Reporting Sources of Error: The United States Experience

D. Kasprzyk, D. Atkinson, L.Giesbrecht, M. McMillen, D.Schwanz, W. K. Sieber
National Center for Education Statistics
555 New Jersey Avenue, NW
Washington, DC 20208
E-mail: Daniel_Kasprzyk@ed.gov

I. Introduction

The United States Office of Management and Budget's Federal Committee on Statistical Methodology (FCSM) has had a leadership role in discussions of the methodology of federal surveys for more than two decades (Gonzales, 1995). In 1996, the FCSM established a subcommittee to review the measurement and reporting of data quality in federal data collection programs. Although data quality is a multidimensional concept that includes issues of accuracy, relevance, timeliness, and accessibility, the subcommittee has focussed its discussions on the issue of accuracy, its measurement and presentation.

The sources of error that affect survey data quality - sampling error, coverage error, nonresponse error, measurement error, and processing error and their measurement are described in a number of texts. Kasprzyk and Kalton (1999, 1997) provide a summary of methods used to measure error sources and examples of their implementation in U.S. data collection programs. This paper focuses on another area of the subcommittee's interests - the reporting and presentation of information on sources of error in several dissemination media (short-format publications, analytic publications, and the Internet). This review is based on the work of the FCSM Subcommittee on Data Quality (McMillen and Brady, 1999; Atkinson, Schwanz, and Sieber, 1999; Giesbrecht, Miller, Moriarity, and Ware-Martin, 1999).

II. Reporting Sources of Error in Reports

Reporting formats vary from a press release of one or two pages of analysis on a specific question to detailed complex analytic modeling exercises designed to test specific hypotheses. Although there is considerable recognition about the need for reporting the extent and nature of the sources of errors, there is considerable uncertainty about the amount of detail that ought to be provided. Details reported ought to depend on the length of the report, its intended use, the nature of the data (one-time survey vs. continuing), the survey budget, and agency policy.

The subcommittee conducted two studies to help characterize current practices for reporting sources of error in reports (McMillen and Brady, 1999; Atkinson, Schwanz, and Sieber, 1999). The first study reviewed publications and data releases of ten pages or less ("short-format" reports) from twelve U.S. statistical agencies. The second study reviewed selected "analytic publications" from seventeen agencies. These are printed reports resulting from a primary summarization of a one-time survey or an on-going series of surveys. Publications that compile data from many sources and reports designed primarily to study error sources are not included in the study.

II.A Reporting Sources of Error in Short-Format Reports

A total of 454 short-format publications were reviewed for their treatment of survey design, data quality, and measurement error. Virtually all of the reports included some information on how to learn more about the data reported. These sources of information ranged from names and phone

numbers, e-mail and web site addresses, to printed reports. A small percentage (3 percent) of the short reports did not discuss the study design, data quality, or survey error, but did include a reference to a technical report. About two-thirds (69 percent) included either a reference to a technical report or some mention of the study design, data quality, or survey error. Close to one-half (47 percent) of authors of short reports included at least one piece of information describing the purpose of the survey or analysis, the key variables, the survey frame, and/or key aspects of sample selection. Slightly less than one-half of these cases included the sample size (20 percent of all reports), and one-fifth described the mode of data collection (10 percent of all reports). References to weighting and estimation were rare - only 2.4 percent of all reports.

Following the error classification schema mentioned above, the reporting of error sources was reviewed in these reports. About one-fifth of the reports included some reference to sampling error. Most references acknowledged the presence of sampling error, and a few noted that data came from universe data collections and were not subject to sampling variability. About one-third included a reference to at least one type of nonsampling error. Even though nonresponse error is the most visible and well known source of nonsampling error and easily obtainable, only 13 percent included any reference to response rates, to nonresponse as a potential source of error, or to imputations. Measurement error was cited as a source of error in about 22 percent of the reports, but no mention of specific analyses or measures of measurement bias or measurement variance were mentioned. Processing error was cited as an error source in about 16 percent of the reports. Finally, only about 9 percent reported coverage rates or mentioned coverage as a potential source of error.

II.B Reporting Sources of Error in Analytic Publications

In a second study, 49 analytic publications produced by 17 agencies were reviewed. The publications selected for review were a nonrandom convenience sample designed to cover the major statistical agencies as well as some smaller agencies conducting surveys. The review was based on the completeness of background information describing the survey and on the reporting of error sources. Evaluation criteria consisting of desirable information that ought to be available in analytic reports were established - 16 criteria for the reporting of background survey information and 35 criteria for the reporting of information on sources of sampling and nonsampling error.

The review revealed substantial variability in the information provided across and even within agencies. The three top-rated publications contained qualifying information on 35 of 51 criteria, while the lowest rated publication qualified on only 8 criteria. The average number of criteria on which the publications qualified was 21. Background survey information was fairly prevalent in the publications reviewed. In particular, several publications provided excellent descriptions of the sample design, data collection, and estimation procedures. Descriptions of the comparability of survey results with other data sources were somewhat less commonly provided; for example, 19 reports (39 percent) described in detail the survey changes that affect comparisons of the survey's data over time.

The study reviewed the information available on sources of error in analytic publications and found sampling error the error source most often discussed in publications (92 percent); it was presented in 59 percent of the reports reviewed. Nonresponse, however, was not always mentioned as a potential error source and response rates were frequently not reported. Seventy-one percent mentioned unit nonresponse and only 59 percent indicated an overall nonresponse rate. Forty-nine percent of the publications mentioned item nonresponse and only 22 percent presented item response rates. Measurement error was mentioned in 67 percent of the reports, and specific sources were described and defined in 51 percent. Reinterview, record-check, or split sample studies were mentioned in 18 percent of the studies. Processing error was mentioned as an error source in 78

percent of the reports, but very few reports had any detail. Finally, coverage error was specifically mentioned as a possible source of nonsampling error in 49 percent of the reports, but only 16 percent provided coverage rates.

III. Reporting Sources of Error on the Internet

The Internet has become the principal medium for dissemination of data products for many federal agencies. The third study (Giesbrecht et al., 1999) reviewed guidelines and practices for reporting error sources over the Internet. Some federal agencies have written standards for web sites, but these generally focus on web site design, layout, and administrative access. A few agencies, such as the Census Bureau, have begun the process of developing standards for providing information about data quality over the Internet (U.S. Bureau of the Census, 1997). This document gives details on the kind of data quality information that ought to be provided to the user, but does not require or suggest the use of Internet features for making information more accessible. Generally, standards documents related to Internet practices reiterate standards for printed documents (see, for example, United Nations Economic and Social Council, 1998).

This study reviewed the accessibility of data quality documentation on current Internet sites of 15 U.S. statistical agencies. It found that online documentation was available for nearly all the sites visited. For about one-third of the sites, offline documentation was referenced as well. Most agencies seem to upload their printed reports and documentation in the form of simple text or Adobe Acrobat format files. Summary charts that provide study results on the availability of online documentation, printed documentation, online technical support, and agency web site addresses can be found in Giesbrecht et al. (1999). The study also noted a few best practices as found on the visited web sites; 1) the availability of pop-up windows providing definitions of column and row headings in tables; 2) links to send e-mail messages to technical specialists; 3) links to "survey methodology" and "survey design" information; 4) explicit directions to users about errors and comparability issues; and 5) links from one agency's home page to another and common access points to statistical information.

The study found current Internet standards for data quality information echo the standards for printed reports and statistical tables. More explicit standards for how the advantages of the Internet media should be employed to make data quality information more easily accessible do not exist.

IV. Conclusion

Users of survey data need information about a survey's quality to properly assess survey results. Standards adopted by many statistical agencies specify users should be informed of survey quality. There are many dimensions to survey quality and the measurement and presentation of this information is no easy task. Report formats, dissemination media, agency policies and practices vary. The studies reviewed in this paper illustrate the range of agency practices in reporting information on error sources in surveys. The studies suggest that U.S. statistical agencies not merely define policy and standards in the reporting of such information, but monitor the implementation of this policy. Some observers argue a common template for information about error sources in data collection programs would help raise awareness of this need. Others suggest that ongoing data collection programs should develop quality profiles, a report that systematically gathers information about survey procedures and sources of error. In any case, data users are best served when information about survey procedures and sources of error are readily available to them to help the interpretation of the analysis. In the case of the U.S. experience, more emphasis and interest in this topic is desirable.

REFERENCES

Atkinson, D., Schwanz, D., and Sieber, W.K. (1999), "Reporting Sources of Error in Analytic Publications," *Statistical Policy Working Paper: Seminar on Interagency Coordination and Cooperation*, Washington, DC: U.S. Office of Management and Budget.

Giesbrecht, L., Miller, R., Moriarity, C., and Ware-Martin, A. (1999), "Reporting Data Quality on the Internet," *Statistical Policy Working Paper: Seminar on Interagency Coordination and Cooperation*, Washington, DC: U.S. Office of Management and Budget.

Gonzales, M.E. (1995), "Committee Origins and Functions: How and Why the Federal Committee on Statistical Methodology Began and What it Does," *Proceedings of the Section on Government Statistics, American Statistical Association*, 262-267, Alexandria, VA.

Kasprzyk, D., and Kalton, G. (1999), "Measuring and Reporting Data Quality in Federal Data Collection Programs," *Statistical Policy Working Paper: Seminar on Interagency Coordination and Cooperation*, Washington, DC: U.S. Office of Management and Budget.

Kasprzyk, D., and Kalton, G. (1997), "Measuring and Reporting the Quality of Survey Data," *Proceedings of Statistics Canada Symposium 97: New Directions in Surveys and Censuses*, 179-184, Ottawa: Statistics Canada.

McMillen, M., and Brady, S. (1999), "Reporting Sources of Error in Short Format Publications," *Statistical Policy Working Paper: Seminar on Interagency Coordination and Cooperation*, Washington, DC: U.S. Office of Management and Budget.

United Nations Economic and Social Council (1998), *Guidelines for Statistical Metadata on the Internet*, paper contributed to the Conference of European Statisticians, Forty-sixth Plenary Session, Paris, May 18-20, 1998.

U.S. Bureau of the Census (1997), *Draft Survey Design and Statistical Methodology Metadata IT Standards*, Washington, DC: U.S. Department of Commerce.

RÉSUMÉ

La politique de la transparence en ce qui concerne la description complète des données, des méthodes, des hypothèses et des sources d'erreurs, est celle universellement adoptée par les bureaux statistiques. Quoi qu'il en soit, la mise en application de cette politique varie grandement. Cet article résume les résultats de trois études qui passent en revue les pratiques de présentation de la qualité des données pour des programmes de collecte de données organisés aux États-Unis. Ces études menées par un sous-comité du Comité fédéral pour les méthodologies statistiques auprès du Bureau américain du budget et de la maîtrise des ressources, montre le degré de communication aux utilisateurs des informations concernant l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage par rapport à différents moyens de présentation: des publications courtes ou analytiques et via Internet.