Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 553–554

# Rejoinder

*Ting Yan, Frauke Kreuter, and Roger Tourangeau*

We thank the Editors-in-Chief of the Journal of Official Statistics, the reviewers, and the discussants for their comments on and discussion of our article. We especially appreciate it that they all point out, in various ways, the difficulties and challenges of conducting comparative studies like ours. We took a first attempt (perhaps an imperfect one) at it because we believe that difficulties and challenges are not an excuse for not trying, and that an imperfect attempt is better than no attempt at all.

One challenge with a comparison of question testing methods is the large differences between the different question evaluation methods. We decided on a basic metric – whether an item was classified as problematic – that could be easily implemented and compared across different question evaluation methods. This metric may not fully capture the products of a particular evaluation method. But we think many questionnaire designers sort draft items in a similar way, deeming some items as needing more work and other items as ready for administration. In addition, the use of this metric allows readers to easily connect our findings back to the existing literature on methods for testing questionnaire items. In the spirit of advancing research on question pretesting and evaluation, we encourage researchers to build on this simple metric and to propose other criteria that better capture the unique contribution of each question evaluation method. We are happy to make our data available to researchers who are interested in seeing whether alternative schemes for classifying our items would have produced different conclusions.

We do not necessarily disagree with the thinking that convergence should not be expected from these very different question evaluation methods. However, simply dismissing the convergence as a criterion for evaluating different question testing methods does not, it seems to us, push the science further. As we mentioned in our article (and we reiterate here), "the answers to the questions of whether converging conclusions should be expected and how to cope with diverging conclusions about specific items depends in part on how researchers conceive of the purpose of the different evaluation methods." In this regard, we agree with the discussants that the next steps for continuing this research is to outline circumstances under which convergence (or divergence) should be expected, and to identify circumstances under which each of the different methods is likely to be useful. Still, we continue to think it was quite reasonable for us to start with the assumption that the problems detected in cognitive interviews and those pointed out by expert reviewers *should* be related to the item's validity and reliability. If the "problems" detected by a given method are unrelated to whether the item produces reliable and valid answers, it is not clear to us what the value of the method is for evaluating questionnaire items.

We did not intend to criticize any qualitative question evaluation methods and we do not endorse any quantitative evaluation method either. However, we do think it is important for

future research that those who advocate the use of a particular qualitative method make it clear what unique insights or contributions this method is supposed to provide so that these claims can be evaluated. For instance, one discussant pointed out that cognitive interviewing is practiced in various forms. A critical question then becomes what insights cognitive interviewing offers when the goal is to understand survey questions better, and what insights cognitive interviewing provides when the goal is to detect problems with a particular survey question and to fix those problems. We think it is equally important that advocates of each quantitative method make it clear what assumptions are required to apply the method and to specify the circumstances under which the method may fail because the assumptions are not met. In our examination of latent class analysis, we have demonstrated empirically that when the local dependence assumptions are violated or when the model-identifying assumptions are not met, the latent class method can yield inaccurate estimates of error rates and very implausible results about the differences across different modes of administration (Kreuter, Yan, and Tourangeau 2008; Yan, Kreuter, and Tourangeau 2012).

To advance research on question pretesting and evaluation and to enrich survey literature, we believe that the field needs more studies that include solid measures of validity and reliability on the one hand, and that employ multiple question evaluation methods on the other. In this way, question evaluation methods can be compared on questions with known psychometric properties. This is probably too ambitious a goal for one study. However, as studies and evidence cumulate over time, it will strengthen research on question testing and evaluation in particular and on survey research in general. Good examples of accumulating evidence from question evaluation studies include QBANK started by the National Center for Health Statistics (NCHS) in the United States (http://wwwn.cdc.gov/qbank/Home.aspx), QDDS in Germany (http://www.qdds.org/. See also Schnell and Kreuter 2001), and SQP (http://www.sqp.nl/. See also Saris et al. 2011). We advocate similar efforts to start accumulating experiments and other studies comparing different evaluation methods. Our main point is that we cannot simply continue to take it on faith that the methods we use for evaluating survey questions actually yield helpful insights.

## References

Kreuter, F., Yan, T., and Tourangeau, R. (2008). Good Item or Bad – Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions. Journal of the Royal Statistical Society, Series A, 171, 723–738.

Saris, W.E., Oberski, D., Revilla, M., Zavala, D., Lilleoja, L., Gallhofer, I., and Gruner, T. (2011). Final report about the project JRA3 as part of ESS Infrastructure. Available from: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf.

Schnell, R. and Kreuter, F. (2001). Neue Software-Werkzeuge zur Dokumentation der Fragebogenentwicklung. ZA-Informationen, 48, 56–70.

Yan, T., Kreuter, F., and Tourangeau, R. (2012). Latent Class Analysis of Response Inconsistencies across Modes of Data Collection. Social Science Research, 41, 1017–1027.

Received October 2012