DISCUSSION

Daniel Kasprzyk[1]

This session is about the "quality" of survey data, an important topic for National Statistics Offices, particularly over the last dozen years or so. Attending any conference on "data quality" or even a session focused on "data quality" usually results in hearing the litany of dimensions of data quality (Brackstone 1999; Arondel and Depotout 1998; Andersson, Lindstrom, and Lyberg 1997: Collins and Sykes 1999; Statistics Canada 1998; Statistics New Zealand 1998). Depending on the speaker, you will hear quality described as having four to seven dimensions:

Accuracy—an important and visible aspect of quality of concern to statisticians and survey methodologists for many years. It relates to the closeness between the estimated and true (unknown) values. For many, accuracy means the measurement and reporting of estimates of sampling error, but, in fact, the concept is much broader, taking in nonsampling error as well.

Relevance—the idea that the data collection program measures concepts that are meaningful and useful to data users. Does the concept implemented in the data collection program fit the intended use? For example, concepts first measured in a continuous sample survey program 20 years ago may be inapplicable in current society.

Timeliness—can refer to either the length of data collection's production time—the time from data collection until the first availability of a product. Fast release times are without exception looked upon favorably by end users. Timeliness can also refer to the frequency of data collection. Timeliness can be difficult to characterize, since the characteristics of the data collection can affect the availability of data.

Accessibility—is the ability of data users to obtain the products of the data collection program. Data products have their most value and are most accessible when they are easily available to end users in the forms and formats desired.

Arondel and Depotout (1998) and Brackstone (1999) suggest three other characteristics of data quality: comparability of statistics, coherence, and completeness. **Comparability** of statistics refers to the ability to make reliable comparisons over time; **coherence** refers to the ability of statistical data programs to maintain common definitions, classifications, and methodological standards when data originate from several sources; and **completeness** is the ability of the statistical data collection to provide statistics for all domains identified by the user community.

Survey data quality is a concept with many dimensions and each dimension linked with others. In the abstract, all dimensions of data quality are important, but in practice, it is usually not possible to place high importance on all dimensions. Thus, with fixed financial resources, an emphasis on one dimension will result in a decrease in emphasis in another. More emphasis on accuracy can lead to less emphasis on timeliness and accessibility; or an emphasis on timeliness may result in early/preliminary data of lower accuracy. Each dimension is important to an end user, but each user may differ in identifying the most important priorities for a data collection.

This session consists of three papers concerned with data "quality"—each paper treating the topic from a different perspective. The Haworth and Martin paper provides the broadest and most general view of the topic. They describe the approach adopted in the United Kingdom, some work being done to deliver and measure data quality, and a few specific projects to improve data quality and provide better measures of process quality. This paper focuses on the macro aspects of the quality movement. The Sadeghi and Clayton paper acknowledge the multidimensional aspects of quality, but focus on the difficult problem of trading off results in one dimension of data quality to improve results in another dimension, a situation that does not have a single answer, but ultimately relies on the needs and judgments of the principal users of the data. The third paper is more narrowly focused on a single

[1] Daniel Kasprzyk, Mathematica Policy Research, Inc., 600 Maryland Avenue, SW, Washington, D.C., USA, 20024-2512

aspect of the data collection approach—interviewer characteristics and effort and their effect on the data. This paper is, of course, quite different from the two preceding papers in that its focus is solely on accuracy and, in particular, a component within the overall dimension of accuracy. So the papers actually have a hierarchical order as each relates to data quality - from the very general to the very specific.

Haworth and Martin

Haworth and Martin describe the significant work and improvements that have transpired during the last decade in the United Kingdom. I will briefly review what has taken place and provide a few comparative examples from the U.S. A three-pronged strategy has been adopted in the United Kingdom—identifying protocols/standards, identifying quality management principles, and implementing national statistics quality reviews. What is interesting about the approach is that the U.K. has attempted to codify and establish some consistency and uniformity in a significant number of aspects of statistical work across departments and agencies, something in the U.S. that is accomplished inconsistently and certainly not uniformly.

Standards/protocols/guidelines are undertaken to promote consistency among studies and agency practices, and to promote documentation of methods and principles used in collection, analysis, and dissemination. In the U.S. few agencies actually have produced formal written standards/guidelines (U.S. Department of Energy 1992; U.S. Department of Education 1992). Some agencies in the U.S. take a different approach to standards focusing more on the establishment of policies and procedures for reviewing reports, determining sampling errors, and testing hypotheses (Sirken, Shimizu, French, and Brock 1974) or presenting information concerning sampling and nonsampling error (Gonzalez, Ogus, Shapiro, and Tepping 1975; updated by Bailar 1987). My point is that there appears to be two very different approaches that operate in the two countries. A similar situation exists with respect to the production of "quality reports." It appears the U.K. will have a well-defined and consistent policy with respect to the production of quality reports—or in the U.S. jargon, quality profiles. This is quite different than the U.S., where each data collection program is left to its own judgment concerning the development of quality profiles. Recently, however, Census Bureau staff have recommended the routine development of survey quality profiles (Doyle and Clark 2001; U.S. Bureau of the Census 2000) for surveys they conduct. This is more in the spirit of the U.K. model, but it remains to be seen whether the Census Bureau's survey sponsors will adopt this recommendation.

With respect to specific initiatives, the U.K. seems to have worked exceptionally hard on the "coherence" aspect of data quality—harmonizing concepts, definitions, and classifications. They have reached agreements with a large number of stakeholders, and this is very impressive. Here in the U.S., harmonization of few major classification systems, such as "industry and occupation," is done, but for other concepts such as disability, harmonization can be difficult and maybe impossible. This might be a good outcome. The idea of using standard definitions and questions across different data collections is extremely appealing for some analyses and comparisons and extremely efficient for questionnaire design; however, users/sponsors often have different ideas of the analyses desired and the concepts required to answer their specific questions. My point is that rigid adherence to the concept of coherence may lead to analyses that are neither as complete nor as thorough as some may want or need.

The other initiatives Haworth and Martin identify are important and follow prevailing wisdom—cognitive research on questionnaire design, improvements in editing efficiency, communicating quality of statistics, the need to develop better measures of process and output quality. To that end, they have made communication about data quality (i.e., accuracy) a priority. They have a developed a Statistical Quality Checklist (United Kingdom Government Statistical Service 1997) that provides examples of documentation for various aspects of survey processes and have developed guidelines for quality reports (Eurostat). In the U.S., this cross-agency activity occurs through the Office of Management and Budget's (OMB) Statistical Policy Office. In this regard, many in the U.S. statistical system expect OMB's Interagency Council on Statistical Policy (ICSP), that is the statistical agency heads, to play an important leadership role in fostering cross-agency activities, if these activities are thought to be important.

The U.K's ambitious plans end with staff addressing the standardization of nonresponse definitions. The work completed by the American Association of Public Opinion Research (2000) serves as a useful tool, and the U.K. takes advantage of this work. Within the U.S., initiatives exist within agencies, such as the National Center for Education Statistics (McMillen 2001) and across agencies through subcommittees on household and establishment survey nonresponse. Ramirez (2000), reporting for an interagency workgroup on establishment nonresponse,

reviewed issues associated with unit nonresponse in Federal establishment surveys. Shimizu (2000) reported a lack of consistency in reporting response rates in U.S. establishment surveys. Seastrom (2001) found considerable variation in defining and reporting response rates across survey programs at the National Center for Education Statistics. The Federal Committee on Statistical Methodology also corroborated varying reporting practices across the U.S. agencies and the varying pieces of information made available to the data user (Kasprzyk, Atkinson, Giesbrecht, McMillen, Schwanz, Sieber, 1999; Kasprzyk and Giesbrecht 2001; Atkinson, Schwanz, and Sieber 1999). Atrostic, Bates, Burt, and Silberstein (2001), reporting for an interagency household nonresponse group, review practices and provide recommendations for household survey response rate calculations. So the U.S., while not having a concentrated or coordinated effort, has had several initiatives take place that ought to lead to more consistent reporting across a few agencies.

Sadeghi and Clayton

The Sadeghi and Clayton paper explores "data quality" by accepting data reported earlier than the due date and comparing them to the "due date" data— the so called "truth." This paper is a good example/good test of the timeliness/accuracy tradeoff that is discussed, but rarely—it seems—acted on. An extended discussion of how the timelines for reporting data were chosen would be helpful. The Sadeghi presentation made it clear shortened timelines were motivated by a Bureau of Labor Statistics customer, and this consideration is important. However, to someone not familiar with the data requirements and timelines, the two week improvement does not seem compelling. Is this really a breakthough? Is the uncertainty caused by the early data not important when weighed against data availability two weeks earlier? The fundamental data assumption made is that editing and data cleanup in the last two weeks before submission was minimal. This may be true and ANY shortening of time to data release may be looked upon favorably, but would it not have made sense to research reasonable cut-off times for consideration by a consensus of key data users and then acted on their input?

The authors selected employment numbers and wages—differences between early and final submissions, and the level and number of imputations—as critical variables for this test. As a non-user of these data, it is difficult to judge the results. If statistics for "all states combined" are the key statistics, then the overall employment change between the early and the on-time submission is only .5 percent, .47 percent and .56 percent for the three quarters in question. These differences do not seem important, but I am not familiar with the scrutiny given these estimates. The disturbing aspect of the data is their lack of consistency from time-to-time and from state-to-state. At issue are the states and their updating procedures; there are obviously differences between states in how these updates are done, and how the data are edited and cleaned. For the wage estimates, the estimates that combine states were reasonable, but for individual states, Minnesota, for example, there were differences. I repeat these findings because I want to emphasize that one needs to establish in advance what estimates are important to report and the magnitude of the differences one can accept between preliminary and final.

Imputation ratios and number of edit failures are process statistics used for measuring the quality of the data. These decline, as expected, over the two week period the authors evaluate. The analysis relies on assumptions and understandings a data user would be aware of, but I do not. Because of this, I find the conclusion that accuracy has not been compromised somewhat premature. At the national level, the change may be acceptable, but for estimates at a lower geographic level, there appears to be some large, possibly significant, changes—either with state data or with classification data, such as industry and occupation data.

The authors give no attention to change estimates, perhaps because they are not critical to users, but this assumption should be stated in advance. If, however, change estimates have some importance, are change estimates of interest from month-to-month, quarter-to-quarter or year-to-year? It seems that some additional analysis is necessary—obtaining data from more states and over more points in time. Certainly, a better understanding of the reporting mechanism, the month-to-month changes, is necessary, otherwise each "final" release will require additional research (that is, time, money, and staff work) to understand why estimates change. Without a better understanding of the reporting mechanism, interpreting the early data may be difficult. As the number of states added to the early reporting due date increases, the potential for inconsistent and difficult to interpret data increases.

Bitler and Wolken

The third paper studies the assumption that difficulty in completing an interview (interviewer effort) is correlated with characteristics of the firm in the 1998 Survey of Small Business Finances. This research aims to understand

and improve field operations while also providing more informed understanding of the quality of the data collected. The use of selected interviewers with special characteristics (at converting refusals) is a fairly common method to improve declining response rates; the use of incentives to increase response rates has also become more common in recent years.

The analysis draws several conclusions. First, there was no significant effect between the two pools of "special" interviewers - interviewers expert at converting refusals and interviewers assigned to the specific minority strata - on the distribution of characteristics of reporting firms; and second, that pre-incentives reduce the number of calls to complete the interview. Concerning the first point, the quality of the interviewer, his/her persuasiveness, his/her training, and articulateness are very important considerations in the conduct of difficult to obtain interviews. It seems to me that while there may be some differences between "refusal conversion experts" and "minority pool interviewers" that, in general, their measurable quantifiable characteristics are not all that different. And if an interviewer has particular language expertise, he/she is probably a somewhat better interviewer than the typical interviewer, resulting in reasonable success at converting nonrespondents. With respect to the second point concerning incentives, the analysis suggests what Singer (2001) has recently summarized for household surveys; that is, incentives (pre-incentives, in particular) accomplish the intended effect of increasing response rates, though the size of the effect varies by mode and other factors affecting willingness to respond. And prepaid incentives are more effective than promised incentives. So, the authors' results are consistent with what might be expected by intuition and experience.

Two points are important, but not considered in the paper. First, the paper looks at interviewers in terms of the number of contacts made BEFORE an interview is actually conducted. So, the sample case may be a "difficult" case to convert to respondent status; however, the sample cases are, in fact, converted, and they are, therefore, respondents. A more significant issue, though, concerns cases where substantial effort is expended to convert the case to a "respondent," but the effort is ultimately unsuccessful. These sample cases are likely different from the respondent cases, even the "hard to convert" respondents. This topic is not the topic the authors chose to discuss here, but studying nonrespondent characteristics is an important issue to address. The second point of interest concerns the comparison of "easy to contact" respondents with " hard to convert" respondents on two variables—the length of the interview and the extent of item nonresponse in a questionnaire, using both variables (long interviews and low item nonresponse) as proxies for being a good respondent. The question is: are the "hard to convert" cases who become respondents good respondents who provide complete data?

General Observations

The last paper and its findings lead directly into some general observations concerning the theme of communicating "data quality" information to users of data from National Statistical Offices—a topic to which Haworth and Martin assign high priority. To many, data quality information usually means information about "accuracy," even though the authors of the papers presented today recognize a broader concept. I believe most data users expect statistical and methodological rigor from the National Statistical Office; consequently, high expectations for accuracy exist for data products produced by National Statistics Offices. Because of the expectations of high quality (i.e., accuracy), the availability and timeliness of data and reports may often appear to be more important than accuracy to the data user, even though accuracy is held in high regard.

The high expectations for and assumptions of rigor and accuracy suggest to me that providing information about accuracy **MUST BE** a core or essential product of any data collection program. Haworth and Martin recognize that importance, and U.S. statistical agencies do also, but recognizing the importance of communicating information about accuracy does not always result in action—as the FCSM subcommittee observed in Statistical Policy Working Paper 31(U.S. Federal Committee on Statistical Methodology 2001).

The important words here are "communicating information" - not merely reporting field statistics or operational statistics. A first step is to have such statistics available At a minimum, information about the survey processes and results of survey processes ought to be a product of the survey system, and available for reporting to the data user. But the issue is whether routine operations output can be made acceptable, presentable, and understandable to the many kinds of data users who are at various levels of skill and technical sophistication. Raw process statistics must be available to survey statisticians to manage their programs, but, in general, more synthesis and discussion of these statistics is necessary for them to be useful to the data analyst. Many routine operational statistics relate more or

less to "accuracy" issues, but users may find the interpretation and implications of the statistics difficult to understand. Some effort by data producers to interpret operational statistics for data users is necessary. If systematic, interpretable process statistics are easily and regularly disseminated, the reporting of other aspects of "data quality," coherence and comparability, for example, will be given more emphasis.

Finally, the end-user of the data is important in decisions about the communication of information about "quality." They must be consulted when addressing any issue concerning the general topic of "data quality." End-users should have a hand in describing the information they need about the "quality of survey data." Different users will have different ideas about what information is important, and so representatives of key data stake-holders should be consulted. No one should forget, however, that data producers are also users, and that information about accuracy and the "quality" of the survey operations is necessary for improving the design, management, and operations of the data program. And so it goes—users and producers have a stake in a successful survey program and its output.

The three papers today have given us a multi-tiered view of quality. We should recognize that gauging the worth and success of the topics described in these papers ultimately depends on data user reactions. So as the authors proceed to the next steps of their work, they should seek feedback on their approaches, products, and results and seek advice on priorities for next steps. An iterative consensus-building process with the user community will lead to more useful products. User feedback is critical to successful implementation and acceptance of methods to convey data quality information. Data users, analysts, and policy researchers are our customers. We should never forget this simple but critical point.

REFERENCES

American Association for Public Opinion Research. 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* Ann Arbor, MI: AAPOR.

Andersson, C., Lindstrom, H., and Lyberg, L. 1997. "Quality Declaration at Statistics Sweden." *Seminar on Statistical Methodology in the Public Service.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 26). 131–144.

Arondel, P. and Depoutot, R. 1998. "Overview of Quality Issues when Dealing with Socio-economic Products in an International Environment." Paper prepared for presentation at the XXXth ASU Meeting, May 1998.

Atkinson, D., Schwanz, D., and W.K. Sieber. 1999. "Reporting Sources of Error in Analytic Publications." *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.

Atrostic, B.K., Bates, N., Burt, G., and Silberstein, A. 2001. "Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights." *Journal of Official Statistics.* 17(2): 209-226.

Bailar, B. 1987. "Policy on Standards and Review of Census Bureau Publications." U.S. Bureau of the Census memorandum, June 2.

Brackstone, G. 1999. "Managing Data Quality in a Statistical Agency." *Survey Methodology*. 25(2): 139-149.

Collins, M. and Sykes, W. 1999. "Extending the Definition of Survey Quality." *Journal of Official Statistics*. 15(1): 57–66.

Doyle, P. and Clark, C.Z.F. 2001. "Quality Profiles and Data Users." *Proceedings of the International Conference on Quality in Official Statistics.* Statistics Sweden: Stockholm.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. 1975. "Standards for Discussion and Presentation of Errors in Survey and Census Data." *Journal of the American Statistical Association.* 70(351)(II): 5–23.

Kasprzyk, D., Atkinson, D., Giesbrecht, McMillen, M., Schwanz, D., Sieber, W.K. 1999. "Reporting Sources of Error: the United States Experience." *Bulletin of the International Statistical Institute*. 52: 313-316.

Kasprzyk, D. and Giesbrecht, L. 2001. "Reporting Sources of Error in U.S. Federal Government Data Collection Programs." *Proceedings of the International Conference on Quality in Official Statistics*. Statistics Sweden: Stockholm.

McMillen, M. 2001. "Revising Statistical Standards: An Exercise in Quality Improvement." *Proceedings of the International Conference on Quality in Official Statistics*. Statistics Sweden: Stockholm.

Ramirez, C. 2000. "Report from the Interagency Group on Establishment Nonresponse." *Proceedings of the Second International Conference on Establishment Surveys.* Available on CD-ROM. Statistics Canada: Ottawa.

Seastrom, M. 2001. "Response Rates as an Effective Tool in Managing Data Quality." *Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective.* Statistics Canada: Ottawa.

Shimizu, I. 2000. "Response in Federal Establishment Surveys." *Proceedings of the Second International Conference on Establishment Surveys.* Available on CD-ROM. Statistics Canada: Ottawa.

Singer, E. 2001. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In R. M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A.Little (eds.), *Survey Nonresponse*. New York: John Wiley & Sons. 163-178.

Sirken, M. G., Shimizu, B.I., French, D.K., and Brock, D.B. 1974. *Manual on Standards and Procedures for Reviewing Statistical Reports.* Washington, DC: National Center for Health Statistics.

Statistics Canada. 1998. *Statistics Canada Quality Guidelines*. Ottawa, Canada.

Statistics New Zealand. August 1998. *Protocols for Official Statistics.* Statistics New Zealand.

United Kingdom Government Statistical Service. 1997. *Statistical Quality Checklist.* United Kingdom Office for National Statistics: London.

U.S. Bureau of the Census. 2000. "Standardized Quality Profile Outline for Demographic Surveys at the Census Bureau." Associate Directorate for Demographic Programs and Associate Director for Methodology and Standards. Washington, DC.

U.S. Department of Education. 1992. *NCES Statistical Standards.* Washington, DC: National Center for Education Statistics (NCES 92–021r).

U.S. Department of Energy. 1992. *The Energy Information Administration Standards Manual.* Washington DC: Energy Information Administration.

U.S. Federal Committee on Statistical Methodology. 2001. *Measuring and Reporting Sources of Error in Surveys.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 31).