Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith

by Courtney Kennedy, Andrew Mercer and Arnold Lau

Release date: June 25, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 3-21
Statistics Canada, Catalogue No. 12-001-X

Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith

Courtney Kennedy, Andrew Mercer and Arnold Lau[1]

Abstract

Statistical approaches developed for nonprobability samples generally focus on nonrandom selection as the primary reason survey respondents might differ systematically from the target population. Well-established theory states that in these instances, by conditioning on the necessary auxiliary variables, selection can be rendered ignorable and survey estimates will be free of bias. But this logic rests on the assumption that measurement error is nonexistent or small. In this study we test this assumption in two ways. First, we use a large benchmarking study to identify subgroups for which errors in commercial, online nonprobability samples are especially large in ways that are unlikely due to selection effects. Then we present a follow-up study examining one cause of the large errors: bogus responding (i.e., survey answers that are fraudulent, mischievous or otherwise insincere). We find that bogus responding, particularly among respondents identifying as young or Hispanic, is a significant and widespread problem in commercial, online nonprobability samples, at least in the United States. This research highlights the need for statisticians working with commercial nonprobability samples to address bogus responding and issues of representativeness – not just the latter.

Key Words: Nonprobability; Online surveys; Measurement error; Benchmarking.

1. Introduction

Survey statisticians have long understood that raw samples from commercial online nonprobability panels or marketplaces are apt to be unrepresentative (e.g., Rivers, 2007; Dever, Rafferty and Valliant, 2008). The statistician's task then is combining the respondents' answers with auxiliary information to adjust away differences between the sample and the target population. Various approaches have been developed for this purpose (see Elliott and Valliant, 2017 for a review). These approaches work, at least in theory, when the answers from nonprobability respondents are genuine and reasonably accurate. But there is growing evidence that sizable shares of nonprobability respondents provide bogus data, including on variables statisticians use for adjustment.

1.1 Focusing on commercial, online nonprobability surveys

Nonprobability samples can take many forms, such as people recruited from a social media platform or a snowball sample of a rare, at-risk population. In American public opinion polling, however, one general form of nonprobability sampling dominates. Over 80% of U.S. public opinion polling is currently conducted using commercial, online nonprobability panels (Kennedy, Hatley, Lau, Mercer, Keeter, Ferno and Asare-Marfo, 2021). Public opinion surveys using commercial online nonprobability samples almost never feature

1. Courtney Kennedy, Andrew Mercer and Arnold Lau, Pew Research Center, 1615 L St., NW, Washington D.C., Suite 800, 20036, U.S.A. E-mail: CKennedy@PewResearch.org.

a companion probability-based sample for purposes of calibration, a characteristic distinguishing them from studies statistically blending probability and nonprobability samples (e.g., Elliott and Haviland, 2007).

Since "commercial online nonprobability panel" is cumbersome, we will use the shorthand "opt-in panel". We urge caution in not conflating such commercial data with other, qualitatively different nonprobability sources, such as those in Beaumont (2022) and Li (2022). Commercial opt-in samples have a set of unique issues, which are the focus of this study.

In Section 2 we provide some background on the nonprobability survey benchmarking literature as well as the bogus respondent literature. In Section 3 we present a new benchmarking study comparing average error levels in six different online survey panels. We observe that errors are particularly large for certain subgroups and that those same subgroups are prone to claiming highly unusual characteristics. In Section 4 we present a follow-up study to tease out whether the claims of unusual characteristics are credible or evidence of bogus responding. In Section 5 we discuss some limitations of these data collections. In Section 6 we provide concluding remarks reflecting on the implications of this research.

2. Background

Two lines of methodological research have developed around data quality in opt-in panels. The first is focused on quantifying the average size of errors and comparing them to the levels in probability-based samples. The second line of research is focused on measurement error in individual survey responses caused by fraudulent, mischievous or otherwise insincere respondents in opt-in panels. Our goal in this paper is to connect these literatures using new data and highlight ways that we see them informing each other.

2.1 Benchmarking literature

Many of the studies focused on representativeness have explored the relative accuracy of nonprobability and probability-based survey estimates compared to available population benchmarks. For example, MacInnis, Krosnick, Ho and Cho (2018) found that weighted estimates from a probability-based online sample had significantly smaller root mean square errors than estimates from six different opt-in panels. Overall, the number of studies finding greater accuracy for probability sample estimates (Malhotra and Krosnick, 2007; Chang and Krosnick, 2009; Yeager, Krosnick, Chang, Javitz, Levendusky, Simpser and Wang, 2011; Szolnoki and Hoffmann, 2013; Erens, Burkill, Couper, Conrad, Clifton, Tanton, Phelps, Datta, Mercer, Sonnenberg, Prah, Mitchell, Wellings, Johnson and Copas, 2014; Sturgis, Baker, Callegaro, Fisher, Green, Jennings, Kuha, Lauderdale and Smith, 2016; Dutwin and Buskirk, 2017; Pennay, Neiger, Lavrakas and Borg, 2018) outnumber those finding similar accuracy in nonprobability estimates (Vavreck and Rivers, 2008; Ansolabehere and Schaffner, 2014) by about five to one.

Other studies in this vein have examined the efficacy of different statistical methods for reducing selection bias in estimates from online opt-in samples. One consistent finding in these studies is that even after employing more sophisticated statistical approaches, such as machine learning or doubly-robust methods, there often remain sizable errors in nonprobability survey estimates (Dutwin and Buskirk, 2017; Mercer, Lau and Kennedy, 2018). For example, Dutwin and Buskirk (2017) note that "advanced techniques such as propensity weighting and sample matching did not improve these (nonprobability) measures, and in some cases made matters worse".

One question raised by this literature is why, even with extensive modeling, do opt-in panel estimates often contain large errors? Is it because the models are mis-specified or missing key covariates, or is the opt-in data flawed in ways immune from statistical modeling correction? In addition to leaving open these questions, the benchmarking literature has another limitation. Most benchmarking studies only consider estimates for the full population (e.g., all U.S. adults). They do not explore the possibility that the accuracy of opt-in estimates may vary between major subgroups (e.g., based on age, race, or ethnicity). Our benchmarking study (presented in Section 3) seeks to address this gap by examining subgroup variation and then using those results to better understand a second body of literature on nonprobability surveys.

2.2 Bogus respondent literature

A separate line of research has documented fraudulent, mischievous, or insincere respondents in commercial opt-in panels. The scale of this problem is alarming. The Insights Association (2022) estimates that researchers should anticipate removing 15% to 25% of opt-in completes due to poor data quality. Geraci (2022) put that rate even higher, noting that "just 10 years ago, researchers would need to remove 5%-10% of all interviews from online samples because of poor quality. That proportion is now in the 35%-50% range". This line of research focuses not on whether opt-in respondents are representative of a broader population, but on whether their answers are credible.

It is increasingly clear that bogus respondents are not merely a nuisance (e.g., adding noise to estimates, requiring replacement interviews). Instead, bogus respondents can lead to highly biased estimates and false conclusions. For example, Litman, Rosen, Rosenzweig, Weinberger-Litman, Moss and Robinson (2021) showed that Centers for Disease Control (CDC) reports of high rates of Americans ingesting bleach to protect against COVID-19 were an artifact of bogus respondents in an opt-in sample. Lopez and Hillygus (2018) found that bogus respondents erroneously inflated estimates of public belief in conspiracies by a factor of two. More recently, Westwood, Grimmer, Tyler and Nall (2022) found that bogus respondents erroneously inflated estimates of support for political violence in the United States by over a factor of two.

While data quality concerns with opt-in samples are not a new phenomenon (e.g., Downes-Le Guin, 2005; Baker, Blumberg, Brick, Couper, Courtright, Dennis, Dillman, Frankel, Garland, Groves, Kennedy, Krosnick, Lavrakas, Lee, Link, Piekarski, Rao, Thomas and Zahs, 2010), research about the extent and impact of insincere responding has accelerated in recent years for several reasons. These include concerns about survey bots (Baxter, 2016; Shanahan, 2018; McDowell, 2019; Puleston, 2019; Geraci, 2022), foreign workers concealing their identity to qualify for U.S. surveys (Kennedy, Clifford, Burleigh, Waggoner and Jewell, 2018; Moss, 2018; Ahler, Roush and Sood, 2019), and people taking a survey multiple times on different devices to evade panel security checks (Ahler et al., 2019; Kennedy et al., 2021). While the companies running commercial opt-in panels are aware of and attempt to address many of these challenges, fraudulent respondents continue to make up a sizeable share of cases in online, opt-in samples.

2.3 Connecting the literatures on commercial nonprobability samples

Statistical approaches developed for nonprobability survey data (e.g., Rivers, 2007; DiSogra, Cobb, Chan and Dennis, 2011; Valliant and Dever, 2011; Valliant, 2020) assume that the task at hand is leveraging on the most effective auxiliary variables. In other words, the problem to be solved is addressing the relevant ways in which the opt-in respondents differ from the target population. Well-established theory states that one can remove selection bias by conditioning on the right set of auxiliary variables (Elliott and Valliant, 2017; Mercer, Kreuter, Keeter and Stuart, 2017; Kohler, Kreuter and Stuart, 2019). There is an implicit assumption in this literature that failure to eliminate error in opt-in estimates simply means we haven't found the right auxiliary variables or they are not available. By this logic, progress comes from finding new sources of auxiliary data or developing statistical methods that can better leverage whatever auxiliary data is available. But this logic only holds if measurement error is nonexistent or small, and in general, the conventional wisdom has been that while satisficing – perhaps the most frequently studied source of measurement error – may be somewhat worse in opt-in surveys, it is not a major contributor to error.

The second line of research discussed raises the possibility that bogus respondents actually introduce much larger measurement error than previously thought. When 15% to 50% of the data collected should be discarded for poor quality, then focusing on auxiliary variables risks missing the bigger picture. Sampling quotas and weighting are unlikely to be effective if the variables they use contain large errors introduced on purpose by bad faith respondents.

To be clear, we do not assert that bogus respondents are the only error source for opt-in panels. Instead, it is far more likely that both bogus respondents (i.e., those answering erroneously) and unrepresentative respondents (i.e., those answering genuinely but who, in aggregate, differ from the target population) contribute to error. Our goal in this study is two-fold: (1) to show how the bogus respondent literature helps explain the benchmarking literature, and (2) highlight the need for statisticians working with opt-in samples to address bogus responding *and* issues of representativeness – not just the latter.

3. Benchmarking study

We initially set out to conduct a benchmarking to examine the extent to which the accuracy of online survey estimates differs by subgroups. The results led us to develop a hypothesis about bogus responding. The benchmarking component was not designed to detect bogus responding, and so we designed a follow-up data collection to address that hypothesis directly. We discuss each data collection in turn.

3.1 Sample design for the benchmarking

The benchmarking study featured samples of U.S. adults from each of six online platforms (Table 3.1). Responding sample sizes ranged from 4,912 to 5,147. Three of the samples came from commercial opt-in panels. The opt-in survey vendors, which routinely use quotas, were provided with the quota targets for age × gender, race × Hispanic ethnicity, education. We computed the quota targets from the 2019 American Community Survey.

Table 3.1
Sample sizes and field dates, by source.

Source	n	Field dates
ABS panel 1	5,027	June 14 - 28, 2021
ABS panel 2	5,147	June 14 - 27, 2021
ABS panel 3	4,965	June 29 - July 21, 2021
Opt-in panel 1	4,912	June 15 - 25, 2021
Opt-in panel 2	4,931	June 11 – 27, 2021
Opt-in panel 3	4,955	June 11 – 26, 2021

Note: "ABS" refers to an online panel that is recruited using probability, address-based sampling. "Opt-in panel" refers to a commercial online nonprobability panel or marketplace.

The three other sources are probability-based survey panels that interview online but recruit panelists offline. Most if not all the panelists in these three sources were recruited using address-based sampling (ABS) from the U.S. Postal Service Computerized Delivery Sequence File. Before adopting ABS recruitment, two of these panels recruited offline using random digit dial samples of telephone numbers.

One of the ABS panel samples was used for this methodological study as well as substantive research that is not part of this project. The substantive research required a larger sample size. For that study, the entire panel was invited to participate, out of which 10,606 panelists completed this questionnaire. We did not want this larger sample size confounding the analysis (i.e., by allowing that one sample to be double the size of the other five). To address this, we drew a stratified random sample from the full panel following the panel's standard procedure for a target sample size of 5,000 completes. Only respondents that were selected as part of this subsample were included in this study. These are the cases that would have been obtained if only that subsample had been invited to participate. The subsampling process was conducted independently from any analysis of the data. As reported below, all three ABS panels performed in a similar manner, giving us confidence that subsampling from the larger panel did not impair this study by rendering its performance meaningfully different.

ABS panels 1, 2 and 3 had study-specific response rates of 61%, 90%, and 71%, respectively. The cumulative response rate to the surveys (accounting for nonresponse to recruitment, to the current survey and for panel attrition) was 5% for ABS panel 1, 3% for ABS panel 2, and 7% for ABS panel 3. Comparable response rates cannot be computed for the opt-in samples. Ipsos was the data collection firm. A common questionnaire was administered in English or Spanish for all six samples.

Each survey panel has their own approach for weighting a national survey of U.S. adults. This project is focused on data quality, and so it was necessary to avoid having different weighting protocols confound comparisons between the samples. We employed a standard weighting approach for all six samples. For the

ABS samples, the weighting protocol began with panel base weights adjusting for differential probabilities of selection. Because design weights do not exist for the opt-in samples, these cases were assigned a starting weight of 1 and treated as if they were simple random samples.

The second step was to calibrate the starting weights for each sample to a common set of population control totals. The population targets were age × sex, education × sex, education × age, race/ethnicity × education, born inside versus outside the U.S. among Hispanics and Asian Americans, years lived in the U.S., Census region × metro/non-metro, volunteered in past year, voter registration status, political party affiliation, frequency of internet use, and religious affiliation. The first six benchmarks came from the American Community Survey. The next three benchmarks came from the Current Population Survey supplements, and the last three came from Pew Research Center National Public Opinion Reference Survey.

3.2 Benchmarking analysis

In total, 25 benchmark measurements were used in this study. They touch on many different topics including smoking, military service, vehicle ownership, health care coverage, income, social program participation, household composition, and more. The benchmarks were derived from high quality federal sources based either on national surveys or administrative data. None of the 25 benchmark variables were part of the weighting protocol. The full list of benchmarks and sources is provided in the appendix.

To estimate survey error relative to benchmarks, we computed the mean absolute value of the difference between the weighted online survey estimate and the benchmark. For a categorical benchmark variable Y with categories $c = 1, \ldots, k,$ let $\overline{Y}_c$ denote the "true" population value and $\overline{y}_c$ denote the survey estimate for the share of the population belonging to category $c.$ For a given survey, the mean absolute error (MAE) for Y is

$$\text{MAE} = \frac{\sum_{c=1}^{k} \left| \overline{y}_c - \overline{Y}_c \right|}{k}. \tag{3.1}$$

In other words, each panel's MAE reflect a two-step process. First, we averaged the difference between the survey and the benchmark across all the answer categories for the question. Second, we computed the average of those 25 averages. We did this separately for each of the six online panels. This way, all the benchmarks have equal influence in the analysis. We see no theoretical justification to do otherwise. Refusal and "Not sure" responses are not included as a benchmarking category. These responses are, however, reflected in the denominator of the online survey estimates because that reflects actual practice. It is unusual for public opinion researchers to re-base estimates just on those providing an answer other than "Refused" or "Not sure".

The average benchmark deviations (Table 3.2) reveal familiar findings as well as new ones. Consistent with prior studies, the estimated average absolute error is systematically lower in the probability-based samples (3.0 percentage points) relative to the nonprobability samples (6.8 percentage points). Within those two groupings, the samples performed roughly the same. The average absolute errors ranged from 2.6 to 3.5 among the probability ABS samples, and it ranged from 5.9 to 7.3 among the nonprobability opt-in samples.

Table 3.2
Average absolute error in online survey estimates for 25 benchmarks.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	High school or less	Some college	College grad	White	Black	Hispanic
ABS Mean	3.0 (0.08)	4.0 (0.31)	3.2 (0.11)	3.1 (0.16)	4.0 (0.16)	3.1 (0.17)	2.5 (0.13)	2.7 (0.09)	4.4 (0.35)	4.2 (0.34)
Opt-in Mean	6.8 (0.09)	11.6 (0.27)	7.4 (0.12)	3.3 (0.13)	7.2 (0.14)	6.4 (0.17)	7.1 (0.18)	5.9 (0.11)	7.6 (0.27)	11.5 (0.29)
ABS 1	2.6 (0.10)	2.9 (0.32)	2.9 (0.15)	2.7 (0.18)	3.2 (0.19)	2.8 (0.19)	2.3 (0.18)	2.4 (0.12)	4.1 (0.47)	3.4 (0.32)
ABS 2	3.5 (0.11)	5.7 (0.45)	3.5 (0.16)	3.6 (0.22)	4.6 (0.23)	3.6 (0.22)	2.8 (0.15)	3.1 (0.12)	4.9 (0.42)	4.7 (0.42)
ABS 3	2.9 (0.16)	3.3 (0.50)	3.3 (0.21)	2.8 (0.27)	4.2 (0.34)	2.9 (0.30)	2.5 (0.17)	2.7 (0.15)	4.2 (0.55)	4.4 (0.63)
Opt-in 1	7.1 (0.16)	11.9 (0.45)	8.1 (0.21)	3.4 (0.17)	7.0 (0.26)	7.3 (0.25)	7.9 (0.31)	6.4 (0.17)	8.2 (0.44)	11.7 (0.48)
Opt-in 2	7.3 (0.15)	12.8 (0.47)	8.1 (0.22)	3.5 (0.20)	7.3 (0.25)	6.6 (0.28)	8.5 (0.33)	6.6 (0.18)	8.6 (0.41)	11.7 (0.47)
Opt-in 3	5.9 (0.15)	10.2 (0.53)	6.1 (0.19)	3.0 (0.19)	7.3 (0.25)	5.3 (0.31)	5.0 (0.25)	4.8 (0.18)	6.1 (0.44)	11.2 (0.48)

Note: Standard errors are shown in parentheses. Estimates for White and Black adults are based on those who do not identify as Hispanic.

More novel, and perhaps under-appreciated in the field, is the substantial variation by subgroup. For the opt-in samples, the average absolute error in estimates for young adults (ages 18-29) is more than three times larger than the error for adults ages 65 and older (11.6 versus 3.3 percentage points). For the ABS samples, by contrast, the average absolute error for estimates based on younger and older adults is far more similar (4.0 versus 3.1 percentage points).

Table 3.2 shows a similar pattern with respect to Hispanic ethnicity. For the opt-in samples, the average absolute error in estimates for Hispanics is nearly double that for (non-Hispanic) White adults (11.5 versus 5.9 percentage points). For the ABS samples, by contrast, the average absolute error for estimates based on Hispanic and White adults is more comparable (4.2 and 2.7 percentage points).

Put differently, after extensively weighting each of the opt-in samples, the estimates for young adults and for Hispanic adults were off by more than 10 percentage points on average. Those are large errors, and it is not clear from the literature why they are concentrated in those two groups. The opt-in estimates for adults ages 65 and over, for example, are relatively accurate, departing from the benchmarks by only about 3 percentage points. Other demographic variables could be examined, of course. We also carried out this analysis by gender and educational attainment, but the variance across those dimensions was much more muted than the differences by age and ethnicity, and so education and gender are not considered further in this study. Instead, the remaining analysis considers why these errors are so concentrated among those self-identifying as young or Hispanic.

3.3 Benchmarks with the largest errors

To investigate why commercial online nonprobability survey errors are concentrated in certain subgroups, we take a closer look at variables where the errors are largest. Table 3.3 presents weighted estimates for the share of U.S. adults receiving four different government benefits: Supplemental Nutritional Assistance Program (SNAP), Social Security, Unemployment Compensation, and Worker's Compensation. The key finding is not simply that the survey estimates contain error, but the errors are all in one direction. Namely, the commercial online nonprobability surveys contain proportionately too many respondents claiming to receive these benefits. The same pattern is observed for the ABS samples, but the magnitude of the errors is dramatically different.

Table 3.3
Estimates for receipt of four different government benefits.

	SNAP	Social security	Unemployment Compensation	Worker's Compensation
Benchmark	*11.1%*	*21.8%*	*9.3%*	*0.4%*
ABS 1	14.0%	25.6%	12.4%	1.3%
	(0.57)	(0.55)	(0.52)	(0.20)
ABS 2	19.0%	27.7%	17.2%	2.9%
	(0.71)	(0.60)	(0.68)	(0.35)
ABS 3	18.4%	25.6%	13.0%	1.6%
	(0.83)	(0.63)	(0.72)	(0.31)
Opt-in 1	29.9%	38.7%	18.8%	10.0%
	(0.77)	(0.74)	(0.68)	(0.50)
Opt-in 2	30.0%	37.5%	21.0%	11.8%
	(0.81)	(0.72)	(0.71)	(0.51)
Opt-in 3	21.7%	34.7%	16.9%	7.6%
	(0.67)	(0.73)	(0.65)	(0.47)

Note: Standard errors are shown in parentheses. SNAP is the Supplemental Nutritional Assistance Program.

For example, receipt of Worker's Compensation is an incredibly rare characteristic among U.S. adults. The population incidence is less than 1%. But according to the commercial online nonprobability samples, the incidence is closer to 10%. Similarly, for receipt of nutritional assistance, the commercial online nonprobability samples estimate the incidence at 22% to 30%, while the true population rate is just 11%.

These results suggest that commercial online nonprobability respondents are prone to saying "Yes" they have certain characteristics. To further distill that phenomenon, Table 3.4 presents the weighted share of respondents in each sample who claimed to receive at least three of the four government benefits measured in the survey.

Again, it is important to note that receiving at least three of these benefits is incredibly rare (0.1% population incidence). The probability-based ABS samples reflect these dynamics, with estimates for the share of adults receiving three or more of these benefits ranging from 0% to 1%. According to the commercial online nonprobability samples, however, the incidence ranges from 6% to 11%.

Table 3.4
Percentage of adults who self-report receiving at least three of four different government benefits.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	HS or less	Some college	College grad	White	Black	Hispanic
Benchmark	*0.1%*	*0.1%*	*0.1%*	*0.2%*	*0.2%*	*0.1%*	*0.0%*	*0.1%*	*0.2%*	*0.1%*
ABS 1	0.8%	1.1%	0.8%	0.4%	1.2%	0.9%	0.3%	0.3%	2.7%	1.1%
	(0.16)	(0.51)	(0.20)	(0.17)	(0.33)	(0.33)	(0.11)	(0.09)	(0.96)	(0.50)
ABS 2	1.2%	2.0%	1.2%	0.7%	2.2%	1.2%	0.3%	0.4%	3.1%	2.7%
	(0.24)	(0.78)	(0.28)	(0.39)	(0.55)	(0.38)	(0.10)	(0.15)	(1.08)	(0.86)
ABS 3	1.4%	2.4%	1.4%	0.4%	2.0%	1.7%	0.3%	1.0%	2.3%	2.9%
	(0.30)	(1.08)	(0.39)	(0.18)	(0.68)	(0.63)	(0.21)	(0.32)	(1.17)	(1.33)
Opt-in 1	7.8%	17.8%	6.9%	0.8%	5.1%	7.5%	11.1%	4.3%	12.4%	17.2%
	(0.44)	(1.44)	(0.58)	(0.35)	(0.69)	(0.77)	(0.96)	(0.46)	(1.60)	(1.47)
Opt-in 2	9.0%	18.0%	8.9%	0.2%	6.0%	7.7%	13.7%	6.9%	10.8%	16.9%
	(0.42)	(1.42)	(0.59)	(0.12)	(0.60)	(0.85)	(0.87)	(0.48)	(1.47)	(1.47)
Opt-in 3	5.9%	14.7%	4.9%	0.7%	6.9%	5.5%	5.1%	3.1%	6.8%	18.6%
	(0.41)	(1.60)	(0.45)	(0.25)	(0.81)	(0.77)	(0.62)	(0.37)	(1.36)	(1.78)

Note: Standard errors are shown in parentheses. The four government benefits are the Supplemental Nutritional Assistance Program (SNAP), Social Security, Unemployment Compensation and Workers' compensation. Estimates for White and Black adults are based on those who do not identify as Hispanic.

Table 3.4 shows these estimates for the subgroups with the largest errors: young adults and Hispanics. For these groups, the error in the commercial online nonprobability estimates is staggering. According to the commercial online nonprobability samples, about 20% of young adults and 20% of Hispanics receive at least three of these benefits. These results beg our central question: What is more likely – that these young and/or Hispanic nonprobability cases actually have this rare characteristic, or that these cases are misrepresenting themselves (i.e., providing bogus answers). As mentioned above, the answer is critical because statistical techniques for commercial online nonprobability data are premised on the first explanation, even though it strains credulity.

3.4 Removing apparently bogus respondents from estimates

A natural question is how the accuracy evaluation changes if the apparently bogus respondents are dropped from analysis. To examine this, we repeated the benchmarking exercise in Section 3.2, but this time we removed all respondents who reported receiving the four government benefits measured (SNAP, Social Security, Unemployment Compensation, and Worker's Compensation). We then re-weighted each sample according to the procedure described above. This yielded some accuracy improvement but was far from a panacea. The results are provided in the appendix. The average benchmark deviation for the nonprobability samples improved by 21% (from 6.8 percentage points on average to 5.4). The nonprobability estimates for young adults and Hispanics improved, but the average errors remained higher than for other demographic groups (8.6 and 8.3 percentage points on average for young adults and Hispanics, respectively). Removing the apparently bogus respondents did not close the accuracy gap between the probability and nonprobability samples, as the average benchmark deviation for the probability samples (2.8 percentage points) remained smaller.

In sum, removing respondents with an egregiously suspicious response pattern helps rather than hurts accuracy, but we do not view this as a robust solution. Those claiming to have received four disparate government benefits are just one subset of all the possible bogus respondents in the nonprobability samples. Prior research (e.g., Kennedy et al., 2021) indicates that other tests for bogus responding would flag a different, if partially overlapping, set of problematic respondents. Relying on a single test or single response pattern is unlikely to completely diagnose bogus responding.

4. Examining bogus responding directly

Results from the benchmarking study suggest that some opt-in subgroups are prone to giving data that is not credible. However, the benchmarking study was not designed to distinguish credible from non-credible responses. A more direct test of bogus responding would be needed to definitively determine if the patterns in the benchmarking analysis stemmed from "unusual but genuine" opt-in respondents or from bogus respondents. Again, we feel this distinction matters because statistical approaches for opt-in data assume opt-in respondents may be unusual but are genuine.

4.1 Sample design for the follow-up

In February of 2022 we conducted a short (14 question) survey of $n = 569$ U.S. adults using a different opt-in panel from the three in the benchmarking study. We refer to this as the "follow-up survey" out of recognition that its purpose was to follow-up on and further probe intriguing findings from the benchmarking study. If the patterns from the benchmarking study replicated in this fourth, separate opt-in panel, that would be strong evidence of a systemic problem in the U.S. opt-in panel space. No probability-based samples were included in this follow up because neither the benchmarking study nor other studies (e.g., Kennedy et al., 2021) find meaningful levels of bogus respondents in probability-based samples.

The goal of the follow-up study was to determine if purportedly Hispanic and/or young opt-in respondents are prone to saying "Yes" no matter what is asked. To assess this, we selected questions for which a "Yes" answer is not credible. One question asked, "Are you licensed to operate a class SSGN submarine?" and offered Yes or No as responses. A class SSGN is a nuclear-powered U.S. naval submarine equipped with cruise missiles, of which there are only four in operation by the U.S. Navy. With approximately 425,000 active duty and reserve personnel, the entire U.S. Navy comprises less than 0.2% of the U.S. adult population, making the share of U.S. adults qualified to operate such a vessel is approximately 0%. A separate question in the follow-up survey was formatted as a battery asking "which of the following did you do in the past week? Check all that apply". The list of activities included two common activities (watched TV, read a book) and four extraordinarily uncommon activities (purchased a private jet, climbed a peak in the Karakoram Mountains, learned to cook halušky, and played jai alai).

The goal of the follow-up study was to determine how many opt-in respondents would select the non-credible answers and whether that behavior was concentrated among respondents identifying as Hispanics and young adults, as it was in the benchmarking component. Making inferences about the U.S. public was not the research goal, and so this analysis is not weighted.

4.2 Results of the follow-up survey

The follow-up experiment bore out the post-hoc hypothesis from the benchmarking: opt-in respondents identifying as young or Hispanic were prone to giving bogus responses. They are prone to reporting affirmatively that they have some characteristic when that is simply not possible in the aggregate numbers observed. The first column of Table 4.1 shows that overall, 5.3% of the follow-up survey respondents claimed to be licensed to operate a class SSGN nuclear submarine. Echoing the benchmarking study, the incidence of this bogus claim was particularly high among Hispanics (23.7%) and those under age 30 (12.1%).

The pattern was the same for claims of doing at least one of the extremely uncommon activities in the past week. The share reporting at least one extremely uncommon activity in the past week (buying a private jet, climbing the Karakoram Mountains, learning to cook halušky, or playing jai alai) was significantly higher among those age 18 to 29 than those age 30 and over ($t = 2.99,\ p < 0.01$). Likewise, the share reporting at least one extremely uncommon activity in the past week was significantly higher among Hispanics than non-Hispanics ($t = 5.11,\ p < 0.01$). These findings make clear what the benchmarking suggested: that commercial opt-in respondents in these subgroups are prone to giving bogus answers.

Table 4.1
Commercial online nonprobability estimates of extremely rare population characteristics.

	Licensed to operate a nuclear submarine	**Extremely low incidence behavior**
All adults	5.3%	8.4%
	(0.94)	(1.17)
Ages 18-29	12.1%*	17.2%*
	(3.03)	(3.51)
Ages 30+	3.5%	6.2%
	(0.87)	(1.13)
Hispanic	23.7%*	28.0%*
	(4.41)	(4.66)
Non-Hispanic	1.5%	3.6%
	(0.56)	(0.87)

Note: Standard errors are shown in parentheses. Extremely low incidence behavior was defined as having done any of the following activities in the past week: purchasing a private jet, climbing a peak in the Karakoram Mountains, learning to cook halusky, or playing jai alai. Asterisk (*) indicates that the proportion is significantly higher ($p < 0.01$) than for the complementary group based on 2-tailed t-test assuming unequal variances.

4.3 Implications for statisticians working with opt-in data

These results raise an important question. If Hispanic opt-in cases are clearly answering falsely when self-reporting things like operating a nuclear submarine, might they be answering falsely when they self-report being Hispanic? With opt-in data it is generally not possible to validate respondents' ethnicity. That said, the weight of evidence for bogus responding presented here suggests that all answers from respondents making implausible claims should viewed with skepticism. Indeed, the follow-up survey was a mere 14 questions, eliminating excuses along the lines of "maybe some people get tired near the end of long surveys". That is not what the data show. Instead, we see people claiming Hispanic ethnicity also claiming a series of implausible characteristics. Critically, this replicated in four different commercial nonprobability panels. The simplest explanation, which we also consider to be the most credible, is that some opt-in respondents are prone to saying "Yes" no matter what the question is asking, and this holds for adjustment variables as well as survey outcome variables.

The implication for statisticians working with these data is that some of the variables they use to reduce bias (especially variables measured with a Yes/No format) may contain large errors. Moreover, those errors may be concentrated in certain subgroups, rather than distributed randomly within the responding sample. Consequently, statistical techniques for estimation with commercial opt-in data may not work as well as previously thought. Studies ignoring the existence of bogus respondents in these types of samples are at high risk of overstating the performance of various modeling approaches.

It is less clear why bogus responding is also prevalent among young opt-in respondents. Age is not measured with a Yes/No question, and so we would not expect the same positivity bias that appears to be at play with Hispanic ethnicity. In the benchmarking survey, respondents were asked to select their year of birth from a dropdown menu with years ordered from highest to lowest. In the follow-up study, a binned age variable was provided by the sample vendor. Bogus responders may be simply selecting answers toward the top of the list. It is also possible that the choices are strategic, with bogus responders choosing answers that make them more likely to qualify for a survey or potentially to receive higher incentives. While it seems possible that bogus respondents may in fact skew young, there is little reason to believe that demographics and other auxiliary variables are measured any more accurately than substantive ones. Indeed, the distinction between these two types of variables is only meaningful to statisticians.

5. Limitations

This study addresses one class of nonprobability surveys (i.e., commercial online opt-in panels or marketplaces) in one country (the United States). We would not expect the types of errors observed here to be present in nonprobability samples fielded under qualitatively different circumstances. In commercial nonprobability sources, bogus respondents are rewarded for their bad behavior because they often received incentives with monetary value, but there may be no such reward structure in, for example, a bespoke, offline sample of an at-risk population. We also cannot know from our data whether the findings from this study generalize to commercial nonprobability panels in other countries.

Another limitation of this study concerns the benchmarking analysis presented in Section 3. While benchmarking analysis is useful as a means of evaluating the accuracy of survey estimates, it has its limitations. The benchmarks in this study are drawn from government-funded surveys that are conducted at considerable expense and with great attention to survey quality. But they are surveys nevertheless and subject to some of the same problems facing the online surveys. The surveys used as benchmarks have high response rates, on the order to 60% or more. Accordingly, the risk of nonresponse bias is generally thought to be lower for these surveys, though it still exists. Also relevant is the fact that all surveys, no matter the response rate, are subject to measurement error. Questions asked on government-funded surveys are carefully developed and tested, but they are not immune to some of the factors that create problems of reliability and validity in all surveys. The context in which a question is asked (e.g., the questions that come before it) often affects responses to it. Similarly, all survey items may be subject to some degree of response bias, most notably social desirability bias. Especially when an interviewer is present, respondents may sometimes modify their responses to present themselves in a more favorable light (e.g., by overstating their frequency of voting). All of these factors can affect the comparability of seemingly identical measures asked on different surveys. Assessing the quality of data is an inexact process at best. It is therefore important to bear in mind that benchmarking provides measures of estimated bias and is dependent on the particular set of measures included.

6. Discussion

This study is the first to use benchmarking to identify subgroups where bogus responding is concentrated. It is also the first to demonstrate that high rates of bogus data among young adults and self-identified Hispanics appears to be an industry-wide phenomenon, at least in the United States. Four different commercial nonprobability panels or marketplaces all showed the same pattern. By contrast, opt-in estimates for adults ages 65 and older were relatively accurate (mean absolute error of 3.3 percentage points). This suggests that nonprobability approaches, such as hybrid designs, may be differentially effective depending on the subgroup of interest.

This study also raises the question of whether Hispanic adults are more prone to providing bogus answers than other adults. We do not think that is a credible explanation for the results observed. To our minds, realizing that many of these "Hispanic" cases are, in all likelihood, not actually Hispanic is the single most important finding. Its implications for survey statisticians are profound. The implication is that we should not be relying exclusively on techniques like sample matching, propensity models, or hierarchical regressions to fix errors in commercial nonprobability samples. All such approaches assume that respondents are who they say they are; that their demographic information is measured with little to no error. This study demonstrates that with certain types of nonprobability data (namely commercial, online panels), that is not a safe assumption.

Some researchers working with commercial online data are aware of the threat from bogus respondents and take steps to mitigate it. However, more research is needed around the efficacy of current practices because there is some evidence that they are inadequate. Kennedy et al. (2021) found that 84% of bogus respondents passed an attention check (or "trap" question), 87% passed a check for too-fast response time, and 76% of bogus respondents passed both of those popular data quality checks. More sophisticated detection techniques have been proposed (e.g., Jones, House and Gao, 2015), but they do not appear to be widely adopted.

A related concern is that public reporting of results from commercial nonprobability panels rarely discloses whether and how the threat from bogus respondents was addressed. At a minimum, researchers reporting results based on this type of data should disclose what measures were taken to guard against bogus respondents and to what effect. While some organizations may already provide this information, robust disclosure is far from common. Greater awareness and transparency around the existence of bogus respondents in commercial nonprobability samples may help to reduce instances of erroneous findings (e.g., Litman et al., 2021; Westwood et al., 2022) and promote greater caution when interpreting findings from nonprobability samples.

Acknowledgements

Courtney Kennedy is grateful to the Morris Hansen Award Committee for selecting her to present at the 29th Annual Morris Hansen Lecture. All the authors thank Scott Keeter, the editor, and the referees for their constructive suggestions.

Appendix

Table A.1
Benchmarking variables and source.

Variable	Benchmark Source	Question Wording
English proficiency	2019 American Community Survey	Do you speak a language other than English at home? [Ask if speaks a language other than English at home] How well do you speak English? Very well; Well; Not well; Not at all
Citizenship	2019 American Community Survey	Are you a citizen of the United States?
Parent of child in household	2020 National Health Interview Survey	Are you the parent or guardian of any children under age 18? [Ask if parent or guardian of child under age 18] Are any of those children under 18 now living in your household?
Marital status	2021 Current Population Survey March Supplement	Which of these best describes you? Married; Living with a partner; Divorced; Separated; Widowed; Never been married
Number of adults in household	2019 American Community Survey	How many people, including yourself, live in your household? [Ask if more than one person in household] How many, including yourself, are adults, age 18 and older?
Number of children in household	2019 American Community Survey	How many people, including yourself, live in your household? [Ask if more than one person in household] How many, including yourself, are adults, age 18 and older?
Health insurance	2020 National Health Interview Survey	Are you currently covered by any form of health insurance or health plan?
Retirement account	2021 Current Population Survey March Supplement	At any time during 2020 did you have any retirement accounts such as a 401(k), 403(b), IRA, or other account designed specifically for retirement savings?
Received food stamps	2021 Current Population Survey March Supplement	At any time during 2020, did you or anyone in your household receive benefits from SNAP (the Supplemental Nutritional Assistance Program) or the Food Stamp program, or use a SNAP or food stamp benefit card?
Received social security	2021 Current Population Survey March Supplement	During 2020 did you receive any Social Security payments from the U.S. Government?
High blood pressure	2020 National Health Interview Survey	Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?
Food allergy	2009-2010 National Health and Nutrition Examination Survey	Do you have any food allergies?
Smoking history	2020 National Health Interview Survey	Have you smoked at least 100 cigarettes in your entire life? [Ask if ever smoked 100 cigarettes] Do you now smoke cigarettes... Every day; Some days; Not at all
Vaping history	2020 National Health Interview Survey	Have you ever used an e-cigarette or other electronic vaping product, even just one time, in your entire life? [Ask if ever used e-cigarette] Do you now use e-cigarettes or other electronic vaping products… Every day; Some days; Not at all

Table A.1(continued)
Benchmarking variables and source.

Variable	Benchmark Source	Question Wording
Moved in last year	2021 Current Population Survey March Supplement	Were you living in this house or apartment 1 year ago?
Type of residence	2019 American Community Survey	Which best describes the building where you currently live? (Include all apartments, flats, etc., even if vacant). A mobile home; A one-family house detached from any other house; A one-family house attached to one or more houses; A building with 2 or more apartments; Boat, RV, van, etc.
Home ownership	2019 American Community Survey	Which of the following describes the house, apartment or mobile home where you live? Owned by you or someone in your household with a mortgage or loan (include home equity loans); Owned by you or someone in your household free and clear (without a mortgage or loan); Rented; Occupied without payment of rent
Number of cars	2019 American Community Survey	How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of your household?
Job status last week	2021 Current Population Survey March Supplement	Last week, did you do any work either for pay or profit? [Ask if did not work last week or refused] Last week, did you have a job either full or part time? Include any job from which you were temporarily absent.
Work affected by Covid-19	2021 Current Population Survey March Supplement	At any time in the last 4 weeks, were you unable to work because your employer closed or lost business due to the Coronavirus?
Had a job last year	2021 Current Population Survey March Supplement	Did you work at a job or business at any time during 2020?
Union membership	2021 Current Population Survey March Supplement	Are you a member of a labor union or of an employee association similar to a union?
Received unemployment compensation	2021 Current Population Survey March Supplement	At any time during 2020, did you receive any State or Federal unemployment compensation?
Received worker’s compensation	2021 Current Population Survey March Supplement	During 2020 did you receive any Worker’s Compensation payments or other payments as a result of a job-related injury or illness?
Military/veteran status	2019 American Community Survey	Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?

Table A.2
Average absolute error in online survey estimates for 25 benchmarks, after removing apparently bogus cases.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	High school or less	Some college	College grad	White	Black	Hispanic
ABS Mean	2.8	3.7	3.1	3.0	3.7	2.9	2.5	2.7	4.1	3.8
	(0.08)	(0.30)	(0.10)	(0.17)	(0.16)	(0.18)	(0.13)	(0.09)	(0.37)	(0.32)
Opt-in Mean	5.4	8.6	6.1	3.3	6.2	5.3	5.0	4.9	6.4	8.3
	(0.09)	(0.25)	(0.12)	(0.14)	(0.14)	(0.17)	(0.14)	(0.10)	(0.24)	(0.24)
ABS 1	2.5	2.7	2.8	2.7	3.1	2.7	2.2	2.4	3.7	3.2
	(0.09)	(0.33)	(0.13)	(0.18)	(0.19)	(0.18)	(0.18)	(0.12)	(0.45)	(0.30)
ABS 2	3.3	5.3	3.4	3.5	4.3	3.5	2.8	3.0	4.6	4.3
	(0.11)	(0.42)	(0.16)	(0.22)	(0.24)	(0.23)	(0.16)	(0.13)	(0.38)	(0.40)
ABS 3	2.6	3.2	3.0	2.8	3.7	2.6	2.5	2.6	3.9	3.8
	(0.15)	(0.43)	(0.19)	(0.28)	(0.31)	(0.29)	(0.18)	(0.15)	(0.60)	(0.57)
Opt-in 1	5.8	8.5	6.9	3.3	6.2	6.1	5.5	5.6	6.6	8.2
	(0.15)	(0.39)	(0.21)	(0.18)	(0.26)	(0.23)	(0.25)	(0.17)	(0.38)	(0.40)
Opt-in 2	5.6	9.4	6.4	3.6	6.3	5.3	5.6	5.0	7.4	8.5
	(0.15)	(0.41)	(0.21)	(0.19)	(0.27)	(0.29)	(0.23)	(0.16)	(0.40)	(0.44)
Opt-in 3	4.8	7.8	5.2	3.0	6.2	4.4	3.9	4.2	5.2	8.2
	(0.15)	(0.44)	(0.19)	(0.18)	(0.23)	(0.29)	(0.23)	(0.16)	(0.41)	(0.41)

Note: Apparently bogus cases are defined here as those who claimed to have received all four government benefits measured.

References

Ahler, D.J., Roush, C.E. and Sood, G. (2019). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. Presented at the Annual Meeting of the Midwest Political Science Association, April 6, 2019.

Ansolabehere, S., and Schaffner, B. (2014). Does survey mode still matter? findings from a 2010 multi-mode comparison. *Political Analysis*, 22(3), 285-303.

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. and Zahs, D. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, 74, 711-81.

Baxter, K. (2016). On the internet, nobody knows you're a bot participant: How bots are contaminating online research data and how we can stop them. Medium. Available at https://medium.com/salesforce-ux/on-the-internet-nobody-knows-youre-a-bot-participant-327dd0da5ce7/.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

Chang, L., and Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73, 4, 641-678.

Dever, J., Rafferty, A. and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, 47-62.

DiSogra, C., Cobb, C., Chan, E. and Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, Proceedings of the Survey Methods Section*, 4501-15. Alexandria, VA: American Statistical Association.

Downes-Le Guin, T. (2005). Satisficing behavior in online panels. Presentation at the MRA Annual Conference & Symposium, Chicago, IL, USA. http://www.sigmavalidation.com/tips/05_06_02_Online_Panelists.pdf.

Dutwin, D., and Buskirk, T.D. (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, 213-239.

Elliott, M.N., and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, 33, 2, 211-215. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10498-eng.pdf.

Elliott, M.R., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Erens, B., Burkill, S., Couper, M.P., Conrad, F., Clifton, S., Tanton, C., Phelps, A., Datta, J., Mercer, C.H., Sonnenberg, P., Prah, P., Mitchell, K.R., Wellings, K., Johnson, A.M. and Copas, A.J. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *Journal of Medical Internet Research*, 16(12), e:276.

Geraci, J. (2022). Poll-Arized: Why Americans Don't Trust the Polls and How to Fix Them Before It's Too Late. Houndstooth Press, 153.

Insights Association (2022). Online sample fraud: Causes, costs, and cures. Online, February 11, 2022.

Jones, M.S., House, L.A. and Gao, Z. (2015). Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. *Public Opinion Quarterly*, 79, 687-709.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. and Jewell, R. (2018). How Venezuela's economic crisis is undermining social science research − About everything. *Washington Post*, November 7, 2018. Available at https://www.washingtonpost.com/news/monkey-cage/wp/2018/11/07/how-the-venezuelan-economic-crisis-is-undermining-social-science-research-about-everything-not-just-venezuela/.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J. and Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, 85, 1050-1075.

Kohler, U., Kreuter, F. and Stuart, E.A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, 6, 149-172.

Li, Y. (2024). Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples. *Survey Methodology*, 50, 1, 37-55. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00008-eng.pdf.

Litman, L., Rosen, Z., Rosenzweig, C., Weinberger-Litman, S.L., Moss, A.J. and Robinson, J. (2021). Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data. MedRxiv, DOI: https://doi.org/10.1101/2020.12.11.20246694.

Lopez, J., and Hillygus, D.S. (2018). Why so serious? survey trolls and misinformation. Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

MacInnis, B., Krosnick, J.A., Ho, A.S. and Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82, 707-744.

Malhotra, N., and Krosnick, J. (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(3), 286-323.

McDowell, B. (2019). Minimizing the impact of survey bots. Quirk's Media. Available at https://www.quirks.com/articles/minimizing-the-impact-of-survey-bots.

Mercer, A.W., Kreuter, F., Keeter, S. and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271.

Mercer, A., Lau, A. and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? Available at https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Moss, A. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it, CloudResearch. Available at https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

Pennay, D.W., Neiger, D., Lavrakas, P.J. and Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia (2nd ed.). The Australian National University. http://csrm.cass.anu.edu.au/research/publications/methods-research-papers.

Puleston, J. (2019). Panel Hacking. Presented at the Annual Conference of the Association for Survey Computing. Available at https://ascconference.org/wp-content/uploads/2019/04/11-Jon-Puleston-Panel-hacking-ASC-2019.pdf.

Rivers, D. (2007). Sampling for web surveys. Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA.

Shanahan, T. (2018). Are you paying bots to take your online survey? Fors Marsh Group. Available at https://www.forsmarshgroup.com/knowledge/news-blog/posts/2018/march/are-you-paying-bots-to-take-your-online-survey/.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. and Smith, P. (2016). Report of the inquiry into the 2015 British general election opinion polls. London: Market Research Society and British Polling Council.

Szolnoki, G., and Hoffmann, D. (2013). Online, face-to-face and telephone surveys − Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, 2, 57-66.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-37.

Vavreck, L., and Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion & Parties*, 18(4), 355-366.

Westwood, S.J., Grimmer, J., Tyler, M. and Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119(12).

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A. and Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709-747.