
University of Chicago
5720 South Woodlawn Avenue
Chicago, IL 60637

Benedict W. Lohr
Division of Extramural Research
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Thomas W. Mangione
Center for Survey Research
University of Massachusetts
100 Arlington Street
Boston, MA 02116

Alfred C. Marcus
Director of Program Evaluation
Jonsson Comprehensive Cancer Center
10920 Wilshire Blvd., Suite 1106
Los Angeles, CA 90024

Kent Marquis
Chief, Center for Social Science Research
U.S. Bureau of the Census
FOB 3
Washington, DC 20233

James T. Massey
Chief, Survey Design Staff, ORM
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Kurt Maurer
Chief, Survey Planning & Devel. Branch
Division of Health Examination Statistics
3700 East-West Highway
Hyattsville, MD 20782

Linda McCleary
Office of International Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Sonja McKinlay
Evaluation Director
Pawtucket Heart Health Program
The Memorial Hospital
Prospect Street
Pawtucket, RI 02860

Peter V. Miller
Dept. of Communication Studies
Northwestern University
1815 Chicago Ave.
Evanston, IL 60201

Roberta Balstad Miller
Executive Director
Consortium of Social Science Associations
1755 Massachusetts Ave, NW, Suite 300
Washington, DC 20036

Lois Monteiro
Division of Biology and Medicine
Brown University
Providence, RI 02912

Jean Morton-Williams
Survey Methods Centre
Social/Community Planning Research
35 Northampton Square
London, England EC1V OAX

Janet D. Perloff
American Academy of Pediatrics
Department of Health Systems Res. and Devel.
1801 Hinman Avenue
Evanston, IL 60204

Gail Poe
Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Stanley Presser
Institute for Social Research
The University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Wornie L. Reed
Department of Sociology
Washington University
St. Louis, MO 63130

Dorothy P. Rice
Dir., National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Stewart C. Rice
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Pamela C. Roddy
Health Services Research
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Beth B. Rothschild
Research Group Manager
National Analysts Division
Booz-Allen & Hamilton
400 Market Street
Philadelphia, PA 19106

Beatrice A. Rouse
Epidemiologist
Boston University Medical Center
P.O. Box 842
Rockville, MD 20851

Patricia Royston
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Maurice Satin
Asst. Director, Div. of Mental Health Care Systems
Long Island Research Institute
Stony Brook, NY 11794

Donald W. Schiff
American Academy of Pediatrics
1801 Hinman Avenue
Evanston, IL 60204

Edward Schwartz
Health Scientist Administrator
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Laure M. Sharp
Bureau of Social Science Research, Inc.
1990 M Street, NW
Washington, DC 20036

Eleanor Singer
Sr. Res. Associate, Center for the Social Sciences
Columbia University
420 West 118 Street
New York, NY 10027

Monroe Sirken
Assoc. Director, Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Cecilia Snowden
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Susan A. Stephens
Survey Director, Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08540

R.J. Stimson
Director, Centre for Applied and Survey Research
Flinders University
Bedford Park, So. Australia 5042

Seymour Sudman
University of Illinois
385 Commerce West
Champaign, IL 61820

Carol W. Telesky
UCLA-Jonsson Cancer Center
Division of Cancer Control
10920 Wilshire Blvd.
Los Angeles, CA 90024

Owen T. Thornberry
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Diane Tuteur
Illinois Cancer Council
36 So. Wabash
Suite 700
Chicago, IL 60603

Carmen Noemi Velez
Division of Sociomedical Sciences
School of Public Health
Columbia University
60 Haven Avenue, B-4
New York, NY 10032

Lois M. Verbrugge
Assoc. Research Scientist
Institute for Social Research
P.O. Box 1248
Ann Arbor, MI 48106

Daniel C. Walden
Senior Research Manager
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Elinor Walker
Health Scientist Administrator
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Richard B. Warnecke
Survey Research Laboratory
University of Illinois, Formfit Bldg.

400 South Peoria, Rm. 2300
Chicago, IL 60607

Donna Watts
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Michael Weeks
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Andrew White
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Roy W. Whitmore
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Gail Wilensky
Senior Research Manager
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

P. Douglas Williams
Off. of Interview and Examination Stat.
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Stephen Williams
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Lucy B. Wilson
Vice President, National Analysts Division
Booz-Allen & Hamilton
400 Market Street
Philadelphia, PA 19106

N. N. Woodbury
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Robert A. Wright
Chief, Utilization and Expenditures Statistics
Div. of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

REPORT DOCUMENTATION PAGE	1. REPORT NO. NCHSR 83-45	2.	3. Recipient's Accession No.
4. Title and Subtitle HEALTH SURVEY RESEARCH METHODS FOURTH CONFERENCE AT WASHINGTON, D.C., MAY 2-5, 1982; NCHSR RESEARCH PROCEEDINGS SERIES			5. Report Date May 1982 6.
7. Author(s) Charles F. Cannell and Robert M. Groves (Eds.)			8. Performing Organization Rept. No. --
9. Performing Organization Name and Address University of Michigan Ann Arbor, MI 48109			10. Project/Task/Work Unit No. 11. Contract(C) or Grant(G) No. (C) (G) HS 04569
12. Sponsoring Organization Name and Address DHHS, PHS, OASH, National Center for Health Services Research Publications and Information Branch, Room 1-46 Park Building 5600 Fishers Lane Rockville, MD 20857 Tel.: 301/443-4100			13. Type of Report & Period Covered Res. Proceedings Series 9/1/81 - 12/31/82 14.

15. Supplementary Notes

DHHS Pub. No. (PHS) 83-3346. Library of Congress Card No. 83-600565.

16. Abstract (Limit: 200 words)

This conference report is intended to inform the health research community about recent advances in health survey methods, about continuing concerns of which health survey users should be aware, and about areas requiring further methodological research. The conference concentrated on six major topics: (1) Measures and Correlates of Response Errors; (2) Telephone Survey Methodology; (3) Studies of Survey Measurement Techniques; (4) Use of Records in Health Survey Research; (5) Hiring, Training, and Monitoring Interviewers; (6) Survey Methods for Rare Populations. The conference was supported by conference grants to the Institute for Social Research at The University of Michigan, Ann Arbor, from the National Center for Health Services Research and from the Milbank Memorial Fund, and by services provided by the National Center for Health Statistics.

17. Document Analysis a. Descriptors

NCHSR publication of research findings does not necessarily represent approval or official endorsement by the National Center for Health Services Research or the U.S. Department of Health and Human Services.

Edward Schwartz, Ph.D., NCHSR Project Officer, 301/443-6990

b. Identifiers/Open-Ended Terms

Health services research
Health survey research methods
Health Survey Research Methods Fourth Conference

c. COSATI Field/Group

18. Availability Statement: Releasable to the public. Available from National Technical Information Service, Springfield, VA 22161 Tel.: 703/487-4650	19. Security Class (This Report) Unclassified	21. No. of Pages Est. 400
	20. Security Class (This Page) Unclassified	22. Price

(See ANSI-Z39.18) See Instructions on Reverse

OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Office of the Assistant Secretary for Health
National Center for Health Services Research
1-46 Park Building, 5600 Fishers Lane
Rockville, MD 20857

DHHS Publication No. (PHS) 84-3346