Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 99–102

Comment

Jaki Stanley McCarthy[1]

The goal of National Statistical Agencies is to provide high-quality data to the people of their respective countries. Platek and Särndal ask, ‘‘Can a statistician deliver?’’ Sadly, the answer they seem to suggest is, NO. They contend this is partly due to the fact that the concept of quality is ‘‘nebulous’’ and partly due to a lack of a coherent theory of total survey error.

Platek and Särndal rightly point out that the data user's view is often quite different from the view of those producing the data. The data user simply wants trust in the data, but importantly, data that will suit their purposes. Defining what quality is to a data user, is not a simple task for National Statistical Agencies. In the U.S. in recent years, Federal agencies have been tasked with measuring their customer satisfaction. This has led, in many cases, to the realization that simply identifying who their customers are is very difficult for Statistical Agencies. Official statistics are available to any member of the public, and often anonymously, through libraries, data analysts, the Internet, and other intermediaries. Statistics may be used by many different people, for many different purposes. Each of these people may have very different, yet valid, conceptions of what quality is. Think, for example, of asking people to rate the ‘‘quality’’ of a vehicle. A Rolls Royce may be viewed as a high quality vehicle. But if you want to transport 25 schoolchildren, or 15 bags of cement, or get to work while using the least amount of gasoline, is this a high quality choice?

National Statistical Agencies are trying to satisfy all of these ‘‘drivers,’’ providing each with something they will view as high quality. Because the data users are so numerous and different, this makes defining quality a very difficult task.

Platek and Särndal also point to the lack of a holistic theory behind the production of survey statistics. It is clear this is true, there are currently few areas where a solid theoretical basis exists and error can be predicted or systematically measured. Sampling is one of the few areas that have been the subject of rigorous theoretical development. However, similar progress in developing a theory of nonsampling error, the other half of TOTAL survey error, is woefully lacking. Platek and Särndal point out that sampling error has enjoyed a history with ‘‘major developments and breakthroughs’’ and I doubt many survey statisticians would disagree. But they also point out that many other areas in the production of survey statistics have been neglected. They cite the edit process as one of these areas lacking in theoretical guidance, but I would add many of the other pieces in the process, for example instrument design, data collection, data display, publication and dissemination to this list.

I believe that we will probably never have a theory of nonsampling error that will allow a precise measurement of total survey error. Why? For the simple reason that unpredictable

[1] USDA/NASS, Fairfax, VA, U.S.A.

people are involved at so many points in this incredibly complex process - respondents, interviewers, statisticians, editors, publication authors. The production of official statistics is a complex process which itself consists of a number of complex subprocesses. Scientists have long tried to predict human behavior, but only the most simplistic behaviors can be predicted. People's behavior in complex activities is simply not predictable.

Obviously, some of the subprocesses in the production of survey statistics are amenable to rigorous theoretical treatment, but others are not. This is why much progress has been made in the field of survey sampling and estimation, and not in other areas. Sampling theory rests on a body of mathematics which is easily quantified, measured and modeled. Computer simulations can be run to show the validity of probabilistic predictions. Computers can also be used to compute the immense number of calculations necessary for many complex statistical measures.

Sampling theory has progressed farther than theory in other areas, because these other areas are those where human impact is greatest. Research into the cognitive aspects of survey methods has surged in the past few decades. This research, however, has only provided guidelines for ways to look for potential problems, not predictive power to identify specific problems, nor the ability to measure nonsampling errors. While on the surface, answering survey questions may seem like a simple activity, it is really quite complex. Respondents must hear or see the question, interpret it (correctly), decide whether or not they have the information to answer the question, generate an answer (either from their memory, or other sources), decide whether their answer satisfies the question asked, report their answer, which must then be recorded. Error can arise in any of these stages, and it is surprising the new and different ways in which people are able to generate them. And this is just one of many steps in the production of survey statistics! It is hard for me to foresee the day when we will be able to measure the error generated by an interviewer whose sevens look like ones on the questionnaire, or the one resulting from one respondent defining a ''day'' as daylight hours and another as midnight to midnight, or the error generated because of the decision to edit or not edit a particular data cell, or the error introduced by respondents reporting false data because ''it is none of the government's business.''

The field of psychology has struggled to gain the stature of other sciences like physics, but only the most rudimentary psychological processes have been measured, and those in very artificial experimental settings. The production of statistics is a quantum leap above these settings in complexity and the goal of elevating the production of statistics to a science, despite its being affected by human behavior in so many ways, seems, to me, to be beyond our reach. To use our vehicle analogy, we may have theory on which to base choices of materials to use in drive shafts and the efficiency of fuel combustion or the optimum curvature of the headlight lens, but there is no comprehensive theory of vehicle production which encompasses the entire process of designing and building a vehicle. But is it necessary? Do drivers feel they are not getting high quality vehicles?

Even without a holistic theory of statistics production, we should not give up hope of producing high quality statistics. But the ''collection of practices'' which Platek and Särndal name is likely the most reasonable approach to achieving this quality. Experimental comparisons are a good way to point out the types of problems statisticians should watch out for in the production of statistics. For example, experiments have shown that different methods of asking survey respondents sensitive questions may result in more or less accurate reports. This leads to the general guidance that one should be aware of how sensitive questions are asked, but by no means leads to predictions as to the degree of error in reports of sensitive items (or for that matter, information on what will or will not be considered ''sensitive'').

Similarly, experiments have shown that the order in which questions are asked can affect the answers received. This suggests that questionnaire designers should be aware of this and not ask questions differently with different respondents. But the best use of this information is NOT as a quality indicator, but simply as a reminder to ask questions consistently and in the same order. These experiments have never been intended to measure the error associated with one question order or another, or to indicate which produces ''higher quality'' data.

While we may never have systematic methods of measuring nonsampling error, the good news is that people are very good at detecting the types of errors generated by other people. Because people are involved at so many points in the process, there are also many opportunities to catch errors generated by other people. The formal quality indicators statisticians use, such as nonresponse patterns, imputation rates, reweighting of data, coverage measures, etc., often help direct statisticians to look for the errors that no machine will ever be able to identify. Computer scientists who have tried to simulate human intelligence and decision making have often found that things people find very simple, are incredibly hard (or impossible) to automate. But these skills are the ones which are best suited to understanding the types of nonsampling errors people may introduce into the process.

Platek and Särndal point out there are many different kinds of statisticians involved in the production of statistics, each with their own perspectives and areas of expertise. This is true for statistical agencies throughout the world. Platek and Särndal conclude statisticians deliver quite a bit, but perhaps they are not delivering what the data user wants. Given that a collection of practices is used to deliver high quality statistics, can a statistician deliver? I think the question is rather, ''Can Statistical Agencies deliver?'' Various formal quality indicators are routinely generated and used by statisticians. But these quality indicators are usually used by statisticians to ensure quality in their limited area of responsibility. This does not ensure the final product will be of high quality from the data user's perspective. It is up to the statistical agency to ensure high quality in the process as a whole.

This is why it is critical to have strong leadership in a National Statistical Agency which will take responsibility for putting in place mechanisms to ensure overall quality. This may take many forms but must be designed from a perspective overseeing the ENTIRE ''collection of practices.'' This may mean a collection of quality measures combined with quality reviews from a ''Total Quality Management'' perspective. Statistical agencies often combine narrow quality measures with scheduled quality reviews of the entire process and resulting statistics. This may be done by a permanent staff or it may be done by temporary teams put together to evaluate the quality of specific data products or in response to specific concerns about the quality of an individual data product. In either case, these staffs must be supported by the leaders of their agencies in order to be allowed the resources to assess their products' quality and for their assessments to be taken seriously and recommendations acted upon.

Statistical agencies often do not publish many quality measures or assessments of data quality. But data users usually do not complain about this. The data user simply wants to have, as Platek and Särndal state, a ‘‘feeling of general trust in the agency, in the data it publishes and in its design and process work.’’ An agency can only earn this trust by continuously measuring and maintaining (or improving) the quality of its data based on results from its own internal quality reviews. These need to take place throughout the agency and from a variety of perspectives to ensure quality in the collection of practices used to generate National statistics. This is an essential responsibility of any National Statistical Agency if they want the trust of data users. But in the end this multiple level approach means statistical agencies will be able to deliver.

Received December 2000