When do nonresponse follow-ups improve or reduce data quality? A meta-analysis and review of the existing literature

Kristen Olson, Chun Feng, and Lindsey Witt

University of Nebraska-Lincoln

August 1, 2008 DRAFT

Paper presented at the International Total Survey Error Workshop

June 1-4, 2008

Research Triangle Park, NC

Contact author information:
Kristen Olson
University of Nebraska-Lincoln
703 Oldfather Hall
Lincoln, Nebraska 68588-0324
kolson5@unl.edu

I. Introduction

Do survey respondents, recruited with extraordinary efforts, provide answers of lower quality than respondents who are recruited more easily? This question has worried survey practitioners and analysts alike for at least four decades (Cannell and Fowler 1963; Robins 1963), but an answer is not known. Although no direct relationship exists between response rates and nonresponse bias (Groves and Peytcheva 2008), an open question is the relationship between efforts to increase response rates and other sources of survey error, in particular, measurement error. The common hypothesis is that those who require greater recruitment effort provide answers of lower quality than those who are recruited more readily. A latent cooperation continuum often is posited (e.g., (Mason, Lesser, and Traugott 2002; Cannell and Fowler 1963), such that those who are most difficult to recruit to the sample pool have the lowest motivation and are thus the worst reporters. However, it is not clear if this is generally the case.

This paper reviews existing literature on the relationship between the levels of effort exerted for sample member recruitment and data quality, with a primary focus on item nonresponse. Two methods – a quantitative meta-analysis and a systematic qualitative review – are used to examine 44 articles examining the relationship between levels of recruitment effort and measurement error. These studies use five different measures of levels of effort (number of contact attempts/follow-up reminders, refusal conversion, date of interview, combination of these three, estimated response propensity) and look at multiple measurement error indicators (e.g., item nonresponse, response accuracy, signed deviations, scale reliability, acquiescence, non-differentiation).

This review asks the following four questions.

1. Do respondents recruited with more effort have higher item nonresponse rates than respondents recruited more easily?
2. Are particular study characteristics associated with higher item nonresponse rates among respondents recruited with more effort relative to respondents recruited with less effort?
3. Are particular types of items associated with higher item nonresponse rates among respondents recruited with more effort relative to respondents recruited with less effort?

4. Is there consistent evidence about higher or lower levels of other types of measurement error, and does it vary by the level of effort measure?

II. Background

No single measure has been defined to identify respondents who are more likely to participate in surveys from those who are less likely to participate. One method for identifying such respondents is by using paradata (Couper 1998) about the survey recruitment process to divide cases into those who required little recruitment effort and those that required more recruitment effort. In fact, four measures have been used to define the "level of effort" exerted to bring a sampled person into the respondent pool – the number of calls or follow-up attempts, refusal conversion, the date of interview, a combination measure of date, number of contact attempts and/or refusal conversion. The research question underlying each metric can be phrased as "What would have happened to data quality in this study had we stopped (or continued) after recruitment effort X?," where the definition of X may vary from study to study. For example, the actual number of calls or follow-up attempts may be few (two or three in the case of a mail survey) or many (dozens, in the case of a phone survey), the date of interview may be defined in terms of weeks or months, and refusals may be defined as hard or soft refusals. The level of effort is often used as a proxy of the respondent's motivation to complete the survey (e.g., (Cannell and Fowler 1963). A recent development is the use of predicted probabilities from response propensity models as a measure of the likelihood of a respondent participating in a survey (Yan, Tourangeau, and Arens 2004; Olson 2006; Fricker 2007). These measures are estimated for both respondents and nonrespondents, and are derived from a stochastic view of nonresponse (Lessler and Kalsbeek 1992).

Each level of effort measure reflects organizational procedures in addition to characteristics of the respondent. Studies that use calls as level of effort combine noncontact and refusal nonresponse in the "high" effort group, and reflect influences of the survey protocol (e.g., number of mailings or call scheduler) and the respondent's propensity to participate in the survey. Additionally, the call or follow-up attempt on which the nonrespondents would have participated had recruitment requests continued is necessarily unknown, important for the counterfactual question of "What would have happened had we continued (or stopped) making call attempts?." The second measure, refusal conversion, is clearly indicative of refusal

nonresponse, but can only be used in interviewer administered surveys. Refusal conversion measures also reflect interviewer training in use of result codes, supervisor instructions on when to code a case as a refusal, and general practices in the organization about refusal conversion attempts. As with calls, it is unknown whether the converted refusals are representative of the non-converted refusals; also, no direct inference can be made about noncontacted households. A different refusal conversion measure is that of statements made to the interviewer on the doorstep during the recruitment request. Doorstep statements reflect the reasons that a sampled person does or does not want to participate in the survey, and rely less on organizational practices and result codes and more on the interviewers' ability to adequately reflect the doorstep conversation. These may be recorded for the first or later contacts and are often captured for both respondents and nonrespondents. Although closer to a measure of the respondent's decision processes than general refusal conversion measures, these doorstep statements may reflect other aspects of the recruitment, including conversations with the respondent or with a general householder, and, as a self-administered survey for interviewers, are (likely) subject to the same order effects as found in other self-administered surveys.

The date of the interview is the third metric for judging level of exerted recruitment effort. Usually discussed as "early" versus "late" respondents, this measure combines noncontact and refusal nonresponse, but also reflects sample release dates in studies where not all first contact attempts are made on the first day of the study. The date on which the nonrespondents would have responded, had the data collection continued, is not known. Combinations of these above measures are also used. Finally, predicted probabilities from response propensity models use covariates available for both respondents and nonrespondents to estimate individual likelihoods of participating in a survey. Persons with higher predicted propensities are those who are more likely to participate in the survey; conversely, persons with lower predicted propensities are less likely to participate. Predicted propensities are useful because they are directly translatable for both respondents and nonrespondents and are used in weighting methodologies; however, as with any model-based effort, conclusions may be sensitive to model specification.

We expect that the difference in data quality between high and low recruitment effort respondents varies by the type of level of effort measure used. Since the underlying phenomenon are similar, we expect studies using the date of interview to be similar to those using follow-up attempts, especially if the sample is released on the same date (e.g., mail surveys). By attempting

to isolate refusal nonresponse, we expect that the difference between high and low recruitment effort is greater in refusal conversion studies than in other studies (Bradburn 1978).

Unknown are the characteristics of studies and of items that are correlated with more data quality problems among high recruitment effort respondents. Although large numbers of design characteristics can be posited, the study's response rate, mode, change in mode, and topic are important study decisions or characteristics. Additionally, the type of question (attitude, behavior, demographics, or income), question burden and sensitivity are also likely to be related to the difference in data quality between high and low recruitment effort respondents.

As with unit nonresponse bias (Groves and Peytcheva 2008), we do not expect a clear relationship between measurement error for high and low recruitment respondents and the study's overall unit response rate. From one perspective, with higher response rates, greater numbers of difficult to recruit respondents are in the sample pool. Those who required the most recruitment effort in higher response rate surveys are "extra" difficult. However, we expect this to be moderated by a number of other factors, including the attractiveness of the initial and later requests, the mode of data collection, the sponsor, and so on. As such, no prediction is made about the relationship between the overall unit response rate and difference in data quality for the two groups.

We do, however, expect that mail surveys will have the largest differences between high and low recruitment effort respondents compared to other modes. Mail surveys have the unique property of the respondent being able to simultaneously evaluate the measurement situation while making their participation decision (Dillman 1978). As such, those who return a mail survey quickly are more likely to be interested in the topic of the survey and in performing the measurement task (Martin 1994; Groves et al. 2006). Later mail respondents are likely to be less interested in the measurement task, and hence, may be less motivated to do the job of being a respondent relative to other survey modes where the measurement task is not made apparent. As refusal conversion can only be observed in interviewer administered survey, we expect that the relationship of mode with the outcomes will vary with the level of effort measure. Similarly, mode switches for sample member recruitment are likely to change the measurement situation (de Leeuw 2005) and thus make a change in data quality for high recruitment effort respondents more likely relative to studies that do not have a mode switch.

We also expect variation over different types of questions. If the common motivational hypothesis is correct, more difficult items (e.g., more burdensome retrieval, more sensitive) should show larger data quality differences between low and high recruitment effort respondents than other types of questions (Krosnick, Narayan, and Smith 1996; Krosnick 2002; Krosnick and Alwin 1987). For example, income questions often have the highest item nonresponse rates in a survey. As a sensitive and burdensome question, we would expect that income questions have higher item nonresponse rates among high recruitment effort respondents than other types of questions. Following the same logic, we would expect other burdensome questions or more sensitive questions to have greater data quality problems than other types of questions.

III. Research Design

Multiple search engines of scholarly works published in the social sciences were used for this review. These included JSTOR, the Current Index of Statistics, Academic Search Premier, EconLit, Communication and Mass Media Complete, International Political Science Abstracts, Sociological Abstracts, and Web of Science. We used Google and Google Scholar for additional literature potentially not included in these databases. We also reviewed the contents of the Proceedings of the Survey Research Methods Section of the American Statistical Association. Finally, we examined the reference list of each identified study that met our criteria for any additional relevant studies that had not been previously identified.

Two lists of search terms were created. Search terms for unit nonresponse included attrition, callbacks, calls, contacts, continuum of resistance, difficult interviews, follow-up, interim distributional bias, late interviews, late/difficult respondents, level of effort, noncontact, noncontact nonresponse, nonresponse follow-up, nonresponse propensity, panel nonresponse, phases, refusal, refusal conversion, refusal nonresponse, reluctant interviews, reluctant respondents, response propensity, response timing, speed of response, and waves. Search terms for measurement error indicators included accuracy, change in responses, coefficient alpha, consistent answers, correct reporting, data quality, don't know, inconsistent answers, item nonresponse, item omission, match rate, measurement error, missing data, missing data rate, missing information, nonresponse within the questionnaire, not reported, refusal, reliability, respondent ability, response consistency, response quality, unreliability, and validity. Each

search contained at least one word from each list, joined with an "and" criterion (e.g., "difficult interviews" and "accuracy").

To be included, studies met the following criteria: (1) focused on surveys of persons, (2) contained a measure of level of recruitment effort, and (3) examined levels of the measurement error indicator over these levels of effort. Studies that examined the overall contribution of nonresponse and measurement errors to mean square error properties of survey estimates, but did not examine these over successive levels of effort, were excluded (Assael and Keon 1982; Biemer 2001; Lepkowski and Groves 1986; Warriner 1991; Van Goor and Verhage 1999; Schaeffer, Seltzer, and Klawitter 1991). We also excluded studies that looked at measurement error indicators between different experimental conditions (e.g., incentive levels) that varied in response rates if they did not also examine measurement error across varying levels of effort (Willimack et al. 1995; Goyder, Brown, and Martinelli 2005; Singer, Van Hoewyk, and Maher 2000; Martin 1994; De Leeuw and Hox 1988). We also excluded studies that looked at item nonresponse or measurement error indicators as predictors of panel attrition (Loosveldt, Pickery, and Billiet 2002; Bollinger and David 1995, 2001), as the causal order of unit nonresponse and measurement error is reversed. This yielded 44 eligible articles.

From the 44 identified articles, three types of level of effort-measurement error analyses were identified, varying in the detail provided in the measurement error analyses. Some articles contained multiple measurement error analyses. The first type of analysis looked at measurement error indicators on individual items over successive levels of effort (n=15 articles with 16 studies). The number of individual items per study ranged from one (often income) to over 20 items, with a total of 178 items. The second type of analysis examined aggregate item nonresponse measures or other measurement error indicators across multiple items (n=29 articles). The number of items ranged from a few related items to the entire questionnaire. The final type of analysis examined the relationship between level of effort and a measurement error indicator, but presented results controlling for other covariates (n=6 articles). We will summarize the item-by-item analyses quantitatively; we will then systematically review findings for the aggregated measures. The analyses corrected for other covariates are not currently included in this review.

III.A. Article Coding

Coding criteria were developed for both study characteristics and for item characteristics (see Appendix for codebook). The studies were coded for publication year, whether they examined panel or cross-sectional surveys (panel studies excluded here), whether the measurement error indicators were calculated at an aggregate level or an item level, the types of measurement error indicators examined, the types of level of effort measures examined, the words that were used to identify the measurement error and response propensity indicators, the topic of the survey, the sample frame, the target population, the sponsor, the data collection organization, the country, the mode of each level of effort, the overall response rate, the overall sample size, the response rate at each level of effort, and the number of respondents and nonrespondents at each level of effort.

Each item was coded for whether it was an attitude, behavior, knowledge, or income question, whether it was open-ended or closed ended, and whether proxy reports were allowed. Each coder made a subjective judgment as to whether the item would be considered sensitive, socially desirable, burdensome, or pose difficult retrieval problems. The outcome variable was coded as to whether the analyses were conducted for the overall study population or for a particular subgroup, the total number of item respondents, item nonrespondents, the type of item nonresponse considered (don't know, refusal, no opinion, or a combined measure), the number of item respondents at each level of effort, the number of item nonrespondents at each level of effort, item nonresponse rate at each level of effort. If a different measure of data quality was considered, the coders recorded the type of measure (e.g., response accuracy, non-differentiation), and reports of this measure over each wave, with standard errors or standard deviations, as reported.

III.B. Outcome Measure

Some articles report more than one call attempt, refusal conversion, or month of data collection (we call these waves). To account for these differences, we calculate two outcome measures. The first outcome measure compares the respondents who were recruited with the least amount of effort to those that were recruited with the immediately adjacent level of effort. For example, if there were 6 follow-up attempts, then the first outcome measure compares respondents who were recruited at the first attempt to those who were recruited at the next

follow-up attempt. These outcomes are denoted "Wave 1 to Wave 2" in the tables. The second outcome measure compares respondents who were recruited at the first attempt to those recruited at the last follow-up attempt, presumably comparing the "easiest" respondents to recruit to the "most difficult" respondents. Again, using the 6 follow-up attempt example, this comparison examines those who were recruited at the first attempt relative to those at the sixth attempt. This outcome is denoted "Wave 1 to Last Wave" in the tables.

The outcome variable of interest for the meta-analysis is the relative difference in item nonresponse rates for a particular question for those who participated with little exerted effort compared to those who participated with greater levels of exerted effort. Since item nonresponse rates are proportions, we use a log-odds ratio (Lipsey and Wilson 2001). For example, when comparing item nonresponse rates for wave 1 to the next recruitment effort (wave 2), the log-odds ratio is:

$$\log(ORItemNR_{wave12}) = \log\left(\frac{INR_{wave1} / (1 - INR_{wave1})}{INR_{wave2} / (1 - INR_{wave2})}\right)$$

A negative value indicates that the item nonresponse rate is higher in the later recruitment wave than in the first wave. A positive value indicates that the item nonresponse rate is higher in the first wave than in the later wave.

We use inverse-variance weights (Lipsey and Wilson 2001) to account for unequal precision levels across items and across studies. The sampling error associated with this log-odds ratio is: $se(\log(ORItemNR_{wave12})) = \sqrt{\frac{1}{n_{itemRwave1}} + \frac{1}{n_{itemNRwave1}} + \frac{1}{n_{itemRwave2}} + \frac{1}{n_{itemRwave2}}}$

In the analyses, we treat studies as independent observations, and account for the clustering of items within studies in all of our analyses using precision-weighted random effects regression models. Coefficients with the value of zero indicate that there is no difference in item nonresponse rates between respondents recruited with little effort and those recruited with extensive efforts. Models with a negative intercept indicate higher item nonresponse rates among the respondents recruited with more effort than those recruited with less effort for the reference category. In these models, a positive coefficient indicates that the difference in item nonresponse rates between the low effort and high effort respondents is reduced relative to the reference

category. A negative coefficient indicates that the difference in item nonresponse rates between the low and high effort respondents is increased relative to the reference category.

Six of the 15 studies (40%) report item-level item nonresponse rates for the same items across multiple level of effort measures (e.g., for both number of follow-up attempts and refusal conversion). We account for this repetition through the clustered analysis and by conducting analyses within each level of effort measure.

IV. Findings

We present our findings in three steps. First, we summarize findings from the 178 questions for which question-level item nonresponse rates are presented. This covers 15 of the 16 articles; we exclude one study of item nonresponse rates across waves of follow-up for the U.S. Census as the large sample size leads to analyses in which this study dominates all analyses (Treat and Stackhouse 2002). Second, we examine the aggregate findings on item nonresponse. Finally, we examine item level and aggregate findings for other measurement error indicators.

IV.A. Overall

Overall, the log-odds ratio, across all levels of effort and all items, is -0.861 (SE=0.208) for wave 1 to wave 2 and -0.90 (SE=0.192) for wave 1 to the last wave. This indicates that respondents recruited with more effort have higher item nonresponse rates than those recruited easily on these items. There is variation in the size of the log-odds ratio across items within the same study and across studies. We now turn to examining study and item characteristics that may account for some of this variation.

IV.B. Level of Effort Characteristics

As expected, refusal conversion studies or combination studies show significantly larger differences in item nonresponse rates between those easy to recruit and more difficult to recruit than studies using follow-up calls. The effect size for studies using follow-ups as the level of effort is -0.305 ($p<.01$), but is -0.974 ($p<.01$) for refusal conversion studies, indicating that item nonresponse rates among late respondents is much greater for refusal conversion studies than other studies (Table 1). There appears to be no statistical difference in item nonresponse rate

differences between studies that use date of interview and follow-up measures as the level of effort.

Refusal conversion studies consist of two separate types. The first type of refusal conversion study looks at general refusal conversion indicators from call records. The second defines resistance not by whether the individual was recorded in the call records as a refusal, but by the types of statements made on the doorstep – primarily statements about being not interested in the survey topic, too busy, or having privacy concerns. Looking only among refusal conversion studies, there are striking differences between the "not interested" and the "general refusal conversion" studies. While the effect size for the general refusal conversion studies was -0.723, indicating higher odds of item nonresponse for reluctant respondents than non-refusers, the "not interested" group had even higher odds of item nonresponse (-1.761). As such, we will examine these two groups separately. All of the studies with "too busy" statements or "privacy" concerns are captured in either the general refusal conversion or in the "not interested."

IV. C. Survey Characteristics

We now examine four characteristics of the survey – the overall response rate, the primary mode, whether a mode switch was used, and the topic of the survey. We examine each characteristic individually due to the small number of surveys available for examination. It is likely that interaction effects hold between these study characteristics on the relationship between the error sources. Only two studies report item-level item nonresponse rates by the date of the interview. As such, no study-level characteristics can be examined here. As with date of interview, there are only two studies that contain item-level item nonresponse rates and look at a combination of number of calls and refusal conversion, so we cannot disentangle characteristics of the studies from the articles themselves.

IV.C.1 Response Rates

There is no clear relationship between the survey's unit response rate and the relative difference in item nonresponse rates for low versus high recruitment effort respondents. This non-significant relationship also holds when looking only at studies that defined effort by the number of follow-up attempts and when looking at wave 1 to wave 2 or at wave 1 to the last wave (Table 2 and Table 3). That is, although additional follow-up attempts resulted in higher

response rates, the final unit response rate was not related to the relative difference in item nonresponse rates for low versus high recruitment effort respondents.

The overall unit response rate is related to the relative difference in item nonresponse rates between the refusal converted respondents and respondents who did not need refusal conversion, defined either generally or by statements of "not interested" on the doorstep. However, the direction varies across these two types of refusal conversion studies and the range of the unit response rates is limited for both types of studies. In the general refusal conversion studies, the higher the unit response rate, the higher the item nonresponse rate among the refusal converted relative to the non-refusal converted. For the general refusal conversion, the six studies have response rates ranging from 42 to 71 percent, but are primarily in the mid-60 percent range. Conversely, in the "not interested" studies, higher unit response rates are related to lower item nonresponse rates among the "not interested" relative to the "interested" group. In these studies, the unit response rates range from 61 to 72 percent. Due to the restricted range of the independent variable, we cannot draw definite conclusions about the effect of converting refusals and unit response rates on item nonresponse.

IV.C.2 Study Mode

The mode of the survey matters for the relative difference in item nonresponse rates between low and high recruitment effort respondents (Tables 2 and 3). Overall of the studies, the difference in item nonresponse rates between the two groups is larger for face to face studies (effect size =1.114, $p<.0001$) than for mail (effect size = -0.598) or telephone surveys (effect size = -0.549). The difference between low and high recruitment effort respondents varies substantially by how recruitment effort is defined. Studies that define recruitment effort as the number of follow-up attempts show little difference between the two groups in face to face studies, but, as expected, much larger differences are seen in mail surveys (effect size = -0.499, $p<.0001$) than in phone studies (effect size = -0.050, $p<.0001$) or face to face studies (effect size = -0.140, $p<.0001$). The studies that used refusal conversion to define recruitment effort are almost exclusively face to face studies; thus, no conclusion about the mode of study for studies of different modes of data collection.

IV.C.3 Mode Switch

Overall, there is no difference in relative item nonresponse rates for low and high effort respondents between the surveys that used a mode switch and those that did not use a mode switch (Table 2 and 3). This varies by the level of effort measure used. Mode switches were almost always examined as the final step in a set of follow-up efforts; none of the studies that looked (explicitly) at refusal conversion reported a switch in mode. The studies which look at follow-ups as the recruitment effort measure show a striking difference in item nonresponse rates between persons recruited with high and low effort; the higher the effort required in studies with a mode switch, the higher the item nonresponse rates among the later respondents (effect size = -0.681, $p<.0001$). The difference between studies with and without a mode switch is greater when examining the last recruitment wave (usually corresponding to the mode switch) than the wave 1 to wave 2 comparisons (wave 1 to last wave effect size = -1.304, $p<.0001$).

IV.C.4 Survey Topic

Due to the small number of surveys available for this analysis, the topic of the survey was grouped into three categories – health, employee and organizational member attitudes, and all other topics (Tables 2 and 3). In general, the difference in item nonresponse rates between high and low effort respondents was greatest in health surveys (-0.94, $p<.10$) compared to other non-health, non-employee attitudinal studies (-0.44, $p<.05$). This generally held overall, in follow-up attempt studies, and in "not interested" refusal conversion studies. No consistent difference was found between employee attitude studies and other topics.

IV.D Question Characteristics

We examine three characteristics of questions – the type of question, whether the question is burdensome and whether it is sensitive. We look at four types of questions – behaviors, attitudes, demographics, and income (Tables 4 and 5).

IV.D.1 Type of Question

Overall, the difference in item nonresponse rates for respondents recruited with high effort compared to those recruited with less effort varies by the type of question (Tables 4 and 5). Among the reported items, the relative difference in item nonresponse rates for low versus high recruitment effort respondents is greater for behavioral questions (-0.927, $p<.001$) than for

income questions (-0.478, p<.05). There is no statistical difference overall between behavioral, attitudinal, or other demographic questions.

The relative difference in item nonresponse rates across types of questions varies by the level of effort measure. No statistical difference was found among different types of question when looking at surveys that use numbers of follow-ups as the measure of level of effort. In contrast, the studies that look at general refusal conversion, the date of interview or a combination measure for the level of effort find differences for income questions relative to behavioral questions. The relative difference in item nonresponse rates for low vs. high effort respondents is greater for behavioral questions than for income questions in general refusal conversion studies and combination studies, but is in the direction of high effort respondents giving worse quality answers. In contrast, when effort is defined by the date of interview, later respondents have higher item nonresponse rates than earlier respondents to income questions (effect size=-0.193, p<.0001) but have lower item nonresponse rates to behavioral questions (effect size = 0.106, p<.0001). That is, although income questions are often identified as the most likely candidate for striking differences in item nonresponse rates between easy and difficult to recruit respondents, there is evidence from these studies that this is not necessarily the case.

IV.D.2 Question Burden

The relative difference in item nonresponse rates for burdensome questions are expected to be greater than for non-burdensome questions. However, this is not found to be the case. In fact, non-burdensome questions show no difference to larger differences between respondents recruited with low vs. higher levels of effort relative to questions identified as burdensome. Overall, the effect size is -.829 for non-burdensome questions and -0.611 for burdensome questions (p<.10). Among studies using “not interested” to define level of effort, the effect size for non-burdensome questions is -1.905 and is -1.494 for burdensome questions (p<.001), indicating higher item nonresponse rates among the “not interested” on non-burdensome questions than on burdensome questions. A similar pattern is found among studies that use a combination “late/difficult” measure.

IV.D.3 Question Sensitivity

Unlike question burden, there is no clear difference among questions varying in sensitivity overall. Among studies that use the date of interview or a combination measure, sensitive questions have smaller differences in item nonresponse rates between the two groups than non-sensitive questions (non-sensitive questions effect size = 0.092, $p<.0001$; sensitive questions effect size = -0.137, $p<.0001$).

IV.E Summary

Overall, item nonresponse rates are higher for respondents recruited with more effort than those with recruited with less effort. There is no clear relationship in relative item nonresponse rates for these two groups of respondents with unit nonresponse rates, but there is with study mode, mode switches, and topic of the survey. Among studies that use follow-ups as the measure of level of effort, mail surveys have greater differences between the high and low recruitment effort respondents than interviewer-administered surveys. Similarly, follow-up studies that use a mode switch have greater differences between the high and low recruitment effort respondents, especially when looking at the respondents who were recruited by the mode switch compared to others. Finally, health surveys tend to have greater differences in item nonresponse rates between high and recruitment effort respondents than other studies.

Question characteristics also matter, but in surprising ways. Although a motivational hypothesis would predict larger item nonresponse rates on difficult questions for the high recruitment effort respondents, the opposite was observed, whether difficulty was measured by burden or sensitivity. Also, income questions showed significantly smaller differences between the two groups than other types of questions, counter expectations from a motivational hypothesis.

V. Findings for aggregate measures

Now we examine three aggregate measures of item nonresponse and other forms of measurement error. Since the aggregate measures vary from study to study, we cannot use meta-analytic methods. The most common aggregate measure is the average respondent-level item nonresponse rates, in which the number of items not answered by each respondent is divided by the total number of items that each respondent was eligible to answer, and then averaged across all respondents. The second most common aggregate measure is the average count of

unanswered items for a constant set of questions asked of all respondents. The final measure is the proportion of respondents for whom at least K (usually K=1) items were unanswered. Other measurement error indicators that have been examined by level of recruitment effort include response inaccuracy, scale reliability, variability of answers given, including non-differentiation, acquiescence, extreme answers, inconsistency of answers across logical reports, and providing more than one answer to a question that only required one answer.

V.A Item nonresponse

Average respondent-level item nonresponse rates tended to be higher for those respondents who required more follow-ups, were converted refusals, who said that they were not interested in the survey, or who had lower estimated response propensity (Table 6). Most of the significant comparisons, however, were for "don't know" or "no opinion" answers in election studies. Both studies that looked at "don't know" or "no opinion" responses in election studies and number of follow-up attempts were statistically different from zero; none of the other comparisons met this significance level. Similarly, although all seven of the reported comparisons for converted refusals and "not interested" respondents were in the hypothesized direction of higher recruitment effort respondents having higher item nonresponse rates than lower recruitment effort respondents and statistically different from zero, two of these studies used the American National Election Study (Couper 1997; Miller and Wedeking 2006). Only two studies looked at aggregate measures of item nonresponse over the dates of the field period, with mixed results. Bilodeau (2006), looking at an Australian election study, indicated that item missing data rates increased slowly over the course of the field period, but Wellman, et al.'s (1980) study on use of Virginia outdoor areas found no significant relationship between item nonresponse to a set of agree-disagree questions or a "perceived problems index" and the date of the questionnaire return. No combination study looked at average item nonresponse rates across respondents. The two estimated response propensity studies (not in table) that examined average nonresponse rates and an estimated response propensity tended to show that lower propensity cases had higher item nonresponse rates (3 of the 5 comparisons) (Yan, Tourangeau, and Arens 2004; Fricker 2007), but the conclusions were sensitive to the response propensity model specification. Thus, while significant relationships were found in these studies, they appear to be primarily in election studies. It is difficult to tell whether this is because the examinations of

election studies looked at a focused set of questions (all about politics and elections) or focused on particular types of nonresponse (don't know or no opinion, rather than all item nonresponse combined).

Virtually no study found significant differences between high and low recruitment effort respondents in the mean count of unanswered items. The only studies that showed significant differences in the hypothesized direction on this measure used refusal conversion as the measure of level of effort, all of which were telephone surveys. Most of these studies examined entire questionnaires or large blocks of variables, a level of aggregation that may obscure differences across the two groups on individual items (Table 7).

Finally, when examining the proportion of persons with at least one missing answer on a set of items, most of the comparisons across all of the studies are in the hypothesized direction, but few are statistically different from zero (Table 8). This aggregate measure tends to be on more focused sets of items (e.g., a set of scale items, two related items). Many of the significant comparisons arise from election studies or studies about politics; variation over items is apparent from these studies as well.

V.B Other Measurement Error Indicators

Seven studies looked at differences in response accuracy between high and low recruitment effort respondents (Table 9). The studies covered topics such as hospital visits, academic performance, dental insurance eligibility, voting behavior, delinquency, and medication use. Since these are record check studies, all of the measures are behavioral. Here, findings vary dramatically by the type of effort measure used. In almost every instance when effort is measured by follow-up call attempts, respondents requiring additional follow-ups had higher levels of inaccurate reports than respondents to the first request, the difference was statistically different from zero (in two studies, significance levels are not reported), and when more than one wave of follow-up was conducted, the last wave has a larger difference than the second wave. However, across 14 comparisons made in two studies using refusal conversion as the effort measure, only one was statistically different from zero, and the differences in response accuracy were equally likely to be positive or negative. In the single epidemiological study that looked at the date of interview (Voigt et al. 2005), there is more measurement error for cases that are recruited later than earlier, but no difference for controls. In sum, there is evidence of less

accurate reporting among higher recruitment effort respondents, but it seems to vary by survey and by the level of effort measure examined.

Measurement error in attitudinal scales is measured by scale reliability and other covariance based measures (acquiescence, middle responses, extreme responses, double responses, and inconsistent preferences). No consistent picture emerges overall or a particular level of effort measure, either in terms of the direction of the difference between low and high recruitment effort respondents or the significance level (Table 10). The only striking differences were found on lack of knowledge of political persons and current affairs in the American National Election Studies among the "not interested" refusers (Couper 1997), but most of the comparisons for low and high effort respondents on these attitudinal measures are not statistically different from zero.

Finally, four studies looked at a function of the variation of responses over follow-up waves, sometimes labeled nondifferentiation (Table 11). Here, there is evidence that respondents requiring more recruitment effort have both more variable and less variable answers than those requiring less recruitment effort. From one perspective, this is an examination of the variance of responses over levels of effort and as such, is a question of nonresponse bias on a variance term. From another perspective, this is a measurement error of nondifferentiation where lower variability is indicative of more measurement error. In either case, there is evidence of changes in the variation of responses over recruitment efforts.

VI. Conclusion

This paper used two methods to review existing literature on the nonresponse and measurement error nexus. A meta-analytic review of the literature showed that there are higher item-level item nonresponse rates for respondents who require more recruitment effort than those brought into the respondent pool more easily. The meta-analysis also showed that this differed when looking at effort as measured by number of contact attempts compared to refusal conversion. When examining the effort as the number of contact attempts, there were bigger differences in item nonresponse rates between high and low effort respondents when the mode was mail or when there was a mode switch. Health surveys tended to show larger differences in item nonresponse rates than other types of surveys.

Question characteristics had surprising effects on the differences between high and low recruitment effort respondents. Although we anticipated income questions, burdensome questions, and more sensitive questions would have larger differences between the two groups of respondents, they were no different and often showed smaller differences than other types of questions. This seems to indicate that a difficult question is difficult for everyone, and not especially difficult for those who required more recruitment effort.

Meta-analytic techniques could not be used for aggregate measures of item nonresponse or other forms of measurement error as the outcome variables were many and varied. In general, there was mixed evidence for worse data quality among higher recruitment effort respondents. When such evidence was found, it tended to be for measures that were more precise (i.e., for a single question, for a block of related questions) rather than for an entire general questionnaire and for behavioral items rather than attitudinal items, although not all studies that met this criteria found significant differences between the two groups. This strongly suggests an item-specific nature of the relationship between nonresponse and measurement errors. Interestingly, although refusal conversion studies of item-level item nonresponse had larger differences than follow-up studies, they were not particularly more likely to show lower data quality than other types of studies.

Examinations of election studies often found differences between high and low recruitment effort respondents, especially on item nonresponse measures. One potential explanation for this is that the recruitment request for an election study is fairly indicative of the task at hand. As with mail surveys, where the entire questionnaire can be reviewed as part of the participation decision, one can speculate that respondents may have a better sense of the questionnaire in an election study than other studies. That is, respondents to an election study can be confident that they will be asked about voting behavior and political attitudes, even before the interview begins. As such, those who are less likely to be interested in answering those types of questions or who have little knowledge about politics or voting may be reluctant to participate.

In the 15 meta-analyzed articles, it was rare for the authors to report item nonresponse rates for items that did not follow the hypothesized outcome in which wave 1 was lower than wave 2, regardless of the type of follow-up procedure studied. This appears to be selection bias in the reported items towards reporting items those with higher item nonresponse rates and reporting items where a significant difference appears between the two groups. The studies that

looked at mean count of unanswered items or item nonresponse rates tended to look at larger numbers of questions or the entire questionnaire and show smaller or no differences between the two groups of respondents. This item-level selection bias could account for the discrepant findings between the two types of studies. This also suggests that, had the studies included in the meta-analysis reported all items, not the more limited subset found here, the overall effect size may be attenuated. This is a file drawer problem common to many meta-analyses (Lipsey and Wilson 2001). Here, the file drawer problem is not with lack of representation of overall studies in the meta-analysis, but with items from the studies included in the meta-analysis.

A careful examination of the causes of both survey participation and measurement error is needed to further understand why variation occurs across studies and across items within studies. Each article reviewed posited a mechanism for the relationship between level of effort or response propensity and measurement error. These mechanisms can be reduced to three "ideal types" (Olson 2007), summarized in Figure 1.

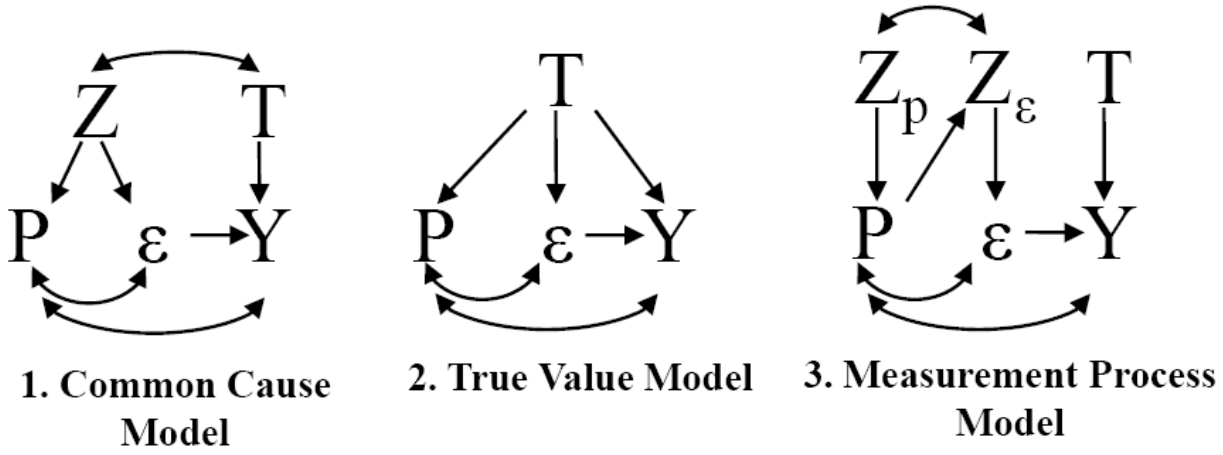

Figure 1: Three "ideal type" models for the relationship between nonresponse and measurement errors

In Figure 1, *P* refers to the likelihood of survey participation, or response propensity. This could be operationalized by any of the level of effort measures examined here (e.g., number of follow-up attempts, refusal conversion) or by an estimated response propensity. ε refers to measurement error. Here, ε was operationalized by item-level item nonresponse and other aggregate measures. *T* is the "true" value for a particular survey question (e.g., the sample member's true income) and *Y* is the report for that survey question (e.g., their reported income). The arrows running from *T* and ε to *Y* indicate that the survey report is a function of the true

value and measurement error. Z_p is a "cause" of response propensity; Z_ε is a cause of measurement error. We are interested in explaining the arrow between P and ε, that is, the relationship between response propensity and measurement error.

The first "ideal type" model is a Common Cause model. This model indicates that the cause of survey participation and of measurement error is the same (Z without a subscript). The most commonly posited Common Cause model is a motivational model. Here, Z is motivation, which decreases the likelihood of survey participation, increases the likelihood of poor data quality, and hence, leads to a relationship between survey participation and data quality. Other common causes between the error sources could include the interest in the topic (Martin 1994), relationship with the sponsor (Tourangeau, et al., 2008), or other design features. Respondent background characteristics such as age or education (as proxies for cognitive ability, motivation, or other factors) may also be common causes. When this model holds, survey variables that are the most related to the common cause should experience the greatest hit on data quality.

The second "ideal type" is the True Value model, a special case of the Common Cause model. In this model, the common cause is the "true" value on the survey variable itself. For example, voters are more likely to participate in an election survey than nonvoters, and are better reporters of voting status (Voogt 2004). Similarly, students with high GPAs are more likely to participate in surveys of students and are better reporters of their GPA than low GPA students (Olson 2007). This model implies that both nonresponse bias and measurement error on analyses using a survey variable Y (e.g., mean of Y) will change with additional recruitment efforts, as persons who are less likely to participate have different Y values and different ε values than more easily recruited individuals.

The final model is the Measurement Process Model. Here, low response propensity leads to a change in the recruitment and measurement protocol, as indicated by the arrow running from P to Z_ε. This could be a mode switch, a change in interviewers, or an increase in incentive levels, for example. The distinction between this ideal type and the other two is important because the relationship between level of effort or response propensity and measurement error is induced by decisions made by the survey organization.

These models imply that the relationship between survey participation and measurement error should be item specific, will vary by the causes of survey participation, and thus will vary across noncontact and noncooperation nonresponse, and will vary according to the measure of ε.

That is, item nonresponse should be different from response accuracy in its relationship with level of effort or response propensity to the extent that the causes of item nonresponse are different from those of response accuracy.

The meta-analysis yields some evidence for the Measurement Process model. Follow-up studies that use mode switches have larger differences in item nonresponse rates between low and high effort respondents than studies that do not use mode switches. The other models cannot be tested in this review as most studies only provide the P-ε relationship overall, and not for particular subgroups. Future research in this area would benefit from explicit testing of at least one of these models for different types of items and different measures of measurement error.

Table 1: Beta Coefficients for Simple Weighted Regression Models Predicting Log-Odds of Item Nonresponse in Wave 2 compared to Wave 1 , All studies, Level of Effort Measures and Type of Refusal Conversion Measure

	Wave 1 to Wave 2		Wave 1 to Last Wave		
	Beta	SE	Beta	SE	# Items
Level of Effort Measure					
Intercept (Follow-ups/calls = Reference)	-0.305**	0.093	-0.430**	0.159	59
Refusal Conversion	-0.669**	0.248	-0.550*	0.270	96
Date of Interview	0.052	0.075	0.050	0.121	9
Combination of Above	-0.515****	0.100	-0.386*	0.163	14
Type of Refusal Conversion					
Intercept (Any RC)	-0.723**	0.227	-0.722**	0.230	55
Not Interested	-0.939****	0.113	-0.939****	0.115	15
Too Busy	-0.189	0.121	-0.190	0.123	14
Privacy	-0.129	0.114	-0.130	0.116	12

Table 2: Beta Coefficients for Simple Weighted Regression Models Predicting Log-Odds of Item Nonresponse in Wave 2 compared to Wave 1 , Unit Response Rate, Study Mode, Mode Switch, and Study Topic

	Overall		Number of follow-ups		General Refusal Conversion		Not Interested	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Unit Response Rate								
Intercept	-0.336	0.486	-0.968	0.392	0.429	0.329	-8.758****	0.690
Response Rate	-0.005	0.006	0.011	0.007	-0.016*	0.007	0.107****	0.011
Study Mode								
Intercept	-1.114****	0.184	-0.050****	0.00	-0.725****	0.153	-1.741****	0.000
Mail	0.516*	0.213	-0.449****	0.095	n/a	n/a	n/a	
Phone	0.565*	0.251	-0.090****	0.018	0.323+	0.196	n/a	
Mode Switch								
Intercept	-0.682*	0.274	-0.249**	0.087	-0.478****	0.133	-1.741****	0.000
Mode Switch	0.067	0.284	-0.432****	0.087	n/a		n/a	
Topic								
Intercept (reference = all other topics)	-0.440*	0.198	-0.071	0.053	-0.59	0.373	-1.139****	0.100
Health	-0.500+	0.303	-0.453****	0.053	0.035	0.416	-0.602****	0.100
Employee/ Organization Attitudes	0.179	0.214	-0.207+	0.108	0.318	0.373	n/a	

Note: +p<.10, *p<.05, **p<.01, ***p<.001, ****p<.0001; n/a indicates no variation on this attribute for these studies.

Table 3: Beta Coefficients for Simple Weighted Regression Models Predicting Log-Odds of Item Nonresponse in Last Wave compared to Wave 1 , Unit Response Rate, Study Mode, Mode Switch, and Study Topic

	Overall		Number of follow-ups		General Refusal Conversion	
	Beta	SE	Beta	SE	Beta	SE
Overall Response Rate						
Intercept	-0.243	0.503	-0.406	0.813	0.423	0.338
Unit Response Rate	-0.008	0.006	0.000	0.014	-0.016*	0.008
Study Mode						
Intercept (reference=face to face)	-1.190****	0.111	0.026	0	-0.724****	0.161
Mail	0.391	0.245	-0.614**	0.212	n/a	n/a
Phone	0.623**	0.2	-0.259****	0.013	0.322	0.202
Mode Switch						
Intercept (reference = No Mode Switch)	-0.715**	0.267	-0.286***	0.079	-0.473****	0
Mode Switch	-0.322	0.33	-1.018****	0.079		
Topic						
Intercept (reference = all other topics)	-0.506*	0.217	-0.064	0.103	-0.564	0.433
Health	-0.44	0.316	-0.460****	0.103	0.01	0.47
Employee/Organization Attitudes	0.088	0.308	-0.406	0.278	0.293	0.433

Note: +p<.10, *p<.05, **p<.01, ***p<.001, ****p<.0001; n/a indicates no variation on this attribute for these studies.

Table 4: Beta Coefficients for Simple Weighted Regression Models Predicting Log-Odds of Item Nonresponse in Wave 1 compared to Wave 2 , Type of question, Question Burden, and Question Sensitivity

	Overall		Number of follow-ups		General Refusal Conversion		Not Interested		Days in the field		Combination	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Type of Question												
Intercept												
(Behavior =reference)	-0.927***	0.262	-0.356*	0.136	-0.724****	0.147	-1.742	--	0.106****	0.012	-1.063****	0.062
Attitude	0.408	0.272	-0.034	0.105	0.300	0.269	0.687	--				
Other Demographics	0.080	0.294			0.012	0.352						
Income	0.449*	0.207	0.152	0.164	0.270+	0.155	0.158	0.173	-0.299****	0.040	0.437****	0.085
Question Burden												
Intercept	-0.829*	0.376	-0.258***	0.093	-0.525*	0.233	-1.905****	0.053	-0.004	0.016	-1.156****	0.017
Burdensome Item	0.218+	0.133	-0.122	0.147	0.083	0.166	0.411***	0.122	-0.011	0.021	0.384****	0.068
Question Sensitivity												
Intercept	-0.782*	0.334	-0.266*	0.126	-0.529*	0.234	-1.833****	0.078	0.092****	0.013	-0.666****	0.038
Sensitive Item	0.129	0.092	-0.062	0.124	0.054	0.125	0.172	0.145	-0.229****	0.033	0.149***	0.044

Note: +p<.10, *p<.05, **p<.01, ***p<.001, ****p<.0001; n/a indicates no variation on this attribute for these studies.

Table 5: Beta Coefficients for Simple Weighted Regression Models Predicting Log-Odds of Item Nonresponse in Wave 1 compared to Last Wave , Type of question, Question Burden, and Question Sensitivity

	Overall		Number of follow-ups		General Refusal Conversion		Not Interested		Days in the field		Combination	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE	Beta	SE
Type of Question												
Intercept							same as wave 1 to				same as wave 1 to	
(Behavior =reference)	-0.941***	0.256	-0.354*	0.169	-0.726****	0.148	wave 2		-0.002	0.005	wave 2	
Attitude	0.254	0.322	-0.369***	0.111	0.303	0.271						
Other Demographics	-0.147	0.384			0.018	0.353						
Income	0.399*	0.196	-0.021	0.14	0.286+	0.167			-0.320****	0.016		
Question Burden												
Intercept	-0.882*	0.364	-0.430*	0.212	-0.521*	0.233			0.259****	0.034		
Burdensome Item	0.223+	0.137	0.024	0.265	0.088	0.163			-0.456****	0.042		
Question Sensitivity												
Intercept	-0.854**	0.314	-0.46	0.293	-0.524*	0.236			0.006	0.009		
Sensitive Item	0.151+	0.086	0.058	0.26	0.055	0.124			-0.281****	0.021		

Note: +p<.10, *p<.05, **p<.01, ***p<.001, ****p<.0001; n/a indicates no variation on this attribute for these studies.

Table 6: Aggregate Item Nonresponse Studies, Average Respondent-level Item Nonresponse Rates

		# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-ups or Calls				
(De Leeuw and Hox 1988)	Overall	Not avail.	(no #) ns	(no #) ns
(Gilbert, Longmate, and Branch 1992)	Overall	46	-4#	-3#
(Helasoja et al. 2002)	Overall, Finland, Men	40	-0.96 ns	
	Overall, Latvia, Men	40	-2.35 ns	
	Overall, Lithuania, Men	40	-3.13 ns	
	Overall, Finland, Women	40	-1.39 ns	
	Overall, Latvia, Women	40	-2.46 ns	
	Overall, Lithuania, Women	40	-1.97 ns	
(Jobber, Allen, and Oakland 1985)	Overall	~50	-2.7 ns	
	Overall	~50	0.8 ns	
	Overall	~50	-1.4 ns	
(Kaminska, Goeminne, and Swyngedouw 2006)	Overall, Don't Know Responses	23	5.1***	-8***
	Overall, Other Missing Responses	55	-3.4 ns	-5.5 ns
(Miller and Wedeking 2006)	7 items, No opinion responses, 2000	7	-3.1*	
	7 items, No opinion responses, 1996	7	(no #) ns	
(Tancreto and Bentley 2005)	Overall, Household respondents	6	-0.1 ns	-1.3****
	Overall, Proxy respondents	6	-2.1 ns	-4.8****
Refusal Conversion				
(Couper 1997)	Political Attitudes, Not Interested	89	-3.7**	
	Nonpolitical Attitudes, Not Interested	10	-1.7**	
	Background items, Not Interested	9	-1.1**	
(Miller and Wedeking 2006)	7 items, No opinion responses, 2000	7	-6.6*	
	7 items, No opinion responses, 1996	7	-3.8**	
(Smith 1983)	Family Income, Refusal	1	(No #)*	
	Overall, "Don't Know"	Not avail.	(No #) ns	
Date of Interview				
(Bilodeau 2006)	20 questions about leaders, issues and other political considerations	20	-0.5 #	
(Wellman et al. 1980)	Agree-disagree battery	Not avail. ('lengthy')	(no #) ns	
	Perceived problem battery	Not avail. ('lengthy')	(no #) ns	

Note: ns not significant, # Implied significance, test or level not stated, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort.

Table 7: Aggregate Item Nonresponse Studies, Average or Median Count of Item Nonresponse Answers

		# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-up attempts				
(Jobber, Allen, and Oakland 1985)	Average number of missing items	~50	-1.27 ns	
	Average number of missing items	~50	0.39 ns	
	Average number of missing items	~50	-0.69 ns	
(Kennickell 1999)	Median number of missing items among dollar variables, area probability sample	Not avail.	0 ns	
	Median number of missing items among dollar variables, list sample	Not avail.	-1 ns	
(Schmidt et al. 2005)	Average number missing items among 75 items, Study 1	75	0.1 ns	
	Average number completed items, Study 2	Not avail.	-0.7 ns	
Refusal Conversion				
(Currivan 2005)	Mean number of don't know answers in questionnaire	Not avail.	0.42*	
	Mean number of refusal answers in questionnaire	Not avail.	-0.18*	
(Keeter et al. 2000)	Average number of missing answers out of 89 questions in questionnaire	89	-0.2+	-0.7*
	Median item-level absolute difference in item nonresponse rates across all items (range -4.2 to +3.9; Median 0.9)	85	0.9	
(Tripplett et al. 1996)	Mean number of missing items among 38 diary questions, adult sample	38	-0.06 ns	-1.32**
(Yan, Tourangeau, and Arens 2004)	Mean number of no opinion answers among 10 questions on people's attitudes towards different science and technology developments	10	-0.01*	
Combined Measure				
(Kennickell 1999)	Median number of missing items among dollar variables, area probability sample	Not avail.	0 ns	
	Median number of missing items among dollar variables, list sample	Not avail.	-1 ns	
(Keeter et al. 2000)	Mean number of missing answers out of 89 questions in questionnaire	89	0.3 ns	0 ns

Note: ns not significant, # Implied significance, test or level not stated, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort.

Table 8: Aggregate Item Nonresponse Studies, Proportion with at least one missing answer on at least two items

		# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-up attempts				
(Donald 1960)	% respondents with item nonresponse on either of two questions in participation index	2	0 ns	-3 ns
(Diaz de Rada 2005)	% respondents with at least one question with item nonresponse in survey	24	-7.7 ns	-13.7 ns
(Korkeila et al. 2001)	% Four or more missing items in the 21-item Beck Depression Inventory	21	-0.8*	
	% Two or more missing items in 7 questions on alcohol use	7	0 ns	
	% One or more missing items in 3 questions on sex life	3	-0.5*	
	% One or more missing items in 3 questions on Hostility	3	0 ns	
(Newman 1962)	% 1+ Missing Own or Expenditures on Hunting Equipment	2	-5 ns	
	% 1+ Missing Own or Expenditures on Still Camera	2	-9.4*	
	% 1+ Missing Own or Expenditures on Movie Camera	2	-1.7 ns	
	% 1+ Missing Own or Expenditures on Movie Projector	2	-9.9*	
	% 1+ Missing Own or Expenditures on Slide Projector	2	-6.9 ns	
	% 1+ Missing Expenditure for Membership in a Country Club	2	-7.6 ns	
	% 1+ Missing Own or Expenditure for Membership at a Boat or Yacht Club	2	-8.9 ns	
	% 1+ Missing Own or Expenditures on Pleasure Boat	2	-5.6 ns	
	% 1+ Missing Own or Expenditures on Fishing Equipment	2	-7.4 ns	
	% 1+ Missing Own or Expenditures on Golf Equipment	2	-11.4*	
Refusal Conversion				
(Blair and Chun 1992)	Percent of questions with statistically significant differences, don't know answers	Not avail.	11% to 36%*	
	Percent of questions with statistically significant differences, refusal answers	Not avail.	Not avail.	
(Campanelli, Sturgis, and Purdon 1997)	% at least 4 Don't know answers, Not Interested, Political Tracking Study	Not avail ('all qn's')	-18*	
	% at least 4 Don't know answers, Too Busy, Political Tracking Study	Not avail ('all qn's')	-1 ns	
	% at least 1 Not answered answers, Not interested, Political Tracking Study	Not avail ('all qn's')	-3 ns	
	% at least 1 Not answered answers, Too busy, Political Tracking Study	Not avail ('all qn's')	-2 ns	
	% at least 1 No opinion answers, Not interested, Political Tracking Study	Not avail ('all qn's')	13***	
	% at least 1 No opinion answers, Too busy, Political Tracking Study	Not avail ('all qn's')	2 ns	
	% at least 1 Don't know answers, Not Interested, FRS Study	Not avail ('all qn's')	-13 ns	
	% at least 1 Don't know answers, Too Busy, FRS Study	Not avail ('all qn's')	-5 ns	
	% at least 1 Not answered answers, Not interested, FRS Study	Not avail ('all qn's')	-3 ns	
	% at least 1 Not answered answers, Too busy, FRS Study	Not avail ('all qn's')	0 ns	
(Couper 1997)	% Can't judge/Don't know, 1+ feeling thermometer items, Not Interested	18	-20**	

Note: ns not significant, # Implied significance, test or level not stated, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort.

Table 9: Response Accuracy Studies Using Follow-ups or Number of Contacts, Refusal Conversion, and Date of Interview as Level of Effort

Study	Question Text	# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-up attempts				
(Cannell and Fowler 1963)	% inaccurate reports, length of hospital stay	1	-7#	-27#
	% inaccurate reports, hospital visits, overall	1	-2#	-19#
	% inaccurate reports, hospital visits, Less than 4 years of high school	1	(no #)	-8#
	% inaccurate reports, hospital visits, 4 years of high school+	1	(no #)	-13#
(Eckland 1965)	% inaccurate reports, Father's occupation	1	-4.7#	-6.9#
	% inaccurate reports, City/farm residence	1	0 ns	1.9#
	% inaccurate reports, Academic failure	1	-10.4#	0.4#
	% inaccurate reports, Earned degree	1	-4 ns	-0.6 ns
(Gilbert, Longmate, and Branch 1992)	% inaccurate reports, VA dental eligibility	1	0.5**	-3.9**
(Robins 1963)	% inaccurate reports, have adult nontraffic arrests	1	-18*	
	% inaccurate reports, ever divorced	1	-6 ns	
	% inaccurate reports, have problem spouses	1	-5 ns	
	% inaccurate reports, did not attend high school	1	7 ns	
	% inaccurate reports, truanted as children	1	-1 ns	
(Voogt and Saris 2005)	% inaccurate reports, Dutch voting in 1998 national elections, Long Q'naire, Voters, Listed Numbers	1	-3.5 ns	-24**
	% inaccurate reports, Dutch voting in 1998 national elections, Long Q'naire, Voters, Unlisted numbers	1	-34.2**	
	% inaccurate reports, Dutch voting in 1998 national elections, Long Q'naire, Non-voters, Listed Numbers	1	-1.1 ns	-22.1**
	% inaccurate reports, Dutch voting in 1998 national elections, Long Q'naire, Non-voters, Unlisted numbers	1	-20.9**	
Refusal Conversion				
(Olson and Kennedy 2006)	Difference in absolute signed deviations, Received D or F	1	2 ns	
	Difference in absolute signed deviations, Recent graduate	1	-1 ns	
	Difference in absolute signed deviations, Received W	1	6 ns	
	Difference in absolute signed deviations, A Grades	1	-2 ns	
	Difference in absolute signed deviations, B Grades	1	2 ns	
	Difference in absolute signed deviations, C Grades	1	-1 ns	
	Difference in absolute signed deviations, Graduate with honor	1	1 ns	
	Difference in absolute signed deviations, Alumni member	1	1 ns	
	Difference in absolute signed deviations, Ever donate to UMD	1	-7 ns	
(Robins 1963)	% inaccurate reports, have adult nontraffic arrests	1	40*	
	% inaccurate reports, ever divorced	1	-1 ns	
	% inaccurate reports, have problem spouses	1	3 ns	
	% inaccurate reports, did not attend high school	1	-17 ns	

Study	Question Text	# items	Wave 1 to Wave 2	Wave 1 to Last Wave
	% inaccurate reports, truanted as children	1	-5 ns	
Date of Interview				
(Voigt et al. 2005)	False positive rate - use of antihypertensive medication, Cases	1	-8*	
	False positive rate - use of antihypertensive medication, Controls	1	-3 ns	
	False negative rate - use of antihypertensive medication, Cases	1	-10*	
	False negative rate - use of antihypertensive medication, Controls	1	-4 ns	

Note: ns not significant, # Implied significance, test or level not stated, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort

Table 10: Aggregate Measures of Measurement Error, Other Measures for Attitudinal Items

Study	Question Text	# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-ups				
(Green 1991)	Scale unreliability (1-alpha), Review of research literature	5	0.05 ns	-0.02 ns
	Scale unreliability (1-alpha), Conduct of research	3	0.04 ns	0.05 ns
	Scale unreliability (1-alpha), Presentation of research at a professional meeting	3	-0.01 ns	0.01 ns
	Scale unreliability (1-alpha), Attitude toward use of research	23	-0.08*	0.09 ns
(De Leeuw and Hox 1988)	Scale unreliability, respondent role anticipation	3	(no #) ns	
(Yan, Tourangeau, and Arens 2004)	Acquiescence, 10 attitude qs towards science and technology	10	-0.02*	
	Middle responses, 10 attitude qs towards science and technology	10	0.00 ns	
	Extreme responses, 10 attitude qs towards science and technology	10	-0.01 ns	
(Armenakis and Lett 1982)	Circular preferences, pharmacy patrons to 10 items, University alumni/university sponsorship	10	2.6 ns	
	Circular preferences, pharmacy patrons to 10 items, University alumni/consulting sponsorship	10	-4 ns	
	Circular preferences, pharmacy patrons to 10 items, General public/university sponsorship	10	-7.4 ns	
	Circular preferences, pharmacy patrons to 10 items, General public/consulting sponsorship	10	-10.9*	
(Diaz de Rada 2005)	% respondents with two responses to questions of one response	24	-2.2 ns	-9.9 ns
(Kaminska, Goeminne, and Swyngedouw 2006)	Combined measure of straight-lining, agreement, and extreme or middle responses	30	0.13*	-0.41*
Refusal Conversion				
(Blair and Chun 1992)	Recency effects	Not avail.	Not avail. Ns	
	Primacy effects	Not avail.	Not avail. Ns	
(Couper 1997)	Don't recognize one or more feeling thermometer names of 6, Not Interested	6	-22.1**	
	Haven't thought much about current affairs for one or more items among 4 items, Not Interested	4	-19.1**	
(Yan, Tourangeau, and Arens 2004)	Acquiescence, 10 attitude qs towards science and technology	10	0.00 ns	
	Middle responses, 10 attitude qs towards science and technology	10	-0.00 ns	
	Extreme responses, 10 attitude qs towards science and technology	10	-0.03*	
Date of Interview				
(Yan, Tourangeau, and Arens 2004)	Acquiescence, 10 attitude qs towards science and technology	10	-0.01 ns	
	Middle responses, 10 attitude qs towards science and technology	10	0.01 ns	
	Extreme responses, 10 attitude qs towards science and technology	10	-0.02 ns	

Note: ns not significant, # no significance levels, significance implied, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort

Table 11: Aggregate Measurement Error Studies, Variation in Responses

Study	Question Text	# items	Wave 1 to Wave 2	Wave 1 to Last Wave
Follow-up attempts				
(Donald 1960)	Standard Deviation, influence of local president	1	-0.007*	-0.231*
	Standard Deviation, influence of local board as a group	1	-0.073*	-0.117*
	Standard Deviation, influence of local membership as a whole	1	0.021*	-0.116*
(Green 1991)	Standard Deviation, Review of research literature	5	0.05*	
	Standard Deviation, Conduct of research	3	0.10 ns	
	Standard Deviation, Presentation of research at a professional meeting	3	-0.03+	
	Standard Deviation, Attitude toward use of research	23	-0.07*	
(Yan, Tourangeau, and Arens 2004)	Non-differentiation, 10 attitude qs towards science and technology	10	0.02*	
(Miller and Wedeking 2006)	Non-differentiation, 7 items, political attitudes, 2000	7	-0.18*	
	Non-differentiation, 7 items, political attitudes, 1996	7	(no #) ns	
Refusal Conversion				
(Yan, Tourangeau, and Arens 2004)	Non-differentiation, 10 attitude qs towards science and technology	10	0.01 ns	
(Miller and Wedeking 2006)	Non-differentiation, 7 items, political attitudes, 2000	7	-0.08 ns	
	Non-differentiation, 7 items, political attitudes, 1996	7	-0.18+	
Date of Interview				
(Yan, Tourangeau, and Arens 2004)	Non-differentiation, 10 attitude qs towards science and technology	10	0.01 ns	

Note: ns not significant, # no significance levels, significance implied, *p<.05, **p<.01, ***p<.001, ****p<.0001. Negative implies higher levels of measurement error among respondents requiring greater levels of effort

Appendix A: Coding Forms

Study Characteristics

Author
Year
Type of survey
(Cross-sectional, Panel, Panel with fresh cross-section sample added)
Level of Measurement Error Calculation
(Aggregate Statistic, Respondent)
Measurement Error indicators
(code up to 2; Item Nonresponse (don't know, no opinion), Response Accuracy (Record – Report = 0), Signed Deviations (Record – Report), Absolute Deviations (|Record – Report|), Scale reliability (coefficient alpha), Variation in responses (e.g., standard deviation of responses), Acquiescence / yea saying, Straight-lining, Middle responses, Inconsistent answers across related questions, Other, Specify)
Definition of Waves, Phases, or Propensity
(code up to 2, Mode Switch, Number of Calls, Refusal Conversion, Combination of Number of Calls and Refusal Conversion, Reminder letters/Follow-up visits, Number of Days in the Field, Date of Interview, Propensity models (Logistic models, Probit models), Experimental comparison of recruitment protocol features, Other (specify))
Words used to describe measurement error
Words used to describe response propensity
Sequence of nonresponse and measurement error in time
(Measurement error occurred before nonresponse (e.g., item nonresponse predicts panel attrition); Nonresponse occurred before measurement error)
Topic of the survey
(code main and secondary; Health (illness, hospitalization); Employee Attitudes, Organization Membership Attitudes; Income, Transfer Payments or Other Financial Information; Crime; Drugs or Alcohol; Demographic characteristics; Transportation; Voting or Elections; Time Use; Contingent Valuation; Other)
Sample
(General population (RDD, area probability); List; Previous survey (panel study))
Population for the survey
(General adults; Adult patients, HMO members, or other users of medical care; Health care personnel; Employees at other organizations; Teachers; Students; Customers; Organizations/ businesses; Respondents to previous study (panel), Other)
Sponsor
(Government, Academic Organization, Business)
Data Collection Organization
(Government; Academic Organization; Private Firm, Government contractor ; Private Firm, Market Research; Other, Specify)
Country
(United States; United Kingdom / England / Great Britain; Canada; Australia; Germany; Other Europe; Other (specify))
Mode 1st Wave
(Code up to 6 waves, Telephone, Face to Face, Mail, Web)
Additional waves?
(Yes, No)
Total Sample Size
ineligible
Known Eligible units
Respondents
Overall Unit Response Rate
complete first – sixth wave
remaining first – sixth wave
First – Sixth wave RR

Characteristics of the items

Question text
Response Options
Number of response options
Type of Question
    (Attitude, Behavior, Knowledge, Demographics, Income)
Type of statistic
    (Mean, Proportion, Median, Correlation, Regression Coefficient, Odds ratio or relative risk, Other (specify))
Open-ended vs. Closed-ended
    (Open ended; Closed ended, Forced Choice; Closed ended, check all that apply)
Proxy reports allowed?
    (Yes, No)
Records available?
    (Yes, No)
Sensitive?
    (Sensitive, Not sensitive)
Socially desirable?
    (Socially desirable, Socially undesirable, Not socially (un)desirable)
Burdensome?
    (Burdensome, Not burdensome)
Difficult retrieval?
    (Difficult Retrieval, Not difficult retrieval)
Subgroup for item
    (Overall, Men, Women, Interested, Disinterested, Other (specify))
Number of item respondents
Number of item nonrespondents
Type of item nonresponse
    (Don't know, Refusal, No opinion, Combined)
Number of items
Number of item respondents first – sixth wave
Number of item nonrespondents first –sixth wave
Item NR Rate first – sixth wave
Other ME (specify type)
    (Item Nonresponse (refusal, don't know, no opinion, combined), Response Accuracy (Record – Report = 0), Signed Deviations (Record – Report), Absolute Deviations (|Record – Report|), Scale reliability (coefficient alpha), Variation in responses (e.g., standard deviation of responses), Acquiescence / yea saying, Straight-lining, Middle responses, Inconsistent answers across related questions, Other, specify)
Other ME, specify
Other ME first-sixth Wave
Standard deviation ME first-sixth Wave
Standard error ME first-sixth Wave
Number of items in entire questionnaire
Number of items used in Nonresponse and Measurement Error analysis

Articles used in the Meta-Analysis

Campanelli, Pamela, Patrick Sturgis, and Susan Purdon. 1997. Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates. London: Survey Methods Centre at SCPR.

Chiu, Pei-Lu, Howard Riddick, and Ann M. Hardy. 2001. A Comparison of Characteristics between Late/Difficult and non-Late/Difficult Interviews in the National Health Interview Survey. Paper read at Proceedings of the American Statistical Association.

Couper, Mick P. 1997. Survey Introductions and Data Quality. *Public Opinion Quarterly* 61 (2):317-338.

Dahlhamer, James M., Catherine M. Simile, and Beth Taylor. 2006. Exploring the Impact of Participant Reluctance on Data Quality in the National Health Interview Survey (NHIS). Paper read at Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health.

Donald, Marjorie N. . 1960. Implications of Nonresponse for the Interpretation of Mail Questionnaire Data. *Public Opinion Quarterly* 24 (1):99-114.

Korkeila, K., S. Suominen, J. Ahvenainen, A. Ojanlatva, P. Rautava, H. Helenius, and M. Koskenvuo. 2001. Non-response and related factors in a nation-wide health survey. *European Journal of Epidemiology* 17:991-999.

Mason, Robert, Virginia Lesser, and Michael W. Traugott. 2002. Effect of Item Nonresponse on Nonresponse Error and Inference. In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York: John Wiley and Sons.

Newman, Sheldon W. 1962. Differences between early and late respondents to a mailed survey. *Journal of Advertising Research* 2:37-39.

Retzer, Karen Foote, David Schipani, and Young Ik Cho. 2004. Refusal Conversion: Monitoring the Trends. *Proceedings of Survey Research Methods Section of the American Statistical Association*:4984-4990.

Schoenman, Julie A., Marc L. Berk, Jacob J. Feldman, and Andrew Singer. 2003. Impact of Differential Response Rates on the Quality of Data Collected in the CTS Physician Survey *Evaluation & the Health Professions* 26 (1):23-42.

Stinchcombe, Arthur L., Calvin Jones, and Paul B. Sheatsley. 1981. Nonresponse Bias for Attitude Questions. *Public Opinion Quarterly* 45 (3):359-375.

Stoop, Ineke A.L. 2005. *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office of the Netherlands.

Thran, Sara, Lorayn Olson, and Richard Strouse. 1987. The Effectiveness and Costs of Special Data Collection Efforts in a Telephone Survey of Physicians. Paper read at Proceedings of the American Statistical Association, Survey Research Methods Section.

Voigt, Lynda F., Thomas D. Koepsell, and Janet R. Daling. 2003. Characteristics of Telephone Survey Respondents According to Willingness to Participate. *American Journal of Epidemiology* 157 (1):66-73.

References

Armenakis, Achilles A., and William L. Lett. 1982. Sponsorship and Follow-up Effects on Response Quality of Mail Surveys. *Journal of Business Research* 10:251-262.

Assael, Henry, and John Keon. 1982. Nonsampling vs. Sampling Errors in Survey Research. *Journal of Marketing* 46 (2):114-123.

Biemer, Paul P. 2001. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics* 17 (2):295-320.

Bilodeau, Antoine. 2006. Non-response Error versus Measurement Error: A Dilemma when Using Mail Questionnaires for Election Studies. *Australian Journal of Political Science* 41 (1):107-117.

Blair, Johnny, and Young I. Chun. 1992. Quality of Data from Converted Refusers in Telephone Surveys. In *Annual Meeting of the American Association for Public Opinion Research*. St. Petersburg, FL.

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 (2):129-141.

Bradburn, Norman. 1978. Respondent Burden. Paper read at Proceedings of Survey Research Methods Section of the American Statistical Association.

Campanelli, Pamela, Patrick Sturgis, and Susan Purdon. 1997. Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates. London: Survey Methods Centre at SCPR.

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 (2):250-264.

Couper, Mick P. 1997. Survey Introductions and Data Quality. *Public Opinion Quarterly* 61 (2):317-338.

———. 1998. Measuring Survey Quality in a CASIC Environment. Paper read at Proceedings of the American Statistical Association, Survey Research Methods Section.

Currivan, Doug. 2005. The Impact of Providing Incentives to Initial Telephone Survey Refusers on Sample Composition and Data Quality. In *American Association of Public Opinion Research*. Miami, FL.

de Leeuw, Edith. 2005. To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics* 21 (2):233-255.

De Leeuw, Edith, and Joop Hox. 1988. Artifacts in Mail Surveys: the Influence of Dillman's Total Design Method on the Quality of Responses. In *Sociometric Research, Vol. 2: Data Analysis*, edited by W. E. Saris and I. N. Galhofer. New York: St. Martin's Press.

———. 1988. The Effects of Response Stimulating Factors on Response Rates and Data Quality in Mail Surveys: A Test of Dillman's Total Design Method. *Journal of Official Statistics* 4 (3):241-249.

Diaz de Rada, Vidal. 2005. The Effect of Follow-up Mailings on The Response Rate and Response Quality in Mail Surveys. *Quality & Quantity* 39:1-18.

Dillman, Don A. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons.

Donald, Marjorie N. . 1960. Implications of Nonresponse for the Interpretation of Mail Questionnaire Data. *Public Opinion Quarterly* 24 (1):99-114.

Eckland, Bruce K. 1965. Effects of Prodding to Increase Mail-Back Returns. *Journal of Applied Psychology* 49 (3):165-169.

Fricker, Scott. 2007. The relationship between response propensity and data quality in the Current Population Survey and the American Time Use Survey, Joint Program in Survey Methodology, University of Maryland, College Park, College Park, MD.

Gilbert, Gregg H., Jeffrey Longmate, and Laurence G. Branch. 1992. Factors Influencing the Effectiveness of Mailed Health Surveys. *Public Health Reports* 107 (5):576-584.

Goyder, John, Steven D. Brown, and Guil Martinelli. 2005. Larger and Smaller Prepaid Cash Incentives on Mailed Questionnaires: A new look. *Canadian Journal of Marketing Research* 22 (1):27-34.

Green, Kathy E. 1991. Reluctant Respondents: Differences between early, late and nonresponders to a mail survey. *Journal of Experimental Education* 59 (3):268-276.

Groves, Robert M., Mick P. Couper, Stanley Presser, Eleanor Singer, Roger Tourangeau, Giorgina Piani Acosta, and Lindsay Nelson. 2006. Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly* 70 (5):720-736.

Groves, Robert M., and Emilia Peytcheva. 2008. The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly* 72 (2):167-189.

Helasoja, Ville, Ritva Prattala, Liudmila Dregval, Iveta Pudule, and Anu Kasmel. 2002. Late response and item nonresponse in the Finbalt Health Monitor Survey. *European Journal of Public Health* 12:117–123.
Jobber, David, Neal Allen, and John Oakland. 1985. The impact of telephone noification strategies on response to an industrial mail survey. *International Journal of Research in Marketing* 4 (2):291-296.
Kaminska, Olena, Bart Goeminne, and Marc Swyngedouw. 2006. Satisficing in Early versus Late Responses to a Mail Survey. In *Midwest Association of Public Opinion Research*. Chicago, IL.
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):125-148.
Kennickell, Arthur B. 1999. What do the "late" cases tell us? Evidence from the 1998 Survey of Consumer Finances. In *International Conference on Survey Nonresponse*. Portland, OR.
Korkeila, K., S. Suominen, J. Ahvenainen, A. Ojanlatva, P. Rautava, H. Helenius, and M. Koskenvuo. 2001. Non-response and related factors in a nation-wide health survey. *European Journal of Epidemiology* 17:991-999.
Krosnick, J. A. 2002. The Causes of No-opinion Responses to Attitude Measures in Surveys: They are Rarely What They Appear to Be. In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York: John Wiley & Sons, Inc.
Krosnick, J., S. Narayan, and W. Smith. 1996. Satisficing in Surveys: Initial Evidence. *New Directions in Evaluation: Advances in Survey Research* 70:29-44.
Krosnick, Jon A., and Duane F. Alwin. 1987. Satisficing: A Strategy for Dealing with the Demands of Survey Questions.
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 (396):930-937.
Lessler, Judith T., and William D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: John Wiley & Sons, Inc.
Lipsey, Mark W. , and David B. Wilson. 2001. *Practical Meta-Analysis*. Vol. 49, *Applied Social Research Methods Series*. Sage Publications: Thousand Oaks.
Loosveldt, Geert, Jan Pickery, and Jaak Billiet. 2002. Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey. *Journal of Official Statistics* 18 (4):545-557.
Martin, Charles L. 1994. The impact of topic interest on mail survey response behaviour. *Journal of the Market Research Society* 36 (4):327-338.
Mason, Robert, Virginia Lesser, and Michael W. Traugott. 2002. Effect of Item Nonresponse on Nonresponse Error and Inference. In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York: John Wiley and Sons.
Miller, Joanne M., and Justin Wedeking. 2006. Examining the Impact of Refusal Conversions and High Callback Attempts on Measurement Error in Surveys. In *Annual Meeting of the American Association for Public Opinion Research*
Newman, Sheldon W. 1962. Differences between early and late respondents to a mailed survey. *Journal of Advertising Research* 2:37-39.
Olson, Kristen. 2006. Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly* 70:737-758.
———. 2007. An Investigation of the Nonresponse/Measurement Error Nexus. Ph.D., Survey Methodology, University of Michigan, Ann Arbor.
Olson, Kristen, and Courtney Kennedy. 2006. Examination of the Relationship between Nonresponse and Measurement Error in a Validation Study of Alumni. Paper read at American Association for Public Opinion Research Annual Meeting.
Robins, Lee N. 1963. The Reluctant Respondent. *The Public Opinion Quarterly* 27 (2):276-286.
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 & Research* 20 (1):30-59.
Schmidt, Jeffery B., Roger J. Calantone, Abbie Griffin, and Mitzi M. Montoya-Weiss. 2005. Do Certified Mail Third-Wave Follow-ups Really Boost Response Rates and Quality? *Marketing Letters* 16 (2):129-141.
Singer, Eleanor, John Van Hoewyk, and Mary P. Maher. 2000. Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly* 64:171-188.
Smith, Tom W. 1983. The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly* 47 (3):386-404.

Tancreto, Jennifer Guarino, and Michael Bentley. 2005. Determining the Effectiveness of Multiple Nonresponse Followup Contact Attempts on Response and Data Quality. Paper read at Joint Statistical Meetings.
Treat, James B., and Herbert F. Stackhouse. 2002. Demographic comparison between self-response and personal visit in Census 2000. *Population Research and Policy Review* 21:39-51.
Tripplett, Timothy, Johnny Blair, Teresa Hamilton, and Yun Chiao Kang. 1996. Initial Cooperators vs. Converted Refusers: Are There Response Behavior Differences? Paper read at Proceedings of the Survey Research Methods Section, ASA.
Van Goor, H. , and A.L. Verhage. 1999. Nonresponse and Recall Errors in a Study of Absence because of Illness: An Analysis of Their Effects on Distributions and Relationships. *Quality & Quantity* 33:411–428.
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 (4):401-402.
Voogt, Robert. 2004. "I'm Not Interested": Nonresponse bias, response bias and stimulus effets in election research. Ph.D. thesis, Department of Communication Studies, University of Amsterdam, Amsterdam.
Voogt, Robert J. J., and Willem E. Saris. 2005. Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics* 21 (3):367-387.
Warriner, G. Keith. 1991. Accuracy of Self-Reports to the Burdensome Survey Question: Survey Response and Nonresponse Error Trade-offs Accuracy of self-reports. *Quality & Quantity* 25:253-269.
Wellman, J.D., E.G. Hawk, J.W. Roggenbuck, and G.J. Buhyoff. 1980. Mailed Questionnaire Surveys and the Reluctant Respondent: An Empirical Examination of Differences Between Early and Late Respondents. *Journal of Leisure Research* 12 (Second Quarter):164-173.
Willimack, Diane K., Howard Schuman, Beth-Ellen Pennell, and James M. Lepkowski. 1995. Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey. *Public Opinion Quarterly* 59 (1):78-92.
Yan, Ting, Roger Tourangeau, and Zac Arens. 2004. When Less is More: Are Reluctant Respondents Poor Reporters? Paper read at Proceedings of the Survey Research Methods Section, Annual Meetings of the American Statistical Association, June 1, 2004, at Toronto.