Discussion

*Phillip S. Kott*¹

Let me start by thanking Mike Brick for his informative guide through the literature, both old and new, on unit nonresponse in household surveys, its impact on the quality of survey estimates, and some of the methods developed to reduce the negative consequences of that impact. Although I have a few quibbles, to which I will get shortly, I nonetheless found the article a treasure trove of useful ideas and references. Moreover, I agree wholeheartedly with many of its conclusions.

Now to those quibbles. We read in Section 3 that “modeling either the response propensity or the outcome variable can be effective for response bias.” Nevertheless, “in our experience most cross-sectional household surveys produce multiple characteristics and there are few auxiliary variables that are related to *any* of these outcomes [*italics added*]. In that situation, response propensity modeling may be the only remaining tool to reduce nonresponse bias.”

Had “any” been replaced by “every” I would be inclined to agree. We only need one model to explain unit response, but we need a separate model for each outcome variable. The single response model either (nearly) holds or does not, while the outcome models can hold for some variables and fail miserably for others.

The author, however, claims that we often can model unit response but do not have the auxiliary variables to model any outcome variable. In my experience, the variables we use to model response can usually be employed to model survey outcome. When we form response homogeneity groups (RHGs) or poststrata, we very often are also creating outcome model groups, the units within each group having a common mean for many survey variables of interest. Ironically, that is why we use the design weights when computing the implicit probability of response within each RHG/poststrata: it reduces or removes the bias of the resulting estimates under the outcome model and thereby reduces the overall mean squared error (see [Kott 2012](#)).

If I may belabor the point a bit, consider a variation of the example in the beginning of Section 7. A simple random sample is divided into initial responders and nonresponders. The latter group is subsampled and extra effort is given to elicit responses from the subsample. Suppose, unlike in the text, the extra effort had been successful. If the responding units from two groups (initial responders and nonresponders) have distinct outcome means, then we need to have separate weights for each. Otherwise, we do not.

In practice, a survey usually has many variables of interest. It is possible that some have a common mean across the two groups while others have distinct means. We do not know.

With full response, we do not have to assume a model of outcome behavior; we can use probability-sampling principles to produce unbiased estimates in some sense by weighting

¹ RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. Email: pkott@rti.org

the groups separately. We often accept additional variance to protect ourselves from potential bias with survey samples because the former can be measured while measuring the latter is illusive. Moreover, sample sizes tend to be large, so bias often dominates variance within overall mean squared error.

Now, suppose there remains some unit nonresponse in the initial-nonresponder group even after expending the added effort. It seems reasonable to treat the units from that group as if they had a common mean for a survey variable, a mean that may be distinct from that of the initial responders. Again, we are trying to protect ourselves from potential bias. We are not, of course, fully protected because we have assumed an outcome model that may be flawed. In particular, we have assumed that the survey variables of respondents and nonrespondents among the initial nonresponders have a common mean. In the very special case of this example, the outcome model is equivalent to the response model, namely that each initial nonresponder is equally likely to respond after the extra effort. This equivalence is not usually the case. Still, I argue that if there are variables on which to build a response model, there are variables on which to build outcome models. Moreover, it is precisely when that is the case that there is nonresponse bias to reduce (see Little and Vartivarian 2005).

Little and Vartivarian also point out that an increase in the variability of the weights does not always lead to higher variances. For my taste, there is an over-emphasis in the text – and in general practice – on weight variability. At the very least, a prudent statistician should conduct sensitivity analyses to assess differences in the resulting estimates for key survey variables caused by using alternative nonresponse models, one simpler and one more complex. Moreover, under the assumption that an estimator using the more complex model is bias free, one can test whether the simpler model leads to a systematically biased estimate for a particular survey variable using a procedure proposed by Fuller (1984) for determining whether survey weights matter in a linear regression. An analogous procedure using replication was suggested by Korn and Graubard (1993).

In the Fuller procedure (the replication version is trivial), each observation is duplicated with one version assigned to Domain A and weighted one way while the other is assigned to Domain B and weighted the alternative way. Recognizing that both versions are from the same primary sampling unit, one can then use design-based software to measure whether the difference in the domain means for the variable are larger than we would expect due to random chance alone. Since we are as much concerned with Type 2 error (ignoring a real bias) as Type 1 (finding a false one), I would argue against accepting the null hypothesis (that the observed difference in the estimates derives from random noise) when the absolute t value of the difference is much greater than 1.

I wish the text had provided a deeper discussion on using calibration weighting to adjust for unit nonresponse. As it correctly points out, calibration can be viewed as a generalization of reweighting using RHGs (also called “weighting classes”) or poststratification depending on whether one is calibrating to the original weighted sample (RHGs) or the population.

With linear calibration weighting, the adjusted weight for respondent i has the form:

$$d_i^* = d_i \left\{ 1 + \left[\mathbf{X} - \sum_{k \in s_r} d_k \mathbf{x}_k \right]^T \left(\sum_{k \in s_r} d_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \mathbf{x}_i \right\} = d_i (1 + \mathbf{g}^T \mathbf{x}_i) = d_i v_i \quad (1)$$

(by the way, I would not call ν_i the “linear regression estimate” as is done in the text). There is no reason why some components of $\mathbf{x}_i$ cannot be related to the efforts employed in eliciting response as advocated by the author. Moreover, as Mike correctly points out, the totals in $\mathbf{X}$ can contain both population totals and weighted sums from the original sample.

Using linear calibration weights, the estimator in Equation (10) of the text ($\hat{y}_{ca} = \sum_{i \in s_r} d_i^* y_i$) is nearly unbiased under the response model:

$$\Pr(R_i = 1 | \mathbf{x}_i) = \frac{1}{\boldsymbol{\gamma}^T \mathbf{x}_i}, \quad (2)$$

where the unknown parameter vector $\boldsymbol{\gamma}$ is estimated by $\mathbf{g}$ in Equation (1). Equation (2) is not a very plausible response model except when the components of $\mathbf{x}_i$ are group-membership indicators, the special case of RHGs/poststrata.

To my way of thinking, a better justification for linear calibration weighting in general is that the survey variable behaves roughly like a random variable with mean $\mathbf{x}_i^T \boldsymbol{\beta}$. In household surveys, many survey variables are binary and cannot be modeled precisely as a linear function of the components of $\mathbf{x}_i$. That is why I added the modifier “roughly.” In some surveys, there may be variables for which the linear outcome model ($E(y_i) = \mathbf{x}_i^T \boldsymbol{\beta}$) does not come close to holding. I suspect that the linear calibration estimates for the totals of these variables will often not be close to unbiased either.

Even though linear calibration can sometimes be effective in reducing or removing nonresponse bias when the response model in Equation (1) is clearly wrong (see [D’Arrigo and Skinner 2010](#)), I prefer using a back-link function that implicitly assumes a more plausible response model. One such is $d_i^* = d_i [1 + \exp(\mathbf{g}^T \mathbf{x}_i)]$, which assumes $\Pr(R_i = 1 | \mathbf{x}_i) = [1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_i)]^{-1}$, a logistic response model with $\mathbf{g}$ in the back-link function being a consistent estimator for $\boldsymbol{\gamma}$ when the assumed response model holds. Calibration weighting using this back-link function provides double protection against nonresponse bias since the estimator is nearly unbiased in some sense when either the response model or the linear outcome model holds (or when both hold; see [Kott and Liao 2012](#)).

Let me also point out that the vector, let us now call it $\mathbf{q}_i$, of variables in the response model need not coincide with the calibration variables in $\mathbf{x}_i$. As a result, although the calibration-variable totals ($\mathbf{X}$ in $\sum_{i \in s_r} d_i^* \mathbf{x}_i = \mathbf{X}$) must be known for either the original weighted sample or the population, some of the component of $\mathbf{q}_i$ need not be.

The response model is now $\Pr(R_i = 1 | \mathbf{q}_i) = f(\mathbf{g}^T \mathbf{q}_i)$ for some back-link function $f(\cdot)$. Since nothing prevents components of $\mathbf{q}_i$ from being survey variables, this version of calibration weighting can be used to treat nonresponse that is not missing at random. To my knowledge, [Deville \(2000\)](#) was the first to point this out. In Deville’s formulation, the number of components of $\mathbf{q}_i$ and $\mathbf{x}_i$ must coincide. [Chang and Kott \(2008\)](#) extend calibration weighting for nonresponse adjustment to allow more calibration variables than response-model variables. Many of the calibration weighting ideas in that article have been incorporated into SUDAAN 11 ([RTI International 2012](#)).

I have concentrated here on relatively small points related to areas of my research in an article that has a much broader sweep. Mike Brick is to be congratulated for producing a fine contribution to the literature.

References

- Chang, T. and Kott, P.S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. *Biometrika*, 95, 557–571.
- D’Arrigo, J. and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. *Survey Methodology*, 36, 181–192.
- Deville, J.-C. (2000). Generalized Calibration and Application to Weighting for Non-Response. *Compstat: Proceedings in Computational Statistics; 14th Symposium Held in Utrecht, The Netherlands*, J.G. Bethlehem and P.G.M. Van Der Heijden (eds). Physica Verlag: Heidelberg, 65–76.
- Fuller, W.A. (1984). Least Squares and Related Analyses for Complex Survey Designs. *Survey Methodology*, 10, 97–118.
- Korn, E.L. and Graubard, B.I. (1993). Response to “Bias in Weighted Versus Unweighted Estimates.”. *American Journal of Public Health*, 83, 1351.
- Kott, P.S. (2012). Why One Should Incorporate the Design Weights When Adjusting for Nonresponse Using Response Homogeneity Groups. *Survey Methodology*, 95–99.
- Kott, P.S. and Liao, D. (2012). Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-weighting Routine. *Survey Research Methods*, 6, 105–111.
- RTI International (2012). SUDAAN Language Manual, Release 11.0. Research Triangle Park. NC: RTI International.