Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

COMMUNICATING QUALITY: DISCUSSION

Daniel Kasprzyk[1]

The word "quality" seems to be everywhere and on everyone's minds. National Statistical Offices have been consumed over the last ten to fifteen years over "quality" -- managing quality, improving quality, motivating quality, measuring quality, communicating quality. While there are many examples to draw on, let me remind you that in 1990 Statistics Canada convened a conference on the "measurement and improvement of data quality," covering topics such as improving the quality of frame data, improving data processing and estimation, improving data collection, improving administrative records, and improving the fitness for use of our statistical products (Statistics Canada 1991). In 1991, the U.S. Federal Committee on Statistical Methodology hosted a "seminar on the quality of federal data," highlighting topics such as survey quality profiles, quality in business surveys, survey coverage, and computer assisted statistical surveys (U.S. Federal Committee on Statistical Methodology 1991). There have been many other venues during the last 10 years, the most recent being the Stockholm Conference in May and now this conference hosted by Statistics Canada.

In 2002, a conference will take place in Copenhagen on "improving surveys" -- that is, improving the quality of surveys (http://www.icis.dk). There have been invited sessions at the last 2 meetings of the International Statistical Institute, invited sessions at the last three U.S. Federal Committee on Statistical Methodology conferences (U.S. Federal Committee on Statistical Methodology 1997; U.S. Federal Committee on Statistical Methodology 1999: U.S. Federal Committee on Statistical Methodology 2000), a number of papers at the meetings of the American Statistical Association, papers recently published (as cited by the authors in this session) and most recently a U.S. Federal Committee on Statistical Methodology working paper (U.S. Federal Committee on Statistical Methodology 2001) on "measuring and reporting sources of error in surveys." The point is that there is an enormous amount of activity surrounding the subject of "data quality."

This session provides three more contributions to the continuing discussion concerning national statistics offices response to the topic of quality -- in particular, the subtopic communicating quality. The three papers make the important and necessary assumption that national statistical offices have an obligation to report limitations of the data; users should know and understand those limitations; and, as a result of understanding the limitations, users ought to be able to determine whether the data are fit for their purposes.

Bill Allen emphasizes the general user, the presentation of information to the general user, and training for data users. He describes prototypes for dissemination of information on data quality. The first is called a "Quality Issue Summary," which is a short summary of issues related to quality; the second prototype is to provide "improved methods" for presentation of meta data information over the Internet. Bill's paper makes a number of points, but one in particular can easily be overlooked, because it is buried within his discussion. Bill asserts that "technical systems must support presentation and dissemination methods." If this is true with respect to information about data quality, then, at a minimum, basic quality information ought to be a product of the survey processing system. This is essentially the point of view adopted by Census Bureau staff when they recommended the routine development of survey quality profiles (U.S. Bureau of the Census 2000) -- a concept that seems very similar to a "survey quality assessment." It remains to be seen whether routine operations statistics can be made acceptable, presentable, and understandable to the data user. I believe, in general, more synthesis and discussion of these statistics is necessary for them to be useful to most data users. Many routine operations statistics relate more or less to "accuracy" issues. Other dimensions -- coherence, completeness, relevance, and interpretability -- need

[1] Daniel Kasprzyk, Mathematica Policy Research, Inc., 600 Maryland Avenue, SW, Washington, DC 20024, USA.

more work or are simply not addressed in discussions of data quality. Future work targets improved access and presentation of quality measures for tabulations and charts and more flexibility regarding the accuracy presentation. These are useful initiatives, and similar ideas have been suggested in the U.S. Office of Management and Budget's Statistical Policy Working Paper 31 (U.S. Federal Committee on Statistical Methodology 2001). With respect to training, he describes the development of a formal course for data users that assists them in defining their requirements, assessing available data sources, while using the quality framework to make a decision about the use of a particular data source. One difficulty here is how best to focus a course when there are multiple levels of user knowledge and sophistication. Another difficulty is user motivation -- are users interested in expending this much effort to obtain data or would they rather just ask knowledgeable data producers? A key goal for such a course is that it must be developed in a way that results in a multiplier effect -- the idea that students attending this course can train others in their own institutions, thereby spreading information to a much larger group of potential users.

The Full, Haworth, and Stephens paper report that the Office of National Statistics (ONS) will follow the expert Leadership Group on Quality and their recommendations for quality management developments within the European Statistical System (ESS). The recommendations include: National Statistical Offices ought to follow a framework of data quality (relevance, accuracy, timeliness, accessibility (clarity of results), comparability, coherence, and completeness), and, consequently, report according to those dimensions on a regular and consistent manner. To meet demands for "quality information," ONS will embark on a program that has many dimensions: consulting with other National Statistics Offices, establishing internal working groups to standardize definitions, developing quality reports and identifying and addressing methodological problems. ONS will analyze current processes for survey programs through the use of a questionnaire designed to help identify strengths and weaknesses of the survey program. They will also develop a number of new indicators of the quality of the survey processes. Finally, research will be initiated on important unresolved issues, for example, some issues related to sampling error. This is a most ambitious program that will take years to implement in current and ongoing survey programs. Standardizing definitions, response rates in particular, is a difficult and time-consuming process. The questionnaire proposed to assess programs seems like a self-audit but with an independent interpretation. I think, however, a quality assessment that actively engages program staff is a more powerful tool for change.

The Paul Johanis paper describes a Statistics Canada database created for the end user -- an integrated meta database containing information about every aspect of the survey program: questionnaire design, the survey instrument, sample design, data collection, estimation, and "quality measures," etc. The database appears to contain a formidable amount of information. The construction of such a database is a significant achievement, especially since the database is the repository of information for 800 Statistics Canada surveys. This is an important achievement and an archive of information about Statistics Canada surveys. I have not had a chance to access the database, but hope that "information" not "data" resides on the database; that is, the data on the database must be able to be understood and able to be interpreted by the user for his/her requirements. More data may not always be better for the user. He/she must be able to understand what is available and use it accordingly. Also, one hopes users can easily find the information they feel they need.

What's important about the ideas found in these papers? First, there is a strong and unambiguous acknowledgement of the importance of the end user of data and the provision of information to that user. So there is an emphasis on caring about the people who use the data. This certainly follows currently accepted approaches that have had a high profile since the "quality movement" gained popular acceptance over a decade ago. Second, the papers seem to acknowledge (by implication) the failure of the National Statistical Offices to provide data users the information they want. Most acknowledge, the National Statistical Offices come up short, and provide information, but that information may not be that helpful or as complete as the user needs. Third, the papers and this conference specifically continue to help maintain the "high level" profile and discussion of "data quality" within National Statistical Offices, thereby, whether expected or not, influencing program managers, data collectors, methodologists, etc. to continue to seek and implement "quality improvements." Finally all papers acknowledge the desirability to adopt consistent and regular approaches to the presentation of information on "quality."

Several issues are not directly addressed in the papers. No paper makes a distinction between large, ongoing and periodic surveys, and one-time surveys. While the papers suggest all surveys are covered, I wonder whether program managers of small one-time surveys may have different priorities for their surveys than program managers of large, continuous or periodic surveys. Financial and staff resources to support the measurement and communication of data quality may not be available for small one-time surveys, unless substantial aspects of the measurement and communication system are routine, systematized operations. Furthermore, no paper acknowledges the different kinds of data products survey programs produce and the different levels of user sophistication. There appears to be a one model fits all users perspective. This may not be what was intended, but it is pretty obvious that press releases and short reports will have less information reported about data quality than a major data report, data compendium or microdata file. Short "quality issue" summaries may fill the need, but it is not obvious these summaries will be included with a press release. It is also not obvious whether other aspects of quality, besides accuracy, are addressed. Users have various levels of sophistication, and it is not obvious how these different levels are addressed.

I recognize that the purpose of the papers is to describe programs aimed at "communicating data quality" information, and that they do not necessarily address constraints; however, the papers do not acknowledge staffing or cost limitations when developing the appropriate materials for communicating data quality information. Most users obviously feel this work should be completed as part of the survey program budget, but if resources are limited, do we know what the data user is willing to pay for to produce additional -- better/more usable -- data quality information.

I had little sense from any of the papers that data users were consulted concerning the various "data user" projects described in the three papers. Obviously, some followup with users of data products should not just be thought of as useful or helpful, but necessary. It is very important that users have a hand in describing the information they need about the "quality of survey data" and react to the various ways that information is conveyed. Having said this, I expect the National Statistical Offices have a traditional role in presenting "accuracy" information and that this information is available in reasonable detail. Reporting the other dimensions of quality is not as obvious, nor is it obvious what users want. For continuous and periodic surveys, this may be less of a problem than for one-time surveys.

Now, I would like to review the issue of communicating data quality information from several perspectives: the survey program manager, the survey methodologist, and the data user/researcher. I admit that my perspective is largely influenced by my experiences in the United States. So my remarks may not reflect well the situation in National Statistical Offices in other countries.

From the Survey Program Manager's point of view:

- Documenting quality and having information documenting the quality of data collection, data processing, statistical procedures, etc. helps ensure the continuity and improvement of operations. Kasprzyk and Kalton (2001) identify the use of the survey quality profile (survey quality assessment or self-assessment questionnaire) as an internal documentation and management tool.
- Measurement of error sources may complicate data collection, possibly taking funds and staff from daily operations and program maintenance (and then there is the issue of reporting); research is inevitably off-line and secondary, may or may not improve the program, and can be costly. Also a formal program to improve the communication of information about sources of error and data quality take away resources from other parts of the survey program.
- There is a tendency when providing information about "quality" to emphasize the negative -- describing error sources, areas where improvement is necessary, reporting nonresponse rates, etc. This can lead Program Managers to feel somewhat defensive about their survey products.
- An investment in "quality" and communicating information about data quality may not be sufficient to keep budget authorities from reducing the program budget (Depoutot 1999). This point was reiterated in several floor comments at a session on "communicating data quality information" at the 1999 meeting of the International Statistical Institute.

- Developing information about "data quality" -- metadata -- may delay release of the report or the microdata.
- Other aspects of "quality" such as comparability, coherence and completeness are usually less important than timeliness and accuracy.
- The reward structure for Survey Program Managers does not necessarily have explicit rewards or incentives for improving communication about "data quality;" rather, in my experience, the reward structure is typically related to the availability of data early or on time and within budget.

From the Survey Methodologist point of view the following observations can be made concerning the communication of information about data quality:

- Documentation of data quality issues allows the methodologist to identify the most critical areas requiring research. The quality profile concept (Kasprzyk and Kalton 2001) or "Quality Assessment" may be particularly valuable in this instance.
- The methodologist must develop indicators of data quality that are meaningful and useful to program staff and data users.
- The methodologist must synthesize critical data quality information for data users.
- The methodologist must provide understandable data quality information at various levels of user sophistication, including the nonspecialist.
- Developing indicators may be more interesting than communicating such information.
- Methodologists are often more interested in accuracy than timeliness, interpretability, coherence, and completeness.

From the Data User/Policy Analyst/Researcher perspective the following observations can be made concerning the communication of data quality information:

- Most data users expect statistical and methodological rigor from the National Statistical Office; high expectations of accuracy exist for data products from National Statistics Offices.
- Because of the expectations of high quality, the availability and timeliness of data and reports may often appear to be more important than quality to the data user. For example, the availability of microdata and basic file documentation is far more important to users of data than a survey quality profile, or a report on the survey design and estimation.
- Data users want to know about serious data, statistical, or methodological issues before analyzing the data; that is, issues must be "flagged" in a way a data user can not ignore. The Quality Issue Summary may an appropriate vehicle.
- Data users will use the best available known data. Even if serious quality issues exist and are communicated to data users, they may still go forward with their analysis - for lack of any better data set.
- While data quality information is important, it is not obvious how much a user will pay for better information on quality or, if given a set of trade-offs, that such information will remain a high priority. Rather, easier access to data may be more desirable. In other words, different aspects of "quality" are important to different users, and accessibility may prove to be more critical to some users.

As can be seen by these different perspectives, priorities are not always the same; occasionally, the priorities may be in conflict. As survey program managers and methodologists discuss strategies on best approaches to communicate information about the quality of the data, they must always remember whom they are trying to service -- the principal data users -- and these users come at various skill levels. Trying to address all levels simultaneously is not possible; rather, a multiple component system that builds on successive levels of sophistication is necessary. Full-scale adoption of such a system requires management and funding commitment.

While there are many issues still to address, these three papers describe approaches/models from three National Statistical Offices that should provide substantial "feedback;" first, on their usefulness to the data

user; and second, on the question of whether the manner in which we communicate information on data quality is effective.

These efforts are essential. The difficulties of satisfying many users at different levels of analytic sophistication can be overwhelming. I hope to hear about user feedback to these approaches at future conferences. That feedback is critical to successful implementation and acceptance of methods to convey data quality information. Our data users are our customers. Keep them involved in the projects. They remind us that despite how far we feel we have come, we still have a long way to go!

REFERENCES

Depoutot, R. 1999. "Reporting Sources of Error: the Eurostat Experience." *Bulletin of the International Statistical Institute*. 52:317-320.

Kasprzyk, D. and Kalton, G. 2001 "Quality Profiles in U.S. Statistical Agencies." *Proceedings of the International Conference on Quality in Official Statistics*. Statistics Sweden: Stockholm.

Statistics Canada. 1991. *Symposium 90: Measurement and Improvement of Data Quality*. Ottawa, Canada.

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. 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). Available at http://www.fcsm.gov.

U.S. Federal Committee on Statistical Methodology. 2000. *1999 Federal Committee on Statistical Methodology Conference: Complete Proceedings*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 30). Available at http://www.fcsm.gov.

U.S. Federal Committee on Statistical Methodology. 1999. *Seminar on Interagency Coordination and Cooperation*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). Available at http://www.fcsm.gov.

U.S. Federal Committee on Statistical Methodology. 1997. *Seminar on Statistical Methodology in the Public Service*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 26). Available at http://www.fcsm.gov.

U.S. Federal Committee on Statistical Methodology. 1991. *Seminar on Quality of Federal Data*. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 20). Available at http://www.fcsm.gov.