*Public Opinion Quarterly*, Vol. 76, No. 3, Fall 2012, pp. 555–575

# RESEARCH SYNTHESIS
# A COMPARISON OF ALTERNATIVE INDICATORS FOR THE RISK OF NONRESPONSE BIAS

JAMES WAGNER*

**Abstract** The response rate has played a key role in measuring the risk of nonresponse bias. However, recent empirical evidence has called into question the utility of the response rate for predicting nonresponse bias. The search for alternatives to the response rate has begun. The present article offers a typology for these indicators, briefly describes the strengths and weaknesses of each type, and suggests directions for future research. New standards for reporting on the risk of nonresponse bias may be needed. Certainly, any analysis into the risk of nonresponse bias will need to be multifaceted and include sensitivity analyses designed to test the impact of key assumptions about the data that are missing due to nonresponse.

## Introduction

Nonresponse bias has long been a concern for surveys. It is well known that this bias is the product of nonresponse rates *and* differences between respondents and nonrespondents on survey statistics. Unfortunately, the survey outcome measures are usually not available for nonrespondents. Therefore, assessing nonresponse bias is bound to be fraught with uncertainty. In most cases, we will not know what the nonresponse bias of a given statistic is, and we will be left to speculate on what it might be, using more or less reasonable assumptions.

We do, however, know the response rate. Perhaps this explains why, despite its incomplete nature, the response rate is commonly used as a key measure

James Wagner is an assistant research scientist at the University of Michigan's Survey Research Center, Ann Arbor, MI, USA. The author would like to thank the anonymous reviewers and the editors for their very useful editorial comments. This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development [5R03HD070012-02 to J.W.]. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the National Institute of Child Health & Human Development. *Address correspondence to James Wagner, G373 Perry, 426 Thompson Street, Institute for Social Research, University of Michigan, Ann Arbor, MI 48104, USA; e-mail: jameswag@isr.umich.edu.

doi:10.1093/poq/nfs032 Advance Access publication 10 September 2012

of the risk of nonresponse bias for surveys. A recent meta-analysis of special studies of nonresponse concludes that the response rate is not generally predictive of nonresponse bias (Groves and Peytcheva 2008), calling into question the utility of this key statistic. Without a doubt, alternatives to the response rate as a measure of data quality are needed. There is, however, very little research into alternative indicators. Rather than attempt to summarize or demonstrate all possible indicators, this research synthesis will examine broad categories or types of indicators in order to discuss the characteristics of each type and to suggest avenues for further research. The goals of this research synthesis are the following: 1) to provide a new typology for measures of data quality; 2) to summarize what is known and what is not known about a limited set of statistical measures from each of these types; and 3) to suggest areas for future research.

## Background

Response rates have become a near universal measure of data quality (Biemer and Lyberg 2003). However, there is not necessarily a direct link between nonresponse rates and nonresponse bias. This should be readily apparent since the nonresponse rate is a property of a survey, yet we know that nonresponse bias is a property of a statistic. Assuming that the population consists of two strata—respondents and nonrespondents—there are two components of the bias of a mean: the nonresponse rate and the difference in the population means between respondents and nonrespondents,

$$B\left(\bar{Y}_r\right)=\left(\frac{M}{N}\right)\left(\bar{Y}_r-\bar{Y}_m\right)$$

where $\bar{Y}_r$ is the mean of the respondent population, $\bar{Y}_m$ is the mean of the nonrespondent population, $M$ is the number of nonrespondents in the population, $N$ is the population size, and $B\left(\bar{Y}_r\right)$ is the bias of the respondent mean. Nonresponse bias can also be formulated in a stochastic version (Bethlehem 2002). In either formulation, the bias involves both the response rate and the survey values of the nonrespondents.

Recent empirical evidence indicates that the nonresponse rate may be a poor predictor of nonresponse bias. The previously mentioned meta-analysis of Groves and Peytcheva (2008) included 59 specialized studies of nonresponse bias. They found little correlation between the response rate and nonresponse bias across the many statistics produced by these studies. In addition, several empirical studies have shown instances where lower response rates do not lead to increased nonresponse bias (Keeter, Miller, et al. 2000; Curtin, Presser, and Singer 2000; Merkle and Edelman 2002; Keeter, Kennedy, et al. 2006). In these case studies, for the range of response rates considered, there seems to be very little correlation between the overall response rate and nonresponse bias.

The obvious problem is that while the nonresponse rate is known, the difference between respondents and nonrespondents on a statistic of interest is not usually known. To the extent that response rates are not a good indicator for nonresponse bias, decisions about data-collection activities or post-survey adjustments that are made based on the response rate will be inefficient, biasing, or both. Something is needed to fill this gap between response rates—which are known—and nonresponse biases—which are unknown, but are the thing about which we are really concerned.

Unfortunately, research into alternative indicators is just beginning. Potential indicators have recently been suggested by Schouten, Cobben, and Bethlehem (2009) and Wagner (2010). Very little is known about the ability of these indicators to detect nonresponse bias under various circumstances.

## Typology of Indicators

Given that we are at the beginning stages of identifying alternative measures, it may be useful to develop a typology for those indicators to help organize what is known and to guide future research. One such typology has been proposed as a result of a workshop held in 2007 on the topic of alternative measures. A paper that developed as a result of that workshop (Groves, Kirgis, et al. 2008) described two types of alternative indicators: (1) a single indicator at the survey level; and (2) individual indicators at the estimate level.

This typology does not explicitly differentiate the response rate from other indicators—the response rate is a survey-level indicator. I am proposing an extension of this typology to include the response rate as a separate type. The new typology also involves a conceptual shift that should aid in the description of the strengths and weaknesses of each type of indicator. The typology I propose is the following: (1) indicators involving the response indicator; (2) indicators involving the response indicator and frame data or paradata; and (3) indicators involving the response indicator, frame data or paradata, and the survey data.

This typology separates the response rate from other "survey-level" indicators. The conceptual shift is that the indicators are differentiated based on the types of data used to estimate each. The first type is the response rate. Its calculation involves only the response indicator, a binary variable that indicates whether the sampled unit has responded or not.

The second type involves auxiliary data that are available for both respondents and nonrespondents. These data include data from the sampling frame and paradata (Couper 1998, 2000; Couper and Lyberg 2005). Paradata are data about the process of collecting data. These may include call records, keystroke data, or observations made by interviewers. Subgroup response rates are an example of this second type of indicator. They involve the response indicator and a variable on the sampling frame or from paradata that differentiate the sample into subgroups of interest.

The third type of indicator uses survey data from respondents as well as the data used for the first two types. An example of indicators of this type is the correlation between the nonresponse-adjusted weight and a survey outcome variable.

I will briefly discuss the first type (i.e., the response rate). I will then discuss the latter two types of indicators in turn, and conclude with discussion of areas where additional research is needed.

## RESPONSE RATES

Conceptually, the response rate is a very simple statistic. The AAPOR *Standard Definitions* includes the following definition for response rate: "The number of complete interviews with reporting units divided by the number of eligible reporting units in the sample" (2011, p. 5). In practice, however, there are complications. For instance, if the sampling frame includes ineligible units, then the response rate is not known since the total number of eligible units in the sample may not be known. This has led to research into methods for estimating the number of eligible units among cases for which this status is undetermined (see Smith 2009 for a comprehensive review).

An implicit model is required to relate the nonresponse rate to nonresponse bias. One form that this implicit model can take is the assumption that the higher the nonresponse rate, the higher the risk of nonresponse bias. Or, alternatively, the model might be that higher nonresponse rates lead to higher nonresponse biases. The Office of Management and Budget (OMB) issues guidelines on nonresponse for federally sponsored surveys. Those guidelines suggest that "a survey with an overall unit response rate of less than 80 percent [should] conduct an analysis of nonresponse bias using unit response rates as defined above, with an assessment of whether the data are missing completely at random" (OMB 2006). These guidelines implicitly suggest that the risk of nonresponse bias is extremely low for all cases when the response rate is above 80 percent. The *Journal of the American Medical Association* (JAMA) has a similar standard, specifying that "[s]urvey studies should have sufficient response rates (generally at least 60 percent) and appropriate characterization of nonrespondents to ensure that nonresponse bias does not threaten the validity of the findings" (JAMA 2012).

As previously discussed, theory shows that these assumptions about the relationship between the response rate and nonresponse bias need not be true. Further, we have empirical evidence that the response rate is not necessarily predictive of nonresponse bias (Curtin, Presser, and Singer 2000; Groves and Peytcheva 2008; Keeter, Kennedy, et al. 2006; Keeter, Miller, et al. 2000; Peytcheva and Groves 2009). In general, if the probability of response is unrelated to any particular statistic produced by the survey, then the response rate will not matter. On the other hand, if the protocol of recruitment is more likely to recruit a set of sampled units who are *different* on any particular statistic than those who are less likely to be recruited under that protocol, then the response rate will be predictive of the statistic of interest (Groves 2006). It is

**Table 1. Counts of Sampled Cases by Response Rate Category, January 2006**

| Category | Count |
|---|---:|
| Interview (I) | 296 |
| Partial interview (P) | 1 |
| Refusal (R) | 165 |
| Noninterview (NI) | 14 |
| Other (O) | 30 |
| Unknown household (UH) | 104 |
| Nonsample (NS) | 322 |

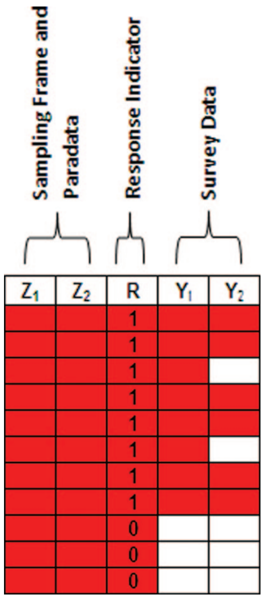
also possible that the response rate might be related to the nonresponse bias of a statistic, but only for certain ranges of the response rate.

Theory also allows that an increase in the response rate could *increase* the nonresponse bias of a statistic. This might occur in a situation where there are equal response rates across a set of subgroups. When the response rate for one subgroup is increased relative to those of other subgroups, the subgroup with the increased response rate will be "overrepresented." If this group differs from other groups on a statistic, then the bias for that statistic will be increased. Merkle and Edelman (2009) provide an example of this.

As an example of response rate calculation, I will use a monthly Random-Digit-Dial (RDD) survey, the Survey of Consumer Attitudes (SCA). The survey conducts about 300 RDD interviews per month. One of the issues when calculating a response rate to an RDD survey is that a proportion of the cases have an undetermined status at the end of data collection. That is, there is a set of telephone numbers for which we do not know whether they are actually households. Table 1 shows how the sampled telephone numbers were classified at the end of the field period in January 2006.

The AAPOR RR1 (AAPOR 2009) for this month is 296/(932 – 322) = 0.485. Under this response rate, all of the undetermined numbers ("UH") are treated as eligible households. If we estimated that 62 percent of those undetermined numbers were actually households (i.e., "$e$" = 0.62), then we could report an AAPOR RR3 of 296/(296 + 1 + 165 + 13 + 30 + 0.62 * 104) = 0.519.

The response rate does have certain strengths. Mainly, it is based on (largely) complete data. Figure 1 shows the response indicator ($R$) as part of a larger data matrix. Each row in the matrix is a data record (i.e., sampled unit), and each column is a variable. The shaded cells are observed elements. The unshaded cells are missing. The rows that have $R = 0$ are unit nonresponse. None of the $Y$ variables are observed for these cases. There are also some cases where $R = 1$, but one of the $Y$ variables is not observed. This is item nonresponse. The response indicator (1 = Respondent, 0 = Nonrespondent) is known for all sampled units. The only model assumption required to calculate the response rate is how to handle the undetermined numbers.



| $Z_1$ | $Z_2$ | R | $Y_1$ | $Y_2$ |
|---|---|---|---|---|
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 1 | 🟥 |  |
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 1 | 🟥 |  |
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 1 | 🟥 | 🟥 |
| 🟥 | 🟥 | 0 |  |  |
| 🟥 | 🟥 | 0 |  |  |
| 🟥 | 🟥 | 0 |  |  |

**Figure 1. Response Indicator, Frame/Paradata, Survey Data.**

Another strength of the response rate is that it is a single-number summary for a survey. This is quite convenient for reporting. In addition, with some context (e.g., mode, sponsor, topic, etc.), survey practitioners have experience with assessing response rates. They know from experience what is a “good” response rate and what is a “bad” response rate—even if the meaning of these evaluations relative to nonresponse bias is unclear.

The response rate is an easy-to-use metric for comparing design options (e.g., different incentive amounts). Again, it should be said that choices made using the response rate as a comparison may not have an effect on the nonresponse bias of a given statistic. Response rates are also useful for evaluating interviewers. Under the assumption of equivalent samples across interviewers, the response rate should be a useful indicator for comparing the effectiveness of interviewers in recruiting sampled units.

Unfortunately, these are only truly “strengths” of the response rate in those situations where it is predictive of nonresponse bias. To the extent that it is a poor indicator of the risk of bias, survey organizations that use it as a guide to data collection may distort their practices. When the response rate guides data collection, the best action is always to target the case that has the highest propensity to respond. If this case is “like” other cases that have already been collected, then we may not be reducing the risk of nonresponse bias.

Heerwegh, Abts, and Loosveldt (2007) describe a situation where data collection is distorted in this manner. They use data from the Flemish Housing

Survey. This survey has on the sampling frame information about whether the household owns or rents the current housing unit. This variable is also collected by the survey. In their analysis, they note that the bias due to noncontact is relatively trivial when compared to the bias due to refusal. Unfortunately, the survey spent more effort (frequent callbacks) addressing the noncontact issue than the refusals. They show that after about 30 percent of the final response had been achieved, the estimates of a key survey statistic had achieved the same level of bias that was achieved when 100 percent of the interviews (not a 100-percent response rate) had been completed. This gives a sense of the extent to which simply adding more relatively easy-to-interview cases may not be the best strategy for minimizing nonresponse bias.

Since the response rate (or its inverse) is a component of nonresponse bias, knowing the response rate allows us to set bounds on the possible nonresponse bias for a proportion. Groves, Presser, and Dipko (2004) derive the maximum possible nonresponse bias for a difference in means. With some assumptions, the response rate can also be used to place bounds on other types of estimates (Nicoletti 2010).

## INDICATORS INVOLVING SAMPLING FRAME DATA AND PARADATA

The second type of indicators includes more data than the response rate. They include auxiliary data from the sampling frame and paradata. These data are largely complete. Figure 1 shows the Response Indicator ($R$) and two sampling frame or paradata variables ($Z_1$ and $Z_2$). We can know how respondents and nonrespondents differ on these variables.

For example, in an RDD telephone survey, the sampling frame might include information about whether the sampled telephone number is listed in a directory. It might also include information from the estimated geography of the telephone number—for example, Census data about the estimated ZIP code, such as the proportion of households that rent, the median income, or the proportion over age fifty (Johnson et al. 2006). Paradata for an RDD telephone survey might include the number of calls made (Couper 2005) and observations from interviewers about contact with informants. An example of the latter type of observations in an RDD survey is information from interviewers about the interactions with households that refuse to complete the survey (Fuse and Xie 2007).

In a face-to-face survey with an area probability sample, the sampling frame might include Census data about the selected census blocks (population and housing counts as well as information about the age, race, and ethnicity distributions of the population) and more detailed data from the American Community Survey about the census block group. In this setting, paradata might include interviewer observations about the characteristics of the neighborhood, observations of the selected housing unit, and observations about interactions with household members (for examples, see Groves and Couper 1998). Paradata in face-to-face surveys can also include interviewer observations that are specifically tailored to the survey. For example, the National Survey of Family Growth (NSFG) Cycle 7 asked interviewers to estimate whether the selected respondent was in a sexually active relationship (Kreuter et al. 2010).

Examples of indicators based on sampling frame, paradata, and the response indicator include subgroup response rates (Groves, Brick, et al. 2008), variance functions of nonresponse weights, variance functions of poststratification weights, variance functions of response rates on subgroups, goodness-of-fit statistics on propensity models (Särndal and Lundström 2008), and R-Indicators (Schouten, Cobben, and Bethlehem 2009). Comparisons of demographic information from the interview to published population totals, as is done in poststratification, rightly fall into this type since they infer combined coverage and response rates for subgroups (Skalland 2011).

The response rate may also be divided into components using paradata about whether there has ever been a contact or refusal at each sampled unit. This allows the evaluation of the sources of nonresponse. The contact and cooperation rates are essentially rates that describe how much of the nonresponse is due to noncontact and how much is due to a lack of cooperation from households that have been contacted. As pointed out by Groves and Couper (1998), these different mechanisms for nonresponse may produce different biases.

These indicators rely on model assumptions that apply to each statistic produced by a survey. A key assumption is that, conditional on the auxiliary variables used to create the subgroups, propensity models, or adjustment weights, the respondents are simply a sample of the sample. Since they may “sample” themselves at different rates across the subgroups, we need to assign different weights to respondents of each subgroup in order to create unbiased estimates. Little and Rubin (2002) describe this as the “Missing at Random” (MAR) mechanism. These mechanisms, like nonresponse bias, may vary by statistic. If this is the mechanism for a statistic and we identify the correct set of subgroups, then we can create unbiased estimates with a nonresponse adjustment. Little and Rubin also describe a “Missing Completely at Random” (MCAR) mechanism where the respondents are a random sample of the sample, without respect to any subgroupings. In such a situation, no nonresponse adjustment would be necessary. Finally, they describe a “Not Missing at Random” (NMAR) mechanism. Under this mechanism, no nonresponse adjustment is available to eliminate the nonresponse bias. These mechanisms can be extended to continuous variables as well.

Type 2 indicators vary in the difficulty with which they can be implemented. Subgroup response rates involve dividing the sample into groups and calculating a response rate for each group. These response rates can be easily calculated during the field period. The variance functions of nonresponse and poststratification weights, on the other hand, require that these weights be created. Often, this is done only after a survey is completed. These comparisons may be helpful in evaluating the survey after the fact. For them to be useful during data collection, these adjustment weights would need to be calculated at regular intervals throughout the field period. This is somewhat more complex than comparison of subgroup response rates. Indicators that use propensity models can be used throughout the data-collection period to monitor data collection and compare strategies (Schouten, Cobben, and Bethlehem 2009). There is no guarantee that these models will provide consistent results over the course of the field period. Future research might consider the impact of sequential estimation of these models using incoming paradata.

One strength of this type of indicator is that they are based on complete data. There are generally no missing values for the variables used in the estimation of these indicators. A second strength of these indicators is that they can be reported at the survey level. This allows for a condensed summary of the survey relative to the risk of nonresponse bias. In addition, indicators of this type should encourage the search for a better sampling frame and paradata. The more predictive these data are of the response indicator and, implicitly, the survey statistics, the better able we are to assess the risk of nonresponse bias.

Another strength of these measures is that when they are used as a guide to data collection, the goal becomes something other than interviewing the case with the highest probability of responding. The goal may be to balance response rates across groups defined by the auxiliary data. Under these measures, cases have differing impacts on the guiding statistic and, therefore, can be given different priorities during data collection.

An indicator of this type is the coefficient of variation of subgroup response rates. The coefficient of variation is $\frac{\sigma_x}{\bar{x}}$. In this case, $\bar{x}$ is the overall response rate (the weighted mean of the subgroup response rates) and $\sigma_x$ is the standard deviation of the subgroup response rates. This is a standardized version of the variance of the response rates. For the RDD survey described above, this statistic is calculated daily. Figure 2 shows the coefficient of variation for the response rates by census region.

The decreases in this indicator show that the variance of the subgroup response rates is decreasing relative to the mean response rate. This indicates that the respondents are becoming a more balanced sample of the original sample on the variable census region. Assuming that census region is predictive of the survey outcome statistics, equalization of these response rates should decrease the bias of the unadjusted statistics.

Another indicator of this type is the "R-Indicator," or representativity indicator (Schouten, Cobben, and Bethlehem 2009). The R-Indicator follows from a definition of representative response. It is intended to be a "measure of the similarity between the response to a survey and the sample or the population under investigation" (Schouten, Cobben, and Bethlehem, p. 101). The R-Indicator is based on the standard deviation of the response probabilities estimated in a logistic regression model. The logistic model is fit with a set

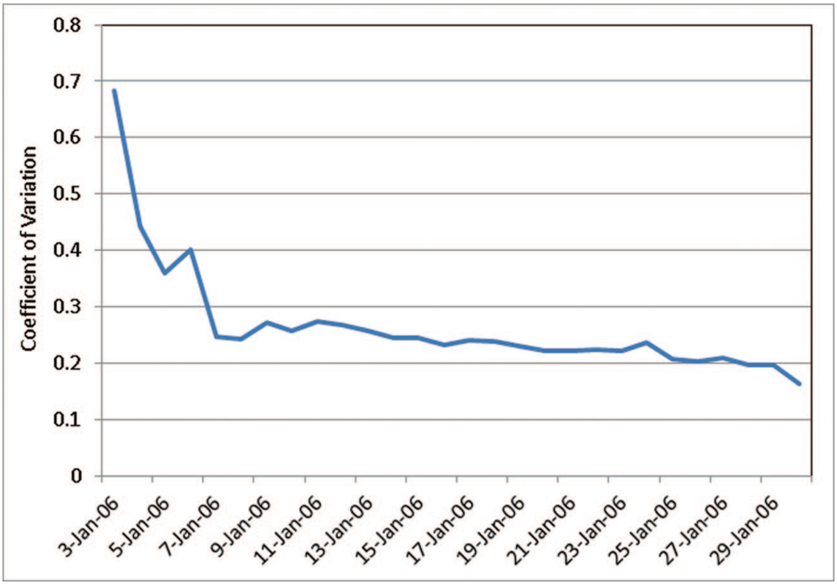


**Figure 2. Coefficient of Variation of SCA Census Region Response Rates by Field Period Day.**

**Table 2. Response Rate and R-Indicator for Alternative Data Collection Strategies on 2005 Dutch LFS[a]**

| Protocol | Response Rate | R-Indicator |
|---|---|---|
| LFS | 62.2% | 80.1% |
| LFS + callback | 76.9% | 85.1% |
| LFS + basic-question | 76.5% | 78.0% |

[a] Adapted from table 4 of Schouten et al. (2009).

of covariates that are available for the entire sample—respondents and nonrespondents alike. These covariates can be drawn from the sampling frame or from paradata. The R-Indicator will be higher when the variability among the estimated response probabilities is lower.

The following example from Schouten, Cobben, and Bethlehem (2009) shows how the R-Indicator can be used to compare data-collection strategies. For the 2005 Dutch Labor Force Survey (LFS), two separate samples of nonrespondents were approached with two different protocols. The first protocol was additional callbacks. The second protocol was the use of a "basic question" approach—that is, using a shortened questionnaire. The response rate for phase 1 and the rates following each of the follow-up methods are presented in

**Table 3. Summary of Typology**

| Data sources | Example indicators | Assessing risk of NR bias | Comparison across surveys | Comparison within surveys across waves/time | Comparison of statistics within survey | Monitoring data collection |
|---|---|---|---|---|---|---|
| 1. Response indicator | • Response rate | • Useful under very strong assumptions<br>• Can set limits on bias<br>• May be useful for MCAR evaluation<br>• Empirical evidence draws utility into question | • Comparable, but requires strong assumptions to say “better” or “worse” | • Comparable, but requires assumptions to say “better” or “worse” | • Single response rate for survey<br>• Item missing indicator for each statistic | • Yes, but all cases have the same value<br>• Strategies can only be compared on this dimension<br>• Leads to a preference for methods that raise response rate |
| 2. Response indicator and sampling frame/ paradata | • R-indicators<br>• Coefficient of variation of subgroup response rates | • Informative under model assumptions that apply to all statistics from a survey<br>• Useful for evaluating composition of response relative to sample | • Comparable when same frame/ paradata are used<br>• Requires somewhat weaker assumptions to say “better” or “worse” | • Comparable<br>• Possible to say whether respondents are becoming more like sample in terms of frame/paradata characteristics | • Not comparable, single model of response across all statistics | • Yes, and cases have different values<br>• Leads to a preference for methods that equalize response rates across subgroups |

*Continued*

Table 3. *Continued*

| Data sources | Example indicators | Assessing risk of NR bias | Comparison across surveys | Comparison within surveys across waves/time | Comparison of statistics within survey | Monitoring data collection |
|---|---|---|---|---|---|---|
| 3. Response indicator, sampling frame/ paradata, survey data | • Fraction of missing information<br>• Mean of survey variable of decile of nonresponse-adjusted weight | • Informative under model assumptions for each statistic<br>• Useful for evaluation of MAR assumption and NMAR evaluation possible with sensitivity analysis | • Comparable for same statistic with same frame/ paradata<br>• Requires some what weaker assumptions and allows for different models across statistics | • Comparable for same statistic with same frame/paradata<br>• Under model assumptions, possible to say whether response is becoming more/less biased | • Comparable since it is estimated at the statistic level<br>• Possible to have separate model for each statistic or use the same model | • Yes, cases have different value but need to account for multiple statistics<br>• Leads to a preference for methods that reduce uncertainty about the impact of nonrespondents on estimates |

table 2. The auxiliary variables in the propensity model used to estimate the R-Indicators are listed in table 3 of Schouten, Cobben, and Bethlehem (2009). These include telephone number listing status, region of the country, age groups, ethnic groups, degree of urbanization, gender, average home value at the ZIP-code level, and at least one household member has a paid job.

By this measure, the "basic-question" protocol brings in "more of the same." The "callback" approach, on the other hand, manages to equalize response rates along the dimensions defined by the predictors in the logistic regression model, presumably by interviewing households that came from groups with lower response rates. On this measure, the "callback" approach produces more representative response.

A key weakness of these indicators is that they are at the survey level. There is an implicit model that the frame data and paradata used to create these indicators are correlated with all of the survey estimates. For instance, in the example from the Dutch LFS, the implicit model is that the variables used in the model (age, gender, ethnicity, region, etc.) are correlated with all of the survey estimates. This assumption is not likely to be true, but is certainly more tenable than the implicit assumption required for the response rate to be a useful indicator of nonresponse bias. There is limited empirical evidence about these assumptions. In a separate study of subgroup response rates, Peytcheva and Groves (2009) performed a meta-analysis of twenty-three specialized studies of nonresponse and found that variation in demographic subgroup response rates was rarely predictive of nonresponse bias.

Another weakness is that the quality of these data varies a great deal from survey to survey. For example, RDD surveys have a very limited set of sampling frame data—usually not more than data describing the estimated geography of the telephone number. Surveys that draw samples from government registries, on the other hand, may have a great deal of useful data available on the sampling frame.

In addition, these indicators may be sensitive to decisions about modeling them. Thus, two different researchers may come up with quite different versions of the "same" statistic if they make different choices in the process of building their models. The ability to say that one survey has done better or worse than another is predicated on the assumption that the model used to create the statistic is correct. The impact of these sorts of misspecification errors needs to be more fully explored. Perhaps using multiple models would be useful in addressing uncertainty about the correct specification.

## INDICATORS INVOLVING SAMPLING FRAME DATA, PARADATA, AND SURVEY DATA

Indicators of this type can include all of the available data: the response indicator, the sampling frame data, paradata, and the observed survey data (i.e., the data for respondents; see variables $Y_1$ and $Y_2$ in figure 1). $Y_1$ and $Y_2$ are missing when $R = 0$. Item-level missing data can be incorporated by producing an item-level response indicator. The potential to use all of these data is a strength of these indicators.

Examples of these types of indicators include correlations between auxiliary data and survey variables (Kreuter et al. 2010; Maitland, Casas-Cordero, and Kreuter 2009), correlations between post-survey weights and the survey variables, the variation of means across the deciles of the survey weights, comparison of respondent means across deciles of estimated response propensities (Olson 2006), comparison of late and early respondents (Dunkelberg and Day 1973; Smith 1984; Lin and Schaeffer 1995), follow-up surveys of nonrespondents (Glynn, Laird, and Rubin 1993; Lynn 2003; Zaslavsky, Zaborski, and Cleary 2002; Peytchev, Peytcheva, and Groves 2010), and the fraction of missing information (FMI) (Wagner 2010; Andridge and Little 2011).

Indicators of this type are more complex to compute than the other types. Since they are at the statistic level, they require an examination of results at this level. This certainly requires more effort. These statistics vary in the form that these models can take. Some of these statistics use the model developed for nonresponse weights and then compare the results of many statistics using this single model. This is the approach taken by correlations between post-survey weights and the survey variables, the variation of means across the deciles of the survey weights, and comparison of respondent means across deciles of estimated response propensities.

Comparison of late and early respondents and follow-up surveys of nonrespondents might be special cases that use as their model subgroups defined by paradata about effort. The follow-up surveys differ from comparison of late and early respondents in that they normally involve a different recruitment protocol that has been altered to bring in sampled units that were reluctant to respond under the original protocol.

Since nonresponse bias occurs at the level of the statistic, the fact that these indicators are estimated at that level is a strength. If the model assumptions are good and there are data that are correlated with the survey outcome variable and the response indicator, then it is also possible to begin to say something more directly about the bias. Of course, it will not usually be known that the assumptions are good.

An additional strength of these measures is that they encourage the development of paradata. Sampling frame data are usually very general. Therefore, designing paradata that are predictive of key survey variables becomes important for these types of indicators to be successful. In fact, although prediction of the response propensity is useful, the ability to predict survey outcome variables may be more useful (Little and Vartivarian 2005; Kreuter et al. 2010).

In addition, if these measures were to guide data collection, then survey organizations would not simply be interested in *how many* cases they are able to interview, but also in *which* cases are interviewed. All cases contribute equally to the response rate. This is not the case for the type 2 and 3 indicators. As a result, there is a reason to prioritize some cases over others.

A weakness of these indicators stems from the fact that they are estimated at the statistic level. Given that most surveys have multiple objectives, there would be multiple indicators that lead to potentially different conclusions about data-collection strategies. Sample design faces the same issue. Much of sampling theory is developed for a single statistic (usually a mean or total), whereas most surveys actually have multiple purposes. Methods have been recommended for multipurpose sample design (for example, Kish 1988; Valliant and Gentle 1997) that may apply to these indicators as well.

Another weakness of these indicators is that they involve model assumptions about the relationship of the frame and paradata to the survey variables (possibly separately specified for each survey variable) that allow us to "fill in" the missing survey data—through either weighting or imputation. This model is, however, explicitly specified. Statements about the risk of nonresponse bias are based on the assumption that these models are correctly specified. As with indicators involving frame and paradata, misspecification of these models could lead to the incorrect conclusion.

A further weakness of these indicators is that they depend on good auxiliary data. Data that are related both to the response propensity and to the survey outcome variables are difficult to find, and this is especially true for RDD telephone surveys.

An example of this type of indicator is the variation of means of survey variables across the deciles of the survey weights. If the means vary across the weights, then the weights could make a difference in the analysis. Since the weighted and unweighted means are going to be different, we have to make some assumptions in order to say why we prefer the weighted mean.

Figure 3 shows the mean Index of Consumer Sentiment (ICS) by deciles of the survey weights (which include nonresponse and poststratification adjustments). The weights are increasing from left to right. The mean for each group is indicated by a plus sign, while the 25th and 75th percentiles are the top and bottom of each box. It should be clear from this figure that the means of the ICS are not significantly different across the deciles of the weights. As a result, the weights will not change the estimate very much. From this, we can conclude either that the unadjusted mean is relatively free of bias or that the model used to create the weights is wrong and both the weighted and unweighted means may still be biased.

The fraction of missing information is another example of this type of indicator. The concept was developed as part of a framework for statistical analysis in the presence of missing data (Dempster, Laird, and Rubin 1977; Rubin 1987). The fraction of missing information is a measure of our uncertainty about the values we would impute for missing elements. The most straightforward method for estimating this statistic is creating multiple imputations under an assumed model. The model can take the form of a "hot deck," regression,

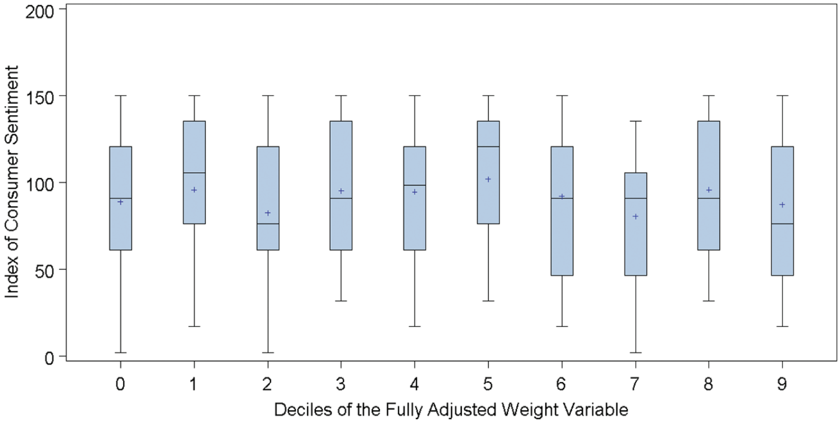


**Figure 3. Variation of Mean ICS by Deciles of the Survey Weight.**

or other stochastic imputation model. If the variation between fully imputed estimates is low relative to the total variance, then we can conclude that correlations between sampling frame data and paradata, on the one hand, and the survey data on the other are allowing us to "recover" some of the information lost due to nonresponse.

Wagner (2010) discusses the use of this statistic for monitoring data collections. As with many of the indicators described here, Wagner assumes that the data are missing at random. That is, conditional on the covariates, the survey variable is independent of the response indicator. Andridge and Little (2011) extend this approach by proposing methods to explore the implications of situations where the data are not missing at random—that is, when there are no statistical adjustments that rely only on the observed data (including sampling frame or paradata) that can correct this bias.

## Summary and Directions for Future Research

Table 3 summarizes the utility of the three types of measures across five different dimensions: detecting nonresponse bias, comparison across surveys, comparison within surveys (across time and waves), comparison of statistics within same survey, and monitoring data collection.

This matrix highlights some of the trade-offs involved when moving across these types. In terms of utility in assessing the risk of nonresponse bias, moving from type 1 to type 3 decreases the strength of assumptions required. Type 1 indicators require quite strong assumptions. The type 2 indicators model the risk of nonresponse bias very generally at the level of the survey. They require the assumption that this mechanism produces similar biases across all statistics produced by a survey. The type 3 indicators model the risk of nonresponse bias directly for each statistic. Each of these indicators requires an MAR assumption, although for the type 3 indicators even NMAR models are possible.

In terms of comparability, moving from type 1 to type 3 involves decreasing the range of comparability. “Comparability” here is restricted to the extent to which the data, the models, and even the essential survey conditions are the same. In fact, in each case, although the estimated indicators (based on the same models) might be similar across surveys or statistics, the nonresponse biases can still be very different. Apart from the treatment of units with undetermined eligibility, response rates have standard calculations across surveys and are, therefore, comparable in the limited sense just described. The type 2 indicators can be estimated in the same way across surveys with the same paradata and sampling frame data. They do not rely on the surveys producing the same statistics. Type 3 indicators are comparable across the same statistics produced by different surveys. Again, underlying the standard of comparability is an assumption that being similar on these measures indicates similar risks of nonresponse bias. This assumption may not hold for any of the types of indicators.

There is very little research into the set of alternative indicators mentioned here. There are at least five broad areas of research that would extend our knowledge in this area. A first area of research has to do with model specification. Since all of these indicators have some model—implicit or explicit—that allows us to speculate about what the unobserved values might be, future research should focus on the consequences of misspecifying these models. The consequences of misspecification in regression are well known. For methods that rely on regression—as is the case with regression-based methods for imputation—this knowledge can be directly applied to these measures. For other indicators, the consequences of misspecification may need to be explored using simulation. Brick and Jones (2008), for instance, explore the implications for misspecification of models used in calibration weighting. There is also a need to explore the use of multiple models to incorporate uncertainty about model specification.

A second area of research has to do with the consequences of nonresponse that is not missing at random—that is, even after conditioning on covariates, the survey outcome variable still depends upon the response indicator. In this situation, stronger assumptions are required in order to address nonresponse bias (Little 1993; Andridge and Little 2011). It may be useful to explore the impact of this type of missing data mechanism on all the indicators identified here. Are some of them more robust to this situation than others?

An assessment of the impact these measures may have on data-collection practices is needed. I have speculated on how a different set of guiding indicators might lead to changes in data-collection practices. However, these changes need to be carefully evaluated. We may not fully understand the implications of changing current practices. Great care must be taken if we begin to target other kinds of cases and even possibly reduce response rates. Experimental methods and specialized studies that have "true values" for respondents and nonrespondents alike are needed in this evaluation. This process needs to be carefully controlled so that we do not develop new procedures that may have unintended, negative consequences on total survey error.

Related to this, we need methods to identify cases that should be prioritized in order to maximize each of the indicators. For a measure like variation of subgroup response rates, it should be fairly obvious which cases, if interviewed, would increase the measure the most. For the R-Indicators, Schouten, Shlomo, and Skinner (2011) have defined univariate "partial R-Indicators" that consider the impact of a single categorical variable on the overall R-Indicator. Multivariate approaches will be needed that allow prioritization based on a vector of covariates.

A fourth area of research is incorporating new knowledge about the risks of nonresponse bias into a total survey error perspective. Monitoring the risk of nonresponse bias may be very expensive. Already, OMB guidelines require that a nonresponse analysis be conducted for federally sponsored surveys with response rates lower than 80 percent. These types of analyses require effort from methodologists, statisticians, and substantive area experts. If these analyses are to be more commonly carried out across the many types of surveys, then new infrastructure may be needed across survey data-collection organizations. However, it may be that expending resources elsewhere is actually more productive in minimizing total error. The proper balance needs to be identified.

Finally, the utility of the type 2 and type 3 indicators relies upon having a high-quality sampling frame and paradata. These data should be correlated with the response indicator *and* with the survey data we aim to collect. Building enriched sampling frames is certainly one area of research. In addition, new methods for designing and building paradata, particularly "proxy-*Y*" variables, are needed.

Each of these areas of research may require examination of mode-specific issues when examining surveys across Web, telephone, mail, and face-to-face modes.

In addition to these areas of research, the field needs new reporting guidelines. AAPOR has worked very hard for many years to develop standards for reporting response rates. This has helped clarify reporting. But the reality of nonresponse bias is that there will need to be multifaceted examinations of its potential impact on surveys. There can be no cookbook for this examination, as every survey faces its own unique problems. No single indicator is likely to replace the response rate. It is necessary to look at the problem from many angles, with different assumptions, in order to provide a plausible case that nonresponse biases for the many statistics produced by any given survey have been diagnosed and remedied. Of course, there is no room for this type of discussion in most articles intended for academic journals. But at least they should make reference to such an analysis which has been done and made public in some format.

In the interest of transparency, our field also needs standards for the release of data that are relevant for probing this problem. For example, most data sets that are publicly released include only records for respondents. An analysis that attempted to address the risk of nonresponse bias for a particular statistic would also need the data on nonrespondents to be made available.

In short, there are important questions that require new research. There are also important questions that require agreement on standards or best practices.

## References

American Association for Public Opinion Research. 2011. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th ed. Available at http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156.

Andridge, Rebecca R., and Roderick J. A. Little. 2011. “Proxy Pattern-Mixture Analysis for Survey Nonresponse.” *Journal of Official Statistics* 27(2):153–80.

Bethlehem, Jelke G. 2002. “Weighting Nonresponse Adjustments Based on Auxiliary Information.” In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 275–88. New York: Wiley.

Biemer, Paul P., and Lars Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

Brick, J. Michael, and Michael E. Jones. 2008. “Propensity to Respond and Nonresponse Bias.” *Metron—International Journal of Statistics* 66(1):51–73.

Couper, Mick P. 1998. “Measuring Survey Quality in a Casic Environment.” *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 41–49.

——. 2000. “Usability Evaluation of Computer-Assisted Survey Instruments.” *Social Science Computer Review* 18(4):384–96.

——. 2005. “Technology Trends in Survey Data Collection.” *Social Science Computer Review* 23(4):486–501.

Couper, Mick P., and Lars Lyberg. 2005. “The Use of Paradata in Survey Research.” *Proceedings of the International Statistical Institute Meetings*, 1–5.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. “The Effects of Response Rate Changes on the Index of Consumer Sentiment.” *Public Opinion Quarterly* 64(4):413–28.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society, Series B (Methodological)* 39(1):1–38.

Dunkelberg, William C., and George S. Day. 1973. “Nonresponse Bias and Callbacks in Sample Surveys.” *Journal of Marketing Research* 10(2):160–68.

Fuse, Kana, and Dong Xie. 2007. “A Successful Conversion or Double Refusal: A Study of the Process of Refusal Conversion in Telephone Survey Research.” *Social Science Journal* 44(3):434–46.

Glynn, Robert J., Nan M. Laird, and Donald B. Rubin. 1993. “Multiple Imputation in Mixture Models for Nonignorable Nonresponse with Follow-Ups.” *Journal of the American Statistical Association* 88(423):984–93.

Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly* 70(5):646–75.

Groves, Robert, J. Michael Brick, M. Couper, William D. Kalsbeek, Brian Harris-Kojetin, Frauke Kreuter, Beth-Ellen Pennell, Trivellore E. Raghunathan, Barry Schouten, Tom W. Smith, Roger

Tourangeau, Ashley Bowers, Matt Jans, Courtney Kennedy, Rachel Levenstein, Kristen Olson, Emilia Peytcheva, Sonja Ziniel, and James Wagner. 2008. “Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools.” *Survey Practice* (October 30):14–22.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, Robert M., Nicole Kirgis, Emilia Peytcheva, James Wagner, William G. Axinn, and William D. Mosher. 2008. “Responsive Design for Household Surveys: Illustration of Management Interventions Based on Survey Paradata.” *Proceedings of the Conference on Health Survey Research Methods*, 1–24.

Groves, Robert M., and Emilia Peytcheva. 2008. “The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis.” *Public Opinion Quarterly* 72(2):167–89.

Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. “The Role of Topic Interest in Survey Participation Decisions.” *Public Opinion Quarterly* 68(1):2–31.

Heerwegh, Dirk, Koen Abts, and Geert Loosveldt. 2007. “Minimizing Survey Refusal and Noncontact Rates: Do Our Efforts Pay Off?” *Survey Research Methods* 1(1):3–10.

Johnson, Timothy P., Young I. K. Cho, Richard T. Campbell, and Allyson L. Holbrook. 2006. “Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey.” *Public Opinion Quarterly* 70(5):704–19.

Journal of the American Medical Association. 2012. “Instruction to Authors.” http://jama.jamanetwork.com/public/instructionsForAuthors.aspx.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey.” *Public Opinion Quarterly* 70(5):759–79.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. “Consequences of Reducing Nonresponse in a National Telephone Survey.” *Public Opinion Quarterly* 64(2):125–48.

Kish, Leslie. 1988. “Multipurpose Sample Designs.” *Survey Methodology* 14(1):19–32.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivellore E. Raghunathan. 2010. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys.” *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 173(2):389–407.

Lin, I. Fen, and Nora Cate Schaeffer. 1995. “Using Survey Participants to Estimate the Impact of Nonparticipation.” *Public Opinion Quarterly* 59(2):236–58.

Little, Roderick J. A. 1993. “Pattern-Mixture Models for Multivariate Incomplete Data.” *Journal of the American Statistical Association* 88(421):125–34.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley.

Little, Roderick J. A., and Sonja Vartivarian. 2005. “Does Weighting for Nonresponse Increase the Variance of Survey Means?” *Survey Methodology* 31(2):161–68.

Lynn, Peter. 2003. “PEDAKSI: Methodology for Collecting Data about Survey Nonrespondents.” *Quality and Quantity* 37(3):239–61.

Maitland, Aaron, Carolina Casas-Cordero, and Frauke Kreuter. 2009. “An Evaluation of Nonresponse Bias Using Paradata from a Health Survey.” *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 370–78.

Merkle, Daniel M., and Murray Edelman. 2002. “Nonresponse in Exit Polls: A Comprehensive Analysis.” In *Survey Nonresponse*, edited by R. M. Groves, pp. 243–57. New York: John Wiley & Sons.

———. 2009. “An Experiment on Improving Response Rates and Its Unintended Impact on Survey Error.” *Survey Practice* (March):6–10.

Nicoletti, Cheti. 2010. “Poverty Analysis with Missing Data: Alternative Estimators Compared.” *Empirical Economics* 38(1):1–22.

Office of Management and Budget. 2006. “Standards and Guidelines for Statistical Surveys.” Available at http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Olson, Kristen. 2006. “Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias.” *Public Opinion Quarterly* 70(5):737–58.

Peytchev, Andy, Emilia Peytcheva, and Robert M. Groves. 2010. “Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences.” *Public Opinion Quarterly* 74(2):319–27.

Peytcheva, Emilia, and Robert M. Groves. 2009. “Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates.” *Journal of Official Statistics* 25(2):193–201.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Särndal, Cark-Erik, and Sixten Lundström. 2008. “Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator.” *Journal of Official Statistics* 24(2):167–91.

Schouten, Barry, Fannie Cobben, and Jelke G. Bethlehem. 2009. “Indicators for the Representativeness of Survey Response.” *Survey Methodology* 35(1):101–13.

Schouten, Barry, Natalie Shlomo, and Chris Skinner. 2011. “Indicators for Monitoring and Improving Representativeness of Response.” *Journal of Official Statistics* 27(2):231–53.

Skalland, Benjamin. 2011. “An Alternative to the Response Rate for Measuring a Survey’s Realization of the Target Population.” *Public Opinion Quarterly* 75(1):89–98.

Smith, Tom W. 1984. “Estimating Nonresponse Bias with Temporary Refusals.” *Sociological Perspectives* 27(4):473–89.

———. 2009. “A Revised Review of Methods to Estimate the Status of Cases with Unknown Eligibility.” Report prepared for the AAPOR Standard Definitions Committee. Available at http://www.aapor.org/AM/Template.cfm?Section=Do_Response_Rates_Matter_1&Template=/CM/ContentDisplay.cfm&ContentID=4682.

Valliant, Richard, and James E. Gentle. 1997. “An Application of Mathematical Programming to Sample Allocation.” *Computational Statistics & Data Analysis* 25(3):337–60.

Wagner, James. 2010. “The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data.” *Public Opinion Quarterly* 74(2):223–43.

Zaslavsky, Alan M., Lawrence B. Zaborski, and Paul D. Cleary. 2002. “Factors Affecting Response Rates to the Consumer Assessment of Health Plans Study Survey.” *Medical Care* 40(6):485–99.