sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 631–645, http://dx.doi.org/10.2478/JOS-2020-0032

The Effects of Nonresponse and Sampling Omissions on Estimates on Various Topics in Federal Surveys: Telephone and IVR Surveys of Address-Based Samples

Floyd J. Fowler[1], *Philip Brenner*[1], *Anthony M. Roman*[1], *and J. Lee Hargraves*[1]

With declining response rates and challenges of using RDD sampling for telephone surveys, collecting data from address-based samples has become more attractive. Two approaches are doing telephone interviews at telephone numbers matched to addresses and asking those at sampled addresses to call into an Interactive Voice Response (IVR) system to answer questions. This study used in-person interviewing to evaluate the effects of nonresponse and problems matching telephone numbers when telephone and IVR were used as the initial modes of data collection. The survey questions were selected from major US federal surveys covering a variety of topics. Both nonresponse and, for telephone, inability to find matches result in important nonresponse error for nearly half the measures across all topics, even after adjustments to fit the known demographic characteristics of the residents. Producing credible estimates requires using supplemental data collection strategies to reduce error from nonresponse.

Key words: Mixed modes; address-based samples.

1. Introduction

The theory behind making estimates from sample surveys is fairly straightforward. Find a list or other way to give most people in your study population a chance to be selected. Draw a probability sample of people in the population, then find a way to get a high percentage of them to answer some survey questions. If that is done, the statistics based on answers that the surveyed sample gives should do a good job of describing the entire population.

Since the 1940s, and perhaps before, the gold standard for how to do this in the United States was to draw an area probability sample of housing units and send interviewers in person to the selected households to do surveys. This is still the approach used for important federal surveys such as the Current Population Survey, the National Crime Victimization Survey, and the National Health Interview Survey. However, such surveys are comparatively expensive. For many years, an acceptable alternative for many purposes was to do telephone surveys based on random-digit dialing. In the 1980s and 1990s, over 90% of housing units in the United States had telephone service, and techniques were

[1] Center for Survey Research, University of Massachusetts Boston, 100 Morrissey Blvd, Boston, MA 02125, U.S.A. Emails: floyd.fowler@umb.edu, philip.brenner@umb.edu, anthony.roman@umb.edu and lee.hargraves@umb.edu

Acknowledgments: This work was funded by the National Science Foundation, Grant No.1424433.

© Statistics Sweden

developed to efficiently sample housing units by randomly sampling telephone numbers (Waksberg 1978). Telephone interviewers could obtain response rates that, while usually lower than in-person interviewers, were considered quite respectable, and the resulting data were quite comparable to those from in-person interviews (Groves and Kahn 1979; Groves et al. 1988).

Then, in the last 20 years, two fundamental changes made telephone surveys more problematic. On the sampling side, the growth of cell phones and decline in the use of landlines made sampling people or households via sampling telephone numbers much more complicated. Blumberg and Luke (2016) provide a recent estimate of the distribution of land lines and cell phones in the United States showing that a majority of American households now have only cell phones. In parallel, a variety of factors led people to be much less willing to answer their telephones when they receive calls from "unfamiliar" numbers. As a result, response rates to telephone surveys plummeted (Curtin et al. 2005; Tourangeau and Plewes 2013; Kohut et al. 2012), while concerns about the comprehensiveness of random-digit-dialed based samples grew. These trends, in turn, have led researchers to explore alternative ways to survey general populations (AAPOR Task Force 2017).

One approach is to go back to address-based samples, which give almost everyone living in a housing unit a chance to be selected, and then to experiment with ways other than in-person interviews to collect data. Approaches to data collection can include mailing respondents requests to return mail questionnaires, asking them to go on the internet to complete surveys, or asking them to call an 800 telephone number to provide answers to an automated interviewer (Interactive Voice Response, IVR). Another approach is to try to find a landline or mobile telephone number that matches selected addresses and to conduct interviews at selected households by telephone. All of these approaches can be used in various combinations in mixed-mode surveys (Groves et al. 2009; De Leeuw et al. 2008; Dillman et al. 2014).

The purpose of this study was to learn about the effects of nonresponse on estimates on various topics when data are collected using two of these approaches: by having interviewers call telephone numbers matched to the extent possible with households in an address-based sample and when data are collected by mailing a request to those in an address-based sample to call a telephone number to do an interview with an automated interviewer (IVR). Note that asking households in an addressed-based sample to call in to do interviews is a different methodology from having a computer-based interviewing system calling telephone numbers, asking for an interview. For the telephone survey, error can stem from both an inability to match a telephone number with an address and from an inability to complete interviews by telephone when targeted respondents refuse or simply never answer the telephone. For the IVR mode, error stems from the fact that some people choose not to call the IVR number and do the survey.

Studies of the effect of nonresponse on survey estimates have shown that it is inconsistent (Groves 2006; Groves and Peytcheva 2008; Keeter et al. 2000; Kohut et al. 2012; Keeter et al. 2006). Some estimates from surveys with low response rates are quite biased, while others from the same surveys are similar to estimates from more reliable sources. Thus, one of the important questions for any survey designer is whether or not low response rates are likely to affect estimates in the particular subject area that is the focus of

the survey. These issues are particularly important to those designing US federal surveys who may be considering the value and problems associated with alternative ways of collecting data from household samples.

The primary goal of this study was to assess the extent to which nonresponse and, in the case of telephone surveys, the inability to match telephone numbers to addresses affected estimates using the telephone or IVR to survey addressed-based samples. In addition, we wanted to assess the extent to which any nonresponse effects were or were not related to common topics covered in major government surveys.

2. Methods

2.1. The Sample

Two unclustered probability samples of 1,500 residential addresses each were drawn by Marketing Systems Group (MSG) from an address-based sampling frame covering five Boston area communities. These contiguous communities, including three neighborhoods of Boston (Dorchester, Jamaica Plain, and Mattapan) and two of Boston's immediate suburbs (Milton and Quincy), were chosen based on their demographic diversity. The samples were drawn proportionate to the size of each of the five neighborhoods and cities.

2.2. Data Collection Protocols and Results

One of the samples was devoted to studying nonresponse to efforts to do telephone interviews and the other was devoted to studying nonresponses to IVR invitations. All survey response rates are based on AAPOR definitions (AAPOR 2016).

The sample provider, MSG, attempted to match either a landline or cell telephone number to each selected address in the telephone half of the sample. They were successful for about 60% of the addresses. About 97% of the numbers that were matched were landlines.

Addresses that were matched to a telephone number were sent a letter, accompanied by a 2-USD cash incentive, explaining the background and purposes of the survey, assuring confidentiality, and informing them that an interviewer would be calling in the next few days. Soon thereafter, professional interviewers, working from a central phone facility, called each household. If there was more than one person 18 or older living in the household, the interviewer followed a randomized protocol that chose either the oldest or youngest to be the designated respondent. Each telephone number was called a maximum of 12 times; the median and modal number of calls was six. Calls were placed on various days of the week, primarily during evenings and on weekends. All telephone interviews were conducted in English. In this part of the subsample, 143 respondents (20 percent response rate, AAPOR RR 3) completed the survey.

For the IVR part of the study, a sample of 1,500 addresses received a letter, accompanied by a 2-USD cash incentive, inviting residents to call a toll-free number to complete an IVR survey with an automated interviewer. Invitation letters were printed in English and addressed to the household by name (e.g., "The Smith Household") with "or current resident" added to the addressee in case the matched name was out-of-date or otherwise incorrect. These materials described the topic of the survey as focused on "health and our community" and advised residents that it would take approximately 10–15

minutes to complete. In parallel with the telephone protocols, invitation letters included wording that randomly selected either the youngest or oldest adult 18 years of age or older in the household if there was more than one potentially eligible adult. In this subsample, 148 respondents (10% response rate, RR1) completed the survey.

There were three groups of nonrespondents that we then attempted to interview in person:

1. Those from the telephone sample for whom we had numbers but were unable to interview (whom we refer to as telephone nonrespondents),
2. Those initially selected for the telephone sample for whom we could not find a telephone match and hence could not even attempt a telephone interview, and
3. Those who were invited to complete an IVR survey who did not do so (whom we refer to as IVR nonrespondents).

For each of these groups, the protocols were quite similar.

Approximately half of the addresses for which there was a telephone number (N = 350) but that did not respond to telephone interview efforts were randomly selected for nonresponse follow-up in-person interviews. They were sent a second letter informing them that an interviewer would be visiting their home to complete a personal interview and promising a 20-USD post-paid incentive upon completion of the interview. Each address was visited a maximum of 12 times; the median and modal number of visits was six. In this nonresponding telephone subsample, 128 respondents (43 percent response rate, RR3) completed the survey.

All but a small sample of the addresses that were not matched to a telephone number received a letter, generally similar to the telephone interview letter, informing them that an interviewer would be visiting their homes to complete a personal interview. A 20-USD post-paid cash incentive was promised upon completion of the interview. Addresses were visited a maximum of 13 times; the median and modal number of visits was six. All personal interviews were completed in English. The same within household selection protocol was used as in the telephone interviews. In this subsample, 166 respondents (41 percent response rate, RR3) completed the survey.

The protocol for the 336 addresses selected from the IVR nonrespondents was virtually the same as for the telephone nonrespondents, and the level of interviewer effort was similar as well. For this subsample, it was estimated that a total of 262 were in fact eligible for an interview, and 124 interviews were completed, with a response rate of 47% (RR3).

Data collection began in the fall of 2015 and was completed in April 2016. All the data collection procedures were approved by the university's Institutional Review Board. Table 1 summarizes the samples and these results.

2.3. *The Survey Questions*

All of the questions in the survey were drawn from ongoing US federal surveys: The National Health Interview Survey (NHIS), the Behavior Risk Factor Surveillance Survey (BRFSS), the Health Information National Trends Survey (HINTS), The Current Population Survey (CPS), the National Crime Victimization Survey (NCVS), and the American Community Survey (ACS). The goal was to assess the extent to which a propensity for nonresponse error was related to the topic of the survey questions. In all, there were 35 estimates that we made from the surveys, which we categorized as follows:

Table 1. Data collection results by sample.

	Telephone sample with telephone number match	Telephone sample with no telephone number match	Interactive Voice Response (IVR) sample
Number of initial addresses sampled	921	579	1500
Estimated eligible*	727	NA	1500**
Number of responses by initial mode (Telephone interview or IVR)	143	NA	148
Initial response rates (AAPOR RR3)	20%	NA	10%**
Number assigned for in-person interviews	350	537	336
Estimated eligible*	294	409	262
Number of in-person interviews completed	128	166	124
Response rate for in-person interviews (AAPOR RR3)	43%	41%	47%

*Ineligible addresses included those that were non-residential or that were vacant plus those in which all adult residents were found to be away for an extended period of time or in which no adult could be interviewed in English.

**Address eligibility could not be evaluated, since there was no contact other than a recruitment letter with the addresses during the IVR phase of the survey. As a result, the calculation of the response rate was AAPOR RR1, since there could be no adjustment for eligibility.

- **Health status**: six chronic health conditions plus BMI, self-rated health, and total number of chronic conditions reported
- **Health services received**: doctor visits, dentists' visits, flu shots,
- **Health risks**: health insurance, getting enough sleep, smoking,
- **Giving**: to charity, donating blood, volunteering,
- **Use of technology**: use of phones and computers,
- **Work and income**: Employment status, number hours worked, work loss due to health, on welfare,
- **Household characteristics**: type house, own/rent, number of vehicles owned,
- **Crime**: victim of crime in last 12 months.

The exact wording of the questions, and their sources, are provided in the online Supplemental material.

2.4. Analysis

The sampling methodology used in this study was an application of one originally proposed by Hansen and Hurwitz (1946). An original data collection methodology was performed uniformly for all eligible sample cases. All respondents to that original

methodology were considered to be part of one stratum in which cases were selected with certainty. All nonrespondents (either due to nonresponse or, in the case of the telephone sample, inability to match a telephone number to an address) were then considered part of additional strata. Simple random samples with known probability of selection of these three groups of nonrespondents were then selected with face-to-face interviews being attempted. Weights were constructed for all sample cases that took into account the original probabilities of selection, as well as all subsampling probabilities of selection. Therefore, variable weights were obtained dependent on whether the sample case responded originally or through the follow-up procedures.

First, we looked at how the demographic characteristics of the various samples (respondents and interviewed nonrespondents) compared with Bureau of the Census estimates based on the American Community Survey (ACS) data for the population living in the study area. We also put the data from the test mode and the in-person interviews with nonrespondents together, weighted to adjust for probabilities of selection, to see how the estimates derived from the combined data from the samples targeting telephone and IVR respectively compared with the ACS estimates.

Then, we took two basic approaches to looking at the effects of nonresponse and sample frame limitations on the estimates.

The first analysis (shown in Table 3) addressed three questions about how the substantive survey results compared:

1. How do telephone respondents compare with telephone nonrespondents, those for whom we had telephone numbers but were unable to complete an interview by telephone but whom we were able to interview in person?
2. How do those interviewed who lived at addresses for which there was a telephone number match compare with those interviewed who lived at addresses for which no telephone match was found?
3. How do the IVR respondents compare with the IVR nonrespondents who were interviewed in person?

To make these comparisons, we did *t*-tests or Z-tests on each of the 35 estimates from the survey to see if the estimates were or were not different ($p < .05$). In this case, we did not adjust for demographic differences between the groups because we wanted to see how those who could be surveyed by telephone or by IVR compared with those who could not in the various areas covered in the survey. Demographic differences do not necessarily translate into particular substantive differences. These comparisons also allowed us to separately see the effects of not responding when we had a telephone number and of not being able to match a telephone number to an address on the estimates based on telephone interviews.

The second analysis (shown in Table 4) addressed the question of whether or not estimates based on either the telephone interviews or the IVR responses lay within the 95% confidence intervals around the best estimates we could make using all the data we had collected. The best estimate for the telephone sample combined all the results from the telephone interviews, the in-person interviews with nonrespondents and the in-person interviews in households for which there was not a telephone match. To make these estimates, we, of course, weighted to adjust for the different probabilities of selection for the various strata as described above. After those weights were constructed, a

Table 2. Demographic characteristics by sample type for respondents and for population as a whole.

Demographic characteristics	Telephone respondents	Telephone non-respondents interviewed in person	Households with no telephone match interviewed in person	Combined telephone samples	IVR respondents	IVR non-respondents interviewed in person	Combined IVR samples	ACS estimates for population
Percentage white, non-Hispanic	70%***	56%**	38%	56%***	67%***	50%	52%**	44%
Percentage college graduates	51%***	50%***	40%	45%***	68%***	43%*	46%***	35%
Percentage female	61%*	48%	52%	51%	62%*	56%	57%	53%
Percentage never married	26%***	27%***	47%	36%***	31%***	41%	40%*	46%
Percentage 65 or older	56%***	29%***	14%	26%***	22%**	16%	17%	15%

Note: Statistical significance was calculated by Z-test on differences between estimates from each sample and the estimates based on the American Community Survey (ACS) for the study area: $*p < .05$, $**p < .01$, $***p < .001$

post-stratified adjustment was also performed at the community level using age, race/ethnicity, education, gender and marital status in order to make weighted estimates agree with known demographic profiles for the area under study derived from the American Community Survey data.

We did the same analysis for the IVR respondents. However, in this case we did two comparisons: 1) How the IVR results compared with the estimates based on the combination of responses for the IVR results plus the interviews with the IVR nonrespondents; and 2) how the IVR responses compared with the estimates from the telephone sample, combining data from all three components of the data collection for that sample.

3. Results

Table 2 presents the results of the analysis of the demographic comparisons. The telephone respondents are very different from the population. They are much more likely to be non-Hispanic white, to have graduated from college and particularly to be over age 65; they are much less likely to have never married. The telephone nonrespondents who were interviewed in person tended to differ from the population in the same ways, but the differences were much smaller. In contrast, those interviewed in person at households for which there was not a telephone match look quite similar to the population as a whole; there were no statistically significant differences between those who lacked a telephone match and the ACS estimates. When we put the data from all three sources together, the combined estimates from the telephone sample differ significantly from the ACS estimates on four of the five demographic characteristics: race/ethnicity, education, marital status and age.

Turning to the IVR data, the IVR respondents differ from the ACS estimates in ways that are similar to the telephone respondents in being more likely to be non-Hispanic white, less likely than the population to have been "never married" and they are even more likely than the telephone respondents to be college graduates. In contrast, IVR respondents look more like the population than telephone respondents with respect to age, although they are also significantly more likely to be 65 or older than the ACS estimate.

The IVR nonrespondents who were interviewed in person are more like the population than the IVR respondents with respect to all the demographic characteristics shown in Table 2. When the data from respondents and nonrespondents are combined, however, the estimates are significantly different from the ACS estimates with respect to race/ethnicity and education, but are similar in other respects.

Table 3 summarizes the comparisons between respondents and nonrespondents, and those for whom there was and was not a telephone match across the various survey topics. The bottom line is that phone respondents differed significantly from telephone nonrespondents who were interviewed in person on nine of 35 measures; those with phone matches differed from those without telephone matches on eight measures. There was a little overlap on which measures were affected, but not much. Overall, 13 of the 35 variables had statistically significant differences between the telephone respondents and those interviewed in person from either the telephone nonrespondents or those in households for which there was not a telephone number match, or both.

Table 3. Number of statistically significant differences ($P < .05$) between respondents and nonrespondents by mode and topic, without demographic.

Topic	Telephone respondents versus telephone nonrespondents interviewed in person	Telephone respondents plus telephone nonrespondents interviewed in person versus households with no telephone match interviewed in person	IVR respondents vs IVR nonrespondents interviewed in person	Total number of items
Health conditions	2	1	1	9
Health services received	1	0	2	3
Health risks	0	1	0	4
Giving	1	1	1	3
Use of technology	3	2	2	6
Work and income	2	1	0	6
Household characteristics	0	2	0	3
Crime	0	0	0	1
Total	9	8	6	35

Note: Estimates from the different groups were compared by *t*-tests. Details of estimates and results by individual items are in online Supplemental material Table A1.

IVR respondents differed significantly from nonrespondents who were interviewed in person on only six of the 35 measures.

Table 4 puts all these effects together. It addresses the question of how the best estimates from the telephone interviews or the IVR respondents alone would compare with an estimate that included data from the in-person interviews with nonrespondents. The first column shows the number of estimates that lie outside the 95% confidence interval of the best estimate when the telephone interviews alone are compared with the estimates when data from the telephone interviews are combined with the data from the in-person interviews with the telephone nonrespondents and with those in households for whom there was not a matched telephone number. The second column compares the IVR estimates with the estimates one would make combining the IVR responses with the data from the in-person interviews with IVR nonrespondents. The third column compares the IVR estimates with the estimates from the combined estimates from those we tried to interview by telephone. All the estimates in the table are adjusted to match the characteristics of the population as a whole in the study area based on ACS data for age, education, race/ethnicity, gender and marital status.

Because the data from the original mode are embedded in the combined estimates, a direct test of statistical significance is not advisable. Instead, the criterion in the table for a

Table 4. Number of estimates that lie outside the 95% confidence interval of best estimate by mode, adjusted for demographics to match population characteristics.

Topic	Telephone interviews versus telephone interviews plus in person interviews with telephone nonrespondents and those in households with no telephone match	IVR responses versus IVR responses plus in-person interviews with IVR nonrespondents	IVR responses versus telephone interviews plus in person interviews with telephone nonrespondents and those in households with no telephone match	Total number of items
Health conditions	3	0	4	9
Health services received	1	2	3	3
Health risks	3	2	2	4
Giving	1	1	0	3
Use of technology	5	1	1	6
Work and income	2	0	4	6
Household characteristics	3	2	3	3
Crime	0	0	0	1
Total	18	8	17	35

Note: For all 35 items, two estimates were created: One combined data from the telephone interviews with the in-person interviews with telephone nonrespondents and those at addresses for which there was not a telephone match. The other combined data from IVR responses and the in-person interviews with IVR nonrespondents. When combining data, weights were applied for different probabilities of selection and number of adults in the household. The estimates from the telephone interviews and the IVR responses, as well as the two combined sets of estimates, were all adjusted to match the age, education, race/ethnicity, gender and marital status of the adult population living in area according to the ACS. Then the adjusted estimates from the telephone and IVR respondents were compared with the confidence intervals around the two combined estimates. The counts in the table are the number of items, by type, that fell outside two standard errors around the combined estimates. The details of these analyses are in online Supplemental material Table A2.

"difference" is whether or not the point estimate from the telephone interviews or the IVR responses, when adjusted for demographic differences, lies within the 95% confidence interval of the combined comparison estimate. Since point estimates generally are presented and most likely used in practice, we felt this was an acceptable manner to determine if there were errors in the estimates that should be of concern.

For the telephone interviews, 18 out of the 36 estimates lay outside the confidence interval around the estimate based on the combined data. The IVR estimates only differed from the combined estimates that included data from the in-person interviews with IVR

nonrespondents on eight of the 36 estimates. However, when we compared the IVR estimates to the combined estimates from the telephone sample, (Column 3 of Table 4) 17 of the 36 estimates were outside the confidence interval.

The higher rate at which IVR estimates differ from the telephone combined estimates (17 times) as compared to how they differ from combined IVR estimates with nonrespondents included (eight times) is primarily due to the higher telephone sample sizes and hence smaller 95% confidence intervals around those estimates. It is not because the combined estimates for the telephone samples and the combined IVR samples are fundamentally different. In fact, when we did *t*-tests to compare the two combined estimates, there were only four of 35 estimates that were different at $p < .05$, and two of those (cell phone ownership and landline ownership) were likely related. It should be noted that when conducting 35 tests, all at $p < .05$, on average one would expect about two tests to show up as significant by chance (see online Supplemental material, Table A2). Thus, the two independent samples of the same population produced approximately the same overall estimates even though the methodologies used to collect the data differed.

Finally, when we look at the topics on which differences were observed, it is clear that they occur across all the topics. It would be very difficult from the results in Table 4 to conclude that there is any topic that is immune to the effects of nonresponse. Moreover, when we examined the details of the differences, which are presented in Tables A1 and A2 in the online Supplemental material, the patterns seem to be quite unpredictable.

Compared to our best estimates that included telephone interviews combined with data from in-person interviews, the telephone respondents reported higher BMIs, more diabetes, more work missed due to illness and more uninsured but fewer had depression, had gotten flu shots, reported getting enough sleep recently, or reported smoking. They were more likely to report giving to charity, less likely to be on welfare, more likely to live in a single-family house, less likely to rent and reported having more cars. With respect to technology, they reported more landlines, fewer smart phones, less use of phones for e-mails and more computers at home.

The IVR respondents, compared to our best combined estimates, reported lower BMI, fewer chronic conditions, less lung disease and less depression. They reported more visits to dentists, but fewer visits to doctors. They reported fewer flu shots and less smoking. They were more likely to be students, a lower percentage reported working and those who did worked fewer hours; they also missed fewer work days due to illness. Finally they were more likely in single-family homes, less likely to be renters, had more cars and more cell phones.

4. Discussion

This study found that telephone interviews matched to an address-based sample and IVR interviews conducted by asking sampled individuals to call in to an IVR number are biased both in terms of demographically representing the population of interest and in the accuracy of estimates of characteristics of the population, such as health and health care, use of technology, and employment. Clearly, the estimates based on telephone interviews at households matched to an address-based sample are likely to be problematic. About half the measures we tested were not good. The error was driven about equally by those who did not respond and by the fact that those for whom telephone number matches could not

be found were different from those for whom telephone matches could be found. The differences persisted after matching demographic characteristics to the ACS. The differences cut across all the topics we covered.

The IVR results are not dissimilar. Like the telephone respondents, the IVR respondents overrepresented non-Hispanic whites and college graduates. However, profiles for age and marital status were more similar to the population. The substantive results comparing respondents and nonrespondents appeared to be less problematic than the telephone results, as there were only six significant differences. Of course, a big advantage for IVR results is that they included addresses with and without matched telephone numbers, whereas telephone estimates were missing those for whom there was not a telephone match with the selected address. Moreover, the apparently higher response rate for the telephone of 20%, compared with the 10% rate for IVR, was actually not an advantage. When the fact that only 60% of households had a chance to be interviewed is considered, the effective telephone response rate was only 12%.

We found that only eight estimates from the IVR responses lay outside the confidence interval for the adjusted estimate when data from interviewed nonrespondents were added in. However, when we compared the IVR responses to the same estimates as we used for the telephone estimate, we found that 17 of 35 estimates fell outside the confidence interval–almost the same as the 18 from the telephone interviews.

Thus, from both samples the clear conclusion is that even after adjusting for demographic differences in respondents, close to half the estimates from both the IVR and telephone interviews lay two standard errors or more from our best estimates that reduced the effects of nonresponse. Moreover, the estimates of the number of items "seriously" affected by nonresponse is conservative, as there were additional items that were very close to the edges of the confidence intervals that are not in our counts.

It should be pointed out that our "best" estimates included a good deal of nonresponse; the effective response rates when all data are combined were not much over 50%. Indeed, when we combined data from the primary modes with data from the in-person interviews with nonrespondents, there remained significant differences in demographic characteristics between the survey-based estimates and the ACS estimates. However, even though we did not have external data other than demographic to directly evaluate the survey estimates, it would be difficult to argue that the follow-up interview data collected from nonrespondents and households without telephone numbers, which clearly moved the demographics of the samples closer to the ACS estimates, did not move all the estimates closer to the true values when they raised the percentage responding from about 10% to over 50%.

One could ask if changing modes from the primary mode to in-person interviewing could have affected the comparisons. That seems unlikely. The most likely sources of mode effects are that people would be more likely to give socially desirable answers to an actual interviewer than to an automated interviewer, which would show up in comparing IVR responses to responses to the follow-up interviewers. Of the four significant differences between IVR respondents and in-person interviews with nonrespondents that we thought might include an element of social desirability, three of them had those reporting to an automated interviewer giving fewer socially desirable answers: they reported worse health, fewer dentist visits and less giving to charity. The only answer that went the other way was they reported more flu shots.

Another possible limitation of the study is that we focused on one community area that may or may not produce more broadly representative results. The area was chosen for being diverse with respect to ethnicity, age and education. Nonetheless, we certainly urge others to try to collect data to assess the effects of nonresponse for different data collection modes, different research topics, and different populations.

We think these results contribute to our understanding of nonresponse in several ways. They certainly further strengthen the argument for the importance of using multiple data collection strategies to reduce nonresponse associated with a primary data collection mode. As has been shown before, nonresponse does not affect every survey estimate, but it affects many estimates across an array of subject areas in ways that are very difficult to anticipate (Groves 2006; Groves and Peytcheva 2008; Kohut et al. 2012; Keeter et al. 2006). Moreover, obvious adjustments for demographic anomalies in who responds do not do much to make the estimates better. There may be further adjustments that could have been done that would improve the estimates. However, given the heterogeneity of the effects observed and the lack of a gold standard for estimates other than demographics, we think the potential of adjustments of these data or the data in most surveys to eliminate nonresponse error is limited.

A second contribution is to specifically assess the importance of addressing both survey nonresponse and limitations in the sample frame when doing telephone surveys of addressed-based samples. Although nonresponse seems to have been the most important source of error, the differences between those who did and did not have telephone numbers matched to their addresses contributed important error as well.

Third, although the potential biases in who will actually do a telephone survey and which addresses can be matched to telephone numbers have both been observed, providing information showing that nonresponse to IVR has similar levels and types of nonresponse error is important. On the other hand, these data may also provide some encouragement to try IVR as one of several modes to collect data from addressed-based samples, providing that a nonresponse follow-up is planned.

In short, at a time when response rates are dropping for traditional survey protocols (De Leeuw and De Heer 2002; Tourangeau and Plewes 2013; AAPOR Task Force 2017), there is great pressure to accept low response rates and to use imperfect sample frames in the hope that the estimates will be "good enough". These data are another contribution to the argument that for many survey purposes there is no substitute for comprehensive sample frames and mixed-mode efforts to have a high percentage of a target population represented in survey results. We hope these results will further stimulate research on how best to obtain responses from high percentages of selected survey samples.

5. References

AAPOR. 2016. "Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys (Revised 2016)." American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed September 2019).

AAPOR Task Force. 2017. "The Future Of U.S. General Population Telephone Survey Research." 2017. Available at: https://www.aapor.org/Education-Resources/Reports/

The-Future-Of-U-S-General-Population-Telephone-Sur.aspx (accessed September 2019).

American Community Survey (ACS) Available at: https://www.census.gov/programs-surveys/acs/ (accessed June 2020).

Behavioral Risk Factor Surveillance Survey (BRFSS). Available at: https://www.cdc.gov/brfss/index.html (accessed June 2020).

Blumberg, S.J. and J.V. Luke. 2016. "Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, July-December 2016." Available at: https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201705.pdf (accessed September 2019).

Current Population Survey (CPS). Available at: https://www.census.gov/programs-surveys/cps.html (accessed June 2020).

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69(1):87–98. DOI: https://doi.org/10.1093/poq/nfi002.

De Leeuw, E.D., D.A. Dillman, and J.J. Hox. 2008. "Mixed-Mode Surveys: When and Why." In *International Handbook of Survey Methodology*, edited by Edith D. de Leeuw, Joop J. Hox, and Don A. Dillman, 299–311. New York: Erlbaum. Also European Association of Methodology. Available at: https://www.eam-online.org.

De Leeuw, E.D. and W. de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, edited by Groves Robert M., Don A. Dillman, John L. Eltinge, and Roderick J.A. Little: 41–54. New York: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons., Inc.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys: What Do We Know about the Linkage between Nonresponse Rates and Nonresponse Bias?" *Public Opinion Quarterly* 70(5):646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg, eds. 1988. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72(2):167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Groves, R.M. and R.L. Kahn. 1979. "Surveys by Telephone; a National Comparison with Personal Interviews." New York, N.Y.: Academic Press.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association*, 41(236):517–529. DOI: https://doi.org/10.1080./01621459.1946.10501894.

Health Information National Trends Survey (HINTS). Available at: https://hints.cancer.-gov/ (accessed june 2020).

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70(5):759–779. DOI: https://doi.org/10.1093/poq/nfl035.

Keeter, S., C. Miller, A. Kohut, R.M. Groves, and S. Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64(2):125–148. DOI: https://doi.org/10.1086/317759.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, and L. Christian. 2012. "Assessing the Representativeness of Public Opinion Surveys." *Pew Research Center, Washington, DC*. Available at: http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/ (accessed September 2019).

National Health Interview Survey (NHIS). Available at: https://www.cdc.gov/nchs/nhis/index.htm (accessed June 2020).

National Crime Victimization Survey (NCVS). Available at: https://www.census.gov/programs-surveys/ncvs.html (accessed June 2020).

Tourangeau, R. and T.J. Plewes. 2013. *Nonresponse in Social Science Surveys: A Research Agenda – National Research Council, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Panel on a Research Agenda for the Future of Social Science Data Collection*. Washington, DC: The National Academies Press.

Waksberg, J. 1978. "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association* 73(361):40–46. DOI: https://doi.org/10.1080/01621459.1978.10479995.

Received August 2018
Revised May 2019
Accepted September 2019