OBSERVATIONS ON EVALUATION OF FAMILY PLANNING PROGRAMS

HYMAN GOLDSTEIN, PhD

Several months ago I was asked to prepare a course dealing with evaluation of maternal and child health programs, including family planning programs, for our Master of Public Health students. Since I was relatively new to the family planning evaluation field, I thought that exposure to the concepts, methodology, and procedures of family planning evaluation centers would help me to better understand the problems in this area. The Ford Foundation provided a travel grant, and I visited 14 well-known evaluation centers in September and October 1971. These visits gave

Dr. Goldstein is a research biostatistician, maternal and child health program, School of Public Health, University of California, Berkeley, and adjunct professor of biostatistics, department of obstetrics and gynecology, University of California School of Medicine, San Francisco. *A grant providing funds for travel was awarded Dr. Goldstein by the Ford Foundation. This article is based on a speech he gave at the annual meeting of the Association of Teachers of Maternal and Child Health, Ann Arbor, Mich., April 16–18, 1972.* *Tearsheet requests to Dr. Hyman Goldstein, School of Public Health, Earl Warren Hall, Berkeley, Calif. 94720.*

me a rare opportunity to see what the people in these centers were thinking and doing.

Eight of the centers emphasize family planning programs in foreign countries and six emphasize such programs in the United States. These evaluation centers are assessing the effectiveness and efficiency, or both, of efforts being made by governments, clinics, and local physicians in making family planning services available to people who need them. There is considerable activity, sometimes on a broad front, with varying degrees of sophistication, going on at all centers. Much of the work, in my opinion, is well conceived and well done.

Many of the persons I contacted indicated that it was easier to evaluate family planning programs than those in other health areas. Some persons stated that, in their opinion, greater progress has been made in family planning program evaluation than in evaluating programs in other health areas. Many of the difficulties in evaluating programs discussed in this paper, however, were voiced by persons with whom I spoke. The great increase in recent years in family planning projects and in evaluating units and agencies is related to some of these difficulties. Some of the problems are equally common in the evaluation of health programs other than family planning.

Evaluation must answer at least the following three questions.

1. Are the persons administering the family planning programs able to accomplish what they set out to do?
2. Are they reaching people that would, in the absence of the program, not be reached without implications as to the qualitative adequacy of the services?
3. What effect, if any, has the program had on fertility not only of women coming to the clinic but also on the fertility of the women in the community?

Operators of family planning programs have to make positive and conscious decisions about their programs. How do they arrive at such decisions? Is the decision-making process related to objectives? The main concern in evaluation should be to provide information to those persons who are making decisions for implementation of family planning programs. What elements of information are needed so that they will be able to determine whether or not progress is being made toward achieving the objectives of the program?

Information Needed

The great need is for consistent, standardized, common definitions, criteria, concepts, and procedures so that there may be ways of consistently relating inputs to outcomes, particularly in programs where the objectives are identical or similar. There should be greater agreement on basic essential definitions, such as ways of counting things, among different family planning agencies. For example, the number of living children is sometimes not distinguished from parity—occasionally parity is confused with gravidity and "ever-married" with "husband present." The term "new acceptor" is not defined uniformly. Does it refer to first acceptance of contraception, first acceptance of a particular contraceptive method, or first acceptance of a particular family planning program? The same three ambiguities apply to the terms "continuer" and "noncontinuer."

The Population Council's International KAP Guidelines were an effort to get agreement on basic KAP (knowledge, attitudes, practices) questions and to present a list of such questions for use in doing KAP surveys for comparative purposes—a large step forward. But for comparative purposes a greater uniformity in sampling design and definitions is also needed. Many investigators are trying to aid the effort to obtain agreement on certain basic definitions and procedures by producing evaluation manuals, but their manuals do not always agree.

Evaluators must agree on specified criteria for evaluation

of family planning programs which have identical or similar objectives for meaningful comparison or pooling of data among areas and populations. How many or how few indicators is not important, but, wherever possible, the indicators should be the same or closely comparable for the same objectives. The following are some of the criteria used: (*a*) waiting time of patients, (*b*) service time of patients, (*c*) acceptor rates, (*d*) retention rates, (*e*) ratio of these two rates, (*f*) dropout rates, (*g*) excess fertility data, (*h*) age-parity ratios, (*i*) fertility rates, and (*j*) number of IUDs inserted. If all these may be considered criteria for evaluating effectiveness, then the following questions arise. Are they all of equal importance? What degree of association is there among them? What degree of change in any, in some, or in all of them denotes acceptable progress toward specific objectives of the family planning program?

Within a set of criteria relevant to specific objectives, some are more costly than others when obtaining data and some are more important than others for the objectives in mind. Therefore, it would seem wise to establish a listing of criteria in order of priority, by a consensus of experts in this field, so that if funds are limited for evaluation, there would be some guidelines as to which criteria of the same order of importance would offer greater cost-effectiveness returns.

The goal of standardizing definitions and criteria for evaluation is to be able to make necessary comparisons using the same measuring instruments from place to place and from time to time. This standardization would enable one to place the greatest emphasis where it belongs in evaluation, namely determining which program is most effective, most efficient, and offers most cost benefits. By using uniform methodology, the funding agency can arrive at crucial decisions and can make the wisest investments. The public dollar spent by an evaluating agency is just as scarce as that spent by a family planning agency.

Evaluation designs should be developed for most recurring programs or where programs are slated for replication which can be applied, with minor changes if necessary, to each recurring program with the same objectives. These designs would facilitate the task of evaluation and provide comparable results for projects in different areas, for different populations, at different points in time, and at different stages of family planning program development. In the area of data collection, what is needed to satisfy the requirements of validity and reliability of data? The quality of the evaluation is no better than the data upon which it is based.

Quality of Data

The real difficulties in evaluating family planning programs are the reliability and the validity or their lack in the recorded data, including the indexes (criteria) of evaluation. These indexes have to be objective, measurable, and recorded. Just having tons of data without much evidence of accuracy and consistency is not of much help. The staffs of most service programs are so busy with the service angle that the need for such evidence is given low priority.

The evaluator must be concerned with what a couple is doing, not with what they say they are doing. Are they actually practicing contraception? A representative sample, for example, of women using IUDs might be called in to determine whether the IUDs are still in utero. We need to develop better fertility data, especially marital-age-specific fertility rates, and there is as yet no objective way to ascertain unwanted fertility. The validity and consistency of answers given by women when surveys on fertility have been made are of doubtful value.

In any evaluation a good record system is essential. But even good record systems need proper maintenance and careful attention to detail. The quality of the data secured has to be appropriate to the objectives of the study. Not all studies need the same quality of data, but, among studies with the same or similar objectives, the information supplied is of variable and often inadequate quality and probably varies greatly from one family planning program to another. For this reason, it is difficult to achieve comparability. Comparability over time is as necessary as comparability over place. Trend information is difficult to come by because of such incomparabilities.

Problems in evaluation lie in the (*a*) currently inadequate data base (denominators) and (*b*) retrieval of data to get needed indexes of evaluation (numerators).

No really good evaluation of a program can be made in the absence of rates, although numerator analysis is an attempt to provide alternatives. Rates need denominators, and this is one of the urgent needs of evaluation. The population eligible for family planning services within the defined catchment areas is most often an unknown quantity, subject to the hazards of unvalidated

estimates. An estimate of women needing family planning services within a given area should be tested by sample surveys so that a measure of the standard error of estimate may shed some light on the confidence limits of the estimate. Vital events, such as fetal deaths, births, and infant deaths also need denominators applicable to a defined catchment area during a defined time period, and this also often represents an unknown that is subject to guesswork.

The best evaluation has the nearest approaches to an experimental design. Meaningful evaluation uses the methodology of research. However, not always being able to compare the study community group with a control community group is a problem.

As a result, few evaluators attempt to get a control group. It is often possible, however, to find a comparison group to substitute for a control group that was impossible or impractical to procure. Such comparison groups might be chosen from communities that lack adequate family planning services but are similar in other respects to communities having these services. But, it is essential to document the extent of compromises, as well as when and why they were necessary. Evaluations can be made without absolute purity, but without documentation they cannot contribute anything. The limitations must be adequately described.

A "before and after" evaluation needs a good data base and good baseline data. The evaluation plan should incorporate the collection of baseline information. This collection should start functioning before wrong procedures become imbedded in administrative practice; however, this is rarely done in practice.

An effort to get uniform data was started in May 1969 with the initiation of the National Center for Health Statistics Reporting System for Family Planning Services. Uniform data for the system are collected on some 1,500 participating family planning clinics funded by the National Center for Family Planning Services and the Maternal and Child Health Service of HSMHA, the Office of Economic Opportunity, and Planned Parenthood-World Population. At the end of October 1971, data on 917,627 patients had been collected. Instructions are available to participating clinics with respect to definitions (for example, an initial visit, revisit, and family planning services), administrative procedures, and completing the clinic visit record. Uniform reporting is a step in the right direction. More, much more, however, is needed.

Community Needs

The neighborhood and its needs rather than a head count of clinic patients should be the focus of family planning clinics. Is the clinic located in an area with excess fertility women? Is it attracting these women and, if so, to what degree? The number of family planning acceptors—a self-selected sample—has little value for evaluation. When related to those in the community needing family planning by age and parity it is a more important index.

The major fallacy of family planning evaluation is to focus only on the clientele of the clinics. The population at large of women in the community, with or without family planning services, should be studied. One should relate the amount of input into the total population to outcome. We have to know about old acceptors as well as new acceptors and about the continuation experience if we want to get a prevalence-of-use figure. To do this and get knowledge of the impact of the program, we need community surveys.

The highly fecund woman is most likely to come for family planning services, and this may also be the reason that it is difficult, if not impossible, to obtain adequate control groups in the same service area from which acceptors are drawn. What about the rest of the community needing family planning services? If we generalize from the acceptors, we overestimate the impact of the program on the community. The characteristics by age, parity, and history of pregnancy of women who accept and retain the IUD should be compared with those of women who expel it, who have it removed, who use some other contraceptive, or who use no contraceptive though potentially in need of one.

Comprehensive Evaluation

Most family planning programs do not have comprehensive evaluation programs. They usually do only one thing, perhaps analyze characteristics of acceptors and nothing else. They do the same thing again and again. Needless to say, they often do not worry about cost effectiveness or efficiency.

Making an evaluation based on one criterion is always hazardous. For instance, if the acceptance rate is the only basis on which evaluation is to be made, a low retention rate for a given family planning method might easily be overlooked. In other words, medical salesmanship might increase the acceptance rate but patient distress might push the dropout rate just as high.

Mere ascertainment of the dropout rate is not enough. At what level are they dropping out? Why are they dropping out? What are the variables? Are all dropouts pregnant, or have they switched to another service for contraception, or have they dropped contraception entirely? Are they dropouts from the clinic, the method, or from family planning entirely? The reasons for dropping out can supply clues for the low continuation rate and provide the means for raising it. What has been the dropout's experience with pregnancy since dropping out for comparison with the nondropouts? A family planning program without followup is ineffective.

It is most important that the evaluation determine not only whether the program is progressing but also how and why it is succeeding or failing. This information can usually be obtained by a critical look at the progress or lack of it with respect to the specific intermediate and immediate objectives.

Although there is no single method of evaluation that can adequately measure the accomplishment of a family planning program, it is essential that the methods used are adequately described with clear-cut definitions and objectives and measurable inputs and outcomes. The entire procedure must be replicable if it is to add to our knowledge of evaluation.

In evaluation there is a need for (*a*) analysis of clients' records obtained at clinic registration, (*b*) service statistics on accomplishment of the clinic and of the effectiveness and efficiency experience, which, in turn, should lead to cost-effectiveness analysis, (*c*) sample surveys of the catchment area about every 5 years to get information about the people not coming to the clinic, (*d*) followup surveys every few years of a sample of clients, including dropouts, and (*e*) demographic analysis of the birth rate and the impact of the program on fertility. Properly organized, it has been estimated that meeting these five needs should cost no more than 5 percent of most family planning program budgets. It is my understanding that as much as 5 percent of such budgets is rarely, if ever, allocated for these needs.

Essentially, there are four parts to an evaluation system: (*a*) data collection from study group and control (or comparison) group before and after intervention, (*b*) decision making based on the data and feedback, (*c*) implementation of these decisions (unless implementation takes place, the evaluation is an exercise in futility), and (*d*) reevaluation at a later time to determine whether the implementation effected improvement of the program, using the same criteria for evaluation as previously. Evaluation is still an art, and it is essential that we determine whether the "diagnosis" made and "treatment" prescribed were correct. Only by learning from its mistakes can the field of evaluation move forward.

Conclusions

In my opinion, there has not been enough communication among the evaluation agencies. Conferences on the methodology of evaluating family planning programs are both necessary and valuable. They contribute to a concern for good data and data collecting techniques, no matter what agency conducts the family planning program or makes the evaluation.

Many ideas and suggestions could come from such conferences and, it is hoped, some expression of agreement and ways to implement some uniform and standard procedures, including criteria for evaluation, sampling, collection, and analysis of data. In this way it may become possible to make comparisons in relation to various types of rates and other parameters without worrying over the degree to which variations in definitions, criteria, and procedures have produced the apparent differences among programs.

Representatives of agencies in the field of evaluation of family planning programs have had some meetings, but not nearly enough to produce a more unified approach to evaluation or to establish the criteria for quality of evaluation. That such meetings have been held, however, indicates that a need has been felt and it is hoped that this need will produce the essential action.

The title of this paper mentions "Observations on Evaluation," not "Panaceas for Evaluation." I think that the various agencies evaluating family planning programs should agree to agree upon some essential definitions, objectives, and criteria for evaluation of objectives at different levels. Agreement by consensus would pose no threat to anyone, but it would give the funding agencies some way to compare programs and to compare evaluations, which they should have. Comparison is the heart of evaluation. It is a sad omission when we can evaluate a program using comparison with some control or comparison group, but we are unable to compare family planning programs because of the noncomparability of the evaluating procedures.