Journal of Official Statistics, Vol. 29, No. 3, 2013, pp. 329–353, DOI: 10.2478/jos-2013-0026

Unit Nonresponse and Weighting Adjustments: A Critical Review

J. Michael Brick[1]

This article reviews unit nonresponse in cross-sectional household surveys, the consequences of the nonresponse on the bias of the estimates, and methods of adjusting for it. We describe the development of models for nonresponse bias and their utility, with particular emphasis on the role of response propensity modeling and its assumptions. The article explores the close connection between data collection protocols, estimation strategies, and the resulting nonresponse bias in the estimates. We conclude with some comments on the current state of the art and the need for future developments that expand our understanding of the response phenomenon.

Key words: Response propensity; bias; data collection; calibration.

1. Introduction

This article critically reviews aspects of unit nonresponse in sample surveys, where unit nonresponse is defined as the failure to obtain a valid response from a sampled unit. We emphasize the consequences of unit nonresponse and methods of adjusting for it in circumstances that are typical of cross-sectional household surveys. Establishment surveys and attrition nonresponse in panel surveys are also subject to unit nonresponse, and issues reviewed here pertain to these surveys. However, the data collection design options, reasons for nonresponse, and auxiliary data available for adjustment differ dramatically across types of surveys. Because these features are critical to dealing with nonresponse and nonresponse bias, we have chosen to focus on situations frequently arising in cross-sectional household surveys.

Unit nonresponse is just one form of missing data in surveys. Other types of missing data include incomplete coverage of the target population, item nonresponse, and partial nonresponse such as wave nonresponse in panel surveys and failure to obtain second-phase responses in two-phase surveys. While these are all important, they are beyond the scope of this review.

Most surveys, especially government surveys, employ large sample sizes and design-based theory to make inferences from the sample to the target population. This theory assumes complete response. While surveys employ methods to minimize nonresponse and its effects on estimates, in almost every survey some sampled units do not respond.

[1] Westat, 1600 Research Blvd, Rockville, MD 20850, U.S.A. Email: MikeBrick@westat.com
Acknowledgments: We would like to thank Graham Kalton, Sharon Lohr, and Douglas Williams, who provided helpful comments on earlier drafts.

Model assumptions and adjustments are made in an attempt to compensate for missing data. Because the mechanisms that cause unit nonresponse are almost never adequately reflected in the model assumptions, survey estimates may be biased even after the model-based adjustments. Nonresponse also causes a loss in the precision of survey estimates, primarily due to reduced sample size and secondarily as the result of increased variation of the survey weights. However, bias is the dominant component of the nonresponse-related error in the estimates, and nonresponse bias generally does not decrease as the sample size increases. Thus, bias is often the largest component of mean square error of the estimates even for subdomains when the sample size is large.

The classification of nonresponse by reason is important because the effects and methods of dealing with nonresponse may be directly tied to the reason (Lin and Schaeffer 1995; Steele and Durrant 2011). Reasons for unit nonresponse are usually classified as the failure to contact the sampled unit, the inability to persuade the sampled unit to respond, and other reasons (Brick and Montaquila 2009). Noncontact or inaccessibility nonresponse may occur for a variety of reasons. For example, the sampled unit may not be at home during the times the data collector visits or calls, the survey schedule may limit the number of contact attempts, or data to locate the sampled unit may be incorrect or out of date. Refusal nonresponse may occur because the sampled person does not wish to participate in the particular survey, or because someone else such as a gatekeeper refuses to provide access to the sampled person. For example, in a telephone survey the person answering the telephone may not be willing to give the telephone to the sampled person. While noncontact was a larger component of total nonresponse in earlier times, refusals now constitute the majority of total nonresponse in most surveys (Atrostic et al. 2001; Brick and Williams 2013). The other nonresponse category includes assorted reasons such as language problems and health problems that may prevent the sampled unit from responding. These other problems are typically a small proportion of the total nonresponse in a survey, but may be important in some cases (see Feskins et al. 2011; Brick et al. 2012).

2. Background

Unit nonresponse has been recognized as a potential problem since the early days of probability sampling. Colley (1945), Hansen and Hurwitz (1946), Ferber (1949), Yates (1946), and Deming (1953) are examples of early research that examined data collection and weighting methods to deal with nonresponse[2]. As research on nonresponse and its effects accumulated and worries about increasing nonresponse rates were expressed, the Committee on National Statistics in the United States convened a Panel on Incomplete Data in 1977 to consolidate this research and develop new approaches. The Panel's work resulted in a three-volume set in 1983 (Vol. 1 edited by Madow, Nisselson, and Olkin; Vol. 2 edited by Madow, Olkin, and Rubin; Vol. 3 edited by Madow and Olkin) that was the first monograph dedicated to nonresponse in surveys. Around the time of the Panel, the way nonresponse was conceived and adjustments were motivated began to shift to treat

[2] The references in this section are examples and useful summaries of a body of work and are not intended to assign precedence for ideas.

response as a random rather than fixed outcome. In our review, several references to chapters from one of the three volumes reflect some of these changes.

In the years following the Panel's meetings, several published books were devoted largely to survey nonresponse. These include Kalton (1983), Goyder (1987), Brehm (1993), Groves and Couper (1998), Tourangeau et al. (2000), Groves et al. (2002), Särndal and Lundström (2005), Stoop (2005), Stoop et al. (2010), and Bethlehem et al. (2011). Journals have dedicated special issues to survey nonresponse, including the *Journal of Official Statistics* and *Public Opinion Quarterly*. International workshops and symposiums have been also been held; the Groves et al. (2002) monograph is a product of one of these.

To provide some context for this research, we identify three major themes in nonresponse research (although there is considerable overlap among them). One theme is the study of the response mechanism that causes nonresponse. This research seeks to understand important psychological and sociological factors that dispose some units to respond and others to fail to respond. Goyder (1987) is an example of this work that takes a sociological perspective on the causes of nonresponse; Tourangeau et al. (2000) is an example taking the psychological view. Most of the psychological and sociological research examines the willingness or amenability of the sampled unit to participate in the survey by looking at factors such as the interviewer, the survey materials, and the characteristics of the respondent that might influence response.

A second theme is data collection methods to reduce nonresponse. Dillman's (1978) tailored design method illustrates one branch within this theme. He offers general approaches to the design of data collections to increase cooperation rates and improve the chances of reaching respondents to deliver the survey request. The literature on incentives is another such example (Singer 2002). The other branch within this theme describes a set of methods for following up nonrespondents; survey methods to gain the cooperation of those who refuse the initial survey request or who are never contacted are important topics in this area. Switching modes for nonresponse follow up is an example within this area (Dillman et al. 2009).

Statistical adjustment of the survey weights to adjust for survey nonresponse is a third theme while retaining the design-based mode of inference. Särndal and Lundström (2005) is an example. They examine statistical models to adjust the estimates from the survey after the nonresponse has been realized. The aim of all of this research is to reduce the level of nonresponse and develop methods to minimize nonresponse bias in the estimates.

For many years, nonresponse bias and response rates were often treated as equivalent, or at least surveys with low response rates were thought likely to have the potential for high nonresponse bias in the estimates. Data collection efforts that increased response rates were assumed to reduce nonresponse bias. This presumed relationship is especially pronounced in the literature on incentives, where effects of incentives on response rates are carefully described and nonresponse bias is often not assessed directly (Singer and Ye 2013). The reasons for this assumption are easy to understand. Response rates are easy to compute, provide a single measure for an entire survey, and have face validity.

A spate of articles in the last decade forced researchers to reconsider this presumption. These articles show that the empirical relationship between response rates and nonresponse bias is not strong (e.g., Keeter et al. 2000; Curtin et al. 2000;

Groves 2006). Of course, even long ago we knew that a single measure like a response rate could not be used to predict nonresponse bias. Ferber (1949, p. 672) noted “The problem of response bias must be considered with specific reference to a particular question or characteristic. The presence of bias in one question does not mean *a priori* that the replies to other questions on the same questionnaire are also biased.”

Falling response rates in most countries across the developed world, especially in the past few decades, are documented in various reviews (e.g., Stoop 2005; Steeh et al. 2001; Atrostic et al. 2001; de Leeuw and de Heer 2002; Smith 1995; and Synodinos and Yamada 2000). Furthermore, the trend toward lower response rates is happening despite additional procedures aimed at increasing response in many surveys. Some of these procedures are designed to increase contact rates and others are aimed at reducing refusals. However, none of these methods appear to be capable of reducing the level of nonresponse, and reliance on adjustments to the survey weights is increasing.

Although response rates may not be predictive of nonresponse bias, the declines in response rates have raised the level of concern among survey methodologists and prompted new developments. Some debate whether low response rate probability samples are qualitatively different from nonprobability samples; others have sought to find different measures that are more predictive of nonresponse bias. Schouten et al. (2009) propose R-indicators to serve as a substitute for response rates. These indicators attempt to measure how similar the respondents are relative to the full sample by estimating the variability in the estimated response propensities, where the response propensity (ϕ_i) for every sampled unit i is its probability of responding to the survey. Schouten et al. (2009) define the R-indicator as

$$R(\phi(\mathbf{x})) = 1 - 2S(\phi(\mathbf{x})), \tag{1}$$

where $S(\phi(\mathbf{x}))$ is the population standard deviation of the response propensities and $\mathbf{X}$ is a vector of auxiliary variables known for the full sample. If the R-indicator is close to unity, the respondent set is more ‘representative’ of the target population, at least as measured with respect to $\mathbf{X}$, and has a lower potential for nonresponse bias. Schouten et al. (2011b) extend these results.

Särndal and Lundström (2005) and Särndal (2011a) propose what they refer to as balance indicators that are intended to measure the similarity between the respondents and the sample. Some of these indicators are like the R-indicators in that they measure variation in subgroup response rates, where the subgroups are formed based on auxiliary variables. Wagner (2010) proposes using the fraction of missing information as an alternative to the response rate because this measure permits the inclusion of auxiliary variables in the determination of the influence of the missingness on the estimate.

All of the alternatives for response rates are only able to measure representativeness of the respondents in relation to $\mathbf{X}$, the auxiliary variables available. Different choices of $\mathbf{X}$ lead to different values of the indicators. Although using these data is an improvement over response rates that do not consider any auxiliary data, the measures are only useful when powerful auxiliary variables for the specific estimates are available.

Some of these measures were influenced by the desire to continually monitor the data collection process for responsive designs (Groves and Heeringa 2006). Responsive and adaptive designs are two data collection approaches that have been proposed as a way to reduce nonresponse bias. Responsive design makes changes to data collection strategies during data collection when one recruitment protocol is no longer successful in getting responses from sampled units, especially units with differing characteristics (Groves and Heeringa 2006). For responsive designs, data for making these decisions must be collected and analyzed rapidly during the field period. Adaptive design is similar, but the analysis of response patterns may be done from previous or similar collections (Schouten et al. 2011a). Both responsive and adaptive designs contemplate data collection strategies that are tailored for specific sampled units, whereas the standard data collection procedure for many years has been to apply a single protocol to all units.

3. Bias Representations

The rationale for the design, data collection, and estimation approaches mentioned above is based on models of nonresponse bias. Two models dominate the way we think about nonresponse bias. The models are most often presented in terms of the bias of an unadjusted estimator of the mean, where unadjusted implies using the full sample estimator with just the respondent data. The unadjusted Horvitz-Thompson estimator of the total is

$$\hat{y}_{un} = \sum_{i \in s_r} d_i y_i, \tag{2}$$

where d_i is the inverse of the probability of selection of unit i and the sum is over s_r, the set of respondents. The ratio mean is $\hat{\bar{y}}_{un} = \hat{y}_{un} \Big/ \sum_{i \in s_r} d_i$.

The deterministic representation of bias partitions the population into respondent and nonrespondent strata (Cochran 1977), and nonresponse bias is then a function of the nonresponse rates and the characteristics of the units in these strata. In the deterministic approach, response is a fixed outcome of the survey (and the procedures used in data collection) and is not subject to random variation other than the variation due to sampling the response strata. The nonresponse bias of the unadjusted estimator of the mean is

$$bias\left(\hat{\bar{y}}_{un}\right) \approx (1 - P)\left(\overline{Y}_r - \overline{Y}_m\right), \tag{3}$$

where P is the proportion of units in the respondent stratum, $\overline{Y}_r$ is the mean in the respondent stratum, and $\overline{Y}_m$ is the mean in the nonrespondent stratum (Thomsen 1973). The expression shows that bias depends on the response rate and the distribution of each characteristic as discussed by Ferber (1949). However, a difficulty with Expression (3) is that the response strata definition is *post hoc* so it is difficult to use this in advance of data collection.

The alternative stochastic model has become more popular since the late 1970s, although its origins go back as early as Politz and Simmons (1949) and Hartley (1946). It assumes that response is a random variable and the probability of response is like the probability in an additional phase of sampling, but the probability of response for every unit i in this phase is unknown.

The nonresponse bias of an estimated ratio mean under the stochastic model is

$$bias\left(\hat{\bar{y}}_{un}\right) \approx \bar{\phi}^{-1} \sigma_\phi \sigma_y \rho_{\phi,y}, \tag{4}$$

where $\bar{\phi}$ is the population mean of the response propensities, σ_ϕ is the standard deviation of ϕ, σ_y is the standard deviation of y, $\rho_{\phi,y}$ is the correlation between ϕ and y, and $\phi_i > 0$ for all i (Bethlehem 1988). The estimated respondent mean is unbiased if ϕ and y are uncorrelated.

The two expressions are appropriate for the Horvitz-Thompson of the unadjusted mean, but different relationships hold for totals, correlations, and other statistics as well as for different estimators. Brick and Jones (2008) extend these results to other types of statistics and some calibrated estimators.

Both models are useful for estimating the potential bias under particular circumstances. For example, if data are available for all units in the population, then the bias can easily be computed using (3) or (4) after data collection is complete. Both bias expressions are equivalent in this case. The two models also lead to similar conclusions about how to attempt to adjust for biases due to nonresponse. We find the stochastic model to be generally more helpful when speculating about the potential magnitude of bias prior to data collection. It expresses bias in terms of a correlation so it is bounded, and correlations computed from other surveys may be useful guides for speculating about the magnitude of correlation.

Thus far, we have discussed bias in the simple situation in which no other information is known about the sampled units. In practice, we often have other data available for either the sampled units or the entire population. Thus, the expressions given above can be revised slightly to account for the auxiliary information. For example, the response propensity can be written more formally as

$$\phi_i = \phi(\mathbf{x}_i) = \Pr(R_i = 1 | \mathbf{X} = \mathbf{x}_i), \tag{5}$$

where $\mathbf{X}$ consists of the set of variables known for the full sample and $R_i = 1$ if unit i responds (Rosenbaum and Rubin 1983). The bias expressions for both the deterministic and stochastic models can also be modified to account for auxiliary data. For example, suppose auxiliary data are available and used for poststratification. The stochastic expression for the bias of the poststratified estimator of the mean is

$$bias\left(\hat{\bar{y}}_{ps}\right) \approx N^{-1} \sum_h \bar{\phi}_h{}^{-1} \sigma_{\phi_h} \sigma_{Y_h} \rho_{\phi_h, Y_h}, \tag{6}$$

where h denotes the poststratification classes. See Kalton (1983), Brick and Kalton (1996), and Bethlehem et al. (2011) for such expressions and their implications.

The auxiliary variables are very valuable for adjusting the design weights to account for nonresponse. Kalton (1983, p. 63) states: "Among the potential variables for use in forming weighting classes, the ones that are most effective in reducing nonresponse bias are those that are highly correlated both with the survey variables and the (0,1) response variable." Both (3) and (4) explicitly contain the characteristic being estimated, suggesting that adjustments could be developed by modeling the distribution of the characteristic.

Two types of auxiliary variables can be used: if the auxiliary variables are known for all sampled units, then the adjustment is called sample-based or Info-S; if they are known for the entire population, the adjustment is population-based or Info-U (Kalton and Kasprzyk 1986; Lundström and Särndal 1999). The population-based adjustment is especially useful when characteristics for the entire sample are not available but the population totals are known, because these adjustments only require capturing the data from the respondents. Population-based adjustments may also reduce noncoverage error and sampling error. Sample-based adjustments need data for the full sample but do not require knowing control totals for the entire population. Sample-based and population-based adjustments are equally effective for dealing with nonresponse bias (Särndal and Lundström 2005; Brick and Jones 2008).

4. Modeling and Missing Data Mechanisms

As noted above, modeling either the response propensity or the outcome variable can be effective for reducing nonresponse bias. Nevertheless, this section discusses only response propensity modeling, for two reasons. First, modeling outcomes and using design-based calibration estimators like the generalized regression estimator can be extremely valuable for improving the precision of the estimates even when there is full response. Ratio and regression estimators were originally developed exactly for these reasons. These estimators are also beneficial at reducing nonresponse bias when the same variables are correlated to response (e.g., Bethlehem 1988; Fuller et al. 1994). Our perspective is that powerful auxiliaries for key outcomes should be included in the estimator when they are available, irrespective of their relationship to response.

Second, 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. In this situation, response propensity modeling may be the only remaining tool to reduce nonresponse bias. It has the potential to reduce bias for variables that cannot be modeled directly because powerful correlates of the variable are not available. Of course, this approach is not a panacea by any means. Often, bias is reduced by response propensity weight adjustments, but only partially, as shown by Micklewright et al. (2012).

We also concentrate on nonresponse where the data are missing at random (MAR). In our notation, the missing data mechanism is MAR (see Rubin1976; Little and Rubin 2002) when

$$\Pr(R_i = 1|Y_i, \mathbf{X}_i) = \Pr(R_i = 1|\mathbf{X}_i) \tag{7}$$

for all sampled units. Roughly speaking, under the MAR assumption the missing data mechanism may depend on observed data but not on unobserved data. When (7) does not hold, the missing data mechanism is called not missing at random (NMAR). Although this dichotomy is useful, in practice it is not possible to assess whether the data mechanism is MAR or NMAR without obtaining additional data for the nonrespondents.

Two approaches have been proposed for handling nonresponse when researchers assume the mechanism is NMAR. The first is called the selection model approach; it postulates a model that relates the missing data to the distribution of the outcome. Heckman (1979) is probably the best-known example of an explicit selection model. Greenlees et al. (1982) also use this approach. A second approach is the pattern mixture model (PMM), where the distribution of Y is conditioned on the missing data and mixed or averaged over different populations (Little 1993). Andridge and Little (2011) have recently expanded on the PPM approach using a proxy variable. Nearly all researchers using NMAR models strongly urge sensitivity analyses to determine whether the estimates are robust to the modeling assumptions, since generally there is no other way to assess these assumptions.

Molenberghs et al. (2008) show that for every NMAR model there is a MAR counterpart that has an equal fit to the observed data. This means that the NMAR model cannot be distinguished from its MAR counterpart based on the observed data. Even though they have equal fits, the models do not necessarily produce the same estimates. In a similar vein, David et al. (1986) re-examine the NMAR approach of Greenlees et al. (1982) using a MAR model and find that the MAR model is adequate. Molenberghs et al. (2008) show an example where the estimates from the NMAR models and their MAR counterparts are very different. They use a series of MAR counterparts corresponding to NMAR models for sensitivity analysis. Since MAR models are usually easier to understand and describe, in the following sections we generally restrict our attention to MAR models. We will return to this concept later.

5. Response Propensity Weight Adjustment

One approach to weight adjustment is to model the response propensities for the sampled units individually, and the adjustment factor is the inverse of the estimated propensities of the respondents. The idea is to replace the unknown probability of response by an estimate. The propensity-adjusted estimator of the total is

$$\hat{y}_{rp} = \sum_{i \in s_r} d_i \hat{\phi}_i^{-1} y_i \tag{8}$$

where $\hat{\phi}_i$ is the estimated propensity for unit i where i is a respondent. The $\hat{\phi}_i$ are usually estimated by logistic regression, but probit and nonparametric methods are also used (Little 1986; Da Silva and Opsomer 2009; Phipps and Toth 2012).

As mentioned above, Politz and Simmons (1949) pioneered thinking about stochastic response models when they estimated propensities by collecting data on how often the respondent would be at home on different days. These data provide a basis for estimating contact propensities to account for noncontact nonresponse. Related methods such as those proposed by Bartholomew (1961) and Dunkelburg and Day (1973) have not generally proven to be effective, especially as contact rates have risen due to increased data collection efforts.

Rather than estimating individual response propensities, the approach most surveys use is to form groups and adjust the weights in each group by the inverse of the observed group response rate. Särndal et al. (1992) describe these as response homogeneity groups (RHGs). Weighting classes is an alternative term that has been used for decades. Important outcome statistics or domains may also be considered when forming RHGs. If all the units within an RHG have the same response propensity so that MAR holds, any nonresponse bias is eliminated (see Da Silva and Opsomer (2004) for extensions). In this case, (8) is a weighting class estimator and can be written as

$$\hat{y}_{wc} = \sum_{g} \sum_{i \in s_{r_g}} d_{gi} \hat{\phi}_g^{-1} y_{gi} \tag{9}$$

where $g = 1,2,\ldots,G$ are the RHGs, $i \in s_{r_g}$ is a respondent in RHG g, and $\hat{\phi}_g$ is the estimated response propensity in g. One issue that often arises with weighting class estimators is the need to have large enough respondent counts in each cell to avoid unstable estimates. For this reason, Little (1986) proposes using cells based on the estimated propensity scores rather than individually estimated propensities.

A third general approach is to use calibration estimation (Deville and Särndal 1992) for adjustment. Lundström and Särndal (1999) extend calibration estimators to encompass estimators to include both sample-based and population-based information for nonresponse adjustment. The calibration estimator is

$$\hat{y}_{ca} = \sum_{i\in s_r} d_i^* y_i, \tag{10}$$

where the sum is over the respondents, d_i^* is the adjusted weight that satisfies the calibration equation $\sum_{i\in s_r} d_i^* \mathbf{x_i} = \mathbf{X}$, $\mathbf{x_i}$ is a vector of auxiliary variables, and $\mathbf{X}$ is a vector of totals (sample based, population based, or a combination of the two) of those auxiliary variables. Since the weights are not uniquely defined by these conditions, other constraints may be imposed, such as $d_i^* = d_i v_i$, where v_i is a linear regression estimate (Bethlehem 2002; Särndal and Lundström 2005). A wide variety of nonresponse adjustment estimators are in this class, including poststratification, raking, and generalized linear regression estimators. Lumley et al. (2011) give insight into the relationship between calibration estimators and nonresponse bias for different estimators.

Poststratification is a simple calibration estimator that has a single dimension and has been used for decades (Holt and Smith 1979). Assume that poststrata are defined by the number of persons in age categories (N_h) and that N_h is known for the entire population. In this case, (10) simplifies to $\hat{y}_{ps} = \sum_h \frac{N_h}{\hat{N}_h} \sum_{i\in s_{r_h}} d_i y_i$, where $\hat{N}_h = \sum_{i\in s_{r_h}} d_i$ and the sum is over the respondents in poststratum h. The calibration equation forces the estimator for the age groups to match the known population total for that group.

It is easy to see that the weighting class estimator given by (9) is a sample-based calibration estimator – the calibration equation in this case forces the adjusted weight to reproduce the weighted (using d_i) distribution of the weighting classes from the sample. A related estimator that only uses $\mathbf{x_i} = 1$ for all $i \in s$ is called the primitive estimator by Särndal (2011b) and is given by

$$\hat{y}_{pr} = \left(\sum_{i\in s} d_i\right) \left(\sum_{i\in s_r} d_i\right)^{-1} \left(\sum_{i\in s_r} d_i y_i\right). \tag{11}$$

Estimators of this nature have a substantial effect on the bias of the estimated total but have no effect on the ratio mean.

Details on specific nonresponse adjustment techniques are covered in several articles and texts, including Särndal and Lundström (2005), Kalton and Flores-Cervantes (2003), Chang and Kott (2008), Brick and Montaquila (2009), and Bethlehem et al. (2011). Generally, the specific form of the adjustment is not highly related to the bias reduction, except when the form limits the ability to take advantage of all the information in the auxiliary data. For example, poststratification may be less effective than linear calibration or raking when many variables are available because poststratification has one dimension.

In addition, any method that results in large variability in the nonresponse adjustments due to instability in the estimated adjustments should be avoided since that may increase the variance of the estimates without further reducing bias.

The basic theory underlying the adjustment methods described above is formalized by Cassel et al. (1983), who treat response as an additional phase of "sampling" (see also Oh and Scheuren 1983). According to this theory, the adjusted estimator should have desirable statistical properties such as unbiasedness and consistency when expectations are taken over both sampling and response mechanisms, provided that the response propensities can be adequately estimated. Suppose that RHGs are formed and the adjustment to the sampling weight is the inverse of the response rate in the RHG, $\hat{\phi}_g^{-1}$. The heuristic interpretation is that each respondent in an RHG g "represents" $\left(\hat{\phi}_g^{-1}-1\right)$ nonresponding units in the group. Within this framework, the goal is to identify groups of units with the same probability of responding to the survey at the end of data collection, so that the MAR assumption is satisfied. The methods employed to create the RHGs and the choice of variables for creating these groups are an essential feature of nonresponse weighting.

6. Choosing Auxiliaries and Alternative Metrics

Traditionally, auxiliary variables and weighting classes were developed based on the availability of variables and the judgment of the statisticians (Madow, Nisselson, and Olkin Vol. 1, ch. 4, 1983). Predictors of response, key outcome statistics, and domains are considered in this process. Demographic variables such as age, sex, race, and geography were, and still are, frequently chosen even though they may not be effective in reducing bias (Peytcheva and Groves 2009). Many of these are population-based adjustments using data from a recent census for the controls. When the number of respondents in a cell of the cross-classification of the variables is below a threshold set for the survey, then cells are collapsed to avoid large adjustment factors.

When many variables are available, other methods of choosing which variables to include are needed. Search algorithms and regression models are sometimes used in this setting (Brick and Kalton 1996). These methods divide the sample into cells that discriminate between response and nonresponse or variables correlated with key outcome variables. The main advantage of these methods, especially the search algorithms, is the ability to identify interactions among the variables that may be important for nonresponse reduction. Regression models can also be used to examine interactions, although practitioners often rely on main effect models. Brick and Jones (2008) show the importance of interactions in some situations.

New methods for choosing auxiliary variables to reduce nonresponse bias in the estimates have been recently developed. Schouten (2007) and Särndal and his colleagues (Särndal and Lundström 2005, 2008, 2010; Särndal 2011a) suggest two approaches. These approaches do not assume that the data are missing at random, but to be effective they do require powerful predictors of the response mechanism. The methods are also described in terms of searching for main effects and including or excluding variables. Extensions are needed to deal with interactions among the variables.

Schouten et al. (Schouten 2007; Schouten et al. 2009) use indicators for choosing variables for weighting that are related to R-indicators. Schouten (2007) gives a forward-backward selection strategy for choosing variables, similar to stepwise regression. He starts with the variable that minimizes an estimate of maximal bias (which is linked to the R-indicator) and iteratively adds and removes other variables. The maximal bias is computed based on a generalized regression estimator.

Särndal and Lundström (2008) approach the choice of auxiliary variables by focusing on the estimation phase, although they are explicit about the importance of the design and data collection stages also (see Särndal and Lundström 2010; Särndal 2011a). Särndal and Lundström (2010) propose survey-specific indicators that account for the sample design, the set of observed respondents, and the specific calibration estimator. Their indicators choose auxiliaries based on the distance between the calibrated estimator and the primitive calibration estimator ($\hat{y}_{ca}$ and $\hat{y}_{pr}$) and may be outcome specific or generic. These authors describe an "all vectors procedure" that chooses the auxiliaries that are in the list of vectors that has the highest indicator. They also offer a "stepwise" procedure that builds the vector one variable at a time.

Särndal and Lundström (2010) compare the two approaches and find that they do not always include the same set of auxiliary variables in the estimator. They attribute some of the difference to the different perspectives, especially the fact that Schouten's (2007) approach uses population-level measures while theirs are sample-level. When choosing among many possible auxiliary variables to include or exclude in the estimation phase, the indicators of Särndal have the advantage of assessing improvements in estimators for the specific sample.

In some countries, especially in northern Europe, population registers may provide the types of data needed for using these methods. However, in household surveys in countries like the United States and Canada, these methods are less pertinent because there are few powerful auxiliary variables. When the information available for the sample does not predict response well, researchers have resorted to creating paradata from the survey itself (Beaumont 2005; Bates et al. 2008). The use of paradata is a rapidly developing area, but initial findings reveal that this may be a difficult task (Kreuter et al. 2010).

7. Response Propensity Models in Surveys

Because response propensity scores play such a large role in nonresponse adjustment methods, we describe the underlying theory and assumptions here. We begin with a few observations. First, response propensities are unknown, unlike probabilities associated with an additional phase of sampling. In fact, they are latent variables and cannot be observed directly – we observe only the binary outcome of response or nonresponse. Second, we assume that $\phi_i > 0$ for all i. Deming (1953) explicitly considers units with zero response propensities. He calls those that never participate "permanent refusers." Third, as Brick and Montaquila (2009) note, response propensities are specific to both the units sampled and the survey conditions. The same units may have different response propensities depending on key survey conditions. The survey conditions may be manipulated to increase response rates during data collection.

Rosenbaum and Rubin (1983) provide the framework for the application of propensity scores in observational studies for estimating causal effects. In observational studies, propensity scores are used to approximate unbiased estimates of the average effect of a treatment (the difference in outcomes between those subject to a treatment and those not treated) when the treatment assignment is not randomized. Rosenbaum and Rubin show that the propensity score is the coarsest balancing score and that, at any value of a balancing score, the average treatment effect can be estimated without bias when certain assumptions hold.

Response propensity theory has been used in a wide variety of applications, including survey nonresponse adjustment. Little (1986) applied propensity score theory to surveys, primarily utilizing the property that propensity score is the coarsest balancing score. In surveys, all sampled units are subject to a data collection protocol – as opposed to the observational setting where units are subject to more than one treatment (one of which may be the null treatment). In surveys, the response propensities are primarily used to form groups to satisfy the MAR. In terms of propensity scores, MAR implies that

$$\Pr(R_i = 1|Y_i, \mathbf{X}_i) = \Pr(R_i = 1|\phi(\mathbf{X}_i)). \tag{12}$$

Thus, by conditioning on the groups based on estimated response propensities, we hope to be able to justify the assumption that missing data are independent of the outcome characteristic. The response propensity score is just the dimension-reducing function that facilitates using multiple auxiliary variables in forming groups.

David et al. (1983) outline a structure using the framework of Rosenbaum and Rubin (1983) and define the treatment as the survey response and the outcome as the characteristic being estimated. In observational studies, we are interested in differences in outcomes when subjects self-select into different treatments and outcomes are observed for those with different treatments. In surveys, we do not observe outcomes for those who do not respond. Despite this difference, David et al. (1983) use this structure only to take advantage of theorems of Rosenbaum and Rubin (1983) showing that the propensity score has the dimension-reducing property.

Two assumptions in Rosenbaum and Rubin's (1983) development are the strongly ignorable treatment assignment assumption and the stable unit treatment value assumption (SUTVA). The strongly ignorable assumption roughly translates into the MAR assumption in the survey context, and it is considered in most applications of propensity scores in nonresponse adjustment. In many cross-sectional household surveys, the lack of powerful predictors means that the strongly ignorable or MAR assumption is tenuous. Of course, the effectiveness of propensity scores to satisfy the MAR assumption is bounded by the power of the auxiliary data used to create the score. Researchers appreciate this limitation and have sought to find better variables or to collect them using paradata.

The second key assumption in propensity score theory, SUTVA, is rarely discussed in the nonresponse adjustment literature. In observational studies, SUTVA is sometimes summarized as a lack of interference between units. One way to translate this into the survey situation is to state that the response propensities of the sampled units are not affected by those of other units, at least within the subsets or groups of units used to estimate the propensities. The typical approach to estimate propensities is to assume that the response for a sampled unit is independent of responses for other units. For the multistage, clustered samples used in many household surveys, this practice seems problematic. For example, interviewers in face-to-face surveys are typically clustered in areas to reduce travel costs. There is ample evidence showing that interviewers and supervisors may influence response (Groves and Couper 1998). Skinner and D'Arrigo (2011) use multilevel models and find some bias in estimates of response propensities that ignore clustering. They suggest using conditional maximum likelihood for estimating propensities rather than the standard logistic modeling. They see the problem as a failure to satisfy the strong ignorability assumption rather than the SUTVA. Other examples are clearer failures of SUTVA, such as when sampling more than one adult per household or multiple teachers from a school. In this case, the sampled units may influence other sampled units directly.

Finally, an issue we think is likely to have even greater importance is related to the definition of the propensity in the nonresponse setting. The propensity is often treated as a fixed attribute of a sampled unit. This conceptualization of response propensities prompted Dalenius (1983, p. 412) to take a "dim view" on estimating response propensities because "it appears utterly unrealistic to postulate fixed response probabilities which are independent of the varying circumstances under which an effort is made to elicit a response." In large measure, we agree and believe a more refined definition of response propensities is needed.

We prefer to express the propensity so that the survey conditions are explicit, such as

$$\phi_i = \phi(a_i, \mathbf{X}_i) = \phi(a_{i1}, a_{i2}, \ldots, \mathbf{X}_i) = \Pr\left(R_i = 1 | \mathbf{a}_i, \mathbf{X}_i\right), \tag{13}$$

where the effort or activity vector (**a**) indicates the relevant data collection activities. The components of the activity vector encompass all forms of data collection, such as the number of call attempts, the use of incentives, the modes of data collection, and refusal conversion attempts. Schouten et al. (2011b) and Olsen and Groves (2012) are also explicit about including fieldwork as well as other variables known for all sampled units when defining the propensity. The quantity that should be estimated to create a nonresponse adjustment factor is $\phi'(\mathbf{a}_i, \mathbf{X}_i)$, where the prime denotes the actual activities at the end of data collection. Defining the propensities as in (13) does not simplify the task, but at least it better defines the quantity being estimated.

Olsen and Groves (2012) and Schouten et al. (2011b) both postulate that response propensities are dynamic, with the response propensity of a sampled unit varying as the recruitment protocol changes. They show that response propensities are influenced by the data collection protocol. In our terminology, they demonstrate that the response propensities are not constant when at least some components of **a** are altered.

Olsen and Groves (2012) also plot conditional response propensities and show that these decline over the field period during which a stable data collection protocol is in place. They argue that this decline implies that the individual's response propensity decreases over repeated applications of the same recruitment protocol. While their explanation is consistent with our perception and with the discussion in Schouten et al. (2011b), there is an alternative explanation that highlights our concern about the unobservable nature of response propensities. Assume that the persons in the sample are members of two different RHGs, with 70 percent of the sample having fixed response propensities of 0.4 and 30 percent having propensities of 0.2. The dotted lines in Figure 1 show the constant propensities over the data collection (effort) for each of the RHGs, and the solid line shows the decreasing propensity of the entire sample. The solid line approximates the shape observed by Olsen and Groves (2012), suggesting that combining RHGs with different propensities could produce the effect they observed even though the conditional response propensities for individuals are constant. Because the response propensities are unknown even after data collection, it is impossible to assess whether the propensities are changing or whether we are mixing groups with different, but constant, response propensities.

8. Response Propensities and Data Collection

The importance of the connection between data collection and nonresponse adjustments can be illustrated by simple examples. We begin with an example inspired by Olsen and Groves (2012). A sample is selected and a standard data collection protocol is applied to all sampled units; some units respond at the end of the first phase of data collection. For the second phase, a subsample of nonrespondents is selected and given a new protocol (e.g., a large incentive, more highly trained interviewers, a different mode), which increases response. We assume that all the units in the sample have identical response propensities, $\phi\ (\mathbf{a}, \mathbf{X})$, but that only those in the subsample are given the additional effort.

One approach to estimation (Approach A) is to exclude those units not in the second-phase subsample; weight the first-phase respondents by the inverse of their selection

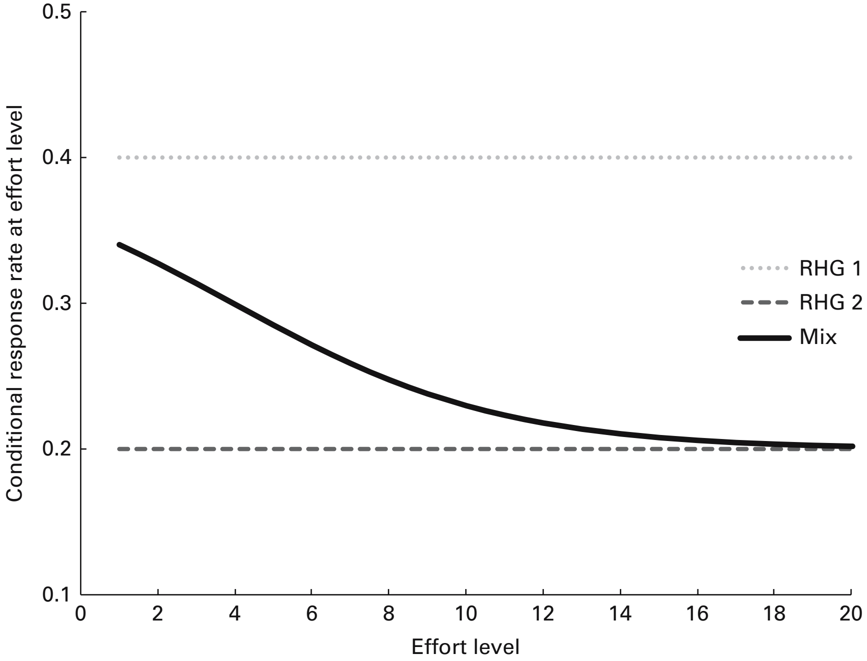

Fig. 1. Observed response propensities for a sample composed of two RHGs

probabilities, d_i; and weight the second-phase respondents by d_i times the inverse of the product of the subsampling rate and the response rate within the subsample. For example, if half of the nonrespondents are subsampled and 40 percent of these respond, the weight for the second-phase respondents would be $5d_i$ ($5 = .5^{-1} \cdot .4^{-1}$ is the adjustment factor). Under the sample-response mechanism, this estimator is unbiased.

In practice, it may be tempting to use an alternative Approach B that computes the nonresponse adjustment across respondents to both the first phase and the second phase to reduce the nonresponse adjustment factor and its impact on the variance of the estimates. In this case, all respondents get the same final weight – d_i times the inverse of the response rate, where the response rate is computed over the entire sample rather than the subsample. Essentially, this estimator ignores the subsampling. The Approach B estimator is biased if the characteristics of the second-phase respondents differ from those of the first-phase respondents. The problem is that the Approach B estimator combines two groups that have different response propensities at the end of data collection. In other words, while all the sampled units have the same ϕ ($\mathbf{a}$, $\mathbf{X}$), they have different values of $\phi'(\mathbf{a}, \mathbf{X})$ because the activity vectors are not identical for the first- and second-phase units. The MAR assumption holds only when the groups are defined by the data collection activity.

Now consider a slightly revised example with the same structure. Suppose we want to estimate the proportion with a characteristic ($y_i = 1$), and assume that the units with $y_i = 1$ have a response rate of 60 percent at the end of the first phase while units with $y_i = 0$ have a first-phase response rate of 40 percent. This is a classic example of topic salience bias. We assume that no auxiliary data are available to identify those with and without the characteristic. A second-phase protocol is implemented by giving *all* nonrespondents an incentive, and the conditional response rate for the second phase is 60 percent for those with $y_i = 1$ and 50 percent for those with $y_i = 0$. The two adjustment methods used above are applied; Approach A computes the nonresponse adjustment factor over just the second-phase respondents (there is no subsampling here); Approach B computes it over all respondents. Figure 2 shows the bias associated with two adjustment approaches. Neither method eliminates the bias completely because the additional phase does not eliminate the difference in the response rates between units with $y_i = 1$ and $y_i = 0$. Thus, this is an example of NMAR. However, Approach A produces estimates that are less biased in this situation because the difference in rates or response propensities is reduced by the second phase of data collection. This result is not always obtained, as discussed below.

In both examples, the data collection activities applied to the units affect the response propensities at the end of data collection. In the first example, the response propensities for all the units are identical but the adjustment groups must be defined by phase for MAR to hold. In the second example, the response propensities differ for those with and without the characteristic, and we must "know" that the incentive applied at the second phase reduces the differences in response rates for these groups to justify using Approach A. We can observe that the percentage with $y_i = 1$ is greater in the second phase than the first phase, but there is no test to show that this reduces bias (an example below has the opposite effect). Of course, the rationale for the second-phase incentive "should" have been that it would reduce bias, otherwise it is hard to justify its application to the second-phase data collection protocol. Unfortunately, in many surveys these factors are not fully considered in data collection, and the main concern is increasing the overall response rate.

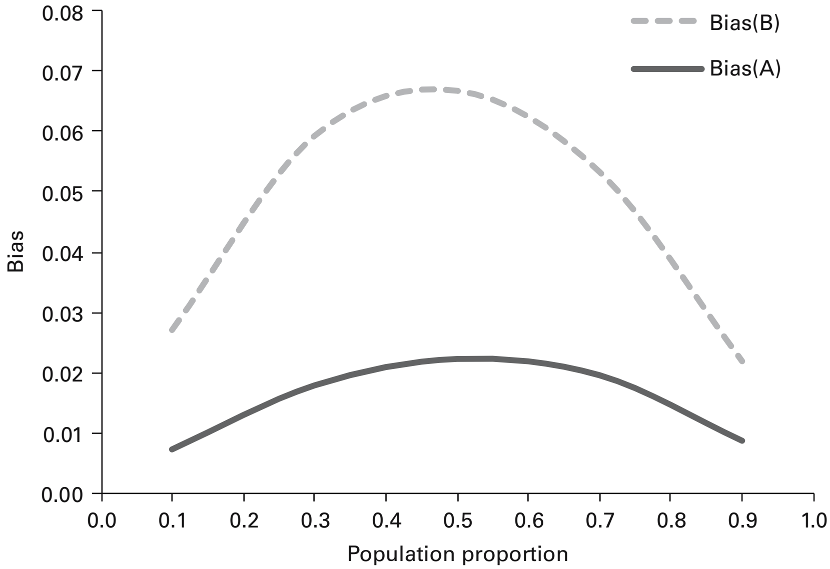

Fig. 2. Bias in estimated proportion using two adjustment methods

These hypothetical examples may seem overly simple or unrealistic, so some examples from real surveys are presented. We begin with two examples with desirable outcomes. The first is taken from Mohadjer et al. (1997). They show that providing incentives in an adult literacy survey improved responses more from low education and minority adults, resulting in reduced bias in key outcomes such as literacy scores by race and education level. A second example is provided by Groves and Heeringa (2006), who offered incentives in the second phase of a responsive design. They too show differential improvements in response rates and reduced nonresponse bias for some statistics.

Examples of changes in data collection protocols that have little or no effect on the estimates appear to be more numerous. This suggests that these results are surprising because publishing null results is generally difficult. One of the first of the recent examples of this genre is Keeter et al. (2000), who substantially increased response rates in a telephone survey by increasing the level of effort (number of call attempts, length of data collection period, etc.). Despite the higher response rates, however, almost none of the estimates from the survey changed significantly. Similar outcomes have been observed numerous times, for example by Curtin et al. (2005), Haring et al. (2009), and Ingen et al. (2009). While there are several possible explanations for the lack of an effect on the estimates, these examples point out gaps in our understanding of the effects of data collection efforts on biases.

There are also examples of data collection efforts that increase both response rates and nonresponse bias. Wetzels et al. (2008) document a survey where incentives increased response rates and had little effect on most estimates. They also report that response rates of non-Western foreigners did not increase with the use of incentives, possibly increasing the biases of estimates related to this subgroup. Merkle et al. (1998) describe an experiment with incentives in an exit poll survey where increased response rates were accompanied by increased nonresponse bias. They suggest that the incentives appealed differentially to voters by party. Schmeets (2010) examines changes in data collection procedures to increase response rates for the Dutch Parliamentary Election Study. He concludes that the changes increased the survey response rates but also might have increased the bias for some of the estimates.

These examples lead us to consider how we can use effort data from the survey to form RHGs for adjustment purposes. Clearly, the data collection activities do not have to be the same for all units; rather, the objective is to classify units with the same final response rate into a RHG irrespective of how they get to this final state.

The traditional approach to forming groups is to use the auxiliary data to identify groups with different response propensities by logistic regression models. Olsen and Groves (2012) suggest using discrete hazard models because response propensities vary over the data collection period. Because the goal is to identify groups of sampled units that have the same value of $\phi'(\mathbf{a}_i, \mathbf{X}_i)$ at the end of data collection, we believe hazard models might be valuable only when the sequencing of data collection activities is important to the response process. The Skinner and D'Arrigo (2011) findings indicate that conditional maximum likelihood estimation might account for clustering.

A perhaps more important realization is that, for most surveys, regression models may not be useful in assigning sampled units to RHGs based on data collection effort. For example, suppose all the units in the sample have the same response propensity for a three-contact data collection protocol. Some units respond at each contact level and some do not respond after all three contacts. If we model response based on the number of calls it took to get a response, we would form RHGs giving different adjustment factors to the respondents by the number of calls it took to respond. These RHGs would only increase the variation in the weights and could, in some situations, introduce bias. Contrast this with the first hypothetical example given above, where bias is reduced by weighting only those units given additional effort. Why should we not adjust the weights only for the cases that responded on the third call? The difference is that we assume in the three call example that the response propensities at the end of the protocol are the same regardless of when the unit responded. The data themselves do not inform us which assumption is correct. Modeling of effort will not reveal this. We would argue that if we subsampled nonrespondents at the end of the first contact, then forming RHGs based on effort would be appropriate in most surveys. The most troubling fact is that the real examples cited above show that we do not always know which assumptions are most reasonable. Although forming RHGs with logistic regression models based on $\mathbf{X}$ is valuable, modeling based on data collection activity may not be as effective without a more complete theory of response.

9. Discussion

As we have mentioned several times, there is a substantial literature that shows the effectiveness of data collection strategies for enhancing response rates. Such strategies include changing modes of data collection, providing incentives, and converting reluctant respondents. When these strategies reduce nonresponse bias, however, is less clear. Without a better understanding of these effects, it is difficult to design effective data collection and estimation strategies to combat nonresponse bias for surveys.

Responsive and adaptive designs have been proposed as a way to reduce nonresponse bias, but these are predicated on making changes to data collection strategies either during data collection (Groves and Heeringa 2006) or from analyses of response patterns in previous collections (Schouten et al. 2011a). Because these designs implement data collection protocols that may vary at the sample case level, they require a refined understanding of the effects these data collection protocols have on nonresponse bias. These types of designs have the potential to increase nonresponse bias if the design, data collection, and estimation stages are not fully integrated.

For example, suppose increased effort is given to some sampled units identified during data collection based on paradata collected in the initial contacts. How should this be handled in forming RHGs? Should units getting extra effort be identified as separate RHGs, or should we assume that the extra effort for those units equalizes response rates so that separate groups are not needed? The answer depends on the assumptions made about the effect of the efforts on nonresponse bias. Surveys that use responsive or adaptive designs need to explain the rationale for their nonresponse adjustment procedures sufficiently so that others can assess the assumptions underlying their estimation methods.

The central problem, in our opinion, is that even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level. This may be a harsh critique given all the progress we have made in many areas. We better understand methods to reduce nonresponse due to noncontact in surveys and have made substantial strides in this area. We also have a much better understanding of correlates of nonresponse. Over time, studies have replicated the correlations between demographic and geographic variables and nonresponse rates (e.g., Groves and Couper 1998; Stoop et al. 2010). These are important developments but have not led to a profound understanding of the causes of nonresponse.

Stoop (2005) reviews some of the areas of research on noncooperation in surveys, but her review shows few lessons that can be generalized and used to reduce nonresponse bias. For example, some research has shown that certain types of people – outgoing and altruistic people – seem to cooperate in surveys more than others. However, utilizing these findings to mitigate nonresponse bias remains a challenge. Another example is the practice of asking people why they refuse to participate in surveys. These requests produce uninformative responses such as being “too busy,” and the distribution of these responses has been constant for years (Brick and Williams 2013). Even though we know that sampled units will never be able to answer our analytic questions about the response process directly, we continue to ask these questions. To better understand the response process we need to reformulate our approach, use less direct questions, and ask both respondents and nonrespondents similar items to support comparative analysis (Singer and Ye 2013).

Some research approaches do appear to have promise and could lead to improvements in our practices and our understanding. For example, if we can increase the perceived value of the survey to the respondent and make the response process simple and enjoyable, then we could potentially lower nonresponse bias (e.g., Dillman et al. 2009). Additional research into ways of increasing the value and making the process more enjoyable is needed. Another promising development is by Groves et al. (2006), who report on an innovative approach to try to generate nonresponse bias in surveys by manipulating factors thought to be related to nonresponse bias. Many practitioners were surprised that their results showed less bias than might have been expected. The idea of prospectively manipulating factors in a controlled manner could increase our understanding of the response mechanism.

One of the difficulties preventing a deeper understanding of nonresponse in surveys is the complexity of the survey process. Many factors in a survey contribute to complexity and may affect nonresponse. These factors include the target population, sponsorship, survey content, interviewer training and experience, mode of data collection, incentives, length of interview, the available field period, and regulatory limitations. Complex systems are inherently more difficult to analyze than simple ones.

One of the ways that other sciences have made progress in studying complex systems is to conduct basic research, often in a laboratory setting, to isolate important main effects. Survey research seems to lack that type of basic research. The exception is statistical design and estimation work that is not as constrained as data collection. Nearly all survey research is empirical, and most of our knowledge comes from experiences in specific surveys. This makes it harder to generalize the findings.

Cognitive research methods were originally introduced into surveys with some of these issues in mind. Over time, this movement has largely devolved into a set of tools to improve questionnaires. Tanur (1999) reviews the origins and evolution of cognitive research in surveys. Today, there are few, if any, settings or laboratories where survey methodologists and psychologists can postulate and explore response theories without being tethered to the needs of a particular survey. The reasons that the cognitive movement has gone in this direction seem clear in hindsight: The research is situated in survey organizations, and those organizations need to justify the allocation of scarce resources. As a result, the application to specific surveys is a higher priority than basic research.

Perhaps the time is ripe for new approaches to the vexing and important question of why people do and do not respond to surveys. Interdisciplinary and basic research may prove profitable if the structural issues can be addressed. But substantive progress cannot be guaranteed by any single approach. Research on making the process more respondent friendly, experiments to induce nonresponse bias, and comparative analysis of respondents and nonrespondents using indirect assessments of attributes of response may have merit. Until we have methods to better understand the relationships between survey requests and response, we are unlikely to be able to structure survey designs, data collection protocols, and estimation schemes to minimize nonresponse bias.

10. References

Andridge, R.H. and Little, R.J. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. Journal of Official Statistics, 27, 153–180.

Atrostic, B.K., Bates, N., Burt, G., and Silberstein, A. (2001). Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights. Journal of Official Statistics, 17, 209–226.

Bartholomew, D.J. (1961). A Method of Allowing for 'Not-at-Home' Bias in Sample Surveys. Applied Statistics, 10, 52–59.

Bates, N., Dahlhamer, J., and Singer, E. (2008). Privacy Concerns, too Busy, or Just not Interested: Using Doorstep Concerns to Predict Survey Nonresponse. Journal of Official Statistics, 24, 591–612.

Beaumont, J.F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment. Survey Methodology, 31, 227–231.

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Bethlehem, J.G. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook in Nonresponse in Household Surveys. New York: Wiley.

Brehm, J. (1993). The Phantom Respondents: Opinion Surveys and Political Representation. Ann Arbor: University of Michigan Press.

Brick, J.M. and Jones, M.E. (2008). Propensity to Respond and Nonresponse Bias. Metron-International Journal of Statistics, LXVI, 51–73.

Brick, J.M. and Kalton, G. (1996). Handling Missing Data in Survey Research. Statistical Methods in Medical Research, 5, 215–238.

Brick, J.M. and Montaquila, J.M. (2009). Nonresponse and Weighting. Handbook of Statistics. Sample Surveys: Design, Methods, and Applications, D. Pfeffermann and C.R. Rao (eds). Vol. 29A. Amsterdam: Elsevier-North Holland, 163–186.

Brick, J.M., Montaquila, J., Han, D., and Williams, D. (2012). Improving Response Rates for Spanish-Speakers in Two-Phase Mail Surveys. Public Opinion Quarterly, 76, 721–732.

Brick, J.M. and Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. The ANNALS of the American Academy of Political and Social Science, 645, 36–59.

Cassel, C., Särndal, C.-E., and Wretman, J. (1983). Some Uses of Statistical Models in Connection With the Nonresponse Problem. Incomplete Data in Sample Surveys, W.G. Madow and I. Olkin (eds). Vol. 3. New York: Academic Press.

Chang, T. and Kott, P.S. (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika, 95, 557–571.

Cochran, W. (1977). Sampling Techniques, (3rd edition). New York: Wiley.

Colley, R.H. (1945). Don't Look Down Your Nose at Mail Questionnaires. Printers' Ink, March, 16, 21–108.

Curtin, R., Presser, S., and Singer, E. (2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment. Public Opinion Quarterly, 64, 413–428.

Curtin, R., Presser, S., and Singer, E. (2005). Changes in Telephone Survey Nonresponse Error Over the Past Quarter Century. Public Opinion Quarterly, 69, 87–98.

Da Silva, D.N. and Opsomer, J.D. (2004). Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism. Survey Methodology, 30, 45–55.

Da Silva, D.N. and Opsomer, J.D. (2009). Nonparametric Propensity Weighting for Survey Nonresponse Through Local Polynomial Regression. Survey Methodology, 35, 165–176.

Dalenius, T. (1983). Some Reflections on the Problem of Missing Data. Incomplete Data in Sample Surveys, W.G. Madow and I. Olkin (eds). Vol. 3. New York: Academic Press, 411–413.

David, M., Little, R., Samuhel, M., and Triest, R. (1983). Nonrandom Nonresponse Models Based on the Propensity to Respond. Proceedings of the Business and Economic Statistics Section of the American Statistical Association, 168–173.

David, M., Little, R.J.A., Samuhel, M., and Triest, R. (1986). Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association, 81, 29–41.

De Leeuw, E. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 41–54.

Deming, W. (1953). On a Probability Mechanism to Attain an Economic Balance Between Resultant Error of Response and the Bias of Nonresponse. Journal of the American Statistical Association, 48, 743–772.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Dillman, D. (1978). Mail and Telephone Surveys: The Total Design Method. New York: Wiley.

Dillman, D., Smyth, J., and Christian, L. (2009). Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, (3rd edition). New York: Wiley.

Dunkelburg, W. and Day, G. (1973). Nonresponse Bias and Callbacks in Sample Surveys. Journal of Marketing Research, 10, 160–168.

Ferber, R. (1949). The Problem of Bias in Mail Returns: A Solution. Public Opinion Quarterly, 12, 669–676.

Feskins, R., Hoop, J., Lensvelt-Mulders, G., and Schmeets, H. (2011). Collecting Data Among Ethnic Minorities in an International Perspective. Field Methods, 18, 284–304.

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994). Regression Weighting for the 1987-88 National Food Consumption Survey. Survey Methodology, 20, 75–85.

Goyder, J. (1987). The Silent Minority: Nonrespondents on Sample Surveys. Boulder, CO: Westview Press.

Greenlees, J., Reece, W., and Zieschang, K. (1982). Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed. Journal of the American Statistical Association, 77, 251–261.

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly, 70, 646–675.

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: Wiley.

Groves, R.M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P., and Nelson, L. (2006). Experiments in Producing Nonresponse Bias. Public Opinion Quarterly, 70, 720–736.

Groves, R., Dillman, D., Eltinge, J., and Little, R. (2002). Survey Nonresponse. New York: Wiley, 41–54.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society, Series A, 169, 439–457.

Hansen, M.H. and Hurwitz, W.N. (1946). The Problem of Non-Response in Sample Surveys. Journal of the American Statistical Association, 41, 517–529.

Haring, R., Alte, D., Völzkea, H., Sauer, S., Wallaschofski, H., John, U., and Schmidt, C. (2009). Extended Recruitment Efforts Minimize Attrition but not Necessarily Bias. Journal of Clinical Epidemiology, 62, 252–260.

Hartley, H.O. (1946). Discussion of "A Review of Recent Statistical Developments in Sampling and Sample surveys.". Journal of the Royal Statistical Society, 109, 37–38.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–162.

Holt, D. and Smith, T.M.F. (1979). Post-Stratification. Journal of the Royal Statistical Society, Series A, 142, 33–46.

Ingen, E., Stoop, I., and Breedveld, K. (2009). Nonresponse in the Dutch Time Use Survey: Strategies for Response Enhancement and Bias Reduction. Field Methods, 21, 69–90.

Kalton, G. (1983). Compensating for Missing Survey Data. Ann Arbor: University of Michigan Press.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. Journal of Official Statistics, 18, 81–97.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. Public Opinion Quarterly, 64, 125–148.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. Journal of the Royal Statistical Society, Series A, 173, 389–407.

Lin, I.-F. and Schaeffer, N.C. (1995). Using Survey Participants to Estimate the Impact of Nonparticipation. Public Opinion Quarterly, 59, 236–258.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R.J.A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association, 88, 125–134.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis With Missing data, (2nd edition). New York: Wiley.

Lumley, T., Shaw, P., and Dai, J. (2011). Connections Between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. International Statistical Review, 79, 200–220.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. Journal of Official Statistics, 15, 305–327.

Madow, W.G., Nisselson, H., and Olkin, I. (1983). Incomplete Data in Sample Surveys, Vol. 1. New York: Academic Press.

Madow, W.G. and Olkin, I. (1983). Incomplete Data in Sample Surveys, Vol. 3. New York: Academic Press.

Madow, W.G., Olkin, I., and Rubin, D.B. (1983). Incomplete Data in Sample Surveys, Vol. 2. New York: Academic Press.

Merkle, D., Edelman, M., Dykeman, K., and Brogan, C. (1998). An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error. Paper presented at the Annual Conference of the American Association for Public Opinion Research, St. Louis, MO.

Micklewright, J., Schnepf, S., and Skinner, C. (2012). Non-Response Biases in Surveys of Schoolchildren: The Case of the English Programme for International Student Assessment (PISA) samples. Journal of the Royal Statistical Society, Series A, 175, 915–938.

Mohadjer, L., Berlin, M., Rieger, S., Waksberg, J., Rock, D., Yamamoto, K., Kirsch, I., and Kolstad, A. (1997). The Role of Incentives in Literacy Survey Research. Adult Basic Skills: Innovations in Measurement and Policy Analysis, A. Tuijnman, I. Kirsch, and D. Wagner (eds). Creskill, NJ: Hampton Press.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every Missingness not at Random Model has a Missingness at Random Counterpart With Equal Fit. Journal of the Royal Statistical Society: Series B, 70, 371–388.

Oh, H.L. and Scheuren, F.J. (1983). Weighting Adjustments for Unit Nonresponse. Incomplete Data in Sample Surveys, W.G. Madow, I. Olkin, and D.B. Rubin (eds). Vol. 2. New York: Academic Press, 143–184.

Olsen, K. and Groves, R.M. (2012). An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period. Journal of Official Statistics, 28, 29–51.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Phipps, P. and Toth, D. (2012). Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data. Annals of Applied Statistics, 6, 772–794.

Politz, A. and Simmons, W. (1949). An Attempt to Get "Not at Homes" Into the Sample Without Callbacks. Journal of the American Statistical Association, 44, 9–31.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–55.

Rubin, D.B. (1976). Inference and Missing Data (with discussion). Biometrika, 63, 581–592.

Särndal, C.-E. (2011a). Morris Hansen Lecture: Dealing With Survey Nonresponse in Data Collection, in Estimation. Journal of Official Statistics, 27, 1–21.

Särndal, C.-E. (2011b). Three Factors to Signal Non-Response Bias with Applications to Categorical Auxiliary Variables. International Statistical Review, 79, 233–254.

Särndal, C.-E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester, UK: Wiley.

Särndal, C.-E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Journal of Official Statistics, 4, 251–260.

Särndal, C.-E. and Lundström, S. (2010). Design for Estimation: Identifying Auxiliary vectors to reduce nonresponse bias. Survey Methodology, 36, 131–144.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Schmeets, H. (2010). Increasing Response Rates and the Consequences in the Dutch Parliamentary Election Study 2006. Field Methods, 22, 391–412.

Schouten, B. (2007). A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption. Journal of Official Statistics, 23, 51–68.

Schouten, B., Calinescu, M., and Luiten, A. (2011a). Optimizing Quality of Response Through Adaptive Survey Designs. The Hague: Statistics Netherlands, Available at: http://www.cbs.nl/NR/rdonlyres/2D62BF4A-6783-4AC4-8E4512EF20C6675C/0/2011x1018.pdf. (Accessed May 24, 2013).

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Measures for the Representativeness of Survey Response. Survey Methodology, 35, 101–113.

Schouten, B., Schlomo, N., and Skinner, C. (2011b). Indicators for Monitoring and Improving Representativeness of Response. Journal of Official Statistics, 27, 231–253.

Singer, E. (2002). Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 163–177.

Singer, E. and Ye, C. (2013). The Use and Effects of Incentives in Surveys. The ANNALS of the American Academy of Political and Social Science, 645, 112–141.

Skinner, C.J. and D'Arrigo, J. (2011). Inverse Probability Weighting for Clustered Nonresponse. Biometrika, 98, 953–966.

Smith, T.W. (1995). Trends in Non-Response Rates. International Journal of Public Opinion Research, 7, 157–171.

Steeh, C., Kirgis, N., Cannon, B., and DeWitt, J. (2001). Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century. Journal of Official Statistics, 17, 227–247.

Steele, F. and Durrant, G.B. (2011). Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal. International Statistical Review, 79, 70–91.

Stoop, I.A.L. (2005). The Hunt for the Last Respondent: Nonresponse in Sample Surveys. The Hague: Social and Cultural Planning Office.

Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response: Lessons Learned from the European Social Survey. Chichester: Wiley.

Synodinos, N.E. and Yamada, S. (2000). Response Rate Trends in Japanese Surveys. International Journal of Public Opinion Research, 12, 48–72.

Tanur, J. (1999). Looking Backwards and Forwards at the CASM Movement. Cognition and Survey Research, M. Sirken, D. Hermann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds). New York: Wiley, 13–20.

Thomsen, I. (1973). A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Nonresponse When Analyzing Survey Data. Statistisk Tidskrift, 11, 278–285.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). The Psychology of Survey Response. New York: Cambridge University Press.

Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. Public Opinion Quarterly, 74, 223–243.

Wetzels, W., Schmeets, H., Van den Brakel, J., and Feskens, R. (2008). Impact of Prepaid Incentives in Face-to-Face Surveys: A Large-Scale Experiment With Postage Stamps. International Journal of Public Opinion Research, 20, 507–516.

Yates, F. (1946). A Review of Recent Statistical Developments in Sampling and Sample Surveys. Journal of the Royal Statistical Society, 109, 12–43.

Accepted March 2013