Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 371–374, DOI: 10.2478/jos-2013-0031

Rejoinder

J. Michael Brick

I want to thank the editors of the *Journal of Official Statistics* for inviting me to prepare this article and for obtaining such a distinguished set of discussants for it. As I prepared the article, I quickly realized that reviewing the massive literature on nonresponse and nonresponse bias is a daunting exercise. It has given me even greater appreciation for those who have done such excellent research in this area.

I would also like to thank all the discussants. Their comments give valuable insights into nonresponse bias and I found their remarks very stimulating. I would also like to thank the discussants for pointing out issues in my initial draft; their suggestions helped me improve the quality of the article.

The diversity of the discussants' comments and concerns highlight some of the challenges we face dealing with nonresponse in surveys. Below, I briefly address some key similarities and differences I have with comments provided by each of the discussants.

Professor Loosveldt perceives my review as being pessimistic, and I understand this reaction. My review tried to paint the challenges of nonresponse as starkly as possible. However, I share his optimism about making progress, but only if we face the complex issues associated with nonresponse. Our field has many skilled and innovative methodologists and, if they work on nonresponse diligently, then I believe we will see significant improvements in our understanding and methodology.

Loosveldt notes that nonresponse has more levels and complications than are discussed in my review. I fully agree and would like to add to those he mentions factors such as the mode of data collection and 'house' or organizational effects. The effects of these factors can be substantial. He also mentions that the survey climate interacts with response propensities. I again agree, but note that thus far we have not done well in specifying exactly how the interaction works (Brick and Williams 2013). As he describes, the entire system, including the data collection process and other cultural factors, could be critical to nonresponse bias and we need to better understand these effects and how national statistical organizations can influence them.

Professor Loosveldt's extension to include cross-country comparisons provides a nice perspective on the nonresponse problem. His idea of capturing data in a doorstep interview seems to be an option worth pursuing. This idea appears to be related to those of Kulka et al. (1982) and Lynn (2003).

A final point of clarification is that I did not intend to suggest nonresponse was the respondent's responsibility. Rather, I was trying to urge those of us who mount surveys to take a more respondent-friendly orientation. In the early days, respondents may have been more willing to do any survey, but those days (if they ever existed!) are gone. It is the survey's responsibility to be more respectful in asking respondents to spend their time on and give attention to a task they did not initiate.

Dr. Kaminska explains several of the complex issues related to nonresponse bias in ways that I found informative. I especially enjoyed her discussion of overall and conditional response propensities. She clarifies when and why each type of propensity is important to researchers. I agree with her that at the data collection stage the conditional propensities are essential, while at the weighting stage only the overall response propensities are important.

She also offers fresh ideas for weighting in a two-phase design. Her Method C seems very reasonable, and she clearly points out difficulties that might arise associated with computing the response propensities required for her method. I encourage her to pursue the research necessary to evaluate the statistical properties of her proposed method because I believe it has promise.

In describing her proposed method, she states that we should include the probability of selection for the second phase in the weights. I agree with her, but suspect others might not. As a design-based survey statistician, I would accept the inclusion of these selection probabilities in the weights as a default proposition and require strong evidence of their ineffectiveness before dropping them. This is related to a comment by Dr. Kott on accepting some increase in variance to reduce the potential for bias.

Dr. Kott notes that sometimes the same variables that affect response are related to the outcome variables and could be modeled for this purpose. I agree, and the paper by Micklewright et al. (2012) is an excellent example of this. His quibble about my use of the word 'any' is an important one. In the example I was trying to point out that if we had this information for any important outcome variables we should include it in the estimator regardless of whether it is related to response. Dr. Kott correctly points out that we would need this information for 'every' important outcome to support the robustness goal when modeling outcomes and using a design-based adjustment framework.

As mentioned above, I also agree with Dr. Kott on the importance of bias in the large sample sizes that are common in national statistical office surveys – and share the position that we should take actions to reduce the potential for bias even if it incurs some additional variance. Although Dr. Kott felt the text overemphasized variability of the weights, I did not intend that. I agree that there may be too much emphasis on this point in general, since with reasonable precautions adjustments for nonresponse rarely substantially increase the variance of the estimates.

Some of the differences Dr. Kott mentions may be a manifestation related to what Särndal (2007) referred to as a difference between calibration and GREG "thinking." Dr. Kott prefers a model justification for the linear calibration estimator, while I think in terms of restoring balance in the calibration variables. He prefers a logistic response model (Kott and Liao 2012), while I often choose raking. In practice, the differences are often not very substantial.

I am skeptical of the use of survey variables in calibration advocated in Chang and Kott (2008) and Kott and Chang (2010). Specifically, I worry that the procedure might induce substantial bias if the statistician makes a poor choice of survey variables for calibration.

I have not investigated this myself but look forward to more research to clarify the robustness of the procedure.

Professor Little states nonresponse requires modeling assumptions and I fully agree. The models I discussed differ from those he advocates – those I describe primarily model the response mechanism while he prefers modeling the relationship between response and the outcome – but modeling is essential. Furthermore, I agree that modeling the relationship between response and outcomes can be useful, especially if the result can be implemented in such a way to produce a general purpose nonresponse adjustment. As I noted in the article, I prefer that "powerful auxiliaries for key outcomes should be included in the estimator when they are available, irrespective of their relationship to response." The rationale is that such modeling reduces variance. If this procedure is followed, then residual nonresponse adjustment must be primarily based on a response propensity model. If the modeling of the outcomes is not done in advance, then modeling outcomes and response propensities is valuable.

Micklewright et al. (2012) is an example of where modeling the outcome led to adjustment related to response propensities. The adjustment they applied reduced the variance of the specific outcomes modeled as well as reducing nonresponse bias in the outcome and other statistics. If the auxiliary data they used had not been related to the specific outcome but was still related to response propensities, it is the type of general purpose nonresponse adjustment I would propose even though it would not reduce the bias for the specific outcome. Unfortunately, there may be no auxiliary data available that are strongly correlated with response propensities, and in this case response propensity adjustments are ineffective.

Professor Little restates his opinion that design-based inference is flawed and needs to be replaced by model-based approaches (Little 2012). I, on the other hand, find the design-based approach and nonresponse adjusted weights to be a valuable tool. Lohr (2007) gives some properties of weights that are desirable. Of those she describes, the properties of robustness, internal consistency of the estimates, and objectivity are critical in my assessment. Model-based estimates, as currently proposed, do not fully satisfy all of these properties. Related issues were raised by Brion, Smith, and Beaumont in their discussion of Little (2012).

The design-based procedures I described are general purpose, simple to use, and accessible for a wide variety of users. This means users can access the data set and produce an estimate without modeling a specific estimate. They can obtain the same estimate as the data set producer. The estimated totals they produce for subsets (e.g., males and females) equal the total for all persons when summed. These properties may sound trivial, but they are important to users. Over the years, national statistical offices have translated these user requirements into quality measures (e.g., Statistics Canada 2009). The quality measures include timeliness, accessibility, interpretability, and coherence; these measures are not statistical in the sense of producing minimal mean square error estimates.

I agree with Professor Little that, with sufficient effort, a model-based estimate may give a more efficient statistical solution for a particular estimate than a general purpose, design-based weighting procedure. If an important decision depended on one or a small set of estimates from a survey, it might be prudent to examine alternatives to the general purpose approach to improve the accuracy of the estimates. My inclination would be to seek more efficient design-based alternatives for the specific estimates, but model-based alternatives are another reasonable approach. However, in my opinion the existing model-based methods do not sufficiently address user-oriented quality measures such as timeliness, accessibility, interpretability, and coherence for the vast majority of applications. I doubt model-based methods will be adopted in practice unless they do.

I also have a different perspective on whether "the field is too focused on reasons for nonresponse." Knowing the reasons for nonresponse is essential to design efforts to reduce nonresponse. Similarly, modeling nonresponse appropriately requires understanding the reasons for nonresponse. For example, Lin and Schaeffer (1995) provide compelling evidence that outcomes can be very dependent on the reason for nonresponse.

On the preferences for weighting and imputation, I consider both as methods of implementing an estimation scheme. In some cases, the two are equivalent; for example, hot-deck imputation can be rewritten in terms of item-specific weights for a particular estimate. The choice of whether to use weights or impute is based largely on usability considerations. Imputation is preferable when sufficient data, such as responses to other items by the same respondent, are available. Weighting is preferable when characteristics at the sampled unit level are limited. However, both weighting and imputation are just different tools for accomplishing the same goal.

Finally, I agree with Professor Little that multiple imputation can be valuable, even though it is not the best solution to all nonresponse adjustment problems. Multiple imputation is a form of replication and I, like many design-based statisticians, am fond of replication.

References

Kulka, R., McNeill, J., and Bonito, A. (1982). On the Manifest Designation of Key Items: a Cost Effective Procedure for Improving the Collection and Processing of Survey Data. Paper Presented at the 37th Annual Conference of the American Association for Public Opinion Research. Maryland, USA: Hunt Valley.

Lohr, S. (2007). Comment: Struggles with Survey Weighting and Regression Modeling. Statistical Science, 22, 175–178.

Lynn, P. (2003). PEDAKSI: Methodology for Collecting Data About Survey Non-Respondents. Quality & Quantity, 37, 239–261.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Available at: http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf (Accessed March 2013).