Marketing Systems Group

THE ACCURACY OF MEASUREMENTS WITH PROBABILITY AND NONPROBABILITY SURVEY SAMPLES

MacInnis & Krosnick, et al. "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension" 2019. *Public Opinion Quarterly, 82*(4)

Abstract

Many studies in various countries have found that telephone and internet surveys of probability samples yielded data that were more accurate than internet surveys of nonprobability samples, but some authors have challenged this conclusion. This paper describes a replication and an expanded comparison of data collected in the United States, using a variety of probability and nonprobability sampling methods, using a set of 50 measures of 40 benchmark variables, larger than any used in the past, and assessing accuracy using a new metric for this literature: root mean squared error. Despite substantial drops in response rates since a prior comparison, the probability samples interviewed by telephone or the internet were the most accurate. Internet surveys of a probability sample combined with an opt-in sample were less accurate; least accurate were internet surveys of opt-in panel samples. These results were not altered by implementing poststratification using demographics.

Overview

Inspired importantly by the insights of R. A. Fisher (1925), as described and applied early on by Neyman (1934) and others, probability sampling via ran- dom selection has been the gold standard for surveys in the United States for decades. The dominant mode of questionnaire administration has shifted over time from face-to-face interviewing to random-digit-dial telephone inter- viewing in the 1970s (for reviews, see Brick 2011) to self-administration via the internet (Couper 2011). Most internet surveys today are done with non- probability samples of people who volunteer to complete questionnaires in exchange for cash or gifts and who were not selected randomly from the popu- lation of interest (Brick 2011). Often, stratification and quotas are used to maximize the resemblance of participating respondents with the population of interest in terms of demographics.

The prominence of nonprobability sampling methods today (Brick 2011; Callegaro et al. 2014) represents a return to the beginnings of survey research a century ago and to a method that was all but abandoned in serious work in the interim (e.g., Converse 1987; Berinsky 2006). The transition to probability sampling from quota sampling was spurred by quota sampling's

failure in pre- dicting the 1948 election (Converse 1987, pp. 201–10) and by "new ground in theory and application" in probability sampling (Converse 1987, p. 204). But in recent years, numerous authors have argued that nonprobability sampling can produce veridical assessments and should be the tool of choice for scien- tists interested in minimizing research costs while maximizing data accuracy (e.g., Silver 2012; Ansolabehere and Rivers 2013; Ansolabehere and Schaffner 2014; Wang et al. 2015). Harking back to the early days, many contemporary observers share Moser and Stuart's (1953) view that "statisticians have too easily dismissed a technique which often gives good results and has the virtue of economy" (p. 388).

During the past 15 years, a series of studies have compared the accuracy of probability samples and nonprobability samples. Some of these studies have shown that probability samples have produced accurate measurements, while nonprobability samples were consistently less accurate, sometimes strikingly so. Such studies led the AAPOR Task Force on Online Panels to conclude that "nonprobability samples are generally less accurate than probability samples" (Baker et al. 2010). And the AAPOR Task Force on Nonprobability Sampling concluded: "Although nonprobability samples often have performed well in electoral polling, the evidence of their accuracy is less clear in other domains and in more complex surveys that measure many different phenomena" (Baker et al. 2013).

However, that Task Force also said: "Sampling methods used with opt-in panels have evolved significantly over time and, as a result, research aimed at evaluating the validity of survey estimates from these sample sources should focus on sampling methods rather than the panels themselves. … Research evaluations of older methods of nonprobability sampling from panels may have little relevance to the current methods being used" (Baker et al. 2013).

Some observers have claimed that since the Task Force report was written, response rates of probability-based telephone surveys have continued to decline (but see Marken 2018), making probability sample surveys no better than nonprobability sample surveys.

This paper addresses these concerns by providing new evidence on the topic. We report evaluations of data collected with an array of methods in 2012. These evaluations assess whether probability sampling yielded more accurate measurements than did various types of nonprobability samples and whether accuracy has changed during the years since 2004, when the last of the studies like this was conducted (Yeager et al. 2011). Further, the present study supplements the work of Dutwin and Buskirk (2017) by evaluating a low-response-rate RDD telephone survey.

Comparing Probability and Nonprobability Sample Surveys

Studies have evaluated the accuracy of survey measurements of probability and nonprobability sample surveys by comparing statistics produced by the surveys with benchmarks assessing the same characteristics using methods of high reliability, such as government records (e.g., the State Department's record of the number of passports held

by Americans) and federal surveys with very high response rates. Such studies have found that nonprobability sample surveys yielded data that were less accurate than the data collected from probability samples when measuring voting behavior (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Sturgis et al. 2016), health behav- ior (Yeager et al. 2011), consumption behavior (Szolnoki and Hoffmann 2013), sexual behaviors and attitudes (Erens et al. 2014), and demograph- ics (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Yeager et al. 2011; Szolnoki and Hoffmann 2013; Erens et al. 2014; Dutwin and Buskirk 2017). Furthermore, current methods of adjusting nonprobability sample data have done little or nothing to correct the inaccuracy in estimates from nonprobability samples (Yeager et al. 2011; see Tourangeau, Conrad, and Couper 2013 for a review).

However, another set of recent papers, focused on pre-election polls, sug- gests that nonprobability samples yielded data that were as accurate, or more accurate than, probability sample surveys (e.g., Ansolabehere and Rivers 2013; Wang et al. 2015). And the very low response rates attained by proba- bility-based telephone surveys in recent years have led some to the belief that the theoretical advantages of probability-based surveys no longer obtain. The research reported here adds evidence to the ongoing discussion of probability and nonprobability sample surveys.

Metrics to Assess Accuracy

Past studies have used various different metrics to assess accuracy of meas- urements by com- paring them to benchmarks (see Callegaro et al. 2014). The present study introduces a new metric to this set.

Some studies have characterized the accuracy of a single measurement. Malhotra and Kros- nick (2007), for example, examined the absolute deviation of the percent of respondents choosing each response category in a survey from the percent of people in the population in that response category. Yeager et al. (2011) computed the absolute deviation of the percent of respondents choosing the modal response category in a survey from the percent of people in the population in that modal category. Walker, Pettit, and Rubinson (2009) and Gittelman et al. (2015) compared the percent of respondents choosing one response category (sometimes the modal category, sometimes a non- modal category) in a survey to the percent of people in the population in that category (without explaining why the particular response category was cho- sen). Kennedy et al. (2016) computed the absolute deviation of the percent of respon- dents choosing one response category or the combination of two response categories (without explaining why the particular response category or categories was/were chosen or combined) in a survey from the percent of people in the population in that category or categories.

Other studies have examined multiple measurements in comparing the accur- acy of probabil- ity and nonprobability surveys. Ansolabehere and Schaffner (2014) and Sturgis et al. (2016) computed the average absolute deviation of the percent of respondents choosing every response category in a survey from the percent of people in the population in those catego- ries. Dutwin and Buskirk (2017) constructed all possible cross-tabulations of pairs of variables

(using a set of four variables) and computed the average absolute deviation of the percent of survey respondents in each cell from the percent of people in that cell in the population. Blom et al. (2017) computed the average (across all response categories for a measure) of the ratio of (1) the deviation of the survey estimate of the percentage of respondents in each category from the percent of the population in that category to (2) the percent of the population in the category. Finally, a number of these investigations combined accuracy metrics for single measures across a set of measures to yield an overall esti- mate of measurement accuracy for a data provider. Yeager et al. (2011), Blom et al. (2017), Ansolabehere and Schaffner (2014), Kennedy et al. (2016), and Dutwin and Buskirk (2017) did so by averaging their accuracy metrics across measures.

We used a slightly different approach. Following Yeager et al. (2011), we first computed the deviation of the percent of survey respondents in the modal category from the percent of the population in that category. Then we aggre- gated across measures by computing the root mean squared error (RMSE). The RMSE is the square root of squared errors (deviation of the percent of respondents in a modal survey category from the percent of the population in that category) averaged across measures. Unlike the simple averaging done in many studies in the past, the RSME penalizes large errors more than small ones.

This approach is valuable for the following reason. Consider two surveys with the identical mean absolute error. One survey has a few very large errors, a few very small errors, and otherwise moderate errors. Another survey has errors that are about equal to one another across comparisons. The RMSE for the former survey will be much larger than that for the latter survey. Extreme errors in a few measures can be especially costly for a researcher. This approach was also used recently by Shirani-Mehr et al. (2018) when aver- aging errors across various surveys. We used this approach instead to average across a set of measures for each survey individually.

The Present Investigation

We applied this measure to data from various survey firms. Data collection with identical questions was accomplished by (1) random-digit-dial (RDD) telephone interviewing via landlines and cellphones, (2) a probability sample internet survey, (3) internet surveys of probability samples combined with opt- in samples with no weighting to match the two, (4) internet surveys of opt-in panel samples who were rewarded with cash or gifts, and (5) an opt-in sample internet survey with the incentive of a charitable contribution made on behalf of the respondent.[1]

---

1. The companies that provided data for this study were promised that their identities would not be revealed. This same promise was made to the firms that provided data for the similar, earlier comparison by Yeager et al. (2011) and was also made by the Advertising Research Foundation to the companies that provided data for its methodology comparisons (Walker, Pettit, and Rubinson 2009; Gittelman et al. 2015). Thus, in such large-scale comparisons, it has been standard practice to promise anonymity in the interest of maximizing participation by as many firms as possible.

RDD telephone surveys of landlines and cellphones remain popular with the nation's leading news media organizations and academics. Inclusion of this methodology allows assessment of frequent claims by advocates of non- probability sampling that response rates for RDD surveys are so low as to completely undermine their accuracy. Data collection from probability sam- ple internet panels has been growing in popularity—it was pioneered in the United States by the company originally called Intersurvey and now called GfK Custom Research, and similar panels have been built by the National Opinion Research Center (in its AmeriSpeak project), the Pew Research Center's online panel, and other organizations. And opt-in online panels, river sampling, and routers are generating a huge amount of data for American sur- veys (see Callegaro et al. 2014). Thus, all of these methods merit investigation

Accuracy was assessed using benchmarks from high-quality federal face- to-face surveys with very high response rates. Assessments were made using three categories of variables: primary demographics, secondary demograph- ics, and nondemographics. Primary demographics are the variables survey firms used in selecting people to invite or to accept to complete the internet surveys or the variables survey firms used in computing poststratification weights. Secondary demographics are other demographics that were not used in sampling or weight construction. Nondemographics are all other vari- ables, including characteristics of housing structures, consumption behavior, economic expenditures, health quality, health-related behaviors, and health care utilization. Accuracy assessed using these three types of measures was examined in two ways—without and with poststratification weights. A total of 38 benchmark variables were examined, substantially more than any other investigation of this sort. For example, Yeager et al. (2011) examined a total of 18 benchmark variables.

Methods

COMMISSIONED SURVEYS

Each of eight survey data collection firms administered two different question- naires (called Questionnaire 1 and Questionnaire 2) to separate samples of the target population of adults, 18 years old and older, residing in the United States.[2] Questionnaire 1 was administered to 10 samples, and Questionnaire 2 was administered to nine of the samples.[3]

---

2. Administering all questions used in this study with a single sample would have made the ques- tionnaire quite long, so the measures were split across two different questionnaires. Questionnaire 1 included measures of primary demographics and some secondary demographics, and was administrated by all data providers. Questionnaire 2 included measures of primary demographics, some secondary demographics, and all nondemographics and was administered by all online data providers (for a list of the measures asked in each of the two questionnaires, see the Appendix). Some primary demographic measures were included in both Questionnaire 1 and Questionnaire 2 but with different wordings and were therefore included in the analyses twice.

3. One firm fielded the probability internet survey, nonprobability internet survey 2, and nonprob- ability internet survey 4. See table 1.

Table 1 displays methodological details, including the numbers of people invited to complete the question- naires, the numbers of people who completed the questionnaires, the dates of fielding the data collections, whether the sampling process involved by design unequal probabilities of selection from a population or pool, whether quotas were used when potential respondents sought to complete the questionnaires, and what incentives were offered for participation (see the Appendix for more details).

Table 1. Sample description information for Questionnaire 1 and Questionnaire 2

Sample	Invitations	Responses	Response rate or completion rate	Field dates	Unequal probability of invitation?	Quota used?	Incentives offered
Questionnaire 1							
Probability samples							
RDD telephone	19,585	805	15.3%[a]	November–December 2012	No	No	$10 to reluctant cell phone respondents
Internet	2,320	1,135	2.0%[b]	November–December 2012	No	No	Free computer and internet access (for some), cash
Combined samples							
1	Unknown	1,075	Unknown	October–November 2012	Yes	Yes	Cash
2	1,204	1,020	84.7%[c]	November–December 2012	Yes	Yes	Free computer and internet access (for some), cash
Nonprobability samples							
1	20,908	1,070	5.1%[c]	October, 2012	Yes	Yes	Cash, prizes
2	41,070	1,030	2.5%[c]	November 2012	Yes	Unknown	Incentives provided but not revealed to the researchers
3	85,506	1,513	1.2%[c]	November 2012	Yes	Yes	Cash, prizes
4	Unknown	1,021	Unknown	November 2012	Yes	Unknown	Incentives provided but not revealed to the researchers

Continued

Table 1. *Continued*

Sample	Invitations	Responses	Response rate or completion rate	Field dates	Unequal probability of invitation?	Quota used?	Incentives offered
5	Unknown	979	Unknown	October–November 2012	Yes	Yes	Cash
6	4,702	1,057	22.5%[c]	October–November 2012	Yes	Yes	Donations to charity
Questionnaire 2							
Probability samples							
Internet	2,318	1,143	2.0%[b]	November–December 2012	No	No	Free computer and internet access (for some), cash
Combined samples							
1	Unknown	1,043	Unknown	October–November 2012	Yes	Yes	Cash
2	1,200	1,001	83.40%[c]	November–December 2012	Yes	Yes	Free computer and internet access (for some), cash
Nonprobability samples							
1	21,392	1,091	5.1%[c]	October 2012	Yes	Yes	Cash, prizes
2	40,580	1,047	2.6%[c]	November 2012	Yes	Unknown	Incentives provided but not revealed to the researchers
3	60,318	1,129	1.6%[c]	November 2012	Yes	Yes	Cash, prizes
4	Unknown	1,029	Unknown	November 2012	Yes	Unknown	Incentives provided but not revealed to the researchers
5	Unknown	978	Unknown	October–November 2012	Yes	Yes	Cash
6	6,073	1,167	19.2%[c]	October–November 2012	Yes	Yes	Donations to charity

Note.—In combined sample 2, 87 percent of respondents in Questionnaire 1 were from the probability sample and 13 percent from a snowball sample; 86 percent of respondents in Questionnaire 2 were from the probability sample and 14 percent from a snowball sample. Such sample composition information was not provided by the firm for combined sample 1.

[a]AAPOR Response Rate 3 (RR3).

[b]Cumulative response rate 2 (Callegaro and DiSogra 2008).

[c]Completion rate note 1: The RDD survey consisted of 604 landline respondents and 201 cellular respondents.

RDD: Questionnaire 1 was administered via RDD telephone calling to landlines and cell phones, with $10 paid to reluctant respondents interviewed on cell phones only. The AAPOR Response Rate 3 (AAPOR 2015) was 15.3 percent.

Probability sample internet panel: Probability sample internet questionnaires (Questionnaire 1 and Questionnaire 2) were administered to members of a panel of individuals who were recruited by probability sampling methods through RDD and address-based sampling (ABS) mailings and were given computers and internet access if needed. Incentive points redeemable for cash were paid for questionnaire completion. The Cumulative Response Rate 2 (Callegaro and DiSogra 2008) was 2.0 percent for both questionnaires.

Combined probability and nonprobability sample internet panels: For two firms that provided data, their online survey panel was built using two means of selection. Some panel members were recruited by probability sampling, and other panel members were recruited by nonprobability methods (e.g., snowball sampling or convenience sampling via recruitment through website ads, news sites, blogs, and search engines). The panel members invited to complete our questionnaires were mixes of these two types of panel members.

Nonprobability sample internet panels: Data from members of six nonprobability sample panels were evaluated. Each provider sampled individuals from their panels of millions of individuals who had volunteered to complete questionnaires for money in response to online advertising, invitations to members of organizations, and the like. For this study, each firm invited stratified samples based on demographics and imposed demographic quotas to restrict who could complete the questionnaire so that the participating individuals would resemble the target population in terms of the selected demographics.

MEASURES

Shown in the Appendix are the questions measuring the primary demograph- ics, secondary demographics, and nondemographics from the following benchmark surveys: the American Housing Survey (AHS), the Consumer Expenditure Survey (CES), the Current Population Survey (CPS), the National Crime Victimization Survey (NCVS), the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS) (see Section 1 of the online supplementary material for details on the methods, measures, and analyses of the benchmark surveys).

Primary demographics (measured by all survey firms) included sex, age, White race, Black race, other race, Hispanic ethnicity, education, region of residence, and household income, and were used by one or more of the survey firms to compute post-stratification weights. Also included in the category of “primary demographics” is a non-demographic variable, cigarette smoker sta- tus, because it was used by one of the survey firms to construct their post-strati- fication weight. Home ownership was also used by one firm in computing their post-stratification weight and was included with the “primary demographics” in analyses that did not include comparing the RDD survey to other surveys.

In the comparisons across samples involving the RDD survey, ten meas- ures in the "primary demographics" category were employed (sex, age, White race, Black race, other race, Hispanic ethnicity, education, region of residence, household income, and cigarette smoker status). In the comparisons across samples that did not include the RDD survey, 20 measures in the "primary demographics" category were used (2 measures of each of the following 9 var- iables (using different wordings): sex, age, White race, Black race, other race, Hispanic ethnicity, education, region of residence, and household income, plus one measure each of cigarette smoker status and home ownership).

In the analyses of secondary demographics and non-demographics (meas- ured by all firms except the telephone survey firm) that did not involve the RDD survey, 30 measures were used, including: marital status (measured with two different questions), citizenship, having served in the armed forces, and volunteering activities (CPS); their food allergies, walking or bicycling, per- forming vigorous recreational activities, performing moderately vigorous rec- reational activities, and donating blood (NHANES); their body height, body weight, sleep, emergency room visits, asthma, high blood pressure, having surgery, seeing a doctor, medical consultation about diet, checking blood pres- sure, and general health (NHIS); the number of times they had moved in the past five years (NCVS); air-conditioning, fire extinguisher, sink, and repairs and maintenance in their home (AHS); and their grocery-shopping expenses, restaurant-meal expenses, free food, and mass transportation use (CES).

ANALYSIS

Base weights and poststratification weights: The firms that provided prob- ability samples provided base weights reflecting unequal probability of selec- tion, as well as poststratification weights. Some firms that provided internet nonprobability samples did not provide poststratification weights. Other firms that provided internet nonprobability samples provided poststratification weights that they normally provide to clients purchasing data from them, and we assessed accuracy of these firms' data using the weights that they provided. To allow a consistent across-firm comparison of the effect of weights on accuracy, we also generated a set of weights for every dataset using ANESrake (https://cran.r-project.org/web/packages/anesrake/anesrake.pdf). These weights maximized the match of each survey sample with the October 2012 Current Population Survey via raking on the following variables: sex (two groups), age (four groups), white race (two groups), black race (two groups), other race (two groups), ethnicity (two groups), education (four groups), and census region (four groups). Base weights were used as input weights in the poststratification weight computation for the RDD and internet probability sample. Weights were capped at 5 to prevent any respondents from having excessive influences on the sample statistics (see DeBell and Krosnick 2009). For Questionnaire 1, the range of weights was 0.12–5 for the RDD sam- ple, 0.02–5 for the internet probability sample, 0.21–5 for the two internet probability/nonprobability combined samples, and 0.08–5 for the six internet nonprobability samples. The design effect was 1.80 for the RDD, 1.66 for the internet probability sample, 1.25 and 1.67 for the two internet combined samples, and 1.42, 1.43, 1.46, 1.85, 1.65, and 2.72 for the six internet nonprob- ability sample surveys, respectively. For Questionnaire 2, the range of weights was 0.03–1.55 for the internet

probability sample, 0.01–5 for the two internet combined samples, and 0.00 to 5 for the six internet nonprobability samples. The design effect was 1.49 for the internet probability sample, 1.13 and 1.82 for the two internet combined samples, and 1.46, 1.30, 1.41, 1.70, 1.56, and 3.31 for the six internet nonprobability sample surveys, respectively.

Accuracy metrics: For each commissioned firm, the RMSE was calculated in three steps. The first step was to compute the squared error for each measure: the square of the deviation between the percent of respondents in the modal category in the benchmark survey and the percent of respondents in that category in the commissioned survey (the modal categories are listed in column 1 of table S1 in the online supplement). The second step was to compute the mean squared error, which is the sum of the squared errors across measures under assessment, divided by the number of measures. The third step was to compute the square root of the mean squared error. Additional metrics for assessing and comparing accuracy (the largest absolute error observed across all measures and the rank of each commissioned survey in terms of its RMSE) were also computed.[4]

Aggregation: For each commissioned survey, the RMSE was computed for (1) primary demographics only, (2) secondary demographics and nondemographics combined, and (3) all measures combined. The comparison of primary demographics across samples should be viewed with caution, because the commissioned internet surveys used some primary demographics to implement stratified sampling or completion quotas or both, which will enhance the accuracy of those distributions. Secondary demographics and nondemographics were not used by any of the survey firms in their sampling or quotas or in the construction of poststratification weights and therefore offer more diagnostic comparisons of accuracy.

The statistical significance of the differences between survey providers in terms of RMSE was computed by first bootstrapping (Efron and Tibshirani 1986) each commissioned survey's RMSE and then performing a t-test to compare the two RMSEs (see Section 2 of the online supplementary material for a description of the methods used to conduct analyses of the commissioned surveys).

Missing data: The survey firms provided data to us only for respondents who answered at least 85 percent of the questions in a questionnaire (see response and completion rates in table 1). Among these individuals, the percent of respondents who did not answer a benchmark question (any of the primary demographics, secondary demographics, or nondemographics) was less than 0.39 percent on average for Questionnaire 1 and less than 2.35 percent on average for Questionnaire 2. In generating the benchmark estimates from benchmark surveys (e.g., AHS), missing cases were excluded; likewise, missing cases in the commissioned surveys were excluded when generating the survey estimates. The rate of item nonresponse for questions used to assess accuracy was similarly low for the probability samples and the nonprobability samples (the modal rate was 0 percent, and the maximum was 2.9 percent) (see Section 3 of the online supplementary material for item nonresponse rates for the two questionnaires we administered).

---

4. Also computed and reported in the online supplementary material are results based on the absolute value of the deviation between the percent of respondents who gave the modal response to a question in the benchmark survey and the percent of respondents who gave that response in the commissioned survey.

Results

RMSE

Primary demographics without poststratification: When examining the ten primary demographics measured in all surveys, without poststratification, the most accurate surveys were the probability sample internet survey (RMSE was 3.94 percentage points) and the RDD survey (RMSE was 4.29) (see row 1 and columns 1 and 2 in table 2), the accuracies of which were not significantly different from one another ($t(99) = 0.56$, $p > 0.10$).[5]

One of the two combined samples (RMSE = 6.04 and 4.90; see row 1 and columns 3 and 4 in table 2) and four of the six nonprobability sample surveys (RMSE ranged from 4.75 to 9.02; see row 1 and columns 5–10 in table 2) were significantly less accurate than the RDD survey (Combined samples: $t(99) = 3.08$, $p < 0.01$ and 0.92, $p > 0.10$; Nonprobability sample surveys: $t(99) = 2.80$, $p < 0.01$; 1.80, $p < 0.10$; 1.29, $p > 0.10$; 2.21, $p < 0.05$; 0.72, $p > 0.10$; 8.66, $p < 0.001$).

Similarly, one of the two combined samples and five of the six nonprobabil- ity sample surveys were significantly less accurate than the probability sample Internet survey (Combined samples: $t(99) = 3.73$, $p < 0.001$ and 1.46, $p > 0.10$; Nonprobability sample surveys: $t(99) = 3.44$, $p < 0.001$; 2.35, $p < 0.05$; 1.96, $p < 0.10$; 2.79, $p < 0.01$; 1.28, $p > 0.10$; 9.41, $p < 0.001$).

Among the nonprobability sample panels, #6 was significantly less accur- ate than all of the others ($t(99)$ ranged from 6.29 to 9.14, $p < 0.001$), which were not significantly different from one another. Quota sampling, which was employed in the combined samples and nonprobability sample panels, did not fare well in the unweighted analysis. This may be in part because the demo- graphic measures used in quota sampling were different from those under evaluation.[6]

When analyzing 20 primary demographics measures from Questionnaires 1 and 2 that were administered by all online survey firms, the same findings appeared. Without post-stratification: (a) the most accurate measurements were made by the probability sample Internet survey (RMSE was 3.66 per- centage points (see row 2 and column 2 in table 2); (b) the combined samples and the nonprobability sample panels were significantly less accurate than the probability sample Internet survey (combined samples: $t(99) = 7.09$, $p < 0.01$ and 3.64, $p < 0.001$; nonprobability sample panels: $t(99) = 6.73$, $p < 0.001$; 4.48, $p < 0.001$; 4.51, $p < 0.001$; 4.19, $p < 0.001$; 3.07, $p < 0.01$ and 17.61, $p < 0.001$); and (c) among the nonprobability sample surveys, #6 was significantly less accurate than all of the others ($t(99)$ ranged from 11.06 to 13.88, $p < 0.001$), which were not significantly different from one another.

Secondary demographics and nondemographics without poststratification: Examining secondary demographics and nondemographics without poststratification, the probability sample internet survey was the most accurate (RMSE = 5.16; see row 3 and column 2 in table 2). The combined samples (RMSE = 6.20 and 6.59; see row 3 and columns 3 and 4 in table 2) and the nonprobability sample panels (RMSE ranged from 6.26 to 11.86; see row 3 and

5. See the tables in the online supplementary material for a list of test results.
6. Because we used primary demographics to conduct poststratification weights, the commis- sioned surveys and benchmark surveys matched almost exactly in terms of primary demographics when the poststratification weights were used.

columns 5–10 in table 2) were significantly less accurate than the probability sample internet survey (combined samples: t (99) = 2.40, p < 0.05 and 3.55, p < 0.001; nonprobability sample panels: t (99) = 2.91, p < 0.01; 3.23, p < 0.01; 4.37, p < 0.001; 4.78, p < 0.001; 6.30, p < 0.001 and 16.82, p < 0.001). Nonprobability sample panel #6 was significantly less accurate than all other nonprobability sample panels (t (99) ranged from 9.84 to 14.03, p < 0.001), which were not significantly different from one another.

Secondary demographics and nondemographics with poststratification: Examining secondary demographics and nondemographics with the poststratification weights provided by the firms,[7] the probability sample internet survey was again the most accurate (RMSE = 4.62; see row 4 and column 2 in table 2). The remaining surveys had larger RMSEs, ranging from 6.20 to 11.86 (see row 4 and columns 3–10 in table 2).[8]

With poststratification weights that we constructed, the probability sample internet survey was again the most accurate (RMSE = 4.93; see row 5 and column 2 in table 2). The combined samples (RMSE = 6.03 and 6.27; see row 5 and col- umns 3 and 4 in table 2) and the nonprobability sample surveys (RMSE ranged from 6.38 to 9.38; see row 5 and columns 5–10 in table 2) were significantly less accurate than the probability sample internet survey (t (99) = 2.28, p < 0.05 and 3.34, p < 0.001 for the combined surveys, t (99) = 3.08, p < 0.01; 3.54, p < 0.001; 4.01, p < 0.001; 6.23, p < 0.001; 5.46, p < 0.001 and 7.70, p < 0.001 for the nonprobability sample internet panels). Nonprobability sample internet panel #6 was significantly less accurate than all other nonprobability sample panels (t (99) ranged from 2.63, p < 0.01 to 4.72, p < 0.001), which were not significantly dif- ferent from one another. A summary of RMSE is shown in figure 1.

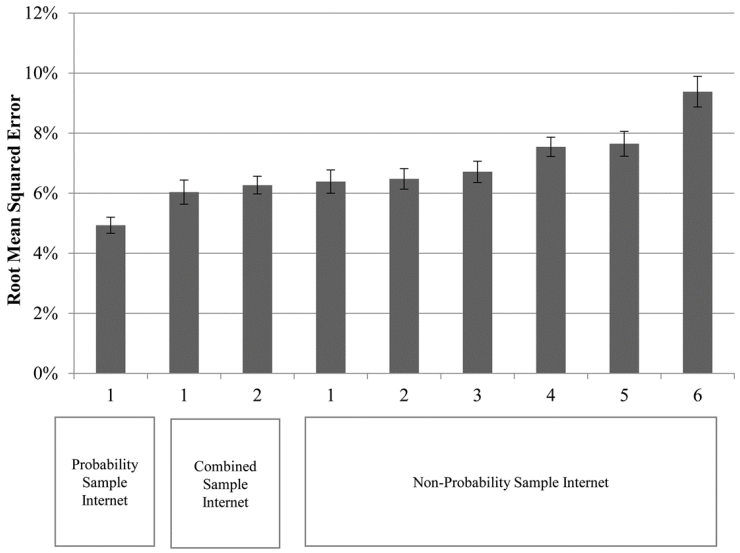


Figure 1. Root mean squared errors for the probability internet sample, the probability plus nonprobability combined samples, and the nonprob- ability samples across secondary demographics and nondemographics, with our poststratification.

---

7. Some firms did not provide poststratification weights, so their data were analyzed without weights when other firms' own weights were used.
8. Almost all the secondary demographics and nondemographic measures were asked in Questionnaire 2, which was not administered via RDD, so no discussion of RDD accuracy is offered here regarding secondary demographics and nondemographics.

Effects of poststratification weights: For the probability sample internet survey data, the firm's weights and our weights were similar (e.g., r = 0.82 for questionnaire 2), improved its accuracy (compare rows 3, 4, and 5 in table 2), and did so similarly well (compare rows 4 and 5 in table 2). Weighting improved accuracy in only 73 percent of the comparisons involving the other surveys in table 2 and decreased accuracy in 27 percent of the comparisons, meaning that poststratification did not consistently improve nonprobability samples' accuracy. The poststratification weights we computed and those the firms provided were similar for some samples and dissimilar in others (correlations of .25, .27, .58, and .98 for Questionnaire 2 in the four samples for which the firms provided weights) but yielded similar accuracy (compare rows 4 and 5 in table 2). This finding resonates with recent studies showing that no single weighting method among raking, propensity weighting, and matching performs consistently better across all measures and all metrics, and that raking, the most basic method and the one we employed, appears to perform better in many cases (Dutwin and Buskirk 2017; Mercer, Lau, and Kennedy 2018).

RANK AND LARGEST ERROR

The nonprobability sample internet surveys were not consistent in terms of their rank order of RMSE—that is, no nonprobability sample internet survey was consistently more accurate than others (see rows 6–10 in table 2). The rank order of the firms in terms of accuracy measuring primary demograph- ics was essentially uncorrelated (r = 0.10) with that in terms of their accuracy measuring secondary demographics and nondemographics. The only exception was that nonprobability sample internet survey #6 was consistently the least accurate.

The same conclusions are reinforced by the largest absolute error produced by each survey (see rows 11 to 15 in table 2). When using the ten primary demographics without poststratification, the smallest of these errors appeared for the RDD telephone survey (7.53) and the probability sample internet sur- vey (8.18). The largest errors for the other surveys were greater, ranging from 10.20 to 23.04. When measuring the secondary demographics and nondemo- graphics without poststratification, the largest absolute error for the probability sample internet survey (13.93) was the smallest among all the commissioned surveys. The remaining largest errors ranged from 16.16 to 40.91. The same conclusion is reached when using the providers' poststratification weights or when using the weights that we constructed.

CONSISTENCY OF ERRORS ACROSS MEASURES

When examining the secondary demographics and nondemographics without poststratification, the errors were more consistently small for the probability sample internet survey than for the nonprobability sample internet surveys. The absolute errors for the individual measures were clustered relatively close to zero for the probability sample internet survey (standard deviation = 3.29; see the left panel of figure 2), and the errors were larger and more widely distributed for the nonprobability sample internet surveys (standard deviation = 5.21; see the right panel of figure 2). The same pattern reemerged when examining primary demographics without poststratification weights, secondary demographics and nondemographics with the firms' poststratifi- cation weights, or secondary demographics and

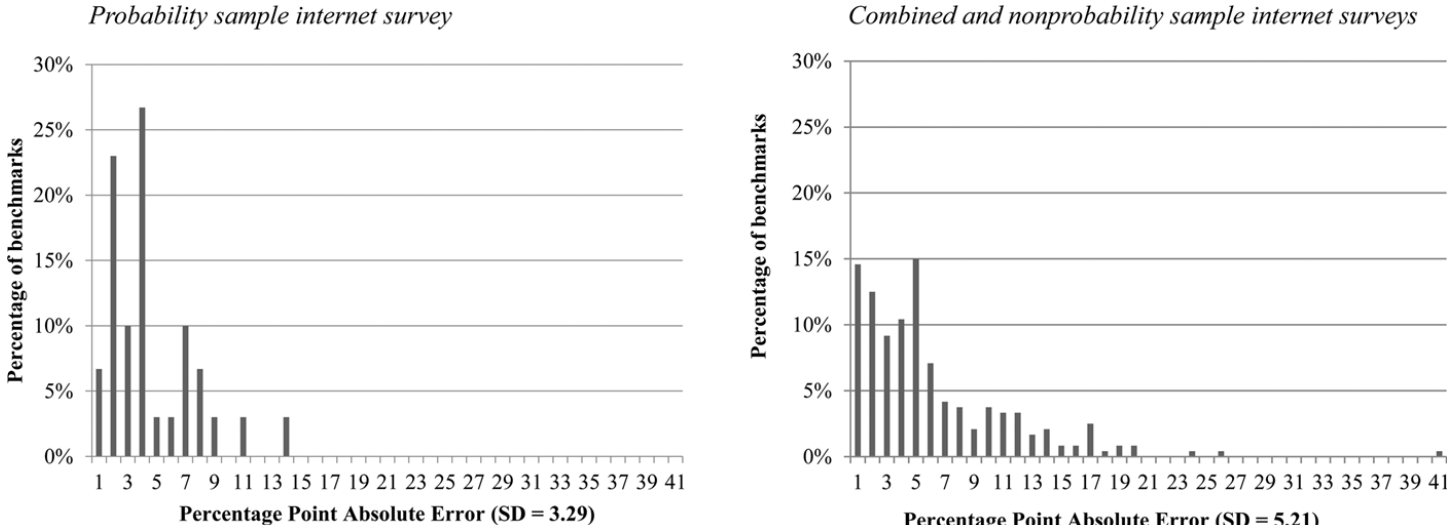
nondemographics with our poststratification weights (see rows 16–17, and 19–20 in table 2).

REPLICATION

Four data providers participated in both the current study and the study described by Yeager et al. (2011): the RDD telephone survey, the probabil- ity sample internet survey, a probability and nonprobability sample combined internet survey, and a nonprobability sample internet survey. The data col- lected via the internet in 2004 and 2012 included the following variables: sex, age, race, ethnicity, education, region, marital status, income, homeownership, and health status. Comparison between the 2004 and 2012 RDD telephone surveys was conducted with a slightly different set of variables: sex, age, race, ethnicity, education, region, and income (see Section 4 in the online supple- mentary material for details on the methods).

From 2004 to 2012, the RDD surveys did not manifest a significant decline in accuracy, and none of the internet surveys manifested significant improve- ment in accuracy. Using each firm's average absolute error in 2004 and 2012 and bootstrapped standard errors, change in average absolute error was 1.00 percentage point for the RDD surveys ($t(99) = 1.46$, $p = 0.15$), –.22 percent- age points for the probability sample internet surveys ($t(99) = 0.41$, $p = 0.68$), 0.02 percentage points for the probability sample and nonprobability sample combined internet surveys ($t(99) = 0.04$, $p = 0.97$), and –0.64 percentage points for the nonprobability sample internet surveys ($t(99) = 1.10$, $p = 0.27$).

Figure 2. Histograms showing absolute errors for secondary demograph- ics and nondemographics without poststratification from the probabil- ity sample internet survey and the combined and nonprobability sample internet survey.



Discussion

This investigation yielded the following findings. First, the most accurate sur- veys were the probability sample surveys (the RDD telephone survey and the probability sample internet survey). Second, the nonprobability sample sur- veys were all less accurate than the probability sample surveys, as were com- binations of probability and nonprobability samples. Third, poststratification weights with primary demographics improved the accuracy of the probability samples but only sometimes improved the accuracy of the nonprobability sam- ples. Furthermore, weighting did not eliminate the superiority of the probabil- ity sample surveys over the nonprobability sample surveys in terms of error rates. Fourth, RDD data were equally accurate in the present study (collected in 2012) as they were eight years before (in 2004, collected by Yeager et al. 2011), despite a 20-percentage-point drop in the survey response rate during this time period (AAPOR RR3: 35.6 percent in 2004, 15.3 percent in 2012). The same consistency in accuracy between 2004 and 2012 was apparent for the probability sample internet survey data, despite a 10-percentage-point drop in response rates (AAPOR CRR1: 15.3 percent in 2004, 4.6 percent in 2012). Finally, accuracy was also no greater in 2012 than it had been in 2004 for the internet data collected from nonprobability samples. This finding chal- lenges the claim that nonprobability sample internet survey procedures may have improved during that time period.

In a few instances, the nonprobability samples were relatively close to the benchmarks: three out of 30 secondary demographics and nondemographics had an error of less than four percentage points in every nonprobability sam- ple (using our poststratification weights): food allergies, purchased/recharged a fire extinguisher during the past two years, and weekly expenses for grocery shopping.

Taken together, this evidence reinforces the claim that probability sampling works well at producing accurate measurements across a wide array of types of measures. This study involved the largest set of benchmark measures and the widest array of sampling methodologies yet evaluated in a single investi- gation. This evidence also resonates with the recent literature indicating that innovation in approaches of recruiting and weighting nonprobability samples has not yet improved the accuracy sufficiently to be at par with probability samples.

Critical voices discrediting conventional survey methodologies in recent years have often asserted that the accuracy and value of RDD telephone surveys have declined to the point of being worthless, because of declining response rates and declining contact rates (e.g., Ferrell and Peterson 2010). The present study, along with previous studies in which benchmark measures were used to evaluate the accuracy of survey measurements, shows that drop- ping response rates in probability sample surveys do not lead inevitably to increasing nonresponse bias (Groves and Peytcheva 2008; Kohut et al. 2012).

IMPACT OF WEIGHTING

Some of the firms that provided nonprobability sample data for this study employed

adjustment methods that they considered optimal for maximiz- ing data accuracy. One of the commissioned nonprobability sample internet surveys employed a propensity score adjustment, but its data were not more accurate than other nonprobability sample surveys. This finding is consistent with recent research that found that propensity score adjustment and other methods of weighting provide limited bias reduction for nonprobability sam- ples (Brick et al. 2015; Dever and Shook-Sa 2015).

SOCIAL DESIRABILITY

Some critics of the current paper's approach to assessing accuracy (i.e., rely- ing on benchmarks from face-to-face surveys with extremely high response rates) have asserted that this approach biases findings in the direction of a close match between telephone and face-to-face survey results, because both are subject to distortion by social desirability bias, and internet surveys are not (e.g., Taylor, Krane, and Thomas 2009). However, the accuracy superiority of the RDD telephone survey over nonprobability sample internet surveys in the present study was illustrated using demographics, and demographic measures seem unlikely to be distorted by social desirability concerns. Furthermore, the superiority of the probability sample internet survey over the nonprobability sample internet surveys was demonstrated on the same playing field with no interviewer involvement. So, the present study's findings do not seem attribut- able to the benchmarks having been collected by human interviewers.

LIMITATIONS

The present study did not involve a random sample of nonprobability sample internet survey providers, nor did it involve a random sample of measures that could have been used as benchmarks. Therefore, it may be most appropriate to view the present results as describing case studies, rather than providing findings that can be generalized to other firms or other measures. This study examined a much larger set of benchmarks than any past investigations and produced results very similar to those seen in the past. Future studies might include other providers of RDD telephone surveys, other providers of prob- ability sample internet surveys, and other providers of nonprobability sample internet surveys to explore the generalizability of the current findings. The present study employed one specific method of poststratification weighting, and thus its findings on the impact of weighting on accuracy may not apply to other methods of weighting.

Conclusion

The findings reported here yielded further evidence that probability sample telephone surveys and internet surveys provide more accurate estimates than nonprobability sample surveys. The present investigation examined a much- expanded set of benchmarks and a telephone survey with a notably lower response rate than were examined in past studies. We look forward to more such investigations in the future.

Appendix
Methods of the Commissioned Surveys

The RDD data were collected as part of another study commissioned by the authors on a substantive topic and paid for by the National Science Foundation. Each company that provided data collected via the internet was contacted by a principal investigator of this project and invited to participate in this study by administering common questionnaires and funding the data collection on their own. No invited companies declined to participate.

PROBABILITY SAMPLE TELEPHONE SURVEY (RDD)

Random digit dialing was implemented to conduct telephone interviews with 604 American adults on landline telephones and 201 American adults on cel- lular phones between June 13 and 21, 2012, in English. The target popula- tion for the study was noninstitutionalized persons aged 18 and over, living in the United States. Persons with residential landlines were not screened out of the cell phone sample. Numbers for the landline sample were drawn with equal probabilities from active blocks (area code + exchange + two-digit block number) that contained one or more residential directory listings. The cellu- lar phone sample was drawn from 1000-blocks dedicated to cellular service according to the Telcordia database.

A maximum of 13 call attempts were made to numbers in the landline and cell phone samples. Refusal conversion was attempted on soft-refusal cases in the landline sample. Calls were staggered over times of day and days of the week. The sample was released for interviewing in replicates. For the landline sample, the respondent was randomly selected from all of the adults living in the household. For the cell phone sample, interviews were conducted with the person who answered the phone. Interviewers verified that the person was an adult and in a safe place before administering the questionnaire. Reluctant respondents among the cellular frame sample were offered a reimbursement of $10 for their participation.

The base weight adjusted for differential probabilities of selection due to the number of adults in the household, the number of voice-use landlines, the number of cell phones, and the multiplicity created by the overlap in the landline and cell phone RDD frames. Sample balancing adjusted for differ- ential response propensities across various demographic groups (age × sex, education × sex, race and ethnicity, and region) using the 2010 ACS one-year estimates as the sample balancing targets, as well as across telephone service type using NHIS estimates as the target, and weights were constrained at a minimum of 0.2 to a maximum weight of 4.0.

PROBABILITY SAMPLE INTERNET PANEL (PROB1)

This probability sample internet panel was recruited using both random digit dialing (RDD) and address-based sampling (ABS) via mailed invitations. The sampling frame of addresses covered approximately 97 percent of US households, including households without landline telephones and internet access. The panel consisted of approximately 55,000 adults. Panel members were recruited via RDD beginning in 1999, and ABS was employed beginning in 2009 to supplement RDD recruiting and eventually replaced RDD recruit- ing. The ABS

sampling of addresses was done from the U.S. Postal Service's Delivery Sequence File. Selected households were first sent a series of mail- ings, and nonresponding households were later contacted by telephone if a telephone number for the address could be obtained through public records. Within the sampled household, a household member was randomly selected to join the panel. New panelists completed a profile questionnaire seeking basic information such as demographics.

Panelists without computers or internet access were given them. The aver- age AAPOR completion rate across surveys was 65 percent.

The base weight of the panel survey accounted for panel recruitment and construction. Since the panel was recruited from two sample frames (RDD and ABS), the construction of the base weight took into account the different designs of the two sample frames, such as the different selection probabilities due to oversampling for minorities. The panel base weight was also adjusted for sampling and nonsampling errors, such as nonresponse to panel recruit- ment and panel attrition among recruited panelists. A poststratification method was used to correct these errors. This adjustment involved poststratification on benchmarks from the Current Population Survey (CPS) and other sources when certain benchmarks were unavailable from the CPS. A study-specific weight was constructed based on the panel base weight after the data of a study sample was compiled. A poststratification process was used to adjust for nonresponse and study-specific sample design.

PROBABILITY/NONPROBABILITY COMBINED SAMPLES

Probability/nonprobability combined sample 1 (COMB1): Respondents of the probability/ nonprobability combined sample 1 (COMB1) were drawn from the members of a panel, most of whom volunteered to complete surveys in exchange for a chance to win prizes. The panel, therefore, was not a representative sample of American adults. The panel members were recruited in several ways. Initially, random digit dialing phone calls were made to invite some American adults to sign up to receive email invitations to participate in surveys. Similar recruitment phone calls were made to professionals working in the information technology sector who were listed in professional directories. These initial panel members (a total of approximately 5,000) were then offered a chance to win cash or gift certificates in exchange for referring other people to join the panel. Referred panel members were offered the same incentives to refer others. Panel members were also recruited through online advertisements (posted on the firm's website, news sites, blogs, and search engines) and through emails sent by businesses and nonprofit organizations with which prospective panelists were affiliated. Panel members were rewarded when one of their referrals, or one of their referrals' referrals, completed a survey. The firm sent an invitation email to panel members. Invitees were selected to maximize the match of the participants to the nation in terms of the distributions of some demographic variables. The firm did not provide the weights.

Probability/nonprobability combined sample 2 (COMB2): The probability/ nonprobability combined sample 2 (COMB2) was a combined sample from a probability sample with a snowball sample. That probability sample covered approximately 5,000 US households.

Participants were recruited from several already-existing probability sample sources, including probability panels that recruited respondents via random digit dialing. In addition to these respondents recruited from already existing probability samples, respondents were recruited via snowball sampling. Respondents were given the opportunity to suggest friends or acquaintances who might want to participate. These people were then invited to participate. No new snowball respondents had been permitted to join since May 2009. For the probability sample part of the survey, respondents were drawn from the members of a panel consisting of more than 5,000 American adults aged 18 and older. Respondents were recruited using probability-based sampling via random digit dialing. If needed, respondents were given laptops and Web-TVs and access to the internet at no cost to allow them to answer questionnaires via the internet. When people joined the panel, they provided demographic information such as sex, age, race/ethnicity, education, and income. Members received emails inviting them to complete the surveys and offering a cash incentive.

NONPROBABILITY SAMPLE INTERNET PANELS

Nonprobability sample internet panel 1 (NONP1): Respondents were drawn from the members of the firm's panel. Most of the members of this panel volunteered to complete surveys in exchange for a chance to win prizes, so this panel was not a representative sample of American adults. Members were recruited through multi-source recruitment. One main source was recruitment via websites. For each recruitment source, the firm used multiple methods of recruitment and reached different types of people through different methodologies, including text advertisements, search engines, banner advertisements, co-registration, and email campaigns. The firm also ran a referral program, inviting current panelists to refer their friends by entering their email addresses. All applicants went through a double opt-in process to join the panel. At registration, panelists completed a profile survey with demographic information, and they were informed that their data would only be used for research purposes and their personal identification information would never be shared with any clients. The firm sent an invitation email to panel members. Invitees were selected to maximize the match of the participants to the national population in terms of the distributions of some demographic variables. By completing the survey, participants received points redeemable for cash and entry to a sweepstakes for prizes like electronics or vacations.

The firm provided poststratification weights to maximize the match of the demographics of the sample to the population targets, generated from the Current Population Survey Annual Social and Economic Supplement admin- istered in March 2010. The set of socio-demographic variables whose distri- butions were matched to produce sample weights for the survey was: sex x race, sex x education, sex x age, and income x number of household members. Sampling weights were generated using an iterative raking algorithm.

Nonprobability sample internet panel 2 (NONP2): The firm contracted with an opt-in sample firm to provide a sample of respondents. The opt-in sample firm sampled from its nonprobability sample-based panel. The details of sampling and how data were collected from this opt-in sample were not provided. No sampling weights were provided by the firm.

Nonprobability sample internet panel 3 (NONP3): Respondents were selected from the firm's panel. The firm employed multiple methods in recruiting potential respondents into the panel, mainly from natural traffic on the website. Four respondent recruitment techniques were employed. It was effectively three, but one had two ways of notifying the respondent that a survey was waiting for them. The most popular technique employed was a direct invitation to the survey. Once a survey invitation was sent to a respondent via email, a notification was also uploaded to an area of the firm's website. The respondent could click on the link in the email invite or click on the link on their website notification. Thus, the same respondent had two different ways of entering the survey. Another way to enter the survey was through the router. A router is a technical device that moves respondents between surveys. Using a router selects a respondent who has taken the time to try to take another survey if that original survey is unavailable or if the respondent entered "through the river" (by clicking on a link on a website). Thus, a respondent who was invited to a survey that was no longer available to them answered a few questions and was randomly routed to a survey for which he or she was appropriate based on their demographic profile or the questions they previously answered. Or, a respondent who clicked on a survey advertisement could be similarly routed to a survey.

Sampling weights were generated to maximize the match of the demo- graphics of the sample to the population targets. Weights were constructed using an RIM weighting scheme including age by gender, education, income, region, smoking status, and race/ethnicity.

Nonprobability sample internet panel 4 (NONP4): The firm contracted with a sample firm that uses routing to provide a sample of respondents. The routed sample drew from a mixture of sources, including opt-in panels, social network samples, and reward-based survey respondents. The sampling details of the routed sample and how data were collected from this routed sample were not provided. Weights were not provided for this sample.

Nonprobability sample internet panel 5 (NONP5): Respondents were selected from the firm's opt-in panel. Potential panelists were invited to join the online opt-in panel via banners, invitations, and messages. The firm used a "blend methodology" to control the quality of its panel by identifying the personality and psychographic traits of the panelists that "impact the way people answer survey questions." The sampling procedure involved a three- stage randomization process. The first step was to randomly select panelists and invite them to participate in a survey. Second, a set of profiling questions for the participants were randomly selected. Third, upon completion of the set of questions in step two, these participants were then matched with a survey they were likely to be able to take, using a further element of randomization. The firm employed a survey router, taking into account factors such as the likelihood that panelists would complete a survey. A wide variety of incentives were provided. Weights were not provided for this sample.

Nonprobability sample internet panel 6 (NONP6): Respondents were selected from the firm's opt-in panel. This nonprobability online panel recruited its panelists by means of website recruitment, online advertisements, and co-registration partners, with the website recruitment method being the primary source of its panel. When joining the panel, panelists

were required to fill out a profile that contained basic demographics and other attributes, such as exercise, phone usage, electronics usage, student status, business owner, employment status, industry focus, job function, gender, age, kids, voting behavior, and income. The study-specific sampling procedure was random sampling based on the number of responses and target demographics provided by the survey creator. The firm randomly selected a group of respondents from the panel and sent to the selected respondents an email invitation that was based on a standard template. Incentives were charitable donations and opportunities to enter a sweepstakes for winning cash awards. Weights were not provided for this sample.

Measures in the Benchmark and Commissioned Surveys

PRIMARY DEMOGRAPHICS MEASURES

Sex (Source: CPS Monthly): "Are you male or female?" For Internet Questionnaire 1: "What is your gender?" (Response options: Male, Female.) For Internet Questionnaire 2: "Are you male or female?" (Response options: Male, Female.) For the telephone survey: Telephone interviewers recorded the respondent's gender as male or female. (Categories used for analysis: Male, Female.)

Age (Source: CPS Monthly): "What is your date of birth?" (Respondents gave open-ended answers.) For Internet Questionnaire 1: "In what year were you born?" (Response options: textbox for year of birth.) For Internet Questionnaire 2: "What is your date of birth?" (Response options: textbox for year of birth.) For the telephone survey: "What is your age?" (Respondents gave open-ended answers.) (Categories used for analysis: 18–29, 30–49, 50–64, 65 and older.)

Region (Source: CPS Monthly): "In what state do you live?" (Respondents gave open-ended answers.) Region was determined by state of residence. For Internet Questionnaire 1 and the telephone survey: "And what is your five-digit zip code at your home?" (Response options: textbox for zip code.) Region was determined by zip code. For Internet Questionnaire 2: "In what state do you live?" (Drop-down menu of the list of US states was shown.) Region was determined by state of residence. (Categories used for analysis: Northeast, Midwest, South, West.)

Hispanic (Source: CPS Monthly): "Are you Spanish, Hispanic, or Latino?" For Internet Questionnaire 1 and Internet Questionnaire 2: "Are you Spanish, Hispanic, or Latino?" (Response options: Yes, No.)

For the telephone survey: "Are you of Hispanic origin or background?" (Categories used for analysis: Yes, No.)

Race (Source: CPS Monthly): Respondents gave open-ended answers categorized into the following: White Only; Black Only; American Indian, Alaskan Native Only; Asian Only; Hawaiian/Pacific Islander Only; White- Black; White-AI; White-Asian; White-HP; Black-AI; Black-Asian; Black-HP; AI-Asian; AI-HP; Asian-HP; W-B-AI; W-B-A; W-B-HP; W-AI-A; W-AI-

HP; W-A-HP; B-AI-A; W-B-AI-A; W-AI-A-HP; Other 3 Race Combinations; Other 4 and 5 Race Combinations. For Internet Questionnaire 1, if NOT Spanish, Hispanic, or Latino: "What race or races do you consider yourself to be?" If YES to Spanish, Hispanic, or Latino: "In addition to being Spanish, Hispanic, or Latino, what race or races do you consider yourself to be?" (Response options for both questions: White, Caucasian; Black, African American, Negro; American Indian, Alaska Native; Asian Indian; Native Hawaiian; Chinese; Guamanian or Chamorro; Filipino; Samoan; Japanese; Korean; Vietnamese; Other Asian; Other Pacific Islander; Some other race. Categories used for analysis: White only, Black only, Asian Only, Other.) For Internet Questionnaire 2: "Here is a list of five race categories. Please choose one or more races that you consider yourself to be: White; Black or African American; American Indian or Alaska Native; Asian; OR Native Hawaiian or Other Pacific Islander." (Response options: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian, Other Pacific Islander.) For the telephone survey, if "Yes" to the Hispanic question: "Are you White Hispanic or Black Hispanic?" If "No" to the Hispanic question: "Are you White, Black, or some other race?" (Categories used for analysis: White only, Black only, Asian Only, Other.)

Education (Source: CPS Monthly): "What is the highest level of school you have completed or the highest degree you have received?" (Respondents gave open-ended answers categorized into the following categories: Less than 1st grade; 1st, 2nd, 3rd, or 4th grade; 5th or 6th grade; 7th or 8th grade; 9th grade; 10th grade; 11th grade; 12th grade; No diploma; High school graduate—high school diploma or the equivalent; Some college but no degree; Associate degree in college—Occupational/vocational program; Associate degree in college—Academic program; Bachelor's degree; Master's degree; Professional school degree; Doctorate degree.) For Internet Questionnaire 1: "What is the highest grade you have completed?" (Response options: Less than high school graduate, High school graduate, Technical/trade school, Some college, College graduate, Some graduate school, Graduate degree; Categories used for analysis: Less than high school, High school graduate, Some college or technical/trade school, College degree, Postgraduate.) For Internet Questionnaire 2: "What is the highest level of school you have completed or the highest degree you have received?" (Response options: Less than 1st grade; 1st, 2nd, 3rd, or 4th grade; 5th or 6th grade; 7th or 8th grade; 9th grade; 10th grade; 11th grade; 12th grade; No diploma; High school graduate—high school diploma or the equivalent; Some college but no degree; Associate degree in college—Occupational/vocational program; Associate degree in college—Academic program; Bachelor's degree; Master's degree; Professional school degree; Doctorate degree). For the telephone survey: "What was the last grade of school you completed? 8th grade or less, some high school, graduated from high school, some college (ask if technical school, if yes, choose 'graduated from high school'), graduated from college, or postgraduate?" (Categories used for analysis: Less than high school, High school degree, Some college, College graduate, Postgraduate.)

Family income (Source: CPS Annual Social and Economic Supplement [ASEC]): "Which category represents the total combined income of all members of your FAMILY during the past 12 months?" (Response options: Less than $5000; 5000 to 7499; 7500 to 9999; 10,000 to 12,499;

12,500 to 14,999; 15,000 to 19,999; 20,000 to 24,999; 25,000 to 29,999; 30,000 to 34,999; 35,000 to 39,999; 40,000 to 49,999; 50,000 to 59,999; 60,000 to 74,999; 75,000 to 99,999; 100,000 to 149,999; 150,000 or more.) For Internet Questionnaire 1: "Was your total income of you and all members of your family who lived with you in 2011, before taxes, less than $50,000, or $50,000 or more?" (Response options: Less than $50,000, $50,000 or more.) IF LESS THAN $50,000: "And in which of the following groups was the total income of you and all members of your family who lived with you in 2011, before taxes?" (Response options: Less than $10,000, $10,000 to $19,999, $20,000 to $29,999, $30,000 to $39,999, $40,000 to $49,999.) IF GREATER THAN $50,000: "And in which of the following groups was the total income of you and all members of your family who lived with you in 2011, before taxes?" (Response options: $50,000 to $74,999; $75,000 to $99,999; $100,000 to $149,999; $150,000 or more.) For Internet Questionnaire 2: "Which category represents the total combined income of all members of our FAMILY during the past 12 months?" (Response options: Less than $5000; 5000 to 7499; 7500 to 9999; 10,000 to 12,499; 12,500 to 14,999; 15,000 to 19,999; 20,000 to 24,999; 25,000 to 29,999; 30,000 to 34,999; 35,000 to 39,999; 40,000 to 49,999; 50,000 to 59,999; 60,000 to 74,999; 75,000 to 99,999; 100,000 to 149,999; 150,000 or more.) For the telephone survey: "Which of the following categories best describes your total annual household income, before taxes, from all sources? Under 20 thousand dollars, 20 to under 35 thousand, 35 to under 50 thousand, 50 to under 75 thousand, 75 to under 100 thousand, or 100 thousand or more?" If "100 thousand or more," ask "Is that 100 to under 150 thousand, 150 to 200 thousand, 200 to under 250 thousand, or 250 thousand or more?" (Categories used for analysis: Less than $20,000; $20,000–49,999; $50,000–74,999; $75,000–99,999; $100,000 or more.)

Living quarters (Source: CPS Annual Social and Economic Supplement [ASEC]): "Are your living quarters owned or being bought by you or someone in your household, rented for cash, or occupied without payment of cash rent?" For Internet Questionnaire 2: "Are your living quarters owned or being bought by you or someone in your household, rented for cash, or occupied without payment of cash rent?" (Response options: Owned or being bought by you or someone in your household, Rented for cash, Occupied without payment of cash rent.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Owned or being bought by you or someone in your household, Rented for cash, Occupied without payment of cash rent.)

Ever smoked (Source: NHIS):[9] "Have you smoked at least 100 cigarettes in your ENTIRE LIFE?" For Internet Questionnaire 1 and telephone survey: "Have you smoked at least 100 cigarettes in your ENTIRE LIFE?" (Response options: Yes, No.) Not asked in Internet Questionnaire 2. (Categories used for analysis: Yes, No.)

SECONDARY AND NONDEMOGRAPHIC MEASURES

Married (Source: CPS Monthly): "Are you now married, widowed, divorced, separated or never married?" For Internet Questionnaire 1: "What is your marital status? Are you…"

---

9. Several benchmark measures were administrated in the commissioned surveys but were not analyzed (see Section 5 in the online supplementary material for the list of such measures).

(Response options: Married/Living as married/ Co-habiting, Separated, Divorced, Widowed, Never married.) For Internet Questionnaire 2: "Are you now married, widowed, divorced, separated, or never married?" (Response options: Married, Widowed, Divorced, Separated, Never married.) For the telephone survey: "Are you married widowed, divorced, separated, or never married?" (Categories used for analysis: Married, Widowed, Divorced, Separated, Never married.)

Citizenship (Source: CPS Monthly): "Are you a citizen of the United States?" For Internet Questionnaire 2: "Are you a citizen of the United States?" (Response options: Yes; No, not a citizen.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Armed forces (Source: CPS Monthly): "Did you ever serve on active duty in the U.S. Armed Forces?" For Internet Questionnaire 2: "Did you ever serve on active duty in the U.S. Armed Forces?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Volunteering (Source: 2012 CPS September Supplement): "We are interested in volunteer activities, that is, activities for which people are not paid, except perhaps expenses. We only want you to include volunteer activities that you did through or for an organization, even if you only did them once in a while. Since September 1st of last year, have you done any volunteer activities through or for an organization?" For Internet Questionnaire 2: "We are interested in volunteer activities, that is, activities for which people are not paid, except perhaps expenses. We only want you to include volunteer activities that you did through or for an organization, even if you only did them once in a while. Since September 1st of last year, have you done any volunteer activities through or for an organization?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Food allergies (Source: NHANES): "Do you have any food allergies?" For Internet Questionnaire 2: "Do you have any food allergies?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Walk or bicycle (Source: NHANES): "Do you walk or use a bicycle for at least 10 minutes continuously to get to and from places?" For Internet Questionnaire 2: "Do you walk or use a bicycle for at least 10 minutes continuously to get to and from places?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Vigorous recreational activities (Source: NHANES): "Do you do any vigorous- intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously?" For Internet Questionnaire 2: "Do you do any vigorous-intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the

telephone survey. (Categories used for analysis: Yes, No.)

Moderate recreational activities (Source: NHANES): "Do you do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously?" For Internet Questionnaire 2: "Do you do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Donated blood (Source: NHANES): "During the past 12 months, that is, since October 2011, have you donated blood?" For Internet Questionnaire 2: "During the past 12 months, that is, since October 2011, have you donated blood?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Height (Source: NHIS):[10] "How tall are you without shoes?" (Respondents gave open-ended answers.) For Internet Questionnaire 2: "How tall are you without shoes?" (Response options: textbox for feet and inches.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: –61, 62–65, 66–68, 69–71, 71– in inches.)

Weight (Source: NHIS): "How much do you weigh without shoes?" (Respondents gave open-ended answers.) For Internet Questionnaire 2: "How much do you weigh without shoes?" (Response options: textbox for weight in pounds.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: –124, 125–149, 150–174, 174–199, 200– in pounds.)

Sleep (Source: NHIS): "On average, how many hours of sleep do you get in a 24-hour period?" (Respondents gave open-ended answers.) For Internet Questionnaire 2: "On average, how many hours of sleep do you get in a 24-hour period?" (Response options: textbox for number of hours.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: 3, 4, 5… 16, 17, 18, 20, 22.)

Emergency room (Source: NHIS): "DURING THE PAST 12 MONTHS, HOW MANY TIMES have you gone to a HOSPITAL EMERGENCY ROOM about your own health? (This includes emergency room visits that resulted in a hospital admission.)" For Internet Questionnaire 2: "DURING THE PAST 12 MONTHS, HOW MANY TIMES have you gone to a HOSPITAL EMERGENCY ROOM about your own health? (This includes emergency room visits that resulted in a hospital admission.)" (Response options: None, 1, 2–3, 4–5, 6–7, 8–9, 10–12, 13–15, 16 or more.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: None, 1, 2–3, 4–5, 6–7, 8–9, 10–12, 13–15, 16 or more.)

Asthma (Source: NHIS): "Have you EVER been told by a doctor or other health professional that you had asthma?" For Internet Questionnaire 2: "Have you EVER been told by a doctor

---

10. With height and other continuous variable measures, we converted the responses into a categorical variable with multiple groups. We created arbitrary groupings to optimize two criteria at once: (1) retaining the general shape of the continuous distribution and (2) equating sample sizes in the groups. Then we picked the modal category, no matter how little its size exceeded the size of other groups.

or other health professional that you had asthma?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

High blood pressure (Source: NHIS): "Have you EVER been told by a doctor or other health professional that you had hypertension, also called high blood pressure?" For Internet Questionnaire 2: "Have you EVER been told by a doctor or other health professional that you had hypertension, also called high blood pressure?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Surgery (Source: NHIS): "DURING THE PAST 12 MONTHS, have you had SURGERY or other surgical procedures either as an inpatient or outpatient? This includes both major surgery and minor procedures such as setting bones or removing growths." For Internet Questionnaire 2: "DURING THE PAST 12 MONTHS, have you had SURGERY or other surgical procedures either as an inpatient or outpatient? This includes both major surgery and minor procedures such as setting bones or removing growths." (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Saw doctor (Source: NHIS): "About how long has it been since you last saw or talked to a doctor or other health care professional about your own health? Include doctors seen while a patient in a hospital." (Respondents gave open- ended answers categorized into the following categories: Never, 6 months or less; More than 6 mos, but not more than 1 yr ago; More than 1 yr, but not more than 2 yrs ago; More than 2 yrs, but not more than 5 yrs ago; More than 5 years ago.) For Internet Questionnaire 2: "About how long has it been since you last saw or talked to a doctor or other health care professional about your own health? Include doctors seen while a patient in a hospital." (Response options: Never; 6 months or less; More than 6 mos, but not more than 1 yr ago; More than 1 yr, but not more than 2 yrs ago; More than 2 yrs, but not more than 5 yrs ago; More than 5 years ago.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Never; 6 months or less; More than 6 mos, but not more than 1 yr ago; More than 1 yr, but not more than 2 yrs ago; More than 2 yrs, but not more than 5 yrs ago; More than 5 years ago.)

Diet (Source: NHIS): "DURING THE PAST 12 MONTHS, has a doctor or other health professional talked to you about your diet?" For Internet Questionnaire 2: "DURING THE PAST 12 MONTHS, has a doctor or other health professional talked to you about your diet?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Checked blood pressure (Source: NHIS): "DURING THE PAST 12 MONTHS, have you had your blood pressure checked by a doctor, nurse, or other health professional?" For Internet Questionnaire 2: "DURING THE PAST 12 MONTHS, have you had your blood pressure checked by a doctor, nurse, or other health professional?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

General health (Source: NHIS): "In general, how would you rate your overall health now?" (Response options: Excellent, Very good, Good, Fair, Poor.) For Internet Questionnaire 2:

"In general, how would you rate your overall health now?" (Response options: Excellent, Very good, Good, Fair, Poor.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Excellent, Very good, Good, Fair, Poor.)

Times moved in past five years (Source: NCVS):[11] "Altogether, how many times have you moved in the last 5 years, that is, since…" (Respondents gave open-ended answers.) For Internet Questionnaire 2: "Altogether, how many times have you moved in the last 5 years, that is, since…" (Response options: textbox for number.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: all valid values from 1 through 96.)

Airconditioning (Source: AHS): "Does this housing unit have central air conditioning?" For respondents asked to verify that the answer to this question is as it was in the previous survey, they were asked: "(Last time) we recorded that your housing unit had central air conditioning. Is this information still correct?" For Internet Questionnaire 2: "Does your housing unit have air conditioning?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Fire extinguisher (Source: AHS): "Is there a fire extinguisher in the home that was purchased or recharged in the last 2 years?" For Internet Questionnaire 2: "Is there a fire extinguisher in the home that was purchased or recharged in the last 2 years?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Sink (Source: AHS): "Does this housing unit have a kitchen sink?" For respondents asked to verify that the answer to this question is as it was in the previous survey, they were asked: "(Last time) we recorded that your housing unit had a kitchen sink. Is this information still correct?" For Internet Questionnaire 2: "Does this housing unit have a kitchen sink with piped water?" (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Repairs and maintenance (Source: AHS): "In a TYPICAL YEAR about how much does your household spend for routine repairs and maintenance, such as painting, plumbing, roofing, or other minor repairs?" (Respondents gave open- ended answers.) For Internet Questionnaire 2: "In a TYPICAL YEAR about how much does your household spend for routine repairs and maintenance, such as painting, plumbing, roofing, or other minor repairs?" (Response options: textbox for number.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: $1–$150, $151–$399, $400–$999, $1000 and more.)

---

11. In NCVS 2012, respondents were asked a few questions before being asked the "times moved in last 5 years" questions. See http://www.bjs.gov/content/pub/pdf/ncvs1_2012.pdf.
Respondents were asked: (33a) "Before we get to the crime questions, I have some questions that are helpful in studying where and why crimes occur. How long have you lived at this address?" (33b) "How many months? (33c) Have you lived here? More than 5 years, less than 5 years but more than 1 year, less than 1 year but more than 6 months, 6 months or less, don't know?" NCVS 2012 conducted a CHECK ITEM A in (33d) "How many years are entered in (33a)? 5 years or more, less than 5 years?" Respondents whose answer in (33d) was "less than 5 years" were then asked the "times moved in last 5 years" question. Responses from (33a), (33b), (33c), and (33d) as well as (33e) were used to calculate the estimates of the number of times one moved in the past five years among all respondents in NCVS; respondents who were not asked (33e) were assigned to the value of 0. In all the commissioned surveys, all respondents were asked this "times moved in last 5 years" question.

Grocery shopping expense (Source: CES): "What has been your or your household's usual WEEKLY expense for grocery shopping?" (Respondents gave open-ended answers.) For Internet Questionnaire 2: "What has been your or your household's usual WEEKLY expense for grocery shopping?" (Response options: textbox for number.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: $1–$499, $500–$999, $1000–$1499, $1500–$1999, $2000 and more.)

Meal expense (Source: CES): "What has been your or your household's usual WEEKLY expense for meals or snacks from restaurants, fast food places, cafeterias, carryouts or other such places?" We divided the values of quarterly expenditure to produce weekly estimates. (Respondents gave open- ended answers.) For Internet Questionnaire 2: "What has been your or your household's usual WEEKLY expense for meals or snacks from restaurants, fast food places, cafeterias, carryouts or other such places?" (Response options: textbox for number.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: $0, $1–$249, $250–$499, $500–$999, $1000 or more.)

Free food (Source: CES): "Have you or any members of your household received any free food, beverages, or meals through public or private welfare agencies, including religious organizations? Do not include free meals in school or preschool programs." For Internet Questionnaire 2: "Have you or any members of your household received any free food, beverages, or meals through public or private welfare agencies, including religious organizations? Do not include free meals in school or preschool programs." (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Mass transportation (Source: CES): "Do you or any members of your household use mass transportation services such as a bus, subway, mini-bus or train? Include all commuter services. Do not include expenses covered by employer-provided transit subsidies." For Internet Questionnaire 2: "Do you or any members of your household use mass transportation services such as a bus, subway, mini-bus or train? Include all commuter services. Do not include expenses covered by employer-provided transit subsidies." (Response options: Yes, No.) Not asked in Internet Questionnaire 1 or the telephone survey. (Categories used for analysis: Yes, No.)

Supplementary Data

Supplementary data are freely available at Public Opinion Quarterly online.

References

American Association for Public Opinion Research (AAPOR). 2015. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys." Accessed April 2016. Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions2015_8thedi

tionwithchanges_April2015_logo.pdf.

Ansolabehere, Stephen, and Douglas Rivers. 2013. “Cooperative Survey Research.” *Annual Review of Political Sciences* 16:307–29.

Ansolabehere, Stephen, and Brian F. Schaffner. 2014. “Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison.” *Political Analysis* 22:285–303.

Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Paul J. Lavrakas, Sunghee Lee, Michael Link, Linda Piekarski, Kumar Rao, Randall K. Thomas, and Dan Zahs. 2010. “AAPOR Report on Online Panels.” *Public Opinion Quarterly* 74:711–81.

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista

J. Gile, and Roger Tourangeau. 2013. “Report of the AAPOR Task Force on Nonprobability Sampling.” Accessed January 2014. Available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf.

Berinsky, Adam J. 2006. “American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data.” *Public Opinion Quarterly* 70:499–529.

Blom, Annelies G., Daniela Ackermann-Piek, Susanne Helmschrott, Carina Cornesse, Christian Bruch, and Joseph W. Sakshaug. 2017. “The Effect of Survey Sampling and Mode on Sample Accuracy and Retention.” Mannheim, Germany: University of Mannheim.

Brick, J. Michael. 2011. “The Future of Survey Sampling.” *Public Opinion Quarterly* 75:872–88.
Brick, J. Michael, Jon Cohen, Sarah Cho, Scott Keeter, Kyley McGeeney, and Nancy Mathiowetz. 2015. “Weighting and Sample Matching Effects for an Online Sample.” Paper presented at the 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL, USA.

Callegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, eds. 2014. *Online Panel Research: A Data Quality Perspective*. West Sussex, UK: John Wiley & Sons.

Callegaro, Mario, and Charles DiSogra. 2008. “Computing Response Metrics for Online Panels.” *Public Opinion Quarterly* 72:1008–32.

Chang, LinChiat, and Jon A. Krosnick. 2009. “National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality.” *Public Opinion Quarterly* 73:641–78.

Converse, Jean M. 1987. *Survey Research in the United States: Roots and Emergence, 1890–1960.* Berkeley: University of California Press.

Couper, Mick P. 2011. “The Future of Modes of Data Collection.” *Public Opinion Quarterly*

75:889–908.

DeBell, Matthew, and Jon A. Krosnick. 2009. *Computing Weights for American National Election Study Survey Data.* ANES Technical Report series, No. nes012427. Ann Arbor and Palo Alto: American National Election Studies. Accessed 2016. Available at https://www.elec- tionstudies. org/wp-content/uploads/2018/04/nes012427.pdf.

Dever, Jill, and Bonnie Shook-Sa. 2015. "The Utility of Weighting Methods for Reducing Errors in Opt-In Web Studies." Paper presented at the International Total Survey Error Conference, Baltimore, MD.

Dutwin, David, and Trent Buskirk. 2017. "Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples." *Public Opinion Quarterly* 81:213–39.

Efron, Bradley, and Rob Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 1:54–75.

Erens, Bob, Sarah Burkill, Mick P. Couper, Frederick Conrad, Soazig Clifton, Clare Tanton, Andrew Phelps, Jessica Datta, Catherine H. Mercer, Pam Sonnenberg, Philip Prah, Kirstin

R. Mitchell, Kaye Wellings, Anne M. Johnson, and Andrew J. Copas. 2014. "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey." J*ournal of Medical Internet Research* 16:e276-1–276-14.

Ferrell, Dan, and James C. Peterson. 2010. "The Growth of Internet Research Methods and the Reluctant Sociologist." *Sociological Inquiry* 80:114–25.

Fisher, Ronald A. 1925. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd. Gittelman, Steven H., Randall K. Thomas, Paul J. Lavrakas, and Victor Lange. 2015. "Quote Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples." *Journal of Advertising Research* 55:368–79.

Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72:167–89.

Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. 2016. "Evaluating Online Nonprobability Surveys." Washington, DC: Pew Research Center. Available at http://www.pewresearch.org/files/2016/04/Nonprobability- report-May-2016-FINAL.pdf.

Kohut, Andrew, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian. 2012. "Assessing the Representativeness of Public Opinion Surveys." Pew Research Center for the People & the Press, May 15. Accessed January 2013. Available at http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf.

Malhotra, Neil, and Jon A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on

Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis* 15:286–324.

Marken, Stephanie. 2018. "Still Listening: The State of Telephone Surveys." *Gallup Methodology Blog,* January 11. Available at http://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. "For Weighting Online Opt-In Samples, What Matters Most?" *Pew Research Center.* Accessed January 26, 2016. Available at http://assets.pewresearch.org/wp-content/uploads/sites/12/2018/01/26170902/Weighting-Online-Opt-In-Samples.pdf.

Moser, Claus, and Alan Stuart. 1953. "An Experimental Study of Quota Sampling." *Journal of the Royal Statistical Society: Series A (General)* 116:349–405.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society 97:558–625*.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel, and Andrew Gelman. 2018. "Disentangling Bias and Variance in Election Polls." *Journal of the American Statistical Association,* March 14. Available at https://pdfs.semanticscholar.org/0e93/bb1c4cbea13d-dd06cb8e44c8fb43a3a8357b.pdf.

Silver, Nate. 2012. "Which Polls Fared Best and Worse in the 2012 Presidential Race." *New York Times,* November 10. Available at http://fivethirtyeight.blogs.nytimes.com/2012/11/10/ which-polls-fared-best-and-worst-in-the-2012-presidential-race/.

Sturgis, Patrick, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Jouni Kuha, Ben Lauderdale, and Patten Smith. 2016. "Report of the Inquiry into the 2015 British General Election Opinion Polls." *Market Research Society and British Polling Council.* Available at https://eprints.soton.ac.uk/390588/1/Report_final_revised.pdf.

Szolnoki, Gergely, and Dieter Hoffmann. 2013. "Online, Face-to-Face and Telephone Surveys—Comparing Different Sampling Methods in Wine Consumer Research." *Wine Economics and Policy* 2:57–66.

Taylor, Humphrey, David Krane, and Randall K. Thomas. 2009. "How Does Social Desirability Affect Responses? Differences in Telephone and Online Surveys." Paper presented at the 64th Annual Meeting of the American Association for Public Opinion Association, Miami Beach, FL, USA.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys.* New York: Oxford University Press.

Walker, Robert, Raymond Pettit, and Joel Rubinson. 2009. "A Special Report from the Advertising Research Foundation: The Foundations of Quality Initiative: A Five-Part Immersion into the Quality of Online Research." *Journal of Advertising Research* 49:464–85.

Wang, Wei, David Rothschild, Sharad Goel, and Andew Gelmnan. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31:980–91.

Yeager, David, Jon Krosnick, Linchiat Chang, Hardold Javitz, Matthew Levendusky, Alberto Simpser, and Rui Wang. 2011. “Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Nonprobability Samples.” *Public Opinion Quarterly* 75:709–47.