VERSITA

Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 425–429, http://dx.doi.org/10.2478/JOS-2014-0025

Discussion

Eva Elvers[1]

1. Introductory Notes

First I would like to thank the Editors for inviting me to contribute to this discussion in the *Journal of Official Statistics*. Their motivation is that I am a person with long experience in Statistics Sweden but without direct involvement in the studied statistical products. Hence, my understanding of the ASPIRE model contains aspects that are both exterior (to the statistical products) and interior (to the organization). The comments are mine and not an official view from Statistics Sweden. They are a selection from my personal thoughts and experiences of the model as described in the article and of the related work at Statistics Sweden. Some of them are a bit provocative to stimulate the discussion.

The work has focused on accuracy and a key set of ten statistical products. These products are indeed diverse with registers and with surveys, including compilations, which use direct data collection, administrative data, other statistics, or a combination. Two of the results so far are: (i) the ASPIRE model, which has been presented at several international meetings and now in the *Journal of Official Statistics*, and (ii) three successive reports to Statistics Sweden with measures, comments, and suggestions. Moreover, and most important, the ten statistical products have made quality improvements. There have been further effects in the organization.

I will largely use the same terms as Biemer et al. and often without explanation. I will, however, use the term estimator (rather than estimate) in case of a random variable. Moreover, I will be a bit Swedish-oriented where I find the distinction important.

2. The Model, Its Ingredients, and Aims

The ASPIRE model involves a lot of information in an often condensed way. It is elegant in several respects. Fully used, all quality dimensions are included. The work for Statistics Sweden has focused on accuracy and so does the article.

2.1 A Common Understanding – Concepts and Terms

Communication between external evaluators and statistical products being evaluated may not be easy. Concepts and terms that are used by the experts in their model should preferably agree with those already used in the organization being evaluated. In this case, with Statistics Sweden, quality concepts and terms of the European and the Swedish

[1] Statistics Sweden, Process Department, P.O. Box 24300, SE-10451 Stockholm, Sweden. Email: eva.elvers@scb.se

statistical systems could have been used. For example, a mapping between ASPIRE and Europe/Sweden could have been made to show differences clearly. The first evaluation period was intensive at a short notice, but further efforts towards a mutual understanding of concepts and terms could have been made successively. Some unnecessary confusion at Statistics Sweden – particularly around specification error and the Swedish template for quality declarations – could have been avoided in such a way.

2.2 *Quality Dimensions*

It was natural to start the evaluation at Statistics Sweden with accuracy. However, in my view Biemer et al. down-weight the other quality dimensions too much when they say in general: "To most statisticians and data analysts, good quality is synonymous with estimates having small mean squared errors (MSEs)". Coherence and comparability, for instance, are also important. Nor do I agree with the cited view on accuracy as "the dimension to be optimized in a survey while the other dimensions (the so-called *user dimensions*) can be treated as constraints during the design and implementation phases of production". Accuracy needs to be balanced with other quality dimensions, for example timeliness. So, further quality dimensions could have been considered together with accuracy to give a more complete picture.

2.3 *Error Risks Are Important and Deserve More Motivation*

Biemer et al. emphasize error risks, which are an important part of their model, and they use two types of risks. They emphasize the difference between *inherent* risk and *residual* risk, which refer to situations without and with the current efforts, respectively. They discuss risks, in principle and as part of the model. They say that risk involves both likelihood and impact. I have two major questions.

Residual risk is described in the text and it is assessed, but it is not really visible in the presentation through tables. Why is the residual risk not given a column in the matrix with quality ratings – does it not deserve a more explicit role in the model than now seen?

The Australian Bureau of Statistics, ABS, has a statistical risk assessment framework, where the building blocks are clearer to me. There are five levels for each of Likelihood and Consequences. These levels are combined, resulting in four levels of risk from low to extreme, as ABS (2010) describes. Would it be possible to clarify or expand ASPIRE in a similar, more explicit, way? Were there specific features of the Statistics Sweden management or statistical production process that would not have fit with the obviously different ABS quality approach (probably well known to the second author)? Does this possibly provide an indication of ways in which ASPIRE may need to be "tuned" to features of other statistical offices?

2.4 *An Informative Matrix – But it is Dangerous to Use Numerical Scores?*

A lot of information is condensed into the ASPIRE matrices. For instance, the matrix in Table 2 shows error sources and evaluation criteria together with risks to data quality and changes from the previous round. Symbols and fonts provide information in a compact way. This table gives a good overview of one statistical product.

However, I find the numerical scores over-simplified and a bit dangerous, for instance for evaluations and priorities for further work. Three simple examples (where the last one includes weighting):

- Biemer et al. give somewhat double messages stating both that the numerical scores "can be used for comparisons across time and products" and that "The interpretation of such comparisons may not be straightforward for several reasons".
- Biemer et al. suggest putting priority on the areas having highest risk and lowest ratings, if other factors are equal. That reasonable guidance is not obtained from the numerical scores alone.
- I heard about a case where the rating of one inherent risk was changed from middle to high. As a result of this higher risk, the overall quality score increased. This was found counter-intuitive for the quality level as such. – It is possible, of course, due to the use of a weighted average of scores.

I would like the authors to clarify the advantages of using numerical scores in addition to categories – and the disadvantages.

2.5 The Evaluation Criteria – Some Questions

As an illustrative example, if the highest ratings for the two criteria Knowledge of Risks and Achievement Towards Mitigation or Improvement Plans are achieved, the MSE is known and under control for the primary purpose of the statistical product, according to Appendix A in the article. This is desirable, and with this achieved the other three criteria seem unnecessary. I would have liked to see more motivations behind the selection of the five evaluation criteria, which have all been included in the model as important factors for product quality.

In particular, I wonder about the criterion Communication with Users, which seems to refer to a one-way communication *to* users. The users will, of course, be able to use the statistics in a better way. Still, that communication does not influence the accuracy per se. Moreover, documentation, quality declarations/reports etc. belong to Accessibility & Clarity. However, a *bi*directional communication is likely to improve priorities and balances between quality dimensions and perhaps also between Accuracy components. Such balances are related to Relevance, too. I would have liked to see Accuracy together with other quality dimensions and components, as already indicated; and also to see process quality more clearly.

A further question relates to the situation where the statistical product is a register: how are the evaluation criteria formulated when it comes to MSE?

2.6 Accuracy and Total Survey Error With Decompositions

The Swedish quality concept and quality declaration – with its hitherto descriptive listing of quality dimensions/components (Statistiska centralbyrån 2001, especially pp. 33–34) – uses the following sources of inaccuracy: sampling, frame coverage, measurement, nonresponse, data processing, and model assumptions. These components are fairly standard, except that "model assumptions" has a pronounced position. It is used for inaccuracy in addition to that from the other sources. The word *error* is deliberately

avoided, whereas Biemer et al. naturally use it in their decompositions of the total survey error in their formula (1) and the surrounding text.

When accuracy is measured by the MSE, both the estimator and the population parameter to be estimated (the target parameter with a Swedish term) play vital roles – and so does the random mechanism, which is not discussed here.

Concepts (specifications) are indeed essential. With a notation similar to that in the article formula (1), a simple survey with a collected/observed variable y uses an estimator $\hat{Y}$ of the target parameter Y. There are other, more complex, cases where the collected variables and the variables of the target parameters do not have simple correspondences. Also, users may desire to interpret or use statistics with the target parameter Y as if they were statistics with a somewhat different target parameter X. In my notation X is not necessarily unobservable. The producer of the statistics then has to be clear about the ingredients of the presented statistics, including the target parameter: whether it is Y or X. The accuracy of one and the same estimator differs, of course, between these two situations with different targets. The producer might prefer to use different estimators, though, in the two cases.

The statistical product Foreign Trade of Goods (FTG) is a bit complex with respect to variables, as Biemer et al. describe. Simply expressed, the collected invoice value, y, is converted into the target statistical value, x, with the aid of a specific sample survey that collects x in addition to y. Hence, the target parameter X is estimated by a direct estimator, which I would prefer to call $\hat{X}$ – not $\hat{Y}$ as Biemer et al. do. The inaccuracy caused by this FTG procedure (which includes the observed y and an estimated conversion model) is put under the heading Model assumptions in the Swedish quality declaration. Biemer et al. instead use two error sources, Specification error and Model/estimation.

As for the article sentence "Some would argue that specification error should be part of the Relevance/Contents dimension", I would say – which is quite different – that the choice of target parameters has its place there, including that this choice influences the relevance of the statistics for a user with a particular interest. As for Accuracy, I already described my view, encompassing the MSE. Dissimilarities between what is observed, targeted, and desired – whether variables or parameters – will, of course, come into play somehow, depending on relationships, modeling etc. The statistical product FTG provides just one example.

Revision error is an unfortunate term, since the revision activity normally means an improvement, where one or more preliminary estimates are successively modified, arriving at the final estimate. Would it not be better to talk about revision size?

3 Some Concluding Remarks

3.1 The Words from Experts Are Heard

Biemer et al. state some advantages with external evaluators in comparison with internal ones. I find it interesting and important to observe also that a suggestion made by an external expert automatically gets attention – more attention than the same suggestion made by somebody internally. It is more likely to be taken as crucial and to lead to activities. As an example, a development project on methods to assess measurement errors

started last year, partly because Biemer and Trewin emphasized this as an error source with high risk in many of the evaluated statistical products.

3.2 Avoid a Strong Person-Dependency

In the short run, it is convenient to have the same evaluators in order to save time and to measure changes in a reliable way. There may still be difficulties, though, as Table 3 indicates: the evaluators have changed some of their own assessments from the previous year. Biemer et al. discuss, similarly, inter-rater variation and ways to reduce such variation. How strong is the current, remaining, person-dependency of the assessments in the ASPIRE model?

In the long run, there are advantages to have further expertise views. When the Scientific Advisory Board of Statistics Sweden discussed the ASPIRE model, there were warnings about measuring the same thing and using the same evaluators over time. The scope may then be narrowed to what is measured and to certain aspects.

3.3 Is the Model Essential?

There are many benefits to Statistics Sweden from the work by Biemer and Trewin. I would say that some major reasons are their (i) expertise, structured discussions about quality, and ways to encourage and note improvements; in combination with (ii) high priority of this work at Statistics Sweden together with internal knowledge – knowledge that has become more visible and also expanded.

I cannot help wondering how essential the model is for the results achieved. Would the same conclusions and improvements have been made if Biemer and Trewin had chosen a different model or route for their work? My guess is "largely yes". This reflection should not be interpreted as a criticism of ASPIRE. It is rather a suggestion to reconsider some ASPIRE ingredients and priorities. Some examples are the numerical scores, the evaluation criteria, and clearer connections to process quality. Some guidance to quality in relation to costs would be interesting but is quite demanding. Such a reconsideration of the model might decrease the person-dependency and broaden the perspective.

In line with this, I would even say that it is useful for an organization to consider its model(s) for quality evaluations with some regularity. There may be good reasons to modify the focus and emphasize, or even add, new priorities. Also, statistical offices may learn from each other.

4. Reference

ABS 2010. Quality Management of Statistical Processes Using Quality Gates, Dec 2010, Appendix ABS Statistical Risk Assessment Framework. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1540.0Appendix1Dec%202010 (accessed June 23, 2014)