Nonresponse Strategies and Measurement Error

Neil Malhotra (corresponding author)
University of Pennsylvania
238 Stiteler Hall
Philadelphia, PA 19104
(408) 772-7969
neilmal@sas.upenn.edu

Joanne Miller
University of Minnesota
Department of Political Science
1414 Social Sciences Building
267 19th Ave. South
Minneapolis, MN, 55455
jomiller@umn.edu

Justin Wedeking
University of Kentucky
Department of Political Science
1661 Patterson Office Tower
Lexington, KY 40506-7040
(859) 257-7040
justin.wedeking@uky.edu

June 20, 2011
ABSTRACT

One of the most daunting challenges facing survey researchers is ensuring high response rates, which have been declining due to societal, political, and technological changes. This paper analyzes whether investing resources into successfully interviewing difficult-to-obtain respondents via nonresponse strategies—thereby reducing nonresponse bias—is counterproductive because these individuals may not be fully engaged with the survey, and therefore might introduce systematic measurement error. We use a novel technique as applied to nonresponse—experimental manipulations of response order—to assess whether strategies to increase response rates increase measurement error in the form of primacy effects. The order in which response options are presented should not influence the responses that are selected. If it does, both bias and error can result. Using three recent datasets from the American National Election Studies (ANES) representing three interview modes (face-to-face, telephone, and Internet), we find that order effects, on balance, are no more pronounced among difficult-to-obtain respondents. This suggests that attempts to reduce nonresponse bias do not increase systematic measurement error in the form of primacy effects.
As recent societal, political, and technological changes have contributed to declining response rates in face-to-face and telephone surveys (de Leeuw and de Heer 2002; Groves and Couper 1998), one of the most daunting challenges facing survey researchers is nonresponse. For instance, technological advances such as caller ID and cellular phones have made it difficult to reach sampled individuals, and people have been increasingly unwilling to participate in interviews when contacted (Groves 1989). Many researchers have become alarmed at these trends because inferential sample statistics assume 100 percent response (Groves 2006; Brehm 1993; Groves and Couper 1998). Consequently, there has been great emphasis on boosting response rates via aggressive callback and refusal conversion strategies (Curtin et al. 2005), along with incorporating various interviewer and respondent incentives (Fowler 1993). Although these procedures effectively decrease nonresponse (albeit not completely), they substantially increase the cost of administration and may compromise data quality by increasing random or systematic measurement error.

This paper uses a novel technique as applied to nonresponse—experimental manipulations of response order—to explore the implications of strategies aimed at increasing response rates for systematic measurement error. Specifically, we examine the impact of doing refusal conversions and making a high number of attempts to contact hard-to-reach respondents on response order effects in Likert survey items. Of course, there is a compelling rationale for engaging in these strategies (see Brehm 1993 for a comprehensive review of the impact of nonresponse on scientific research). Namely, the assumption is that these methods (which we refer to as “nonresponse strategies”) will increase representativeness by increasing response rates (depending on the size of the response rate increase and the characteristics of the respondents added to the sample), which may or may not make the sample more similar to the population of
interest in terms of demographic characteristics and/or the distribution of responses (also
depending on the characteristics and number of respondents who are added to the sample through
these efforts). In this paper, we examine a research question posed by Groves et al. (2004a):
“When efforts to interview reluctant respondents succeed, do they provide responses more
tainted by measurement error?” (195). ¹ This is a crucial part of what Weisberg (2005) terms the
“total survey error approach,” which focuses on the tradeoffs between survey errors and costs.
Thus, our study connects two separate lines of research (representing two distinct sources of
survey error) that have important implications for one another—methods to increase
representativeness and methods to decrease measurement error (with a particular emphasis on
bias, or systematic error).

This paper is organized as follows. The first section describes previous research on
nonresponse and measurement error. The following section discusses the research design, data,
and methods of analysis. The final sections present the results and discuss their implications for
survey research.

**Previous Research**

Previous research on nonresponse has generally studied the effects of increased response
rates on the demographic representativeness of the sample and its effects on the distribution of
responses to survey questions. First, it goes without saying that if refusal conversion attempts
and high callback attempts are successful, then response rates will be higher than if these
strategies were not employed. However, evidence on the impact of these strategies on the
representativeness of the sample is not as clear-cut. As Brehm (1993), Keeter et al. (2000),

¹ Although other scholars refer to such respondents as “reluctant,” we have chosen to refer to them as “difficult-to-
obtain” since we do not have hard evidence that they are, in fact, reluctant responders. They may lack the motivation
to be conscientious but they may also (or instead) lack the ability to be conscientious. Additionally, nonresponse
strategies (such as refusal conversions and multiple callback attempts) do not directly assess a respondent’s stated
willingness to participate in the survey.
Tietler et al. (2003) and others have demonstrated, refusal conversions and high callback attempts can, in fact, increase the representativeness of survey samples (but see Heerwegh et al. 2007 for contradictory evidence). However, the effects are often small even when present. Additionally, as Tietler et al. (2003) suggest, there may exist diminishing returns where at some point increases in effort to obtain a higher response rate no longer correspond to an increase in representativeness:

At very high levels of effort, additional resources had little effect on sample characteristics, both because they yielded a small number of completed cases and because, on many measures, the additional cases more closely resembled the moderate-effort cases than they did the nonrespondents. In other words, the highest effort cases not only failed to reduce nonresponse bias but also are poor proxies for the nonrespondents (135-136; but see Peytchev et al. 2009 for a counter-example).

But even if the demographic representativeness of the sample is improved only slightly (or is unchanged) by the addition of refusal conversions and high callbacks, these methods could still increase the representativeness of the distribution of responses to questions. However, recent research suggests that nonresponse rates have little effect on survey estimates (Curtin et al. 2000; Keeter et al. 2000, 2006; Merkle and Edelman 2002; Tietler et al. 2003). Common empirical strategies employed by these authors include comparing survey estimates to known population benchmarks, and comparing results from low-effort surveys designed to achieve low response rates against high-effort attempts to maximize response rates. Meta-analyses conducted by Groves (2006) and Groves and Peytcheva (2008) revealed instances of positive, albeit inconsistent, relationships between nonresponse rates and nonresponse bias (such that the larger the nonresponse rate, the larger the bias in survey estimates). But in some cases, surveys with low nonresponse were found to have relatively high nonresponse bias (Groves and Peytchva 2008), and the positive relationship between nonresponse rates and nonresponse bias is larger on items related to the survey topic as described to respondents during recruitment (Groves et al.
The research reported in this manuscript tests a third possible effect of strategies aimed at increasing response rates—that systematic measurement error may increase as difficult-to-obtain respondents are introduced into the sample. Specifically, regardless of the effect on the distributions of survey variables, the quality of the data may be compromised as response rates increase. It is possible that nonresponse strategies may bring respondents into the sample who are more likely than their easy-to-obtain counterparts to engage in less-than-optimal question answering strategies, thus increasing measurement error for these respondents.2

Respondents who initially refuse to participate in a survey may agree to participate simply to get the interviewer to stop contacting them. If their goal in assenting to be interviewed is to prevent further bothersome contacts, they may provide “good enough” answers without much cognitive engagement to complete the interview as quickly as possible. Alternatively, they may be more susceptible to memory effects that introduce measurement error. Consistent with this supposition, Groves and Couper (1998) suggest that:

One trade-off decision has been the source of great speculation among survey researchers for some years. Many have suspected that respondents who consent to an interview only after intense persuasion may provide only superficial attention to the respondent task. In short, they may have sufficient motivation to provide the interview (perhaps to avoid continuation of unpleasant encounters with the interviewer), but insufficient motivation to listen thoroughly to the questions, to search their memories carefully for the appropriate answer, and to provide candid replies. This describes a trade-off between nonresponse error and measurement error properties of a survey. The hypothesis of higher measurement error among reluctant respondents deserves more careful thought. The underlying premise of the hypothesis is that lack of desire to grant the interview would continue after a commitment is given to provide the interview. In essence, the respondent would have passed a threshold that separates refusals from acceptances, but would fail to attain a level of motivation to attend carefully to the respondent tasks (271).

---

2 Measurement error can come from a variety of sources, including the interviewer, the mode of interview, the respondent, or the instrument (and such measurement error can be systematic or random). In this paper, we focus on systematic measurement error (also known as bias) due to respondents. This form of measurement error is often referred to as “response bias” (Groves 1989).
Respondents recruited via numerous callbacks may similarly lack sufficient motivation, or, alternatively (or in combination), they may lack sufficient ability (due to time constraints) to optimize. The goal of multiple callbacks is to finally “catch” hard-to-reach respondents at home. These respondents, although they have agreed to complete the interview when they are finally “caught,” might very well still be too busy at the time of the interview to thoroughly engage in the survey.

Consistent with this reasoning, some studies have suggested that measurement error increases as survey organizations engage in more strenuous efforts to increase response rates. In an early analysis, Cannell and Fowler (1963) found that respondents brought into the sample through such strategies were more likely to misreport factual information. Other studies have similarly raised the concern of whether successfully supplementing the respondent pool via persuasive efforts may introduce measurement error (e.g. Biemer 2001; Groves and Couper 1998).

Recent research that has examined the impact of strategies aimed at reducing nonresponse on measurement error have considered the following types of error: item nonresponse, reports of round values, classification errors, inconsistencies between interviews, differences between survey and administrative records data, non-differentiation, no opinion or “don’t know” responses, “straightlining,” and a count of the number of extreme plus middle responses. The extant literature (summarized in Table 1) is quite inconsistent. Some studies have found that nonresponse strategies increase measurement error (e.g. Fricker and Tourangeau 2010; Kreuter, Mueller, and Trappman 2010), others find no relationship (e.g. Sakshaug, Yan, and Tourangeau 2010; Kaminska, McCutcheon, and Billiet 2010), and still others present mixed results (e.g. Olson and Kennedy 2006; Olson 2006). In an attempt to provide coherence to this
area of research, a meta-analysis by Olson, Feng, and Witt (2008) found mixed evidence for the hypothesis that adding “high recruitment effort” respondents would lead to worse data quality, noting it tended to occur in studies with individual measures rather than the entire general questionnaire and for behavioral items rather than attitudinal items. Thus, Olson et al. (2008) suggest this means the relationship is item-specific.

[TABLE 1 ABOUT HERE]

We examine a type of measurement error that has not yet been explored in this literature—primacy effects, or the tendency of respondents to choose response options near the top. Our examination of response order effects through the use of experiments has advantages over prior research. For example, to assess the presence of measurement error, previous studies have either relied on direct evidence (e.g., comparison of survey data to administrative records data) or indirect evidence (e.g., indicators of item nonresponse, don’t knows, etc). The use of direct evidence such as administrative records may be picking up errors in either the administrative records themselves or reporting errors on the part of respondents (due to failing memory or lack of effort in retrieving the information from memory). Indirect measures of error are problematic because it is impossible to determine whether they are, in fact, errors, or whether they reflect true attitudes (e.g. is a respondent who says “don’t know” responding truthfully or lazily?). As we describe below, our approach combines the best of both worlds by directly assessing measurement error but also isolating how respondents are engaging with the survey itself. Moreover, response order effects are valuable to analyze because they are especially pernicious contributors of measurement error. In other words, this study examines the “upper bound” of how problematic nonresponse strategies are in whether they increase measurement error in survey responses.
It is important to note that response order effects can reflect both random measurement error and/or systematic bias. If response order is not randomized, then primacy effects will bias responses towards those that appear first. This systematic bias in responses is distinct from the selection bias in the demographic characteristics of respondents that is more commonly analyzed in studies of nonresponse. Even if order is randomized, the presence of primacy effects will still introduce measurement error as respondents are not reporting their true attitudes. This measurement error, however, will be “noise” or random error that can have significant consequences by attenuating relationships between two variables. If enough random error is present, two variables that normally would be statistically related will no longer be related. Hence, primacy effects introduce a relatively severe form of measurement error that should be mitigated as much as possible in survey administration.

Primacy effects could be due to numerous alternative mechanisms: (1) question difficulty (e.g., Payne 1951) (2) uncrystallized attitudes; (3) forgetting the response options, though the evidence is somewhat mixed (Sudman, Bradburn, and Schwarz 1996; Knauper 1999); (4) the interaction of the serial position of the alternative, the mode (auditory or visual), and the plausibility of the response alternative (Sudman, Bradburn, and Schwarz 1996); (5) satisficing, (e.g., Krosnick and Alwin 1987) which suggests most respondents choose the first acceptable response alternative rather than selecting the optimal alternative; and (6) cognitive ability (Kaminska, McCutcheon, and Billet 2010).³ It is beyond the scope of this paper to distinguish between these various mechanisms that could explain response order effects. Nonetheless, the presence of order effects is a clear and unwanted sign of the presence of measurement error in survey responses that researchers should try to minimize.

³ For a detailed review of these literatures, see Bishop and Smith (2001).
Much of the work on primacy effects in surveys has focused on questions with a longish list of categorical response options from which respondents must choose (e.g., a list of problems facing the nation or desirable child-rearing values). Should we expect primacy effects to occur in Likert-type rating scales, where the response categories are well-known to respondents and the list of response categories is not very long? Several recent studies (Malhotra 2008, 2009; Holbrook et al. 2007) have found significant order effects in Likert scales, administered both visually and over the telephone, even for rating scales with as little as two response options (see also, Payne 1951; Schuman and Presser 1996). As shown below, we replicate these findings and find significant primacy effects in ordinal Likert scales.

We examine the impact of strategies aimed at increasing response rates on measurement error in the form of primacy effects in ordinal Likert scales via surveys that contained response order experiments. In these experiments, respondents are randomly assigned to receive the response options in one order or in the reverse order. Such experiments are an ideal way to examine this question because the order of the response options is exogenously manipulated by the researcher. Hence, the order in which responses are presented should not meaningfully affect reported attitudes. Thus, response order effects represent unambiguous sources of measurement error unlike other measures mentioned above, which may be consistent with respondents revealing their true preferences.

Research Design, Data, and Methods

Data

We leverage three datasets from the American National Election Studies (ANES), representing three interview modes (face-to-face, Internet, and telephone). All three surveys included numerous experimental manipulations of the order of rating scales, and attempted to
reduce nonresponse by converting refusals and making numerous callbacks to contact difficult-to-reach respondents. Thus, we can examine whether the effect of order on responses was different among those who would not have been respondents if not for the strategies implemented, and those for whom nonresponse strategies were not needed.

*2008 ANES Time Series.* The pre-election wave of the 2008 ANES Time Series Study was conducted face-to-face on a nationally representative sample of 2322 U.S. adults between September 2, 2008, and November 3, 2008. Respondents were asked questions in four main ways: (1) orally by the interviewer; (2) orally by the interviewer with the use of a visual aid-booklet for respondents; (3) orally self-administered by computer (through headphones); and (4) visually self-administered by computer. A fresh cross-section of respondents was recruited for the study. The AAPOR RR5 response rate was 78.2%. There were 103 response order manipulations administered to either the entire sample or a randomly-selected portion of the sample. All respondents received advanced mailings with a $25 incentive that was increased to $50 on October 7th. If a respondent initially refused, he/she was sent a specially tailored letter to address the reasons for refusing. Interviewers had two main incentives: (1) interviewers were given $20 per completed interview from September 7-10; and (2) from October 7-November 3 were given amounts ranging from $5-$30 depending on the form of completed interview. The dataset contained 366 refusal conversions, and the range of call attempts was 5-48 (mean: 13.4; s.d.: 5.1). Additional methodological details and full questionnaires for this and other studies can be found at www.electionstudies.org.

*2008 ANES Panel.* We use the first (January 2008) wave of the 2008 ANES Panel Study,

---

4 In all analyses, we excluded response-order manipulations for items into which respondents self-selected. For example, in the 2006 ANES Pilot Study respondents who reported that they believed that “Jesus is the son of God” were then asked how important this belief was to them personally on a five-point scale, for which the order of the response options was manipulated. Although the estimates of order effects among these subsamples are internally valid, they are not externally valid since they apply only to Christian believers, not the general population.
which was conducted over the Internet by Knowledge Networks. Respondents were recruited into the panel in late 2007 over the phone via random digit dialing (RDD). If a respondent agreed to participate in the panel as a result of the phone request, he or she was sent an email notification with a link to the web survey up to a few weeks later. 1623 respondents were interviewed in the first wave. We only analyze the first of the 21 waves to not confound our results with panel attrition. The AAPOR RR5 response rate was 75.0%. There were 47 response order manipulations administered to either the entire sample or a randomly-selected portion of the sample. All respondents were offered $10 a month to complete surveys over a 21-month period. A unique feature of this dataset (compared to the other two) is that there were many opportunities for respondents to opt out of the survey. For example, after joining the panel, respondents may not have responded to email invitations to participate in the survey. Therefore, it is possible that even if a respondent was difficult to recruit into the panel initially, he or she may have been more willing to participate conditional on accepting an email invitation. For the initial RDD recruitment, the dataset contained 47 initial refusals and the range of call attempts was 1-50 (mean: 5.8; s.d.: 6.5). Initial refusers were assigned to “refusal conversion specialists” (see the 2008-2009 ANES Panel Study User Guide for more details).

2006 Pilot. We use data from the 2006 ANES Pilot Study, which was conducted by telephone on a nationally representative sample of U.S. adults between November 13, 2006, and January 7, 2007. The sample consisted of 1212 respondents in the 2004 ANES Time Series Study with 675 completed interviews, yielding a re-interview rate of 56.3%. One limitation of this study to keep in mind is that a fresh cross-section of respondents was not collected, meaning that even the difficult-to-obtain respondents had participated in a previous ANES study. There were 44 response order manipulations that were administered to either the entire sample or a
randomly-selected portion of the sample. Respondents were offered the highest level of incentive they received in 2004 (either $20 or $50), and those that refused after Dec. 15th were offered a higher incentive of $100. The dataset contained 67 refusal conversions and the range of call attempts was 1-94 (mean: 14.1; s.d.: 18.5). This study also recorded whether the respondent broke an interview appointment that needed to be rescheduled. 242 respondents broke at least one appointment and 155 broke at least two appointments. As described below, because the sampling frame (respondents in the 2004 ANES Time Series) is known, the 2006 ANES Pilot Study allows us to estimate a response propensity for each respondent.

All three datasets have several advantageous features. These studies are unique in terms of the sheer number of response order manipulations available to analyze, especially given that many items employed unipolar rating scales. Therefore, unlike studies that rely on a few experiments, we can assess whether findings are consistent and common or rare and infrequent. An example of a typical item is shown in Table 2: “How much do you think people can change the kind of person they are?” This item was administered over the phone as part of the 2006 ANES Pilot. In the “forward” condition, the response options are orally presented in the order they appear in the first column (from top to bottom). In the “reverse” condition, they are presented as they appear in the second column. Although the ANES mainly studies political attitudes, these studies did test various non-political constructs, particularly in the 2006 Pilot. Moreover, in the 2008 Time Series study, respondents (for the first time ever) “were not told they were being interviewed for the ‘National Election Study’ in order to avoid self-selection into or out of the sample based on interest in politics” (2008 Time Series User Guide, 5).

**Method**

[TABLE 2 ABOUT HERE]
Across all three studies, we have 194 tests of whether difficult-to-obtain respondents are more subject to response order manipulations. We estimate the following difference-in-difference (DID):

\[
\left( \bar{Y}_{rh} - \bar{Y}_{fh} \right) - \left( \bar{Y}_{re} - \bar{Y}_{fe} \right)
\]  

(1)

where \( \bar{Y}_{rh} \) represents the average value of the response option (measured on a Likert-type rating scale) for respondents who were “hard” to obtain in the “reverse” order condition and \( \bar{Y}_{fh} \) represents the average value of the response option for respondents who were “hard” to obtain in the “forward” order condition. \( \bar{Y}_{re} \) and \( \bar{Y}_{fe} \) represent corresponding values for respondents who were “easy” to obtain. All responses are coded as in Table 2. In the “forward” order condition, the response options coded with lower numerical values (e.g. “completely”) are presented on top, whereas in the “reverse” order condition, those same response options are coded identically but are presented last. Hence, a positive DID estimate (as represented by equation (1)) indicates that the order effect (our measure of survey satisficing) is higher among hard-to-obtain respondents as compared to easy-to-obtain respondents. If the DID estimate is positive and significantly different from zero (as ascertained by a \( t \)-test), then it suggests that using nonresponse strategies has the tradeoff of introducing systematic measurement error into the data.\(^5\)

Nonresponse Strategies

We analyze several nonresponse strategies explored in previous research (e.g. Curtin et al. 2000). First, we examine a dummy variable indicating whether an interviewed respondent

\(^5\) The DID estimate is equivalent to the coefficient estimate on the interaction term from a linear regression predicting the item response with: (1) the order dummy; (2) the measure of whether the respondent was “hard-to-obtain”; and (3) the interaction between the two. The regression framework also allows us to estimate the effects of continuous measures such as the number of calls. We did not include demographic covariates because we wanted to maximize the chance we observed significant DID estimates and did not want to include variables that might be correlated with the survey administration measures. The results are substantively and statistically similar when controlling for a host of demographics.
initially refused to participate but was then “converted.” Second, we examine the number of calls\(^6\) it took to contact the respondent to complete the interview. Third, we examine the natural log of the number of calls in order to reduce skewness in the data. Fourth, we bifurcate the number of calls at the median and code respondents as requiring a “high” number of calls if they fall above the median.\(^7\)

Fifth, we examine dummy variables indicating whether the respondent broke one or more scheduled appointments (vs. none), or whether the respondent broke two or more appointments (vs. one or none) to complete the interview (only available in the 2006 ANES Pilot).\(^8\)

Sixth, because of the small number of refusals in the datasets, we combine the callback, refusal conversion, and broken appointments dummy variables (using the one or more broken appointments dummy) to create a binary variable indicating whether a respondent was a refusal conversion, required a “high” number of calls, or broke one or more appointments. Seventh, we create a continuous variable indicating the propensity of a respondent to complete an interview (only available in the 2006 ANES Pilot). Using the full sample frame of respondents in the 2004 ANES Time Series Study, we estimated a logistic regression predicting participation in the 2006 Pilot Study using a set of covariates,\(^9\) and then calculated predicted probabilities based on the coefficient estimates. The predicted probabilities are propensity scores representing how likely

---

\(^6\) With respect to call attempts, the 2008 Time Series Study does not distinguish phone calls from house visits. Hence, we use “calls” as a catchall term to refer to contact attempts more broadly.

\(^7\) A low number of callbacks may be due to inefficiencies on the part of the survey organization. Therefore, we include the dichotomized callback variable to capture differences in the number of callbacks that is more likely to indicate a busy, or hard-to-reach, respondent.

\(^8\) Whereas one broken appointment may be legitimate or due to chance, it is less likely that two or more broken appointments by the same potential respondent is legitimate, and likely indicates unwillingness and/or inability to participate in the interview.

\(^9\) To construct the propensity score equation, we selected covariates from the 2004 ANES Time Series based on two criteria: (1) their ability to effectively predict response; and (2) having data on nearly all respondents so as to not truncate our sample due to listwise deletion. Our propensity score equation correctly predicted the dependent variable in 70% of cases. The covariates used in the equation were: gender, refusal conversion, total number of calls, length of interview, interest in campaigns, 2000 turnout, age, highest grade level completed, highest degree completed, employment status, dummy for whether respondent is white, how long respondent has been living in current home, home ownership, party identification, and interviewer assessments (cooperation, level of information, intelligence, suspicion, interest, sincerity).
an individual was to participate in the 2006 study based on observed characteristics.10

Table 3 presents polychoric correlations between the various non-response strategies in the three datasets. Although the different strategies are positively correlated, the correlations are not inordinately high, suggesting that these indicators are picking up different aspects of respondent behavior.

[TABLE 3 ABOUT HERE]

Results

Main Results

Before describing the overall results, it is instructive to describe a statistically significant DID estimate in detail. In the 2008 Panel Study, respondents were asked: “Compared to 2001, would you say the nation’s crime rate is now much better, somewhat better, about the same, somewhat worse, or much worse?” on a five–point scale ranging from “much better” (1) to “much worse” (5). Among respondents who were not refusal conversions, the difference-in-means between the “reverse” condition and the “forward” condition was .05 points (3.42 – 3.37). Among respondents who were refusal conversions, the difference-in-means between the “reverse” condition and the “forward” condition was .65 points (3.78 – 3.13). Hence, the DID estimate was .60 units (.65 – .05), meaning that the primacy effect was significantly larger among refusal conversions ($p = .02$).

Generally, respondents who were obtained through nonresponse strategies were no more likely to exhibit primacy effects in response order manipulations than respondents who were more easily recruited. As a first test, we only examine the 69 items across all studies (or 35% of

---

10 We multiplied the propensity scores by -1 so that this measure was coded consistently with the others, with higher values representing harder-to-obtain respondents.
the full set of 194) which exhibited significant primacy effects in the full sample.\textsuperscript{11} As shown in the first row of Table 4, significant, positive DID estimates (as described in equation 1) appeared in only 2.9\%-11.6\% of cases. Given that we would expect to observe significant effects by chance 10\% of the time,\textsuperscript{12} these findings do not suggest that this type of measurement error is consistently higher among difficult-to-obtain respondents. Moreover, these results may indeed overstate the effects of the non-response strategies. Because we are making multiple comparisons, it is more appropriate to employ a Bonferroni correction, increasing the standard of significance necessary to reject the null hypothesis. For instance, in the 2008 Time Series (39 items), we apply a one-tailed alpha-level of .001282 ($t=3.0$), or .05/39. Applying these corrections, we find zero statistically significant DID estimates in each dataset and overall across all studies. Online Appendices 1-3 present the DID estimates (and associated $t$-statistics) for each individual item.

\textbf{[TABLE 4 ABOUT HERE]}

Of course, even if the main primacy effect is statistically insignificant, the interaction between whether the respondent was difficult-to-obtain and the response order dummy can be statistically significant. Accordingly, we also examined the full set of 194 items. As shown in the second panel of Table 4, we again observed little evidence that primacy effects are concentrated or higher among difficult-to-obtain respondents.

Are the significant difference-in-difference estimates we did observe due to chance alone, or do the items have some shared feature that explains why nonresponse strategies may have mattered in these cases? In an attempt to discern if there was any pattern to the questions for

\textsuperscript{11} Main primacy effects were determined by simply calculating a difference-in-means in response outcomes between respondents in the “reverse” condition and respondents in the “forward” condition.

\textsuperscript{12} Because we have well-specified theoretical reasons to assume a positive DID estimate, we apply the standard $p<.05$ statistical significance threshold using one-tailed tests.
which a significant order effect was obtained in the full sample, or for which the order effect was significantly larger among the difficult-to-obtain respondents, all items with a significant DID estimate were culled from the questionnaires and scrutinized separately. We qualitatively analyzed these questions by looking across the three datasets for possible similarities in: (1) substantive content; and (2) format. In general, there was nothing common across all three surveys in either content or format. The only similarity that we found existed across two of the surveys, but even this did not seem particularly noteworthy given that the overwhelmingly large number of questions we analyzed initially revealed so few significant results. With that said, we feel it is important to highlight this finding on the possibility that it may benefit future researchers. We noticed in the 2008 Panel and 2008 Time Series datasets that a large number of the significant results were “follow-up” questions. For example, question w1p18 in the 2008 Panel dataset asks “How important is this issue to you personally?” The issue was given to the respondent two questions prior. The 2008 Panel contained twelve follow-up questions with significant effects, and the 2008 Time Series contained six follow-up questions that were significant. It is possible that some respondents may be less engaged with follow-up questions due to their repetitive nature. We encourage exploration into this hypothesis in future studies.

One of the benefits of the research design was that it includes three surveys with different modes (phone, face-to-face, web). Hence, we examined the results to see whether there was a pattern to our findings that could be explained by mode of interview. Looking again at the top panel of Table 4, we see that for refusal conversions, the phone survey had the highest percentage of significant difference-in-difference t-tests (11.4%), but that the web (4.3%) and face-to-face (2.9%) percentages were much smaller. However, even this difference is not present when examining the various calls variables. Looking at the combined measure, the Web survey
in this case exhibits the highest percentage of significant DID estimates. In sum, the general conclusion from our analysis is that very little systematic measurement error (in the form of primacy effects on Likert scale items) is introduced by including difficult-to-reach respondents into the survey across any of the three survey modes (face-to-face, phone, and web).

Robustness Checks

Given that we are reporting null findings, we conducted a series of robustness checks to ensure that our DID estimates were consistently and conclusively small and insignificant. One alternative explanation for our null findings is that, because we are analyzing high-quality samples collected at considerable expense, the sample sizes are limited, thus decreasing statistical power. One way to assess power is to analyze the results post hoc and examine items with barely insignificant DID estimates ($t$-statistics ranging from 1.50 to 1.64). We specifically look at refusal conversions since the number of refusal conversion respondents in the datasets is small compared to the number of non-refusal conversion respondents. In the 2008 ANES Time Series Study ($n = 2322$ with 366 refusal conversions), barely insignificant DID estimates were between .4 and .5, representing between 10-13% the length of a five-point Likert scale. In the 2008 Panel Study ($n = 1623$ with 47 refusal conversions), these not-quite-significant estimates ranged between .5 and .6 whereas in the 2008 Pilot Study ($n = 675$ with 67 refusal conversions), the range was between .3 and .4. Based on these calculations, we conclude that there is likely sufficient statistical power to detect moderately-sized DID estimates.

Another potential concern is that many of the items were not administered to the full sample of respondents and therefore those analyses may also not have sufficient statistical power. The third panel of Table 4 presents the results restricting the analyses to items which were administered to all respondents. Although the percentage of significant rejections of the
null hypothesis is a little higher overall, for no nonresponse strategy (across all studies) did the percentage of significant DID estimates exceed 10%.

We also tested whether our DID estimates were higher when respondents were interviewed well before the election, at a time when the salience of the topic of the survey was lower. The 2008 ANES Time Series Study was the only study that allowed us to explore this variability. We found no evidence that interview date moderated the effect of the nonresponse strategies, through analyses that (1) only examined interview cases collected more than 40 days before the election; and (2) estimated three-way interaction terms in a regression framework (see Online Appendix 4).

Finally, one could argue that primacy effects will only manifest themselves in the first response option, which would not be captured by our modeling strategy. As a last robustness check, we re-estimated the results to examine whether difficult-to-obtain respondents were more likely to pick the first category than the other response options. Again, these results are virtually identical to those reported in Table 4 and Online Appendices 1-3. Pooling results across the three datasets, the percentage of significant DID estimates for the various nonresponse strategy measures range between 3.1% and 5.2%. Moreover, for no individual study does the percentage of significant coefficients exceed 10% for any of the nonresponse strategies studied.13

**Moderation by Position in Survey**

Previous research has found that response-level error can increase later in the interview as respondents become more fatigued and motivation decreases (Holbrook et al. 2007; Krosnick et al. 2002). Accordingly, we tested whether DID estimates were larger for items presented later

---

13 Our results are also robust to weighting the data and clustering by sampling unit. We re-estimated DIDs using Stata’s survey commands and the corresponding weights and sampling designs. Our findings remain virtually unchanged.
on in the survey. To calculate an item’s position in the survey, we examined the questionnaire for each survey, identified the first substantive question asked to respondents, and counted each subsequent question as one item. The first substantive item asked would be considered the first item, and the second item asked would be considered the second, and so on. Based on this, we were able to calculate the position of each item within each survey. Table 5 presents OLS regressions (where the unit of analysis is the survey item) predicting the DID estimate with item position. To facilitate comparison across the studies, we divided the item position by the total number of questions asked in the survey instrument. This coding also allows us to easily interpret the regression coefficients as the change in the DID estimate from the first and last items in the questionnaire. In other words, the coefficient captures the relevant counterfactual of moving a question from being at the very beginning of the interview to the very end.

In the 2008 Time Series study, which was conducted face-to-face, the DID estimate associated with refusal conversions was about .5 units higher at the end of interview compared to the beginning (p < .001). Similarly, the DID estimate associated with an additional call attempt on the primacy effect was .015 units higher for questions asked at the end of the interview (p = .03). We also obtained strong moderation by item position in the 2008 Pilot Study, which was conducted by telephone. Although item position did not moderate the effect of refusal conversions, the DID estimate associated with the “high calls” measure was .32 units higher at

---

14 We could not conduct this analysis using the 2008 Panel Study because the rating scales items with randomized response orders tended to be placed in the latter part of the questionnaire.

15 There are a few items to note. First, if the survey employed a split-sample questionnaire design, this was accounted for by ensuring these items were not “double-counted.” Second, the count will be slightly inflated mainly because it treats “follow-up” items that were asked to a subset of respondents the same as questions that were asked of all respondents. This was a necessary sacrifice because it is otherwise impossible to determine where a question occurs in the survey for each respondent, given the high number of_bloc-unit randomizations and how often the respondent was probed. As an approximation, we treated each follow-up question as one item because many follow-up questions are asked to a very large percentage of the respondents.
the end of the questionnaire compared to the beginning ($p = .002$). Additionally, question position positively and significantly increased the DID estimate of breaking two or more appointments ($b = .22, p = .03$). Finally, the coefficient associated with the response propensity variable is also positive and significant, consistent with the notion that difficult-to-obtain respondents evidence more systematic measurement error later on in the interview.

In summary, we find that, in both the phone and face-to-face modes, the effect of nonresponse strategies on primacy effects is more pronounced later on in the interview when respondent motivation might be lower. An implication of this result is that although nonresponse strategies do not seem to be affecting data quality overall, researchers should be cautious of their effects in portions of the questionnaire where satisficing effects might be greater.

**Length of Interview**

We next estimated OLS regressions predicting length of interview (in minutes) to see if respondents obtained via nonresponse strategies completed the questionnaire more quickly (see Table 6). For the 2008 Time Series, we found that while refusal conversion respondents did not significantly differ from other respondents in their completion time, those who required a greater number of calls were significantly quicker in finishing the interview. Each additional call is associated with a reduction in administration time of about 13 seconds. Respondents who required a “high” number of call attempts completed the questionnaire in 2.66 less minutes than other respondents. Even larger effects were observed for the 2006 Pilot Study. Refusal conversions completed the interview over four minutes quicker that other respondents ($p < .001$). Similarly-sized effects were observed for respondents who broke appointments, as well as those that required a large number of call attempts.

---

16 We could not conduct this analysis using the 2008 Panel Study because many respondents appeared to get up from the computer several times during the interview, thereby producing non-sensible completion times. Indeed, the skewness (i.e., third statistical moment) of the completion time variable is 8.68.
Missing Data

We also examined whether the various nonresponse strategies predicted “missingness” in the data. We defined missingness as when a respondent volunteered a “don’t know” answer or “refused to give answer” to a question.\textsuperscript{17} The expectation is that respondents who initially refuse to participate in a survey, or those that require a large number of phone calls to reach would have an increased percentage of “missing” data. For each respondent we summed up all the questions that contained “missing” data, and then divided by the total number of questions in each survey.\textsuperscript{18} We estimated a series of OLS regressions predicting our measure of missingness with the nonresponse strategies as independent variables.

As shown in Table 7, being a refusal conversion respondent increases the level of “missingness” in the data across all three datasets. However, we note that this effect is not substantively large, and the distribution of “missingness” in respondents tends to be highly skewed. In other words, the vast majority of respondents have very little missing data. For example, by our estimation, in the 2006 Pilot data, 96 percent of respondents have less than 2 percent of the questions missing.\textsuperscript{19} Moreover, the calls variables only increased missingness in the 2008 Panel Study, but not the other two surveys. The high calls indicator was significantly negative in the 2008 Time Series Study. Thus, there was no consistent effect of nonresponse

\textsuperscript{17} We examined “don’t knows” and “refused to answer question” as separate indicators of missingness but opted to combine them in a summary measure because the results were similar across the two indicators. Moreover, the 2008 Panel Study, which was the Web survey, did not differentiate between “don’t knows” and “refused to answer question” and simply gave respondents the option of not answering the question. Thus, combining “don’t knows” and “refused to answer question” was more appealing because it allowed for a measure of missingness that was more comparable across surveys and modes.

\textsuperscript{18} Each question counted as one unit, regardless of the type of question. We also tried another measure of “missingness” that divided the total number of missing data by the total number of questions answered, which varied across respondents, and the results are remarkably similar to those we report here.

\textsuperscript{19} In the 2008 Panel Study, 95 percent of respondent had less than 2 percent of the questions missing. In the 2008 Time Series Study, 90 percent of respondents had less than 5 percent of the questions missing.
strategies on missingness across the three datasets.

[TABLE 7 ABOUT HERE]

Discussion

Survey researchers in industry, government, and academia have all been concerned with the trend of declining response rates in all survey modes. Technological and social changes have only exacerbated these concerns, challenging the paradigm of representative sampling. Not surprisingly, these worries have generated a great deal of valuable research assessing the impact of nonresponse rates on bias and measurement error, and potential strategies to improve response rates. However, the effort to increase response rates might be counterproductive because difficult-to-obtain respondents may not be fully engaged with the survey.

Leveraging a novel methodological approach to tackle the question, this paper found little evidence that even though primacy effects were evident in over one third of the Likert scale items we examined, there was little evidence that difficult-to-obtain respondents are more likely than early, enthusiastic respondents to choose the first reasonable response on a rating scale. This was true for both the items for which a significant main primacy effect was found and among the ones for which a main primacy effect was not statistically significant. Using various measures of reluctance and numerous response order manipulations, the relationship between reluctance and primacy effects was rarely found to be statistically significant. These findings have important implications for strategies for tackling non-response. Although agnostic about the question of whether expensive efforts to achieve high response rates effectively increases demographic and response distribution representativeness, our evidence does show that these efforts do not introduce other forms of systematic measurement error. Moreover, although we found that hard-to-obtain respondents were more likely to complete the survey quickly,
nonresponse strategies did not substantially affect item non-response. If there is one area of concern for survey researchers, it is that we did find that DID estimates were stronger in the later portions of the questionnaire. Hence, if researchers are going to implement costly nonresponse strategies in their survey, then they should focus on maintaining respondent motivation as the interview progresses.

A few caveats are in order. First, we have specifically focused on the narrow question of whether nonresponse strategies exacerbate response order effects—a fairly extreme form of measurement error. We look forward to future research that examines the effects of adding difficult-to-obtain respondents into samples on other forms of response error. Moreover, it would interesting to see if random measurement error was affected by refusal conversions and numerous callback attempts. Second, the nonresponse strategies we examined are indirect (and potentially noisy) measures of respondent reluctance. While this paper was mainly concerned with characteristic of survey procedures, future studies can examine respondent-level attributes in more detail. Third, we caution that the bulk (though not all) of our items measure political attitudes on Likert scales. Based on both the response format and subject matter, these types of items may be most subject to comprehension errors. Thus, we must be circumspect in generalizing our findings to other types of questions and topics commonly employed by public opinion researchers. Finally, we look forward to additional research that investigates this question using data from a diverse set of survey sponsors and survey houses.

Given these caveats, we caution against relying on our evidence—that primacy effects are not greater among difficult-to-obtain respondents than their easy-to-obtain counterparts—to definitively conclude that nonresponse strategies have absolutely no adverse measurement effects. Research on the relationship between response rates and measurement error is in its
infancy. Clearly, additional study is needed to more fully understand the tradeoffs between the costs and benefits of achieving higher response rates. Nonetheless, it is comforting to note that efforts to ameliorate nonresponse (i.e., doing refusal conversions and making a high number of attempts to contact hard-to-reach respondents) do not introduce additional complications in the form of increasing the prevalence of response order effects on rating scales.
References


Sakshaug, Joseph W., Ting Yan, and Roger Tourangeau. 2010. “Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of


Table 1: Previous Research on the Relationship between Nonresponse Strategies and Measurement Error

<table>
<thead>
<tr>
<th>Study</th>
<th>Data</th>
<th>Type of Error Examined</th>
<th>Result of test for relationship between nonresponse and measurement error</th>
</tr>
</thead>
</table>
• American Time Use Survey (Jan 2003-Dec 2003) | CPS data  
• item non-response  
• Reports of round values  
• Classification errors between rounds (e.g., changes in reported race)  
• Inconsistencies between interview and re-interview response variance  
ATUS data  
• Total # of activities reported  
• Missing reports of daily activities  
• Reports of “round values”  
• Item non-response | “Although the strength of the relationship varied by indicator and survey, data quality decreased for some indicators as the probability of nonresponse increased” (p. 934). |
| Kreuter, Muller, and Trappmann (2010) | • First wave of German Panel Study “Labor Market and Social Security” (PASS)-December 2006-July 2007 | Differences between respondent’s survey reports and administrative record data for 4 measures  
Examine how components of Mean Square Error are influenced | “Measurement error increased somewhat with increased effort, though total bias was nevertheless reduced” (p. 880). However, one indicator showed mean square error increased despite reduction in nonresponse bias and constant measurement error. |
| Olson (2006)                  | • 1995 Wisconsin Divorce Study                                       | Leverages data from respondents’ survey answers and administrative records  
Examines how the composition of the Mean Square Error changes on 3 statistics as lower propensity respondents are added to survey estimate. | On 2 out of 3 statistics, measurement error bias increased when reluctant respondents were added. However, total bias of all three statistics decreased as a result of including low contact propensity respondents. |
Divides respondents into response propensity classes | “Measurement error, independent of the indicator, does not appear to be related to response propensity” (p. 4187). Though, there are subgroup differences in measurement error. |
| Sakshaug, Yan, and Tourangeau (2010) | • University of Maryland Alumni survey (August-Sept. 2005) | Nine survey questions for which validation was available (e.g., administrative records)  
Focus on how nonresponse error and measurement error contribute to total error | While switching to a self-administered mode can reduce measurement error, they found it substantially increased nonresponse error. But, additional callbacks, while reducing bias from noncontact (but not overall nonresponse error), had “no consistent relation to measurement error” (p. 930). |
| Kaminska, McCutcheon, and Billic (2010) | • European Social Survey (round 3, 2006-2007, separately across Belgium, Sweden, Norway, and the Netherlands) | A measure of satisficing developed with latent class analysis using indicators of:  
• “don’t knows”  
• “straightlining”  
• Inconsistent responses  
• Extreme + middle responses | “The relationship between reluctance and response quality is present but spurious, being completely explained by cognitive ability” (p. 956). Findings are consistent across countries. |
### Table 2: Sample Response Order Manipulation

*How much do you think people can change the kind of person they are?*

<table>
<thead>
<tr>
<th></th>
<th>“Forward” Condition</th>
<th>“Reverse” Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read first</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read first</td>
<td>Completely (1)</td>
<td>Not at all (5)</td>
</tr>
<tr>
<td>A lot (2)</td>
<td>A little (4)</td>
<td></td>
</tr>
<tr>
<td>A moderate amount (3)</td>
<td>A moderate amount (3)</td>
<td></td>
</tr>
<tr>
<td>A little (4)</td>
<td>A lot (2)</td>
<td></td>
</tr>
<tr>
<td><strong>Read last</strong></td>
<td>Not at all (5)</td>
<td>Completely (1)</td>
</tr>
</tbody>
</table>
Table 3: Polychoric Correlations Between Refusal Conversion and Call Measures

<table>
<thead>
<tr>
<th></th>
<th>2008 Time Series</th>
<th>2008 Panel</th>
<th>2006 Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Calls</td>
<td>.35</td>
<td>.13</td>
<td>-.02</td>
</tr>
<tr>
<td>Log Number of Calls</td>
<td>.41</td>
<td>.28</td>
<td>.21</td>
</tr>
<tr>
<td>High Calls</td>
<td>.45</td>
<td>.50</td>
<td>.25</td>
</tr>
<tr>
<td>Broken Appointment</td>
<td>—</td>
<td>—</td>
<td>.28</td>
</tr>
<tr>
<td>Two Broken Appointments</td>
<td>—</td>
<td>—</td>
<td>.31</td>
</tr>
<tr>
<td>Response Propensity</td>
<td>—</td>
<td>—</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note: In 2006 Pilot Study, polychoric correlations between broken appointments and other variables are: .68 (calls), .72 (log calls), .74 (high calls), and -.08 (response propensity). Polychoric correlations between two broken appointments and other variables are: .71 (calls), .82 (log calls), .93 (high calls), and -.17 (response propensity).
## Table 4: Percentage of Statistically Significant Difference-in-Difference Estimates

<table>
<thead>
<tr>
<th>Measure of Responsiveness</th>
<th>Refusal Conversion</th>
<th>Number of Calls</th>
<th>Log Number of Calls</th>
<th>High Calls</th>
<th>Broken Appointment</th>
<th>Two Broken Appointments</th>
<th>Combined Measure</th>
<th>Response Propensity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Items with Significant Main Order Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Studies (69)</td>
<td>2.9%</td>
<td>5.8%</td>
<td>11.6%</td>
<td>7.2%</td>
<td>9.1%</td>
<td>9.1%</td>
<td>10.1%</td>
<td>9.1%</td>
</tr>
<tr>
<td>2008 ANES Time Series (face-to-face) (39)</td>
<td>0.0</td>
<td>5.1</td>
<td>7.7</td>
<td>5.1</td>
<td>——</td>
<td>——</td>
<td>5.1</td>
<td>——</td>
</tr>
<tr>
<td>2008 ANES Panel (Web) (19)</td>
<td>0.0</td>
<td>5.3</td>
<td>15.8</td>
<td>15.8</td>
<td>——</td>
<td>——</td>
<td>21.1</td>
<td>——</td>
</tr>
<tr>
<td>2006 ANES Pilot (phone) (11)</td>
<td>18.2</td>
<td>9.1</td>
<td>18.2</td>
<td>0.0</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1</td>
<td>9.1%</td>
</tr>
<tr>
<td><strong>All Items</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Studies (194)</td>
<td>5.2%</td>
<td>3.6%</td>
<td>5.2%</td>
<td>5.7%</td>
<td>2.3%</td>
<td>4.5%</td>
<td>4.1%</td>
<td>6.8%</td>
</tr>
<tr>
<td>2008 ANES Time Series (face-to-face) (103)</td>
<td>2.9</td>
<td>2.9</td>
<td>4.9</td>
<td>4.9</td>
<td>——</td>
<td>——</td>
<td>2.9</td>
<td>——</td>
</tr>
<tr>
<td>2008 ANES Panel (Web) (47)</td>
<td>4.3</td>
<td>4.3</td>
<td>6.4</td>
<td>6.4</td>
<td>——</td>
<td>——</td>
<td>8.5</td>
<td>——</td>
</tr>
<tr>
<td>2006 ANES Pilot (phone) (44)</td>
<td>11.4</td>
<td>4.5</td>
<td>4.5</td>
<td>6.8</td>
<td>2.3</td>
<td>4.5</td>
<td>2.3</td>
<td>6.8%</td>
</tr>
<tr>
<td><strong>Items with Full Samples</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Studies (78)</td>
<td>6.4%</td>
<td>3.8%</td>
<td>6.4%</td>
<td>6.4%</td>
<td>6.3%</td>
<td>0.0%</td>
<td>7.7%</td>
<td>6.3%</td>
</tr>
<tr>
<td>2008 ANES Time Series (face-to-face) (15)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>6.7</td>
<td>——</td>
<td>——</td>
<td>6.7</td>
<td>——</td>
</tr>
<tr>
<td>2008 ANES Panel (Web) (47)</td>
<td>4.3</td>
<td>4.3</td>
<td>6.4</td>
<td>6.4</td>
<td>——</td>
<td>——</td>
<td>8.5</td>
<td>——</td>
</tr>
<tr>
<td>2006 ANES Pilot (phone) (16)</td>
<td>18.8</td>
<td>6.3</td>
<td>12.5</td>
<td>6.3</td>
<td>6.3</td>
<td>0.0</td>
<td>6.3</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

Note: The numbers in parentheses indicate the number of questions containing a response order manipulation. For example, at the top left, a total of 194 questions across all studies were examined for response order effects. “Combined Measure” for 2008 ANES Time Series and 2008 ANES Panel datasets indicates respondents coded “1” for either refusal conversion or high calls. “Combined Measure” for 2008 ANES Pilot dataset indicates respondents coded “1” for either refusal conversion, high calls, or broken appointment.
<table>
<thead>
<tr>
<th></th>
<th>2008 Time Series</th>
<th>2006 Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refusal Conversion</td>
<td>.495***</td>
<td>-.026</td>
</tr>
<tr>
<td></td>
<td>(.120)</td>
<td>(.175)</td>
</tr>
<tr>
<td>Number of Calls</td>
<td>.015*</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td>(.003)</td>
</tr>
<tr>
<td>Log Number of Calls</td>
<td>.207*</td>
<td>.091**</td>
</tr>
<tr>
<td></td>
<td>(.088)</td>
<td>(.037)</td>
</tr>
<tr>
<td>High Calls</td>
<td>.122</td>
<td>.317**</td>
</tr>
<tr>
<td></td>
<td>(.097)</td>
<td>(.102)</td>
</tr>
<tr>
<td>Broken Appointment</td>
<td>———</td>
<td>.138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.097)</td>
</tr>
<tr>
<td>Two Broken Appointments</td>
<td>———</td>
<td>.223*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.115)</td>
</tr>
<tr>
<td>Response Propensity</td>
<td>———</td>
<td>.97**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.35)</td>
</tr>
<tr>
<td>Combined Measure</td>
<td>.230**</td>
<td>.241**</td>
</tr>
<tr>
<td></td>
<td>(.079)</td>
<td>(.091)</td>
</tr>
<tr>
<td>N</td>
<td>103</td>
<td>44</td>
</tr>
</tbody>
</table>

***p<.001; **p<.01; *p<.05 (one-tailed)
Table 6: Coefficients and Standard Errors from OLS Regressions Predicting Interview Length (minutes)

<table>
<thead>
<tr>
<th></th>
<th>2008 Time Series</th>
<th>2006 Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refusal Conversion</td>
<td>-.47 (1.14)</td>
<td>-4.11***</td>
</tr>
<tr>
<td>Number of Calls</td>
<td>-.21** (.08)</td>
<td>-.11***</td>
</tr>
<tr>
<td>Log Number of Calls</td>
<td>-3.55** (1.22)</td>
<td>-1.84***</td>
</tr>
<tr>
<td>High Calls</td>
<td>-2.66** (.83)</td>
<td>-4.17***</td>
</tr>
<tr>
<td>Broken Appointment</td>
<td>———</td>
<td>-3.50***</td>
</tr>
<tr>
<td>Two Broken Appointments</td>
<td>———</td>
<td>-4.62***</td>
</tr>
<tr>
<td>Response Propensity</td>
<td>———</td>
<td>-4.53*</td>
</tr>
<tr>
<td>Combined Measure</td>
<td>-2.43** (.83)</td>
<td>-4.18***</td>
</tr>
<tr>
<td>N</td>
<td>2310</td>
<td>675</td>
</tr>
</tbody>
</table>

***p<.001; **p<.01; *p<.05 (one-tailed). For 2008 ANES Pilot, interviews longer than 170 minutes were dropped.
Table 7: Coefficients and Standard Errors from OLS Regressions Predicting “Missingness”

<table>
<thead>
<tr>
<th></th>
<th>2008 Time Series</th>
<th>2008 Panel</th>
<th>2006 Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refusal Conversion</td>
<td>.0057**</td>
<td>.0059*</td>
<td>.0026*</td>
</tr>
<tr>
<td></td>
<td>(.0018)</td>
<td>(.0031)</td>
<td>(.0012)</td>
</tr>
<tr>
<td>Number of Calls</td>
<td>-.0001</td>
<td>.0002*</td>
<td>-.00001</td>
</tr>
<tr>
<td></td>
<td>(.0001)</td>
<td>(.0001)</td>
<td>(.00002)</td>
</tr>
<tr>
<td>Log Number of Calls</td>
<td>-.0029</td>
<td>.0013**</td>
<td>-.00006</td>
</tr>
<tr>
<td></td>
<td>(.0019)</td>
<td>(.0005)</td>
<td>(.00030)</td>
</tr>
<tr>
<td>High Calls</td>
<td>-.0035**</td>
<td>.0017*</td>
<td>.00003</td>
</tr>
<tr>
<td></td>
<td>(.0013)</td>
<td>(.0010)</td>
<td>(.00075)</td>
</tr>
<tr>
<td>Broken Appointment</td>
<td>——</td>
<td>——</td>
<td>.0010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0008)</td>
</tr>
<tr>
<td>Two Broken Appointments</td>
<td>——</td>
<td>——</td>
<td>-.0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0009)</td>
</tr>
<tr>
<td>Response Propensity</td>
<td>——</td>
<td>——</td>
<td>.0023</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0021)</td>
</tr>
<tr>
<td>Combined Measure</td>
<td>-.0023*</td>
<td>.0021*</td>
<td>.0012</td>
</tr>
<tr>
<td></td>
<td>(.0013)</td>
<td>(.0010)</td>
<td>(.0008)</td>
</tr>
<tr>
<td>N</td>
<td>2310</td>
<td>1646</td>
<td>675</td>
</tr>
</tbody>
</table>

***p<.001; **p<.01; *p<.05 (one-tailed)