Public Opinion Quarterly, Vol. 70, No. 5, Special Issue 2006, pp. 737–758

SURVEY PARTICIPATION, NONRESPONSE BIAS, MEASUREMENT ERROR BIAS, AND TOTAL BIAS

KRISTEN OLSON

Abstract A common hypothesis about practices to reduce survey nonresponse is that those persons brought into the respondent pool through persuasive efforts may provide data filled with measurement error. Two questions flow from this hypothesis. First, does the mean square error of a statistic increase when sample persons who are less likely to be contacted or cooperate are incorporated into the respondent pool? Second, do nonresponse bias estimates made on the respondents, using survey reports instead of records, provide accurate information about nonresponse bias? Using a unique data set, the Wisconsin Divorce Study, with divorce records as the frame and questions about the frame information included in the questionnaire, this article takes a first look into these two issues. We find that the relationship between nonresponse bias, measurement error bias, and response propensity is statistic-specific and specific to the type of nonresponse. Total bias tends to be lower on estimates calculated using all respondents, compared with those with only the highest contact and cooperation propensities, and nonresponse bias analyses based on respondents yield conclusions similar to those based on records. Finally, we find that error properties of statistics may differ from error properties of the individual variables used to calculate the statistics.

KRISTEN OLSON is a graduate student in the program of survey methodology at the University of Michigan. This article is part of the author's doctoral dissertation research. The material is based on work supported by the National Science Foundation under Grant no. SES-0620228. The author is especially grateful to her committee—Bob Groves, T. E. Raghunathan, Rod Little, Yu Xie, and Norman Bradburn—for discussion and insight into the problem. Frauke Kreuter and Sonja Ziniel provided comments on an earlier draft that improved the article immensely. The author is indebted to Vaughn Call and Colter Mitchell for providing access to the Wisconsin Divorce Study. The Wisconsin Divorce Study was funded by a grant (HD-31035 and HD32180–03) from the National Institute of Child Health and Human Development, the National Institutes of Health. The study was designed and carried out at the Center for Demography and Ecology at the University of Wisconsin–Madison and Brigham Young University under the direction of Vaughn Call and Larry Bumpass. Address correspondence to the author; e-mail: olsok@isr.umich.edu.

doi:10.1093/poq/nfl038

Introduction

Survey response rates in developed countries have fallen over the past three decades (de Leeuw and de Heer 2002). Simultaneously, budgets for surveys have risen dramatically as survey organizations have increased their efforts to counteract this trend (Curtin, Presser, and Singer 2005). Increases in cost and effort have been absorbed because the inferential paradigm of probability sampling demands 100 percent cooperation to guarantee the unbiasedness of a survey estimate. Current best practices argue that researchers should attempt to maximize response rates and to minimize risk of nonresponse errors (Japec et al. 2000). However, recent research (Curtin, Presser, and Singer 2000; Keeter et al. 2000; Merkle and Edelman 2002) has called the traditional view into question by showing no strong relationship between nonresponse rates and nonresponse bias (Groves 2006).

One hypothesis about practices involving nonresponse reduction is that reluctant sample persons, successfully brought into the respondent pool through persuasive efforts, may provide data filled with measurement error (Biemer 2001; Cannell and Fowler 1963; Groves and Couper 1998). Two questions arise when this hypothesized relationship between low propensity to respond and measurement error holds. The first has to do with the quality of a statistic (e.g., means, correlation coefficients) calculated from a survey. That is, does the mean square error of a statistic increase when sample persons who are less likely to be contacted or cooperate are incorporated into the respondent pool? An increase in mean square error could occur because (a) incorporating the difficult to contact or reluctant respondents results in no nonresponse bias in the final estimate, but measurement error does exist, or (b) nonresponse bias exists, but the measurement error in these reluctant or difficult to contact respondents' reports exceeds the nonresponse bias.

The second question has to do with methodological inquiries for detecting nonresponse bias. Although many types of analyses of nonresponse bias can be conducted, four predominant approaches have been used: (1) comparing characteristics of the achieved sample, usually the demographic characteristics, with a benchmark survey (e.g., Duncan and Hill 1989), (2) comparing frame information for respondents and nonrespondents (e.g., Lin and Schaeffer 1995), (3) simulating statistics based on a restricted version of the observed protocol (e.g., Curtin, Presser, and Singer 2000), often called a "level of effort" analysis, and (4) mounting experiments that attempt to produce variation in response rates across groups known to vary on a survey outcome of interest (Groves, Presser, and Dipko 2004). Findings from these studies show that nonresponse bias varies across individual statistics within a survey and is relatively larger on items central to the survey topic as described during respondent recruitment.

The focus of this article is on benchmark comparisons and level of effort comparisons. Benchmark investigations compare a statistic from the survey

with an externally available statistic for the same population, usually from a higher response rate survey or from administrative records. Level of effort analyses investigate the change in a statistic over increased levels of effort, taking change in the statistic to indicate the risk of nonresponse bias, and no change to indicate the absence of risk. But if measurement error is correlated with level of effort (or response propensity), then an observed change or lack of change in the statistic may be due to measurement error and not to nonresponse bias (Groves 2006). Thus, traditional investigations of nonresponse bias based on respondent means may be misleading.

Specifically, in the presence of both measurement error and nonresponse, the bias of a sample mean can be decomposed into a nonresponse bias term and a measurement error bias term. For person i, a survey variable Y_i with true values T_i, the joint effect of nonresponse and measurement error on the respondent mean is $Bias(\overline{y}_r) = \sigma_{pT} / \overline{p} + \sum_{i=1}^{N} (p_i \varepsilon_i / \overline{p})$, where a simple additive error model pertains, $\varepsilon_i = Y_i - T_i$, and σ_{pT} is the covariance of the true values and the response propensity, p (Biemer 2001; Lessler and Kalsbeek 1992). The terms in the equation indicate nonresponse bias and measurement error bias, respectively. There is no nonresponse bias if all sampled units are equally likely to respond, and the only remaining problem is the measurement error bias (Lessler and Kalsbeek 1992). Comparisons of overall nonresponse bias and measurement error bias on survey statistics often show that measurement error bias is at least as large as nonresponse bias, if not larger, and that these nonsampling errors often far outweigh any sampling errors (Assael and Keon 1982; Biemer 2001; Lepkowski and Groves 1986; Schaeffer, Seltzer, and Klawitter 1991).

Similar to analyses described above for nonresponse bias, one approach to studying the joint effects of nonresponse and measurement error is a "level of effort" analysis. Although this method is commonly used to understand nonresponse bias (e.g., Curtin, Presser, and Singer 2000), few studies have jointly examined the change in nonresponse bias and measurement error bias over increasing levels of effort. In this type of nonresponse/measurement error study, survey responses are compared with records for those responses over increasing levels of effort. Such comparisons are rare. Cannell and Fowler (1963) found that the number of hospital stays and length of the stay were misreported more often by those who responded to later follow-ups than to earlier follow-ups. Greater discrepancies for later respondents were found on other topics (Huynh, Rupp, and Sears 2002; Stang and Jöckel 2005; Voigt et al. 2005) and as predictive of sample attrition in panel studies (Bollinger and David 1995, 2001). Each of these studies indicates that measurement error increases for respondents who are more difficult to recruit. Whether this difficulty was due to noncontact or noncooperation, or the relative magnitude of measurement versus nonresponse error over increased levels of effort, is often overlooked in these analyses.

This article provides a first look into these two issues—whether the mean square error of three different statistics changes (and whether the composition of the mean square error changes) as lower propensity respondents are incorporated into a survey estimate. The article also investigates the efficacy of nonresponse bias studies using record data versus respondent reports. A unique data set, the Wisconsin Divorce Study, which used divorce records as the frame, asked questions about information contained on the frame in the questionnaire, and has process data on call outcomes, is used to investigate these issues.

Data

From August 1995 through October 1995 the University of Wisconsin–Madison conducted the Wisconsin Divorce Study. This study was designed as an experimental comparison of mode effects on the quality of divorce date reports. Divorce certificates were extracted from four counties in Wisconsin from 1989 and 1993, and a random sample from each year was selected. One member of the divorced couple was selected at random to be the respondent. Selected persons were randomly allocated to one of three initial modes: CATI, CAPI, and mail. Nonrespondents were followed up in a different mode—CATI and CAPI nonrespondents had a mail follow-up, and mail nonrespondents were followed up by telephone. This article focuses on the CATI with mail follow-up subgroup.

Because of the time lapse between divorce and survey, sampled units were tracked extensively, and addresses were located for 85.2 percent of them. Personalized letters asked the sampled person to participate in the "Life Events and Satisfaction Survey," sponsored and carried out by the University of Wisconsin–Madison. The survey contained questions on satisfaction with life and relationships, marital and cohabitation history, childbearing history, education and work history, satisfaction with current relationships, and demographics. Overall, the response rate (AAPOR RR1) for the CATI with mail follow-up mode was 71 percent, with a contact rate of 80.3 percent and a cooperation rate of 88.3 percent (table 1). Important process data, such as records of the call attempts made by interviewers, were kept for each sampled unit, facilitating our understanding of the participation process and making it possible to disentangle noncontact from refusal nonresponse bias.

Because this survey was not done for the purpose of estimating both nonresponse bias and measurement error bias, the data set has limitations for the present analysis. The most important limitation is that not all variables of interest in the survey are contained in the records. Additionally, records may contain measurement errors, and the construct measured in the survey may deviate slightly from the construct measured in the record. In particular, the frame consists of divorce certificate data on which only the divorce date and child custody arrangements were recorded by an official body; all other

Table 1. Final Disposition of Sample Cases

	n	%
Interviews	523	71.0
Refusal	51	6.9
Contact, no resistance	18	2.4
Noncontact	145	19.7
Total	737	100.0

NOTE.—Nine deceased individuals and one respondent whose gender did not match the frame were removed from the sample.

information was provided by one of the two spouses in the divorcing couple. For this reason, the analyses here largely focus on the statistics calculated using the divorce date, a date used for administrative purposes and probably the least sensitive to measurement error in the record.

FOCAL STATISTICS FOR NONRESPONSE BIAS AND MEASUREMENT ERROR BIAS

Three statistics—all means—are considered in these analyses. First, the length of the marriage is constructed from the difference between the divorce date and the marriage date. The length of marriage is calculated in number of months, the metric in which respondents were asked to report the dates in the questionnaire.[1]

The second statistic is constructed from the difference between the divorce date and the date of the beginning of data collection. This statistic is also measured in months. Thus, two of the three focal statistics use the same variable for these analyses.

Finally, we look at the total number of marriages. Respondents were asked for a count of the number of times they had been married.[2] Marriages that occurred between the divorce in the record and the interview were excluded from this statistic.

Methods

The analyses proceed in four steps. First, we look at overall nonresponse bias by type of nonresponse (noncontact versus noncooperation) and measurement error bias for the three statistics, all sample means, as described above. All estimates of nonresponse bias and measurement error bias are based on differences in

1. The questionnaire asked for each marriage, "In what month and year did your marriage begin?" and, for each divorce, "In what month and year did you get divorced?"
2. The question wording was "How many times have you been married?" and for the month and year of each marriage. Reported marriages that occurred after the divorce date in the record were subtracted from the number of times married.

statistics. The measure of nonresponse bias is the difference between the mean calculated using the records on the entire frame and that calculated using only the respondent pool. Measurement error bias is estimated as the difference between the mean calculated on the complete cases (i.e., those with no item-missing data) from the survey reports and the mean calculated from the record data on all respondents. There is item nonresponse in the survey reports; we take a "naive" analyst approach and ignore the missing data.[3]

Next, we estimate logistic regressions, using available auxiliary data and process data, predicting the probability of being contacted for the survey and the probability of cooperating with the survey request, conditional on contact.

The third step of the analyses examines how nonresponse bias and measurement error bias are associated with response propensity. To do this, we create five roughly equal sized categories or strata from the estimated response propensity scores. Changes in nonresponse bias and measurement error bias for each statistic are examined as lower propensity respondents are incorporated into the estimate of the sample mean (i.e., the cumulative sample mean across propensity strata). Finally, we examine how the total bias and the relative composition of errors change across propensity strata. That is, does the total bias change, and does measurement error bias outweigh nonresponse bias as lower propensity respondents are incorporated into survey estimates?

Findings

NONRESPONSE BIAS: OVERALL

Nonresponse bias of a statistic results when the estimate calculated on the respondent pool differs from the value calculated on the entire population. Table 2 presents the means for the variables available on the frame for five groups: the entire sample,[4] contacts, noncontacts, and interviews and noncooperators (who are mostly refusals) among the contacted. The average length of marriage for the entire frame is 130.29 months, compared with 134.17 months for the respondents, overestimating the population mean by 3 percent.[5]

3. This naive approach, the complete case analysis, has implications for understanding the mechanism behind measurement error and for the estimate of the measurement error itself. Mechanisms behind the misreporting of divorce status, item nonresponse (either don't know or refusal), and inaccurate date reports are confounded in this analysis. Additionally, if the item nonrespondents or the false negatives on divorce status are meaningfully different on the variables of interest, we confound these compositional differences with misreports. However, the naive analyst would not have records at his or her disposal and would not be able to diagnose these problems. Thus, we feel that this complete case analysis is true to the nature of many analyses.

4. One case was excluded because the respondent's gender did not match the gender on the frame.

5. $$bias_{NR} = \left| \frac{\bar{y}_{respondent,record} - \bar{y}_{frame}}{\bar{y}_{frame}} \right|$$

Table 2. Means by Stage of Sample Recruitment

		Length of Marriage (in Months)		Number of Months Since Divorce		Number of Previous Marriages	
	N	Mean	SE	Mean	SE	Mean	SE
Record Value							
Target (full sample)	737	130.29	3.57	49.75	0.90	1.22	0.02
Not Contacted	145	114.46	7.09	48.74	2.07	1.27	0.04
Contacted	592	134.17	4.08	50.00	1.00	1.20	0.02
Contacted, Not Interviewed	69	134.17	13.16	46.68	2.96	1.28	0.07
Interviewed	523	134.17	4.29	50.44	1.06	1.20	0.02
Survey Report (complete cases)	429–520	133.92	4.79	55.74	1.62	1.21	0.02

NOTE.—Variation in *N* for the survey reports due to item nonresponse.

The average length of marriage for noncontacts (mean = 114.46) was significantly ($p = .02$) shorter than the average length of marriage for the interviewed cases, but there was no difference between the interviews and the noncooperators (mean = 134.17, $p = .99$).

Differences between respondents and the frame for the time elapsed between the divorce and the interview are small—49.75 months for the frame versus 50.44 for the respondents, a 1.4 percent overestimate. Both noncontacts and noncooperators were divorced more recently than the interviewed cases (48.74, 46.68, and 50.44 months, respectively), although the differences are not statistically significant. Interviewed cases had slightly fewer marriages than either the noncontacted or noncooperating sample units; the difference between interviews and noncontacts was statistically significant ($p = .06$). Thus, there does appear to be nonresponse bias on the sample means calculated for these estimates, but the overall nonresponse bias is small.

MEASUREMENT ERROR BIAS: OVERALL

Although the frame was constructed such that all selected respondents had been married and divorced, only 98 percent of the respondents reported having been married and 92 percent of the respondents reported being divorced. This, in addition to item nonresponse on the survey, increases the risk of differences between the complete case analysis of the survey reports and the records estimated on the entire respondent pool.

We consider the difference between complete case analyses on the respondents' survey reports and records on the entire respondent pool to be the measurement error bias of the statistic. This difference varies by statistic. For

instance, the difference between the survey report for the length of marriage is 133.9 months versus 134.2 months for the records for all respondents, a relative difference of only 0.2 percent (see table 2).[6] The report of the number of months elapsed between the divorce and the interview is 10.4 percent higher than that calculated from the records (55.7 from survey reports versus 50.4 from the records). The number of marriages estimated from respondent reports is 1.21 marriages, compared with 1.20 estimated from the records for the respondents, a 0.9 percent difference. For two statistics, the measurement error bias is smaller than the nonresponse bias; in the third, the measurement error bias is large relative to the nonresponse bias.

RESPONSE PROPENSITY MODELS

Response propensity is the theoretical probability that a sampled unit will be contacted and will cooperate with a survey request. Many factors in a survey protocol, as well as respondent traits, can influence response propensity. Disentangling these effects requires multivariate modeling. Logistic regression models predicting contactability or cooperation can be used to create summary "response propensity scores" (i.e., the predicted probability from the logistic regression model) that estimate how likely the sampled unit is to participate in the survey, regardless of the actual outcome. Propensity scores have a useful balancing property—conditional on the propensity score, respondents and nonrespondents have equivalent distributions on the observed characteristics entered into the model (Joffe and Rosenbaum 1999; Little 1986; Rosenbaum and Rubin 1984, 1985). Response propensity models are typically estimated when creating weights for postsurvey adjustment. Their use in understanding the risk of nonresponse bias is less well studied.

For these data, we estimate two models—a contact model and a cooperation model, conditional on contact. The dependent variable in the contact model indicates that the sampled case was contacted in the CATI phase or explicitly refused or completed a mail survey. The dependent variable in the cooperation model indicates that the sampled case completed an interview in either phase. These models include three measures of level of effort. First, the number of call attempts before first contact in the CATI phase is available for all sampled cases, measured as the number of calls to first contact for the cases contacted in the CATI phase (mean = 4.29 calls, SE = 0.37) and the total number of calls for the cases not contacted in the CATI phase (mean = 3.54 calls, SE = 0.83).[7]

6. $$relative\ bias_{ME} = \left| \frac{\bar{y}_{respondent,survey\&report\_divorce} - \bar{y}_{respondent,record}}{\bar{y}_{respondent,record}} \right| .$$

7. Virtually all nonrespondents to the CATI phase were sent a mail questionnaire. Disentangling noncontact from refusal in a mail survey is difficult. We consider any case that explicitly returned a mail questionnaire or explicitly refused the mail questionnaire as being a final contact, even if they were not contacted in the CATI phase.

The range of call attempts is quite wide—some cases were never attempted by telephone, only by mail[8]; other cases received up to 102 call attempts. Second, whether a sample case ever refused during the phone attempts (14 percent of contacted cases, SE = 1.44 percent), is available for all sampled cases. Protocol decisions may be made based on both observable characteristics of the respondent, such as age, or on events that occur during the recruitment process, such as persistent noncontacts, in addition to the specified protocol. It is possible that the number of call attempts to first contact reflects both protocol decisions and respondent characteristics. A protocol decision permitted up to two refusals before contact attempts in that mode were stopped. Ever refusing was not included in the contact model, as contact is necessary for a refusal to occur. Finally, all nonrespondents to the phone interview (49.6 percent of the sample cases) were sent a mail questionnaire. Because mail questionnaires followed the phone attempts, they are an indicator of the sampled case having lower contact and lower cooperation propensity (although the mailing itself does not cause these lower propensities).

Additional variables in the propensity models include frame variables that were not used in the construction of the statistics on which nonresponse bias and measurement error bias were measured. These variables include gender (51 percent female, SE = 1.8 percent) and education (some college or more—39.9 percent—versus high school or less—55.4 percent—versus education missing on frame—4.8 percent), whether the sampled person had been married in Wisconsin (74 percent, SE = 1.6 percent), and the number of children in the household at the time of separation (1.05 children, SE = 0.04). These variables are included in the contact and cooperation models.

Clearly, inferences about the relationship between nonresponse bias, measurement error bias, and response propensity are sensitive to the specification of the propensity model. However, level of effort analyses imply a propensity model with one predictor—for instance, the number of call attempts to a sampled household or a mode switch. A typical level of effort analysis implies that respondents with a high number of calls are more like nonrespondents than the rest of the respondents. The models in the present analysis use three measures of level of effort, as well as frame variables, to estimate response propensity, thus making weaker assumptions about the relationship between number of calls and nonresponse bias than a one-variable level of effort analysis. We also estimate noncontact nonresponse propensity separately from noncooperation nonresponse, a separation not typically made in level of effort analyses.

8. While the protocol for the survey was CATI with mail follow-up, about 8 percent (n = 58) of the 737 sample units had no call records, indicating that the case was not called. One case had a result code from the CATI phase of "refusal"; the remainder had a result code from the CATI phase indicating that there was not enough information to contact the case by telephone. Fifty-four of the 58 sampled units without call records were followed up by mail, and 18 returned the mail questionnaire.

Table 3. Response Propensity Models for Contact and Cooperation

	Predicting Contact = 1		Predicting Cooperation = 1, Conditional on Contact	
	Coefficient	SE	Coefficient	SE
Intercept	2.7805****	0.4485	4.1850****	0.6395
Frame Variables				
Married in Wisconsin	0.6031*	0.2435	–0.2038	0.4244
Number of children in household at time of separation	–0.00906	0.1016	0.3835*	0.1793
Some college or more versus high school graduate or less	0.3781	0.2331	0.0530	0.3595
Missing education on frame	–0.1730	0.5205	0.6161	0.8846
Female respondent	–0.1693	0.2126	0.4167	0.3485
Effort Variables				
Sent mail questionnaire	–3.1937****	0.3592	–2.0589****	0.4521
Log (number calls to first contact + 1)	0.4246***	0.1379	–0.1558	0.1986
Ever refused	—	—	–3.0860****	0.3547
N	737		592	
Percent Concordant	82.6		92.8	
Likelihood Ratio Chi-Square	185.87****		187.24****	

+$p < .10$.
*$p < .05$.
**$p < .01$.
***$p < .001$.
****$p < .0001$.

Table 3 provides coefficients from each of these logistic regressions. The strongest predictors are the level of effort variables. The number of calls made before first contact to a household is positively related to contact[9] but not significantly related to cooperation. Interim refusals are significantly less likely to be final interviews than cases that did not refuse. Persons who were sent a mail questionnaire have lower contact and cooperation propensity. Sample persons who were married in Wisconsin are more likely to be contacted than their married-elsewhere counterparts.

9. The relationship between number of calls to first contact and contact propensity is sensitive to the inclusion of the cases whose call records indicate that no calls were made in the CATI phase, but were sent a mail survey. When the cases that received no calls in the CATI phase are excluded, there is no difference in number of calls to first contact between the contacted and uncontacted cases.

Table 4. Response Propensity Strata for Contact and Cooperation Models

	Predicting Contact = 1						Predicting Cooperation = 1, Conditional on Contact					
	Actual Contact Rate		Average Estimated Contact Propensity				Actual Cooperation Rate		Average Estimated Cooperation Propensity			
			Noncontacts		Contacts				Refusers		Cooperators	
Response Propensity Stratum	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*
Low	51.4	148	50.6	72	53.3	76	51.7	118	42.8	57	63.1	61
Group 2	69.4	147	65.6	45	66.2	102	91.6	119	90.5	10	92.0	109
Group 3	83.3	144	81.9	24	88.7	120	99.2	118	96.7	1	98.1	117
Group 4	99.4	154	97.9	1	97.5	153	100.0	121	—	0	98.9	121
High	97.9	144	98.6	3	98.5	141	99.1	116	99.4	1	99.4	115

The number of children in the household at the time of separation is significantly positively related to cooperation.

The predicted propensity scores were divided into five roughly equal-sized groups, ordered from low to high estimated contact or cooperation propensity (table 4).[10] In a perfectly specified response propensity model, the actual response rate and the average estimated propensity for the groups will match. The overall estimated propensities are quite high—the top three groups of contact propensity are above 80 percent estimated likelihood of contact, and the top four groups in cooperation propensity are above 90 percent estimated likelihood of cooperation. Of note, the bottom two contact propensity strata consist entirely of mail respondents and the top two contact propensity strata consist entirely of telephone respondents. Similarly, the bottom two cooperation propensity strata consist almost entirely of mail respondents (at least 88 percent are mail respondents in these strata), and the top two cooperation propensity strata consist entirely of telephone respondents.

RELATIONSHIP BETWEEN LIKELIHOOD OF CONTACT AND LIKELIHOOD OF COOPERATION

The next analyses examine changes in nonresponse bias and measurement error bias by contact and cooperation propensity strata. One question is

10. Five propensity score subclasses are often found to be adequate for removing up to 90 percent of the bias in estimating causal effects (Rosenbaum and Rubin 1984). For the predicted contact propensities, the five groups were calculated on both contacts and noncontacts so that different numbers of contacted cases are in each group. Similarly, the five groups for the cooperation propensity were calculated on both interviews and noninterviews, among the contacted. Thus, there are different numbers of cooperating cases in each group.

Table 5. Distribution of Predicted Cooperation Propensity Strata, Conditional on Contact, by Predicted Contact Propensity Strata among the Cooperators

Predicted Contact Propensity	Predicted Cooperation Propensity						
	Low	2	3	4	High	Total	*N*
Low	16.13	75.81	8.06	0.00	0.00	100%	62
2	37.50	58.33	4.17	0.00	0.00	100%	72
3	18.18	17.17	7.07	13.13	44.44	100%	99
4	3.29	0.66	30.26	32.89	32.89	100%	152
High	0.72	1.45	40.58	42.03	15.22	100%	138
N	61	109	117	121	115		523

whether the respondents in the high contact propensity stratum are also in the high cooperation propensity stratum—that is, are those who are easy to contact also likely to cooperate? If this is the case, then the two sets of analyses will be redundant. There is a relationship between the two propensity strata distributions (table 5, chi-square = 440.34, 16 *df*, $p < .0001$), but it is not a one to one relationship (Spearman correlation = 0.51, asymptotic SE = 0.03). For example, only 16 percent of the respondents in the lowest contact propensity stratum were in the lowest cooperation propensity stratum, and only 15 percent of the respondents in the highest contact propensity stratum were in the highest cooperation propensity stratum.

RELATIONSHIP BETWEEN LIKELIHOOD OF CONTACT, LIKELIHOOD OF COOPERATION, AND NONRESPONSE BIAS

The critical question behind nonresponse reduction efforts is how the nonresponse bias of the estimate changes as respondents with lower propensity are recruited into the survey. That is, do estimates based on the records change over response propensities, and are estimates improved (i.e., lower nonresponse bias) by recruiting lower propensity sampled units into the respondent pool? Figures 1 through 6 present means cumulated over contact and cooperation propensity strata for the respondents. Moving from left to right on each graph indicates how the mean estimated on respondents changes based on adding lower propensity sample units into the respondent pool. The dotted line in each graph represents the target value, that is, the sample mean based on the records. Differences between the solid line (the record mean based on the respondents) and the dotted line indicate nonresponse bias for the unadjusted respondent mean. (The dashed line will be discussed in the next section.)

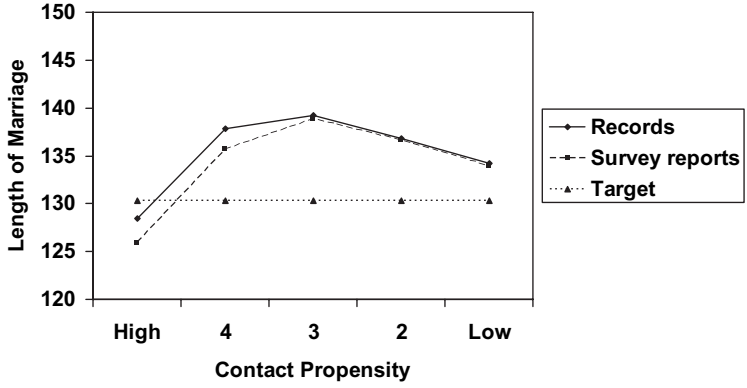

Figure 1. Cumulative mean over contact propensity strata, length of marriage.

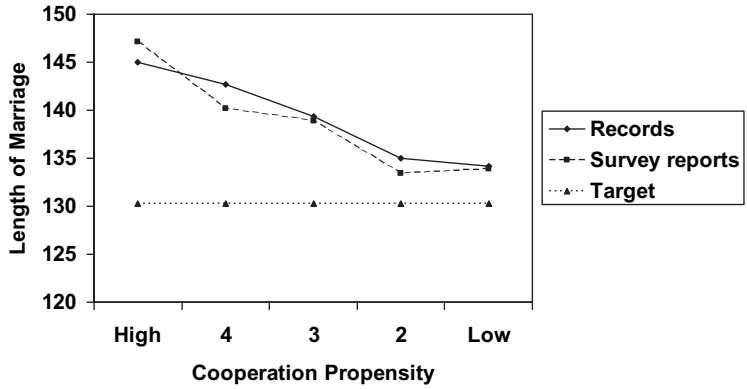

Figure 2. Cumulative mean over cooperation propensity strata, length of marriage.

Three observations can be made from the graphs. First, change in the statistics across contact propensity strata is not the same as change in the statistics across cooperation propensity strata. This makes sense—the relationship between likelihood of contact and cooperation and survey variables is likely to differ if different mechanisms produce contactability and cooperation. For instance, the mean length of marriage has an inverted "U" shape over contact propensity strata. On the other hand, the mean length of marriage calculated over cooperation propensity strata declines, moving closer to the target value.

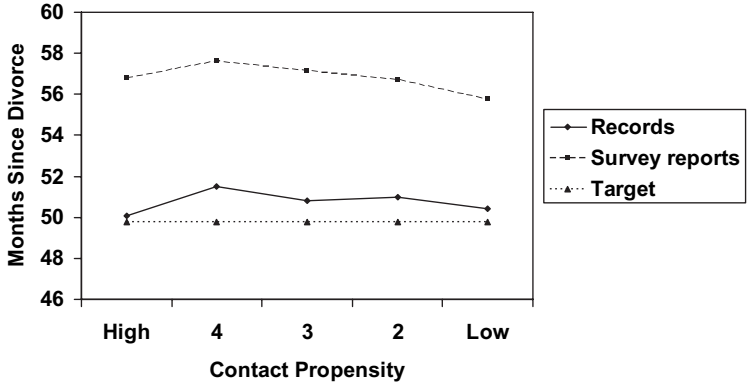

Figure 3. Cumulative mean over contact propensity strata, number of months since divorce.

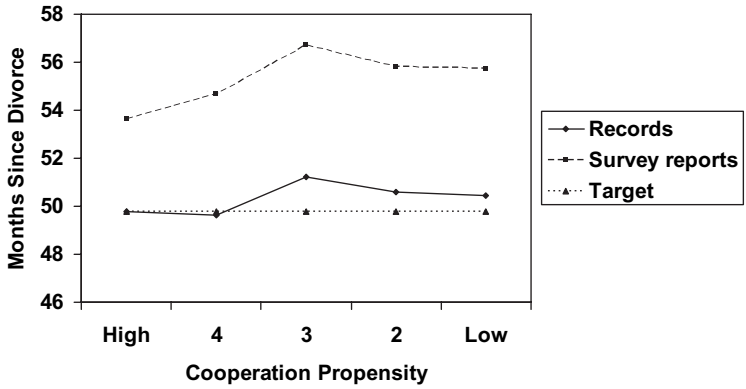

Figure 4. Cumulative mean over cooperation propensity strata, number of months since divorce.

Second, the propensity stratum at which the mean calculated on the respondents is closest to the target value varies by statistic. For example, the nonresponse bias in the mean number of marriages based on respondent reports improves over all contact propensity strata, but the nonresponse bias in the mean number of months since divorce is negligible in almost all cooperation propensity strata. Thus, if these three statistics were being monitored as part of a responsive design (Groves and Heeringa 2006) with phases defined by response propensity, decisions about when

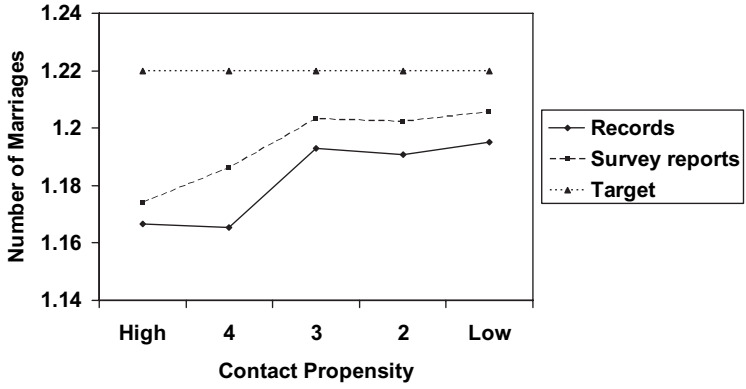

Figure 5. Cumulative mean over contact propensity strata, number of marriages.

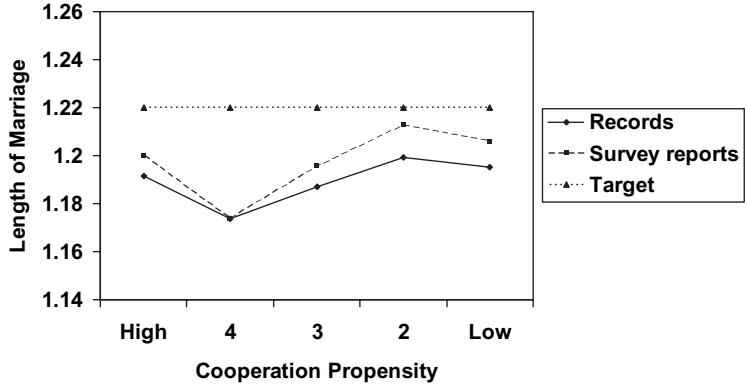

Figure 6. Cumulative mean over cooperation propensity strata, number of marriages.

to adopt a different recruitment strategy would vary depending on the statistic.

Finally, for these statistics, the direction of nonresponse bias (under- or overestimate of the mean) tends to be consistent across response propensity strata. In some cases, the fact that statistics show relatively monotonic trends over response propensity strata (e.g., mean length of marriage for cooperation propensity) can be taken as indication of the statistic's moving closer to the "true" value, although not necessarily reaching the true value. In other cases,

this inference cannot be made (e.g., mean number of months since divorce for cooperation propensity).

EXAMINING NONRESPONSE BIAS USING RESPONDENT REPORTS

Having record values available for estimating nonresponse bias analyses is rare. We now evaluate whether two common approaches to evaluating nonresponse bias based on respondent reports give us the same answer as that using records. The first approach is one in which benchmark data are used to evaluate nonresponse bias properties of a statistic. The second approach is that discussed above, in which movement of a statistic across propensity strata is used to diagnose nonresponse bias. This is the propensity strata equivalent of a level of effort simulation in which respondents recruited with greater levels of effort are removed from the respondent pool, and means from this truncated distribution are compared with the full respondent mean (Curtin, Presser, and Singer 2000).

Assume that the mean for the entire sample based on the records is the obtained benchmark and that the difference between the mean based on respondent reports and the benchmark is ascribed to nonresponse bias. Table 2 shows that for length of marriage, the difference between the "benchmark" and the report-based mean is 3.63 months, compared with 3.88 months when using the records for the interviewed cases. The number of months since divorce shows a difference of 5.95 months when using the survey reports, compared with 0.65 months using the records. The mean number of marriages is 0.01 marriages lower when using the survey reports, and 0.02 marriages lower than the benchmark when using the records. Thus, in two cases, the nonresponse bias estimate is actually smaller when using survey reports instead of records, but in one case, the nonresponse bias estimate is much larger relative to the using the records.

The second scenario is that available to most survey practitioners, in which the change in the respondent mean over different levels of effort is examined. Differences between truncated distributions and the full respondent pool are taken as an indication of nonresponse bias (e.g., Curtin, Presser, and Singer 2000). The dashed line on figures 1–6 represents this respondent report-based mean. As when looking at the record-based means above, as the dashed line moves from left to right on the graph, reports from respondents from lower propensity strata are incorporated into the estimate of the mean.

For all three statistics, the respondent mean calculated from the survey reports tracks quite closely with the respondent mean calculated from the records. Thus, conclusions drawn about whether inclusion of lower propensity respondents improved the nonresponse bias properties of the statistic would be similar, whether or not these estimates were based on respondent reports or record values. Importantly, although the conclusions are similar, the magnitudes

of the estimates differ because the mean is shifted due to measurement error bias in the respondent reports.

CHANGES IN MEASUREMENT ERROR BIAS AND NONRESPONSE BIAS BY LIKELIHOOD OF CONTACT AND COOPERATION

The discrepancy between nonresponse bias estimates based on the survey reports and nonresponse bias estimates based on the records leads to three important questions. First, does the difference between the estimate calculated using the respondent reports and that calculated from records change over response propensity strata? Second, does the total bias change over propensity strata? Finally, does the relative contribution of nonresponse bias and measurement error bias change over propensity strata?

To answer the first question, we calculate the absolute measurement error bias $\left(bias_{ME} = \left|\bar{y}_{respondent,survey} - \bar{y}_{respondent,record}\right|\right)$ for each statistic as cumulated across strata. Columns 1 and 6 of table 6 clearly show that measurement error bias is not constant across propensity strata. For two of the three statistics, measurement error bias decreases as lower contact propensity respondents are incorporated into the sample mean. On the other hand, measurement error bias increases as lower cooperation propensity respondents are incorporated into the sample mean for two of the three statistics, although the increase is not monotonic. For example, the cumulative mean length of marriage, based on the survey reports, decreases in measurement error bias as more reluctant and more difficult to contact cases are included in the estimate of the sample mean. On the other hand, the measurement error bias of the cumulative mean reported number of months since the (last) divorce increases across cooperation propensity strata, but decreases somewhat across contact propensity strata.

To answer the second question, we examine the total absolute bias $\left(\left|\bar{y}_{respondents,records} - \bar{y}_{respondents,reports}\right| + \left|\bar{y}_{sample,records} - \bar{y}_{respondents,records}\right|\right)$. Columns 3 and 8 of table 6 show that the total absolute bias increases between the first and second contact propensity strata, but then decreases across the remaining contact propensity strata for all statistics. The total bias of the overall mean is lower for all statistics compared with the mean for the highest contact propensity stratum. This is not true for cooperation propensity. For mean length of marriage, total bias decreases as lower cooperation propensity respondents are incorporated into the sample mean. For another statistic, the mean time since divorce, total bias increases. Finally, for the mean number of marriages, there is little change in the total bias as lower cooperation propensity respondents are added to the estimate. For these statistics, converting low contact propensity cases appears to contribute more to reduction of total bias than converting low cooperation propensity cases.

Table 6. Measurement Error (ME) Bias, Nonresponse (NR) Bias, and Total Bias for Respondent Means Cumulated over Contact and Cooperation Propensity Strata

	Contact					Cooperation				
	Magnitude of Bias			% Contribution to Total Bias		Magnitude of Bias			% Contribution to Total Bias	
	ME	NR	Total	ME	NR	ME	NR	Total	ME	NR
	1	2	3	4	5	6	7	8	9	10
Length of Marriage										
High	2.50	1.86	4.36	57%	43%	2.09	14.73	16.82	12%	88%
4	2.11	7.52	9.63	22%	78%	2.49	12.43	14.92	17%	83%
3	0.35	8.91	9.26	4%	96%	0.41	9.03	9.44	4%	96%
2	0.21	6.53	6.74	3%	97%	1.52	4.68	6.20	25%	75%
Low	0.24	3.88	4.12	6%	94%	0.24	3.88	4.12	6%	94%
Time Since Divorce										
High	6.72	0.34	7.06	95%	5%	3.86	0.03	3.89	99%	1%
4	6.11	1.76	7.87	78%	22%	5.05	0.12	5.16	98%	2%
3	6.30	1.08	7.38	85%	15%	5.47	1.47	6.94	79%	21%
2	5.71	1.23	6.94	82%	18%	5.23	0.83	6.07	86%	14%
Low	5.30	0.69	5.99	89%	11%	5.30	0.69	5.99	89%	11%
Number of Marriages										
High	0.007	0.053	0.061	12%	88%	0.009	0.029	0.037	23%	77%
4	0.021	0.054	0.075	28%	72%	0.000	0.046	0.046	0%	100%
3	0.010	0.027	0.037	27%	73%	0.008	0.033	0.042	20%	80%
2	0.011	0.029	0.040	28%	72%	0.013	0.021	0.034	39%	61%
Low	0.011	0.025	0.036	30%	70%	0.011	0.025	0.036	30%	70%

Finally, we decompose the total bias within each propensity stratum into the percent contribution due to nonresponse bias $(|\bar{y}_{sample,records} - \bar{y}_{respondents,records}| / (|\bar{y}_{respondents,records} - \bar{y}_{respondents,reports}| + |\bar{y}_{sample,records} - \bar{y}_{respondents,records}|))$ and the percent contribution due to measurement error bias $(|\bar{y}_{respondents,records} - \bar{y}_{respondents,reports}| / (|\bar{y}_{respondents,records} - \bar{y}_{respondents,reports}| + |\bar{y}_{sample,records} - \bar{y}_{respondents,records}|))$. As can be seen in columns 4, 5, 9, and 10 of table 6, the relative contribution of nonresponse bias to the total bias is greater than the relative contribution of measurement

error bias for mean length of marriage and mean number of marriages across virtually all contact and cooperation propensity strata. On the other hand, the relative contribution of measurement error bias outweighs the relative contribution of nonresponse bias for the mean time elapsed since divorce across all propensity strata. Of interest, mean length of marriage and mean time since divorce are two statistics that use the same question, but mean length of marriage is dominated by nonresponse bias and mean time since divorce is dominated by measurement error bias. The contribution of measurement error bias to total bias decreases across contact propensity strata for mean length of marriage and mean time since divorce, but increases for the mean number of marriages across contact propensity strata. There is no difference in the contribution of measurement error bias to total bias among estimates that incorporate the bottom three contact propensity strata. There is no clear trend in change of the contribution of measurement error bias to total bias across cooperation propensity strata for any of the three statistics.

Discussion and Conclusions

This analysis has five main findings. (1) Effects on the nonresponse bias of a survey statistic from turning low propensity sample units into respondents are statistic-specific and specific to the type of nonresponse (contact versus cooperation). This is not a new finding but is worth reiterating. (2) What is new are the findings on how these recruitment efforts are associated with the measurement error bias properties of the same statistics and how measurement error bias affects diagnoses of nonresponse bias. (3) Limited support was found for the suspicion that measurement error bias increases for estimates including reluctant respondents. Such increases were found for two of the three statistics investigated. (4) But, despite the increase in measurement error, total bias of all three statistics decreased as a result of incorporating lower contact propensity cases, and for one statistic as a result of incorporating lower cooperation propensity respondents. (5) Finally, this investigation showed that level of effort analyses came to similar (although not identical) conclusions when based on record data and on survey reports for the statistics and protocol investigated here.

Measurement error bias estimates for these three respondent means differed across contact and cooperation propensity strata. The differences were sometimes small relative to the estimate, and sometimes quite sizable. For two of the sample means, the contribution of nonresponse bias to total bias exceeded that due to measurement error bias over all propensity strata. For one sample mean, the relative contribution of nonresponse bias was much less than the contribution due to measurement error bias. Thus, concerns that the error properties of a sample mean are dominated by measurement error bias after

incorporating low propensity respondents into the sample pool are not consistently borne out. One important caveat is that the total bias changes over propensity strata. Thus, the percent contribution of measurement error bias will not necessarily increase when both measurement error bias and nonresponse bias increase. The relationship between nonresponse bias, measurement error bias, and likelihood of response also clearly depends on which type of nonresponse propensity is considered.

In this study, methods of diagnosing nonresponse bias tended to give similar answers when either records or survey reports were used. The magnitude of the estimate of nonresponse bias differed depending on the data source, but the general direction of conclusions was quite similar for two of the three statistics considered. Replication of this analysis is clearly needed.

The difference between the error properties of a variable in a data set (or question in a questionnaire) and of a statistic such as a mean must be highlighted. Two of the statistics used in this article use exactly the same question—the date of divorce. The length of marriage is the difference between the divorce date and the marriage date. The time elapsed since divorce is the difference between the divorce date and the first day of the field period. Both of these statistics use the divorce date variable but have dramatically different error properties. Mean length of marriage had little measurement error bias, whereas mean time elapsed since divorce was dominated by measurement error bias. One hypothesis is that people are able to retrieve the length of a salient event, such as marriage, but not the individual dates that bound the event. When a questionnaire demands the retrieval of two dates, individuals may recall an approximate date that anchors the beginning of the event (e.g., marriage date), and report a calculated end date (e.g., divorce date) using this retrieved beginning date and length of the event (e.g., length of marriage). Further research on when and how combinations of variables change nonresponse and measurement error structures relative to the original variables is necessary.

One critical element of this analysis is the mixed-mode design. Disentangling whether the mail and phone modes had different nonresponse bias and measurement error bias properties is important for understanding the findings. In this analysis, estimates made on the top two contact and cooperation propensity strata are based solely on telephone respondents. Mail respondents are added into the estimates for the next three strata. Previous research suggests that statistics calculated from self-administered modes may have different measurement error properties than statistics calculated from interviewer-administered modes, at least for sensitive questions (e.g., Tourangeau and Smith 1996). Additionally, mixed-mode surveys are frequently done in the hope that respondents to the second mode will be different from respondents to the first mode on the survey variables of interest (de Leeuw 2005). Thus, one would expect differences in both measurement error bias and nonresponse bias when looking at the two modes individually. We see some hints that this

may be occurring. For example, measurement error bias drops dramatically from mean length of marriage calculated on the telephone respondents alone to the same statistic calculated from phone and mail respondents. Nonresponse bias for mean length of marriage also tends to decrease. However, both measurement error bias and nonresponse bias increase for mean time since divorce as lower cooperation propensity mail respondents are added. Further research into when and how mixed-mode designs are beneficial for mean square error and how error structures change as a result of using more than one mode is clearly needed. The sequencing of modes also is an important question—had this investigation used a mail survey with a telephone follow-up, would similar changes in total error and the composition of error have been observed?

The results of this analysis are conditional on the variables included in the propensity model. Similar analyses were conducted with two other model specifications. One model was identical to the model presented here but excluded the mail questionnaire indicator; the other model replaced the total number of calls to first contact with the total number of calls and replaced the indicator for ever refusing with the total number of refusals. The conclusions from those analyses were similar to those presented here for the two statistics involving dates, but conclusions for mean number of marriages were somewhat more sensitive to model specification. The largest differences between models for all three statistics occurred in the means calculated for the highest contact and cooperation propensity strata. The differences are suggestive of mode differences for the reported number of marriages, but future research on when and why the relationship between total bias, nonresponse bias, measurement error bias, and response propensity changes when different predictors are included in the propensity model is clearly needed.

References

The American Association for Public Opinion Research. 2006. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 4th edition.* Lenexa, Kansas: AAPOR.

Assael, Henry, and John Keon. 1982. "Nonsampling vs. Sampling Errors in Survey Research." *Journal of Marketing* 46:114–23.

Biemer, Paul P. 2001. "Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing." *Journal of Official Statistics* 17:295–320.

Bollinger, Christopher R., and Martin H. David. 1995. "Sample Attrition and Response Error: Do Two Wrongs Make a Right?" University of Wisconsin Center for Demography and Ecology.

———. 2001. "Estimation with Response Error and Nonresponse: Food Stamp Participation in the SIPP." *Journal of Business and Economic Statistics* 19:129–41.

Cannell, Charles F., and Floyd J. Fowler. 1963. "Comparison of a Self-Enumerative Procedure and a Personal Interview: A Validity Study." *Public Opinion Quarterly* 27:250–64.

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

———. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69:87–98.

de Leeuw, Edith. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21:233–55.
de Leeuw, Edith, and Wim de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Perspective." In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 41–54. New York: Wiley.
Duncan, Greg J., and Daniel H. Hill. 1989. "Assessing the Quality of Household Panel Data: The Case of the Panel Study of Income Dynamics." *Journal of Business and Economic Statistics* 7:441–52.
Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Error in Household Surveys." *Public Opinion Quarterly* 70:646–75.
Groves, Robert M., and Mick Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.
Groves, Robert M., and Steven G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, A* 169:439–57.
Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68:2–31.
Huynh, Minh, Kalman Rupp, and James Sears. 2002. *Working Paper 238: The Assessment of Survey of Income and Program Participation (SIPP) Benefit Data Using Longitudinal Administrative Records*. Washington, DC: U.S. Bureau of the Census.
Japec, Lilli, Antti Ahtiainen, Jan Hörngren, Håkan Lindén, Lars Lyberg, and Per Nilsson. 2000. *Minska bortfallet*. Sweden: Statistiska centralbyrån.
Joffe, Marshall M., and Paul R. Rosenbaum. 1999. "Invited Commentary: Propensity Scores." *American Journal of Epidemiology* 150:327–33.
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* 64:125–48.
Lepkowski, James M., and Robert M. Groves. 1986. "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design." *Journal of the American Statistical Association* 81:930–37.
Lessler, Judith T., and William D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: Wiley.
Lin, I-Fen, and Nora Cate Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59:236–58.
Little, Roderick J. A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54:139–57.
Merkle, Daniel, and Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 243–57. New York: Wiley.
Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516–24.
———. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *American Statistician* 39:33–38.
Schaeffer, Nora Cate, Judith A. Seltzer, and Marieka Klawitter. 1991. "Estimating Nonresponse and Response Bias: Resident and Nonresident Parents' Reports about Child Support." *Sociological Methods and Research* 20:30–59.
Stang, Andreas, and Karl-Heinz Jöckel. 2005. "Letter to the Editor: The Authors Reply." *American Journal of Epidemiology* 161:403.
Tourangeau, Roger, and Tom W. Smith. 1996. "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context." *Public Opinion Quarterly* 60:275–304.
Voigt, Lynda F., Denise M. Boudreau, Noel S. Weiss, Kathleen E. Malone, Christopher I. Li, and Janet R. Daling. 2005. "Letter to the Editor: RE: 'Studies with Low Response Proportions May Be Less Biased than Studies with High Response Proportions.'" *American Journal of Epidemiology* 161:401–2.