An Examination of the Bias Effects with a Two-Phase Address-Based Sample

Daifeng Han[1], Jill M. Montaquila[1], Douglas Williams[1], J. Michael Brick[1]
[1] Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

Address-based sampling (ABS) has emerged as a promising alternative to RDD telephone surveys. Little is known about the effectiveness of the various procedures for contacting the households and administering ABS surveys in reducing potential nonresponse bias. This paper is based on an ABS study involving two phases of data collection. Mail was the primary mode of collection for each phase, with limited telephone follow-up and several experiments designed to explore the effect on the response. Using the level-of-effort analysis, we examine how the response rates, demographic and socio-economic characteristics, and key survey estimates change with the screener follow-up attempts. The results compare the extent to which the screener follow-up strategies increased the effective coverage of the target subpopulation and affected the estimates in a two-phase survey setting.

Key words: Two-phase, address-based sampling, nonresponse bias, response rate

1. Introduction

In recent years address-based sampling (ABS) using address lists derived from the US Postal Service (USPS) Computerized Delivery Sequence (CDS) file has emerged as an alternative to telephone random digit dialing (RDD) in response to the declining coverage and response rates of landline RDD surveys (Battaglia et al., 2008; Cantor et al., 2008; Han et al., 2010; Brick et al., submitted). There are multiple approaches to contact the households once the ABS sample has been selected. For example, one option is to mail the survey invitation and/or questionnaire to the sampled addresses. Another possibility is to match telephone numbers to the sampled addresses to the extent possible and attempt telephone contact/interview with the matched cases, and mail to the non-matched cases. Little is known about the effectiveness of the various procedures for administering ABS surveys in reducing potential nonresponse bias. This paper examines the response rates and potential bias effects in a multi-mode ABS study involving two phases of data collection.

Our research uses data from the 2009 Pilot Study of the National Household Education Surveys Program (NHES) redesign. NHES is sponsored by the National Center for Education Statistics (NCES) to complement its institutional surveys; it is the principal mechanism for addressing topics that cannot be addressed in institutional data collections. By collecting data directly from households, NHES allows NCES to gather data on a wide range of topics. The NHES surveys have been conducted by Westat approximately every other year from 1991 through 2007, and each of these administrations used landline RDD sampling with computer assisted telephone interviewing (CATI). The CATI administration allowed within-household sampling of the target subpopulation with an

immediate transition between screening and more detailed topical interview during one telephone contact.

Like many other telephone surveys, NHES experienced a decline in response rates and coverage rates. NHES screener response rates dropped from above 80 percent in the early 1990s to 53 percent in 2007. Declines were also observed in the topical interview response rates during this period. Meanwhile, with the increasing prevalence of households having only cell phone service, landline telephone coverage rates have declined from approximately 93 percent of households in early 2004 to approximately 74 percent of households in the second half of 2009 (Blumberg and Luke, 2010). These changes prompted the need to examine alternatives to the existing landline RDD survey method. As a result, NCES and Westat, with input from a methodological expert panel, collaborated to formulate an alternative approach – a two-phase multi-mode data collection design with an ABS sample. The 2009 Pilot Study was predominantly methodological in nature; it was the initial evaluation of the feasibility of such a design before a large-scale Field Test is conducted in the spring of 2011.

The remainder of this paper is organized as follows: Section 2 describes the sample selection and data collection procedures in greater detail. Section 3 lays out the research questions and analysis methods. In Section 4, we report response rates and compare the extent to which each screener attempt helped reach the target subpopulation of the study. Section 5 assesses the potential impact of the screener follow-up attempts on survey estimates including demographics, socio-economic characteristics and key topical measures. Section 6 summarizes the findings.

2. Study Design of the 2009 Pilot Study

The 2009 Pilot Study targeted households with children aged 0 to 20 who have not yet started kindergarten or are in kindergarten through grade 12. The need to screen for households with eligible children and to sample from the list of enumerated children dictated a two-phase design with separate screening and topical questionnaires. In the first phase, all the sampled addresses were sent a short screening questionnaire. Then only those households that completed the screener and enumerated children were eligible for the second-phase sampling and the topical survey. Mail was the primary mode of data collection in each phase, with a small subsample of cases assigned for telephone follow-up. The Pilot Study also included several other experiments that might affect the cost of and response to NHES, including different screener versions, various incentive levels, and different mailing services (FedEx versus priority mail). The data collection period for the Pilot Study was from early September 2009 through late December 2009.

2.1 Sample Selection

The Pilot Study included three independent samples – a nationally representative sample of 10,200 addresses, a supplemental targeted sample of 800 additional addresses that were each identified by a vendor as containing households with children, and a supplemental sample of 800 addresses drawn from high-density linguistically isolated (Spanish-speaking) areas (census tracts). This research uses the data only from the nationally representative sample. The first-phase/screener sample was drawn from a file of residential addresses (including P.O. Boxes) maintained by a vendor, based on USPS CDS file. To accommodate the use of telephone follow-up, the sampled addresses were reverse-matched to landline telephone directory listings; matches were obtained for 57 percent of the sampled addresses.

For the second/topical phase, one child was sampled from each screener responding household with eligible children. Depending on the age and school status of the sampled child, the household was administered either the Early Childhood Program Participation (ECPP) Survey or the Parent and Family Involvement in Education (PFI) Survey. The ECPP Survey targeted children aged 7 or younger who were not yet enrolled in kindergarten. The PFI population included children and youth enrolled in kindergarten through 12th grade or homeschooled for these grades, with an age limit of 20 years. The topical surveys were administered to a parent/guardian in the household who is knowledgeable about the care and educational experiences of the sampled child.

2.2 Data Collection Approach and Embedded Experiments

This section provides details of the data collection procedures used for the screener and the topical surveys. The screener process is illustrated in Exhibit 1. The initial screener contact was a short mail questionnaire with a $2 cash incentive. Following a thank-you/reminder postcard (sent to all the sampled addresses), a random subsample of the nonresponding cases with matched landline telephone numbers was assigned to receive future screener contact by telephone and the remaining nonresponding cases were treated with a second screener mailing.

If the cases did not respond to the second screener mailing, the third screener contact attempt was made either via FedEx mailing or by telephone. Similar to the second screener attempt, a subsample of the cases with matched landline telephone numbers was assigned for telephone follow-up at this stage. For the remaining cases, the third screener contact was by FedEx delivery.

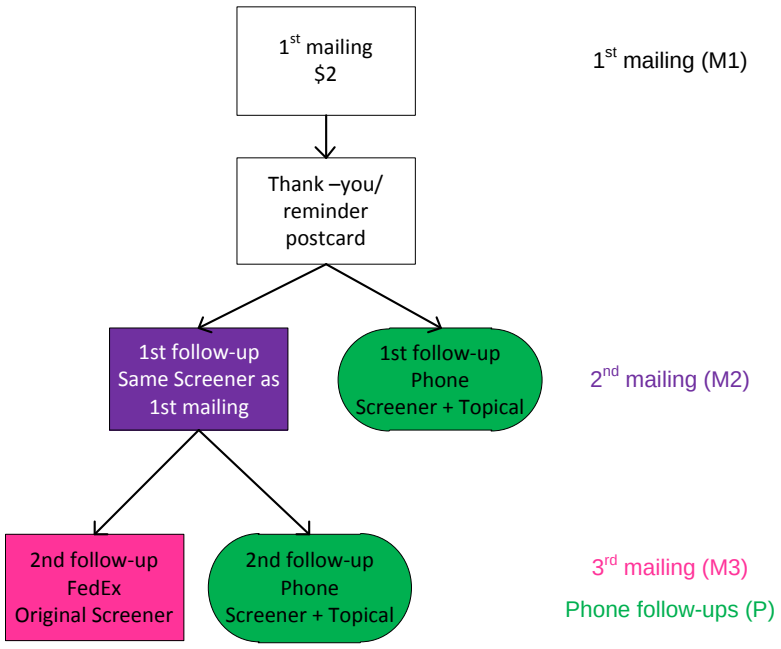

Exhibit 1: Pilot Study Screener Data Collection

The mode assignment for the topical phase depended on how the case had completed the screener. For the cases that completed the screener by telephone, the administration of the topical survey was also by telephone. Otherwise the topical survey was conducted via three mailing attempts and a final telephone follow-up.

Several other experiments were embedded in the data collection procedures, including multiple versions of the screener, different incentive amounts for the topical interview, and delivery methods for the topical survey. Since the cases were randomly assigned into different experimental groups, we will not scrutinize the effect of each experimental condition, but focus on the overall impact of each screener attempt for the purpose of this research.

3. Research Methods

This research is based on the nationally representative sample of 10,200 addresses. Our analysis focuses on the impact of the screener attempts because the response rates for the two phases (see Section 4) indicated that the screener phase was the major source of nonresponse.

As shown in Exhibit 1, three attempts were made in the screener phase. We will refer to the cases in the mailing path as 1st Mailing (M1), 2nd Mailing (M2), and 3rd Mailing (M3) in the sections below. Due to the small sample sizes of the two subsamples for telephone follow-ups, we combined them into one group in the analysis, which will be referred to as Phone (P).

It is worth mentioning that pure categories do not capture the complex of the process. For example, a household might respond to the 1st Mailing after it had been designated for the 2nd Mailing. The designation took place more than two weeks after the 1st Mailing and several days before the 2nd Mailing in order to allow enough time for preparing the mailing packages. In this scenario, the household would be categorized as M2 respondent rather than M1 respondent. That is, the grouping of the cases is based on the assignment of follow-up attempts, not the actual date or mode of response. The same principle applies to later screener attempts and telephone follow-ups. Some cases in the Phone group might return the screener by mail after the case was assigned for telephone follow-up. Then the case would be categorized as screener Phone respondent and the topical interview would be attempted by telephone.

For the bias effect analysis, we used the level-of-effort approach (Curtin et al., 2000) to compare the respondents to 1st Mailing, 2nd Mailing, 3rd Mailing, and Phone. The research attempted to answer the following questions:

- In the screener phase, what was the impact of mail follow-ups? How did the later mail respondents differ from the earlier mail respondents?
- In the screener phase, what was the added value of telephone follow-ups? Did the telephone follow-ups help bring in respondents with different characteristics?
- What was the impact of screener follow-ups on the topical phase? Among the screener respondents obtained with different levels of effort, how did their topical estimates differ?

Approximately 60 variables were examined, including demographics, socio-economic characteristics, and key measures in the ECPP and PFI surveys. The Rao Scott adjusted

Chi-squared test of association was used to detect universal differences across the screener attempt groups. In Section 5 we will present the estimates for a variable if and only if the universal comparison of the estimates across the analysis groups was statistically significant at 5 percent level.

4. Response Rates and Coverage of Target Households

In this section we consider the effects of screener attempts on response rates and coverage of target households. The overall screener response rate was 59 percent, which is 6 percentage points higher than the 2007 RDD screener response rate. The topical response rate among households with eligible children was 75 percent, which is comparable to the 2007 RDD result.

Households with matched telephone numbers responded to the screener at a significantly higher rate than those without, so we weighted the M2 and M3 cases with matched telephone numbers to account for the subsampling for mail follow-ups. The weighting allowed us to examine what the expected response rates and the estimates would have been if all the cases had been followed up by mail.

Table 1 shows the response rates and estimated coverage rates of target households among M1, M2, M3, and P groups. The second and the last columns are the unweighted counts of the fielded deliverable cases and the final topical yield corresponding to a particular screener attempt, respectively. Three interesting patterns can be observed. First, the proportions of households with eligible children were 33.3 percent and 35.9 percent among the M2 and M3 respondents compared to 27.4 percent among the M1 respondents. Mail screener follow-ups picked up a higher proportion of households with eligible children than the initial mailing. Second, mail follow-ups outperformed telephone follow-ups in terms of both screener coverage of target households (33.3 percent and 35.9 percent versus 28.7 percent) and topical response rate (77.4 percent and 61.0 percent versus 44.2 percent). Finally, the "Topical Response Rate" column shows that later screener respondents had significantly lower response rates in the topical phase than the initial screener respondents.

Table 1: Response Rates and Coverage of Households with Eligible Children by Screener Treatment/Response Group

Screener Group	*Number of Fielded Deliverable Cases*	*Conditional Screener Completion Rate**	*Among Screener Responding Households: Proportion with Eligible Children*	*Topical Response Rate*	*Number of Completed Topical Interviews*
1st mailing (M1)	9,121	25.4%	27.4%	87.5%	545
2nd mailing (M2)	6,016	25.2%	33.3%	77.4%	373
3rd mailing (M3)	3,691	30.7%	35.9%	61.0%	204
Phone follow-ups (P)	1,619	28.0%	28.7%	44.2%	57

* Conditional Screener Completion Rate is the proportion of responding households among all the deliverable cases fielded in the Screener Group.

5. Effects of Screener Follow-ups on Estimates

In this section we examine the characteristics of the respondents obtained with different levels of screener effort. Estimates were obtained for four mutually exclusive groups (or combinations of them) – those who initially responded to the 1st Mailing (M1), to the 2nd Mailing (M2), to the 3rd Mailing (M3), and to the Phone (P) follow-ups, respectively. For each analysis group, the respondents were weighted up to represent all the cases that had been fielded for the particular follow-up attempt(s). Section 5.1 compares the respondents in M1, M2 and M3. Section 5.2 combine all the mail respondents into one group and assesses whether the telephone follow-ups helped bring in different types of respondents compared to mail. In Sections 5.3 and 5.4, we address the topical phase and examine the impact of screener follow-ups on the final estimates – household characteristics (Section 5.3), child characteristics and key ECPP and PFI measures (Section 5.4).

5.1 Comparison across Mailing Groups

This section compares the characteristics of later mail respondents to the characteristics of earlier mail respondents under the assumption that the later respondents were more similar to the non-respondents (sometimes referred to as the continuum of resistance model). This comparison gives an indication of potential nonresponse bias. Since the 2009 Pilot Study was methodological in nature, we will not share the actual values of the preliminary estimates, but present the results in a "standardized" form – the ratio of the estimate for a particular group of respondents to the estimate used as the comparison base. In Table 2, the comparison base is the 1st Mailing (M1). We evaluate the impacts of later mailings by examining the ratios of the M2 estimates to the M1 estimates (M2/M1) as well as the ratios of the M3 estimates to the M1 estimates (M3/M1).

Table 2 shows that a lower proportion of later screener respondents had a matched telephone number. This is consistent with the finding that cases with matched telephone numbers had higher overall response rate than those without. Since the reverse-matching could be achieved only for landline telephone numbers, those without matched telephone numbers might be households with only cell phones or might have some other characteristics such as highly mobile associated with response propensity.

Previously we observed that mail follow-ups picked up a higher proportion of households with eligible children than the initial mailing. Table 2 shows that this pattern is reflected in both the ECPP and the PFI domains. For example, the proportion of households with ECPP children among the M3 respondents is 1.55 times the proportion of households with ECPP children among the M1 respondents.

Previous studies have found that household surveys tend to under-represent population subgroups with lower socio-economic status (SES) (Groves and Couper 1998; Picavet 2001). The last three rows in Table 2 indicate that later mail responding households were less likely to be highly educated, more likely to rent, and more likely to have more than 5 persons living in the household. The mail follow-ups might help reduce potential nonresponse bias by converting a higher proportion of less educated households into respondents.

Table 2: Effects of Mail Follow-ups on Screener Estimates

Household Characteristics among Screener Respondents	*M2/M1*	*M3/M1*
Percent of households with matched telephone numbers	0.93	0.90
Percent of households with ECPP children	1.26	1.55
Percent of households with PFI children	1.21	1.27
Percent of households whose highest education was graduate degree	0.94	0.74
Percent of households that were renting	1.24	1.51
Percent of households with 5+ persons	1.14	1.77

5.2 Comparison of Phone Group to Mail Groups

While telephone follow-ups were not as effective as mail follow-ups in increasing response rate and coverage of households with eligible children, here we examine whether telephone follow-ups helped bring in different types of respondents, particularly those with lower SES.

For this analysis, we first generated the so-called "mail-only" estimates using all the mail respondents and their corresponding weights. The mail-only estimates represent the expected results that would have been obtained if the survey had been conducted using a pure mail approach. We compared the characteristics of the respondents obtained via telephone follow-ups to the mail-only estimates (M123). Table 3 lists the ratios between the two sets of estimates (P/M123). Compared to a mail-only approach, adding telephone follow-ups yielded lower proportions of households that had a P.O. Box address, were renters, or had no more than high-school education. Based on these measures, telephone follow-ups did not appear to bring in lower SES groups. Partly this may be due to the reverse-matching process only matching addresses to landline telephone numbers, and thus not covering the cell phone-only population.

Table 3: Effects of Telephone Follow-ups (Relative to Mail-Only) on Screener Estimates

Household Characteristics among Screener Respondents	*P/M123*
Percent of P.O. Box addresses	0.26
Percent of households that were renting	0.70
Percent of households with highest education less than or equal to high school	0.93

5.3 Household Characteristics among Topical Respondents

As shown in Table 4, among the topical responding households, there were significant differences in household size, existence of non-English-speaking adults, and household telephone status across the screener groups (M1, M2, M3, and P). The mail screener follow-ups increased the proportion of children living in larger households and also increased the households with non-English-speaking adult(s) among the topical respondents.

Among the topical telephone respondents, although the proportion of children living in larger households was nominally higher (shown in the last column of Table 4), the difference between the telephone and the mail respondents was not statistically significant. Telephone follow-ups brought in a lower proportion of non-English-speaking households and a larger proportion of landline-mostly households, but this was partially because telephone interviews were conducted only in English and via landline telephone.

The impact of screener follow-up mailings on education and tenure status observed at the screener phase (in Table 3) did not carry forward to the topical phase; the estimates for

these two measures are not presented in Table 4 because the universal difference across the analysis groups was not statistically significant. This may be because a larger proportion of the screener respondents obtained via follow-up mailings did not result in a complete topical interview.

Table 4: Effects of Screener Follow-ups on Topical Household Characteristics

Household Characteristics in Topical Surveys	*M2/M1*	*M3/M1*	*P/M1*
Percent of households with 5+ persons	1.02	1.46	1.56*
Percent of households with non-English-speaking adult(s)	1.64	1.79	0.62
Phone status was "landline mostly"	0.91	0.87	1.53

*The difference between phone estimate (P) and the overall mail estimate (M123) is not statistically significant.

5.4 Child Characteristics and Key ECPP/PFI Measures

The final step of the evaluation focused on child characteristics and key estimates in the topical surveys. This analysis involves the four groups shown in Exhibit 2. $M1_t$ denotes the topical respondents who had completed the 1st screener mailing; $M12_t$ stands for the topical respondents who had responded either the 1st or the 2nd screener mailing; $M123_t$ refers to all the mail respondents; and M&P indicates all the respondents via mail and telephone.

Aggregated Group	*Includes Topical Respondents Who Had Responded to:*
$M1_t$	*1st screener mailing*
$M12_t$	1st screener mailing or *2nd screener mailing*
$M123_t$	1st screener mailing or 2nd screener mailing or *3rd screener mailing*
M&P	Any screener mailing or *telephone follow-ups*

Exhibit 2: Aggregated Groups for Estimating Child Characteristics and Key ECPP/PFI Measures

In order to gauge the extent to which the estimates could change as the result of each additional screener follow-up attempt, we weighted the respondents in the last batch (shown in italicized font in Exhibit 2) of each aggregated group to represent all the cases that had not responded to the previous screener attempt. Often, this weighting method could cause large variation in weights and undermine the power of statistical tests. If the goal of the Pilot Study had been to generate estimates, we would have designed the follow-up treatments differently. However, since the Pilot Study was strictly methodological, this approach allows us to see the potential value of each additional screener attempt. In order to estimate child characteristics, the weights also account for the within-household selection of a child.

As described in Section 2.1, the ECPP Survey was for preschoolers and had key measures such as choices of child care, child health and disability, and literacy and numeracy items (e.g., number of books at home, whether the parents read to the child every day, and whether the preschooler recognized all letters). The PFI Survey was for school-aged children and the estimates included parents' participations in school activities, parent involvement with children's homework, parent involvement in educational activities outside school, parents' satisfaction about the school, and parents' perception of child's academic performance. We examined 15 measures in the ECPP Survey and 25 measures in the PFI Survey, including both demographic and socio-economic characteristics of the

child as well as key statistics from the ECPP and PFI surveys. The comparisons of the four aggregated groups are summarized in Table 5. The three right-hand-side columns correspond to the standardized estimates for $M12_t$, $M123_t$ and M&P using $M1_t$ as the base of comparison. Asterisks denote situations in which the estimate for a particular aggregated group significantly differs from that for the previous aggregated group.

As shown in the column $M12_t$ /$M1_t$ in Table 5, the 2nd screener mailing brought in more minorities and more children who spoke a language other than English. The ECPP children from 2nd screener mailing were less likely to know the alphabet well. For the PFI children, the estimates change significantly as the result of 2nd screener mailing. For example, a lower proportion of children obtained via the 2nd screener mailing had moved to attend the current school, but their parents seemed to engage them with more extracurricular activities; the results give no clear indication of the impact of the 2nd screener mailing on the potential bias.

In Section 4, we noted that the 3rd screener mailing, which used FedEx, had a significant impact on the screener yield. The column $M123_t$ /$M1_t$ in Table 5 shows that the impact of the 3rd screener mailing was attenuated at the topical level; there were some changes in the estimates, but no systematic pattern was found. This could be due to the nature of the very late screener respondents – they were more likely to be nonrespondents in the topical phase. Another hypothesis is that the FedEx delivery might have helped bring in some lower SES groups that would not have been converted through regular mail in the screener phase, but these cases were less likely to complete the topical interview.

Finally, the column M&P/$M1_t$ in Table 5 indicates that telephone follow-ups had no significant impacts on the estimates, partly due to small number of topical completes obtained by telephone.

Table 5: Effects of Screener Follow-ups on Topical Child Characteristics

Child Characteristics in Topical Surveys	*$M12_t$ /$M1_t$*	*$M123_t$ /$M1_t$*	*M&P/$M1_t$*
Demographic/Socio-economic Variables			
Percent of non-Hispanic white alone	0.89*	0.96	0.96
Percent speaking English only	0.96*	0.97	0.98
Percent with a male guardian in household	0.97	1.03*	1.01
ECPP Variables			
Percent able to recognize all/most of alphabet	0.73*	0.99	0.93
PFI Variables			
Percent attending non-public schools	0.88	0.63*	0.79
Percent having moved to attend current school	0.74*	0.81	0.85
Percent feeling very satisfied with current school	0.96	1.09*	1.06
Percent having been contacted due to child doing well/better	1.28*	1.15	1.10
Percent having had 3+ extracurricular activities past month	1.21*	1.11	1.11

6. Conclusions

The 2009 NHES Pilot Study demonstrated that ABS with a two-phase primarily mail-based data collection approach is a promising alternative to landline RDD. Our study focused on the impact of screener effort. It shows that mail follow-up outperformed telephone follow-up in reaching the target subpopulation – households with preschoolers

or school-aged children. Mail screener follow-up also tended to bring in screener respondents with lower socio-economic status, which might help reduce potential nonresponse bias, while landline telephone follow-up did not have this effect. At the topical phase, the impact of screener follow-up was attenuated, yet there were still some significant changes in the key survey estimates as the result of additional screener effort.

The 2009 Pilot Study was the initial test of using a two-phase ABS design for NHES. This paper only covered the exploratory/preliminary evaluation of response rate and potential bias effects. The estimates considered here do not include weight adjustments aimed at reducing bias (e.g., calibration) that would normally be applied to the NHES survey weights. Therefore, these findings should be interpreted as indications of the potential for bias, prior to any reductions in bias realized through weight adjustments.

References

Battaglia, M., Link, M., Frankel, M., Osborn L., and Mokdad, A. 2008. An Evaluation of Respondent Selection Methods for Household Mail Surveys. In *Public Opinion Quarterly* 72 (3). 459-469.

Brick, M., Williams, D., and Montaquila, J. Address-based Sampling for Subpopulation Surveys. Submitted to *Public Opinion Quarterly*.

Blumberg, S., and Luke, J. 2010. Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July-December 2009 (5/12/2010). http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201005.pdf [retrieved September 20, 2010].

Cantor, D., Sigman, R., and Han, D. 2008. Pilot of a Mail Survey for the Health Information National Trends Survey, Paper presented at the 63th Annual Conference of the American Association for Public Opinion Research, May 15-18, 2008, New Orleans, LA.

Curtin, R., Presser, S., and Singer, E. 2000. The Effects of Response Rate Changes on the Index of Consumer Sentiment. In *Public Opinion Quarterly* 64 (4). 413-428.

Groves, R., and Couper, M. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

Han, D., Cantor, D., and Brick, P. 2010. Findings from a Two-Phase Mail Survey for a Study of Veterans. Paper presented at the 65th Annual Conference of the American Association for Public Opinion Research, May 13-16, 2010, Chicago, IL.

Picavet, H. 2001. National Health Surveys by Mail or Home Interview: Effects on Response. In *Journal of Epidemiology and Community Health*, 55. 408-413.