VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 437–442, http://dx.doi.org/10.2478/JOS-2014-0027

Rejoinder

Paul Biemer, Dennis Trewin, Heather Bergdahl, and Lilli Japec

Our sincere thanks go to the discussants for their thoughtful, positive and, in some cases, critical comments. Collectively the comments provide many fruitful ideas for strengthening and improving ASPIRE as the system continues to evolve at Statistics Sweden and at other NSOs who may adopt ASPIRE wholly or partly. We are optimistic that ASPIRE can only improve if it is applied and the results, both positive and negative, are shared frankly and openly. Therefore, we fully endorse Fritz Scheuren's suggestion of establishing an international group to "share knacks and lessons learned" on ASPIRE and other similar approaches. As Statistics Sweden continues to apply ASPIRE, our intent is to continue to report our experiences at conferences, presentations, and in publications.

Fritz Scheuren is absolutely correct in referencing the great quality gurus Juran and Deming regarding ASPIRE. The authors took much inspiration from the work of these pioneering innovators in quality management. As Scheuren suggests, the application of ASPIRE at Statistics Sweden is already having a Kaisen effect in that incremental, continual quality improvement, as promoted by ASPIRE, is becoming engrained in the culture of the organization. Kaisen is definitely taking root there.

David Dolson's remarks clearly illustrate that ASPIRE is just one approach for possibly achieving similar objectives in an NSO. Statistics Canada's Quality Secretariat has been implementing a Quality Review program each year since 2007. Stats Canada's Quality Review program shares some similarities with ASPIRE. For example, like ASPIRE, their program attempts to identify the major risks to data quality and how to mitigate them across multiple programs. However, as Dolson notes, there are key differences. Reviewers are internal to the organization and there are no quality criteria nor are results reported in a numerical format. According to Dolson, the Canadian system does not possess the rigor and comprehensiveness of ASPIRE which, he believes, provides more robustness and greater consistency across products and greater comparability across organizations. However, one aspect of the Quality Review program that we may wish to adopt for ASPIRE is the emphasis on identifying and sharing of best practices across products, not only those under review but across all products in the organization. ASPIRE does this to some degree in its report to management of all products' ratings with their justifications. Also, the reports highlight a number of major "cross-cutting" issues which we know the Executive of Statistics Sweden has found most useful. However, ASPIRE tends to focus on the poorer practices. Drawing out best practices more emphatically and formally could be an important improvement for ASPIRE.

Dolson makes a number of excellent points in discussing the benefits and challenges of using internal and external reviewers. Unlike ASPIRE which uses the same two reviewers

for all products, Stats Canada assigns different, internal review teams for each program. Independence of these reviewers is addressed by ensuring they come from different organizational areas than the programs under review. Although the Stats Canada Quality Secretariat is pleased with the impartiality of the reviewers, we think review objectivity is a difficult attribute to assess and are skeptical that internal reviews are always objective in critical and sensitive situations. Statistics Sweden internal evaluators tended to report no concerns regarding product quality and few areas needing improvement. Quite a contrast to the ASPIRE findings. We agree with Dolson that the use of external reviewers would address any suspicion of partiality of reviewers and would enhance the credibility of the evaluation process.

At this point, we should note that ASPIRE is just one component of Statistics Sweden's quality management system. Because Statistics Sweden is ISO 20252 certified, all statistical products must meet these minimum standards. ISO 20252 provides a quality framework with requirements for numerous processes such as interviewer monitoring, keying, coding, and disclosure control. It requires an ongoing program of internal compliance monitoring which is performed by internal quality auditors who, like at Statistics Canada, are selected from outside of the department being audited. The purpose of these audits is simply to determine whether the ISO 20252 standards and guidelines are being appropriately followed. For this purpose, internal auditors can perform well with a good measure of objectivity. On the other hand, for ten of Statistics Sweden's most important products, ASPIRE strives to achieve a much higher level of quality than is ensured by ISO 20252 alone. As previously noted, attempts at Statistics Sweden to use internal evaluators for this higher purpose have not succeeded and, thus, the external evaluators were called in.

Eva Elvers provides a whole host of comments from someone who has experienced the ASPIRE process first hand at Statistics Sweden and can, therefore, draw upon her experiences with ASPIRE from within the organization. Many of her comments are largely about semantics; for example, the use of the word "error"; whether TSE includes specification error; when to use "estimate" versus "estimator" and issues with other terms we use that may differ slightly from the way some at Statistics Sweden would define them. However, our terminology is consistent with the TSE literature; for example, the term "error" has been used in this literature for more than 70 years.

The entire ASPIRE process including definitions, terms, criteria, and so on has been and continues to be thoroughly vetted at Statistics Sweden. For example, the ASPIRE evaluators have met with the survey methodologists at Statistics Sweden many times (both in Stockholm and in Örebro) over the course of three years to solicit their comments and suggestions on all aspects of the ASPIRE approach. There still remain some issues, particularly regarding terminology, where unanimity was not possible and it was necessary to form a consensus in order for the process to move forward. That there remain lingering questions in this area is neither surprising nor problematic, in our view. For the next round of ASPIRE (Round 4, which will commence in December 2014), we will continue these discussions that we are sure will be both fruitful and enlightening for all involved. Realistically, in any organization of highly intelligent and independent minds, there will always remain areas of disagreement and, thus, consensus must substitute for unanimity to make progress.

Thus, we will not attempt to address semantics here, opting instead to address three of Eva Elvers more substantive questions or comments as follows:

1. Does it make sense to treat user dimensions as constraints when optimizing Accuracy?
2. What are the advantages of using numerical scores?
3. What motivated the five quality criteria used?

Point (1) was initially proposed in Biemer and Lyberg (2003) and further expounded and illustrated in Biemer (2010). Essentially, maximizing accuracy while being constrained by the other quality dimensions simply means that the resources for maximizing accuracy are somewhat constrained by the survey budget once the budget necessary to meet the specifications for the other dimensions has been allocated.

For example, regarding Timeliness and Accessibility/Clarity, the survey design may specify that data collection for the survey should be completed within nine months, and that data files will be released to the public within 15 months. The design may further specify that data files will be provided for download online with full documentation at the time of release. For Comparability, methodologies used in previous implementations of the survey should be continued in the new implementation. Ideally, the survey budget should take into account these objectives in the allocation of resources for the survey.

Now let C_T be the total budget for the survey and C_U denote the combined, estimated costs for achieving the specified objectives for the user dimensions. The remaining budget (i.e., $C_A = C_T - C_U$) is the budget available to maximizing Accuracy. The task for the survey designer is to implement the data collection, data processing, weighting, and estimation phases of the survey to maximize Accuracy, while ensuring that survey costs do not exceed C_A and the time from the start of data collection to the release of data files does not exceed 15 months. Thus, the design specifications for data collection, data processing, weighting, and estimation should ideally minimize TSE subject to these constraints on the total budget. This approach attempts to maximize the total survey quality once the design objectives and specifications under each dimension are set in accordance with both user and producer requirements.

As Biemer (2010) notes, the optimization strategy is likely an iterative process because the designer may realize that the budget, C_A, is inadequate to achieve an acceptable level of Accuracy. If additional resources are not available, then the user dimensions should be respecified in collaboration with users and the budget C_U reduced as necessary to free up resources for Accuracy. Of course, the impact of the budget reallocation on the most important user quality dimensions should be minimized.

John Eltinge provides an excellent point related to (1) in Subsection 3.2 of his comments. He notes that there are no established standards for the user dimensions and the NSO may wish to experiment with alternative specifications of the user objectives to better understand the trade-offs among the quality dimensions as well as between C_A and C_U.

With regard to (2), the Swedish Ministry of Finance directed that the presentation of the results of the quality reviews be concise, transparent, and accessible to administrators and stakeholders who are not familiar with the many, complex details of the statistical production process. The Ministry also placed priority on indicators that reflect quality improvements. Experience with ASPIRE has clearly demonstrated that the numerical

ratings and the graphical displays (particularly, the Harvey balls) satisfy these directives quite effectively. An obvious disadvantage of this simple approach is the risk of oversimplification. For example, a product's ratings may have improved from one round to the next for two high risk error sources, say A and B. However, the improvements for A may have much greater influence on overall data quality than the improvements for B. Of course, this information is not contained in the rating symbols. Digging into the details behind the improvements will reveal the true story, but that will require reading the report rather than simply relying on the ratings matrix.

Another risk of using numerical ratings is that staff may believe the goal is to improve the product's rating rather than to improve the product's quality. This is not necessarily a bad thing as long as improving the product ratings will result in real improvements in product quality. So far, ASPIRE has shown that improving quality will improve ratings and vice versa.

It is worth noting that ASPIRE can be easily customized for to suit the requirements of an NSO. It does not need to be applied precisely as described in the article. Indeed, at Statistics Sweden, there have been some important modifications through the first three rounds in light of experience and a few of these are described in the article. Elvers suggested a different structure for assessing risk. We are not convinced that the additional detail is needed but ASPIRE could easily incorporate this more complex risk assessment structure if it were deemed desirable.

Regarding (3), the five quality criteria were developed after numerous discussions among staff at Statistics Sweden and the evaluators. Together, we believe they span the scope of quality improvement attributes for most products. Knowledge of Risks seems an obvious starting point for quality improvement and its inclusion is well-supported in the literature. As an example, this criterion appears in the evaluation criteria for analytic reports published by the U.S. Office of Management and Budget OMB (2001, pp. 2–6). Further, as Deming famously said "Lack of knowledge . . . .that is the problem" (Deming, n.d.,). Communication with Users (two-way communication implied) is believed to also be essential for improving quality for two reasons: (a) users provide important knowledge about quality that can only be obtained through using the data and (b) users often will ramp up the pressure on an organization to improve quality for a specific product. Such pressure is often needed in organizations where there are few resources for quality improvements and many quality improvement needs. In Round 2, we added "Communication with Providers" (again two-way) to this criterion after realizing that providers of data for a product have a profound influence on product quality and need to be "kept in the loop" regarding how poor quality of the data elements they supply might affect overall product quality.

For quality improvement efforts to be effective, the necessary expertise should be available and applied to the product. Thus, Available Expertise is an important aspect of ASPIRE and may explain why progress on real quality improvement is lacking despite the substantial efforts and resource investments. At a minimum, product design and activities should comply with whatever standards are applicable including national or EU standards as well as the NSOs own standards. However, in ASPIRE, such compliance only rates "Good" on the five-point scale. Compliance with Best Practices raises the bar and is included in order to guide products toward practices that equal or exceed the state of the art

with regard to a particular error source. Finally, no improvements can take place without planning to improve and realize those plans. Therefore, the inclusion of Achievement towards Mitigation and/or Improvement Plans is an obvious and essential criterion that reflects real progress toward error risk reduction.

Elvers raises the question of whether all these criteria are needed. She asks: if a product rates a perfect score on criteria 1 and 5, are criteria 2 to 4 then superfluous? We think not. We believe a product would not be able to attain perfect scores on criteria 1 and 5 much less sustain them, without attending to the other three criteria. Communication with both providers and users, adequate expertise to address quality issues, and attention to standards and best practices are critical and essential attributes for achieving high quality.

She raises a good point regarding the evaluation of registers where the estimation of MSE components, which is ASPIRE's primary metric for estimators, does not apply. Registers, like data sets more generally, are comprised of rows and columns whose intersections create cells that contain values which may be either erroneous or missing. Rather than bias and variance, ASPIRE substitutes more appropriate metrics to describe the error in the register data; in particular, validity and reliability for gauging systematic and variable errors, respectively. These metrics can even be used to capture the error resulting from missing values if the missing values are imputed either using simple approaches such as mean imputation or more complex, model-based approaches, if available. Approaches for assessing the quality of register data are very much in a nascent stage and more work is needed in this area; nevertheless, we believe our classification of error sources for registers is a useful starting point.

We very much appreciate John Eltinge's further elaborations on some of the more challenging concepts in the article. Due to space, we limit our response to two important points that are particularly relevant and have not yet been touched on in this response. First, we agree with his comment in Subsection 2.1 that "quality problems can arise from deeper management issues." This is true for any organization and Statistics Sweden is no exception. Many of these problems relate to communication issues, collaboration barriers, questions regarding responsibility and authority and other problems brought about by organizational "stove piping" (as commonly observed in large-scale statistical organizations), complex management structures, and the ever-changing external environment. Naturally, in the course of conducting in-depth interviews with each product team, ASPIRE identifies such problems and it is completely in the scope of the review to report them to management. For example, in the Round 3 report we noted "a lack of co-operation between the National Accounts staff and data providers," also for "some statistical areas the need to improve the relationship between the IT department and their client areas", and "the lack of succession planning in some statistical areas." Issues of a more sensitive nature were conveyed orally to top management rather than in the written report and there were several of these as well.

Second, in Subsections 2.2 and 2.3, Eltinge rightly notes that it can be quite difficult for an NSO to determine the high risk and high priority areas to address when the budget is inadequate to address them all. An example of the hypothetical situation he posits is measurement error (error source B in his notation) versus household nonresponse (error source C in his notation). Particularly for the LFS, considerable resources have been directed to understanding the causes of nonresponse and reducing its effects on the

estimates. However, in terms of the "quality improvement per monetary unit," the return on investment (ROI) may be quite low relative to the ROI for measurement error for the same expenditure. Possibly redirecting even a fraction of nonresponse reduction resources towards understanding the causes and reducing the effects on measurement error on the estimates might result in a much greater ROI. Unfortunately, the data necessary to compare these two ROIs are often not available but could be obtained through appropriately designed evaluation studies. ASPIRE seeks to promote this view to counter the sentiment that response rates must remain high to ensure confidence in, and credibility of, the survey. Often, the latter view drives the decision to expend more and more resources to incrementally increase response rates, with little or no improvement in TSE.

Decisions on resource allocation for quality improvement are rightly the responsibility of management. We believe that ASPIRE assists them greatly in this important task by identifying those error sources with high risk with relatively low ratings.

We very much appreciate and value the comments of the four discussants and will continue to consider them as we move forward with ASPIRE. They contain many excellent suggestions and ideas for improving ASPIRE and, more generally, for developing better processes for statistical production. Thanks also to JOS for providing this forum and the journal space to fully discuss this important topic for NSOs world-wide.

References

Biemer, P.P. 2010. "Total survey error design, implementation, and evaluation." *Public Opinion Quarterly* 74: 817–848. DOI: http://dx.doi.org/10.1093/poq/nfq058.

Deming W. Edwards. (n.d.). BrainyQuote.com. Available at: http://www.brainyquote.com/quotes/quotes/w/wedwardsd380788.html (accessed July 28, 2014).

Office of Management and Budget 2001. "Measuring and Reporting Sources of Error in Surveys". Statistical Policy Office, Working Paper 31. Available at: https://fcsm.sites.usa.gov/reports/policy-wp/ (accessed August 7, 2014)