Public Opinion Quarterly (2023) **Vol 00 No 0,** 1–25

Are Nonprobability Surveys Fit for Purpose?

Jennifer Jerit[1,*], Jason Barabas[2]

[1]Professor, Department of Government, Dartmouth College, Hanover, NH, US
[2]Professor, Department of Government, and Director of the Rockefeller Center for Public Policy and the Social Sciences, Dartmouth College, Hanover, NH, US

Abstract Social scientists employ survey methods to explore the contours of human behavior. Today there are more opportunities to collect survey data than at any time in recent history. Yet sample quality varies dramatically due in part to the availability of nonprobability samples (NPSs) from commercial survey organizations. While these kinds of surveys have advantages in terms of cost and accessibility, the proprietary nature of the data can be problematic. In this synthesis, we describe situations in which researchers typically employ NPSs and consider whether these data are fit for purpose. Next, we discuss use cases that are not widespread but may be appropriate for these data. We conclude that potential utility of NPSs will remain out of reach unless scholars confront the tension between the operation of online survey organizations and the goals of transparent research.

Of the many research methods employed by social scientists, surveys are uniquely connected to the study of democratic politics (Converse 1987). Today, surveys are the "dominant way" politicians, media organizations, interest groups, and scholars assess public preferences (Berinsky 2017, p. 310; see Herbst 1993 for a historical account). Yet, as Couper observes, "the one constant in survey research seems to be change" (2011, p. 905). A dramatic example is the evolution in survey administration in the modern era, from face-to-face interviewing to random-digit-dial (RDD) telephone interviewing to the self-completion of internet questionnaires. We explore another development with significant consequences for the conduct of survey research: the growing availability of nonprobability samples (NPSs).

An NPS is any sample in which respondents enter the study in a nonrandom fashion, often in response to advertisements, pop-up solicitations, or similar invitations (Lavrakas et al. 2019; Hillygus 2020). It is common to label NPSs "opt-in" because respondents select into the sample. By contrast, in

*Corresponding author: Jennifer Jerit, Department of Government, Dartmouth College, Silsby Hall, Hanover, NH 03755, US; email: jennifer.l.jerit@dartmouth.edu.

https://doi.org/10.1093/poq/nfad037

a probability sample, the researcher "[selects individuals] from a sampling frame that contains all members of a target population" (Callegaro et al. 2014, p. 6). A burgeoning literature investigates features of NPSs, such as representativeness and response quality (e.g., Callegaro et al. 2014; Cornesse et al. 2020; Peer et al. 2021; Ternovski and Orr 2022), even as scholars increasingly rely on these data. Indeed, MacInnis et al. write, "Most internet surveys today are done with nonprobability samples of people who volunteer to complete questionnaires . . . and who were not selected randomly from the population of interest" (2018, p. 708).[1]

It may not be apparent why the growing reliance on NPSs is problematic. After all, their dramatically lower cost (compared to probability samples) makes it possible for more people to collect survey data. But NPSs from commercial vendors have some distinctive features relative to other low-cost survey data. Chief among them is the provenance of the data (Krupnikov, Nam, and Style 2021). When one contracts with a commercial survey firm, the vendor cultivates and manages the sample, often through methods that are not transparent to the client. While there is variation in how specific survey organizations operate, a key commonality is that the data are a culmination of a multistep—and largely invisible—selection process.

The selection process for NPS studies consists of two main parts: ***recruitment***, whereby people "become eligible for inclusion in one or more surveys (e.g., joining a panel)," and ***sampling***, or "the process by which an individual is selected for a particular survey after recruitment" (Mercer et al. 2017, p. 258). In the world of commercial survey organizations, this characterization is itself a simplification since recruitment may take place through several methods. The two basic approaches are recruitment through panels (i.e., lists of people, aka "panelists," who have agreed to take surveys on an ongoing basis) and "the river" (also called "intercepts" because one-time respondents are directed to a survey after clicking through an internet link). Likewise, the designation of a "sampling" stage is a misnomer since the process is closer to "purposive selection" (Mercer et al. 2017, p. 260). For any given study, firms try to meet quotas on specific respondent characteristics (e.g., demographics). Survey companies also may use a "router that 'matches' willing panelists to surveys for which they are likely to qualify" (Unangst et al. 2020, p. 73). With opt-in panels, there is no single sampling method: there

1. Crowdsourced samples (e.g., MTurk) and the typical student sample may be characterized as nonprobability because respondents are not selected from a sampling frame. Here, we focus on NPSs from commercial survey vendors. Among applied researchers, there is a perception that data from a survey vendor is superior to student or crowdsourced samples because the company "often attempts some form of national representativeness by performing deliberate balancing on the types of respondents who are invited to take part" (Krupnikov, Nam, and Style 2021, p. 166). This characteristic, combined with the ease of collecting data from online vendors, has led to a dramatic increase in their use.

are "myriad varied sampling methods" (Baker et al. 2013, p. 92). Unlike probability samples, whose providers can—and routinely do—provide the details of the data-generating process, the path to being in a NPS "is not entirely trackable" (Dutwin and Buskirk 2017, p. 235).[2]

Some of the early concerns with NPSs had to do with the "professionalism" of the people completing surveys. There is evidence, for example, that a small group of frequent survey-takers participate in multiple online panels and that repeat respondents have different characteristics than other respondents (Hillygus, Jackson, and Young 2014). Subsequent research has uncovered dishonest and fraudulent responding in some online samples (Peer et al. 2021; Bell and Gift 2022). In an extensive examination of online sources of polling data, researchers at Pew found that a small, but measurable, percentage of participants (i.e., 4 percent to 7 percent) should be classified as "bogus respondents" (Kennedy et al. 2020). Bogus respondents tend to give affirmative responses to survey questions, which can lead to nonsensical patterns on substantive questions (e.g., endorsing both President Trump and Obamacare) as well as the overreporting of demographic characteristics. Indeed, Kennedy et al. (2020) found that bogus respondents were three times more likely to self-identify as Hispanic or Latino (i.e., because that item has a "yes/no" format).[3] To be sure, there is heterogeneity within the NPS industry, and the 2020 Pew Report does not describe how the organization selected the firms for its analysis. Nevertheless, questions remain about the motives of opt-in respondents and the countless ways they may be distinctive from the population at large.

In the remainder of this article, we describe situations in which researchers typically employ NPSs and consider whether these data are fit for purpose. We then consider other, less common use cases for which NPSs may be well suited. Finally, in the remaining sections we note how the use of NPSs is in tension with the movement toward transparency in social science research and call for discipline-wide changes that could increase the utility of these data.

Are Nonprobability Surveys Fit for Purpose?

In the context of survey research, the phrase "fit for purpose" refers to the notion that "survey data should be evaluated in light of how they are to be used" (Baker et al. 2013, p. 98). Previous American Association of Public Opinion Research (AAPOR) task forces have advanced fit for purpose as a

2. Researchers attempt to account for the self-selection bias inherent in nonprobability sampling through techniques like sample matching, propensity score adjustment, or weighting.

3. Lopez and Hillygus (2018) find evidence of such "mischievous" responding even after controlling for satisficing.

broad, nonstatistical framework that can be used to assess samples (also see Biemer 2010). Here, the central question is whether the data are appropriate *given the inferential goal*. For example, Baker et al. differentiate the purposes of "describers" ("those who use survey data to describe the population") and "modelers" ("those who use the data to describe relationships between variables") (2013, p. 98). In addition to the type of inferences a researcher seeks to make, accessibility of the data (e.g., ease of use/timeliness) and cost (e.g., price) must be considered (Gotway-Crawford 2013; Goel, Obeng, and Rothschild 2021). Probability-based methods are "slow and expensive" (Gelman et al. 2016, p. 89). It takes time—on the order of weeks or months—to identify a sample frame, probabilistically select a sample, and execute a survey. From this standpoint, NPSs offer a low-cost and flexible method for administering surveys. NPSs also can be logistically and financially more feasible than probability samples for studying election campaigns (Gelman et al. 2016), public health crises (Radford et al. 2022), or other fast-changing political phenomena. Thus, while fit for purpose principally has to do with appropriate use of the data, that criterion ought to be balanced against accessibility and cost. Next we consider four situations in which researchers often employ NPSs: (1) estimating treatment effects; (2) describing distributions of variables; (3) modeling relationships between variables; and (4) forecasting elections.

Estimating Treatment Effects

In the domain of experimental research, NPSs are common because the purpose is to estimate a causal effect across randomly assigned groups.[4] Given the history of using nonrepresentative samples in experimental research, there has been sustained discussion of whether the characteristics of study participants affect the generalizability of experimental findings (e.g., Kam, Wilking, and Zechmeister 2007). Today, the accepted wisdom is that a homogeneous treatment effect should be observed among any group of subjects, no matter how unusual or self-selected the sample. In the case of heterogeneous effects, the researcher should recover the treatment effect if there is variance on the moderator of interest and the analytical model includes an interactive specification (Druckman and Kam 2011; see Munger et al. 2021 for an application). In line with this view, experimental effects seem to generalize across a variety of sample types (Weinberg, Freese, and McElhattan 2014; Mullinix et al. 2015; Coppock, Leeper, and Mullinix 2018). Indeed, one study concludes that "even descriptively unrepresentative

4. When people enter an experiment in a nonrandom manner from an undefined population, the researcher estimates the sample average treatment effect, or SATE (Imai, King, and Stuart 2008; Franco et al. 2017).

samples . . . still tend to produce useful estimates not just of the [sample average treatment effects] but also of subgroup [conditional average treatment effects] (Coppock, Leeper, and Mullinix 2018, p. 12445). This conclusion comes from an unusually comprehensive study involving the replication of 27 experiments.[5]

Set against research on the generalizability of treatment effects is a different body of work showing that people who participate in online surveys are systematically different from nonparticipants (Brüggen and Dholakia 2010; Keusch, Batinic, and Mayerhofer 2014; Hargittai and Karaoglu 2018; Schaurer and Weib 2020). Some of the patterns relate to psychological characteristics (e.g., curiosity, need for cognition) and may seem benign from the perspective of heterogeneous treatment effects. But other patterns have political relevance, such as Valentino et al.'s finding that respondents in online samples "were, on average, less open to experience and more politically conservative on a variety of issues compared to their face-to-face counterparts" (2020, p. 446). Several studies suggest that politically engaged respondents are overrepresented in online panels (Karp and Lühiste 2016; also see Malhotra and Krosnick 2007 or Chang and Krosnick 2009). In short, it is no longer safe to assume that the characteristics of opt-in respondents are orthogonal to the phenomena researchers seek to study. However, it has been difficult to study selection into NPSs, given the proprietary nature of panel recruitment and maintenance. Thus, the differences described above may represent the tip of the iceberg.[6]

Despite these limitations, NPSs have a place in social science research. When it comes to theory testing and development, the SATE is a relevant quantity. Indeed, for many researchers the primary goal is to establish a causal relationship between two conceptual variables in any segment of the population. In that situation, a low-cost NPS is an efficient use of resources.[7] In other situations, the researcher may have a "local" experiment but "general aspirations" in the form of a policy goal (Shadish, Cook, and

5. Notwithstanding the breadth of that study, it is important not to overstate the results. The investigation of conditional effects was limited to six observable characteristics (age, education, gender, race, political identification, and ideology), some with "extreme coarsening" (Coppock, Leeper, and Mullinix 2018, p. 12442). Additionally, all the replicated studies relate to persuasion and attitude formation. Yet even in this one substantive area, theory readily suggests other moderators (e.g., political knowledge, processing style; Druckman 2022, p. 76). This observation is not a critique of Coppock, Leeper, and Mullinix (2018) or similar efforts; the point is that knowledge of potential moderators is incomplete and constantly evolving.

6. Druckman says panel conditioning effects (e.g., from participation in multiple online studies) are a "grossly neglected topic," and that "scant research [has explored] the impact on experimental studies" (2022, p. 81). Panel conditioning can take place in probability samples, particularly longitudinal surveys with multiple waves.

7. Comparisons to population benchmarks often are made in this context, but effort could be better spent developing theory about potential moderators (Druckman 2022, p. 73).

Campbell 2002, p. 18). Replications of an experiment—even with nonrepresentative samples—can be useful for accumulating information about effect sizes and scalability. Thus, NPSs can be a useful vehicle for hypothesis testing and theory development, but scholars must remain cognizant of the scope conditions of their findings. In particular, "results should be generalized to the general public with confidence only after those findings have been replicated with representative general public samples" (Malhotra and Krosnick 2007, p. 312).

Making Population-Level Claims about the Distribution of a Variable

Scholars sometimes use a NPS to describe a characteristic of the population, such as the prevalence of an attitude or behavior. This inferential practice is less common than estimating treatment effects because the meaning of standard statistics (e.g., standard errors, confidence intervals) is ambiguous in these data.[8] Nevertheless, the implicit and sometimes explicit goal of a study is to make a population-level claim (see Oliver and Wood 2014; Motta et al. 2021; Reinhart et al. 2021; or Uscinski et al. 2021 for examples).

A common way to validate an NPS in this context is to compare the characteristics of the opt-in sample (usually, demographics) to another, putatively more representative group of respondents. For example, the American National Election Studies (ANES) and the American Community Survey, conducted by the US Census, are common benchmarks. However, there are no agreed-upon protocols about which covariates should be employed in these comparisons, and the contrasts can involve population benchmarks established years prior to the focal survey. In some instances, statistical differences are assessed, but oftentimes the comparisons are illustrative. Even when a sample looks like the target population on a set of observed characteristics, "there are . . . a limited number of benchmarks on which the sample can be compared, so these samples still require the untestable assumption that unmatched characteristics are ignorable" (Hillygus 2020, p. 28; also see Fowler 2014). Such comparisons also may obscure problematic joint distributions.

In studies comparing the accuracy of probability and NPSs in relation to government records, the former consistently outperform the latter (Yeager et al. 2011; Kennedy et al. 2016). One recent analysis (MacInnis et al. 2018)

8. With an NPS, respondents are not sampled in the traditional sense (e.g., random-digit-dial, address-based sampling). There also is no sampling frame that contains all members of the target population. Thus, while it is possible to describe patterns from these data, the population to which these patterns generalize is unknown (Callegaro et al. 2014; Cornesse et al. 2020). With large sample studies such as the Cooperative Congressional Election Study (CCES), it is possible to derive empirical estimates of sampling variability (Ansolabehere and Rivers 2013).

compared the accuracy of probability surveys (RDD telephone and internet), internet surveys that combine nonprobability and probability samples, and internet surveys that are fully nonprobability (but with different incentives). This study is the most extensive of its kind, using a set of 50 measures and 40 benchmark variables from federal face-to-face surveys with high response rates. The analyses showed that probability samples provide more accurate estimates of population quantities than NPSs and samples that combined methods. Moreover, weighting does *not* eliminate this advantage: "poststratification weights with primary demographics improved the accuracy of the probability samples but only sometimes improved the accuracy of the nonprobability samples" (MacInnis et al. 2018, p. 726).[9]

Declining response rates in probability surveys (and potential bias resulting from nonparticipation) often are used as the rationale for abandoning probability surveys. At present, however, there is little evidence of drop-off in the accuracy of RDD surveys even in an era of declining response rates (MacInnis et al. 2018; also see Dutwin and Buskirk 2017). Thus, the recommendation from Yeager et al. remains valid: "... if a researcher's goal is to document the distribution of a variable in a population accurately, nonprobability sample surveys appear to be considerably less suited to that goal than probability sample surveys" (2011, p. 737; also see Cavari and Freedman 2022, p. 337). More generally, relying on data with opt-in respondents puts researchers in an inevitable catch-22. With both weighting and comparisons to benchmarks, variables must be available in the focal survey and in datasets measuring characteristics of the general population. This can result in arbitrary use of variables that are available, but not necessarily diagnostic of the selection process. Berinsky hints at this dilemma when he writes, "Unless we can measure exactly the differences that lead to the self-selection behavior in nonprobability samples, we cannot account for [them]" (2017, p. 316; also see Bethlehem 2016).[10]

We offer an illustration of how NPSs can lead to mistaken population estimates in a public health context. Because bogus (or "mischievous") respondents tend to answer questions in the affirmative (Kennedy et al. 2020), researchers may overestimate the prevalence of rare attitudes and behaviors in the population. In 2020, researchers at the Centers for Disease Control (CDC) reported that 39 percent of Americans engaged in at least one

9. In an earlier study (Ansolabehere and Schaffner 2014), an opt-in internet sample performed well in a comparison to federal benchmarks on 11 items (differing significantly on four questions). Additionally, there were modest differences across sample types in analyses of election polling in the 2020 US presidential election (Clinton et al. 2020). Election forecasting differs from issue polling; hence we discuss that use case in a separate section below.

10. Even if it were possible to identify, measure, and weight on differences, the technique can reduce statistical power (Fowler 2014). Consequently, adjustment methods are vulnerable to threats to statistical conclusion validity (Shadish, Cook, and Campbell 2002).

behavior not recommended for the prevention of COVID (e.g., ingesting a household cleaner; Gharpure et al. 2020).[11] In a replication of that study with a different online sample (and items that identify mischievous respondents), Litman et al. (2021) found that 88 percent of those who reported drinking a household cleaner were problematic respondents. The implications of these overestimates in the context of a public health crisis are grave: in addition to providing an inaccurate picture of the behavior of *millions* of Americans, the publication of the original results could have a self-fulfilling effect (e.g., if the small group of people who are engaging in the unsafe behavior find validation in the belief that others are doing the same). Bogus respondents could contaminate measures of other rare behaviors, such as conspiracy endorsement (Lopez and Hillygus 2018) or political violence (Kalmoe and Mason 2022; Westwood et al. 2022). A flawed statistic may be worse than reporting no statistic for the reason articulated by Lohr: "Bad statistics, once published, can circulate for a long time—even after more rigorous studies show that they are biased" (2022, p. 334).

Making Population-Level Claims about the Relationship between Variables

Another common usage of NPSs is to make claims about associations between variables, in either bivariate or multivariate analyses. Here, questions about fit for purpose have a long history, dating back to the decision of major academic survey organizations (e.g., ANES, British Election Survey [BES]) to employ internet administration as an alternative to traditional face-to-face interviewing (e.g., Malhotra and Krosnick 2007; Sanders et al. 2007; for general discussion see Couper 2000; Tourangeau 2004). Because the ANES and BES are preeminent sources of data for studying political and electoral behavior, a crucial question for researchers is whether differences in sampling/mode "translate into substantive differences in terms of inferences" (Stephensen and Crête 2010, p. 29).

There have been numerous "parallel" studies that investigate the comparability of probability and nonprobability surveys in bivariate and multivariate analyses. The conclusions of these studies are varied, ranging from assessments of modest to no differences (e.g., Sanders et al. 2007; Stephensen and Crête 2010; Ansolabehere and Schaffner 2014) to substantial differences (e.g., Malhotra and Krosnick 2007; Chang and Krosnick 2009; Karp and Lühiste 2016; Pasek and Krosnick 2020) to both patterns (Pasek 2016). These discrepancies are not surprising, given the challenges of conducting a parallel study that isolates the effect due to sampling.

11. Data from Gharpure et al. (2020) come from an opt-in internet panel administered on the Lucid platform.

To do that successfully, the same questionnaire should be administered at the same moment in time on differently sampled respondents (ideally with administration by a single organization to reduce house effects). Few studies meet these conditions. Indeed, in most existing work, it is impossible to disentangle mode from sampling effects because the surveys being compared have different sampling methods (probability vs. nonprobability) as well as different administration modes (e.g., the probability survey is administered via telephone while the nonprobability is administered on the internet).[12] In general, the challenges of isolating sampling differences are steep and likely require the cooperation of online survey organizations.[13]

Aside from the logistics of study design, there are numerous choices when it comes to assessing the comparability of bivariate and multivariate analyses across samples. Unlike distributions of a single variable, which can be compared to a benchmark, there is no "ground truth" for model coefficients. A common approach is to pool the samples and estimate the interaction between a term for the sample and substantive variables (where an insignificant interaction is interpreted as evidence that different samples produce the same results). However, Karp and Lühiste (2016) argue that this empirical strategy obscures the central question—namely, whether the substantive impact of key theoretical variables has changed. They instead estimate separate models for each sample and compare the magnitude of the effects. At present, there is no consensus regarding the best approach for assessing the differences in model parameters across samples. Consequently, one could come to opposing conclusions about fit for purpose simply by referencing different publications. This is an untenable situation for researchers trying to determine whether to invest in a probability sample, as well as the community of reviewers and editors tasked with evaluating such work. A useful area for future research thus involves questions of comparability, with empirical data (e.g., optimally designed parallel studies) or analytical methods (e.g., Monte Carlo analyses).

Election Polling

Like issue polling, there has been an explosion of NPSs in election forecasting, with pollsters conducting "hundreds of preelection surveys" during the typical presidential election (Panagopoulos 2021, p. 214). This trend is

12. Karp and Lühiste (2016) are a notable exception. They analyze the 2012 ANES, administered FTF and online (both probability samples), and the 2010 BES, administered FTF and online (with probability and nonprobability samples, respectively). There is greater correspondence between modes in the ANES versus the BES, which the authors interpret as a sampling effect.

13. One model for such a venture is the "Foundations of Quality" (FoQ) initiative in the domain of advertising research (Walker, Pettit, and Rubinson 2009; Terhanian et al. 2016).

driven by the horse-race news coverage that creates demand for data about candidate preferences. That said, election polling is fundamentally different than issue polling because forecasting an election involves the "notoriously difficult" task of figuring out which survey respondents will turn out to vote (Kennedy 2020). Even probability-based polls have failed to predict the level of support for candidates in particular elections (Panagopoulos 2021; Clinton, Lapinski, and Trussler 2022). Yet the apparent similarity in performance reflects the nature of the task: election polling is inherently uncertain because of the potential for last-minute changes in opinion and/or turnout.[14] At present, there is no evidence that NPSs display greater fitness than traditional, probability-based methods in the domain of election polling. If anything, the practice of herding implies that organizations using NPSs benefit from the (presumed) greater accuracy of probability samples (Silver 2019; also see Langer 2013, p. 134).

Adherents of NPSs maintain that with the "proper statistical adjustments" (Wang et al. 2015, p. 980), it is possible to make accurate election forecasts with nonrepresentative polls, at a fraction of the cost of more traditional survey methods (also see Gelman et al. 2016 or Prosser and Mellon 2018). Others are more doubtful. For example, in their investigation of the 2015 British general election, Sturgis et al. (2018) conclude that "the primary cause of the polling miss was that the samples were unrepresentative of the population of voters. In short, the methods that were used to collect samples of voters systematically over-represented Labour supporters and under-represented Conservative supporters" (p. 777). As of this writing, there remains a lively debate about the utility of NPSs in election forecasting.

When NPSs *May* Be Fit for Purpose

Given the low cost and ease of collecting NPSs, there may be use cases for which these data are well suited. Below we consider four such situations.

Testing Questions and Instrumentation

A basic principle of survey design pertains to the practice of pretesting questions. According to one research methods text, "The best way to know how people interpret the wording of the question is to conduct a pilot test and ask a few people to explain how they interpreted the question" (Jhangiani et al. 2019, p. 194). This advice is the foundation of classic works on

14. Standard margin of error (MoE) calculations only account for sampling error, not other factors such as frame error or nonresponse error. The "true" MoE almost certainly exceeds the magnitude of standard calculations (see Shirani-Mehr et al. 2018).

questionnaire design (Sudman, Bradburn, and Schwarz 1996; Presser et al. 2004). Likewise, in experimental research, pretesting can provide information about the effectiveness of a manipulation (Mutz 2011). Crowdsourced respondents (e.g., MTurk) may be used for both purposes because the data are inexpensive and more heterogeneous than student subjects. However, scholars have documented the nonnaivete of MTurk respondents (Chandler, Mueller, and Paolacci 2014; Krupnikov and Levine 2014) as well as the potential for fraudulent responding (Chandler and Paolacci 2017). There also can be variation in response patterns based on features of a study (e.g., time of day it is posted; Casey et al. 2017). Given these idiosyncrasies, it may be worth the added cost to test instrumentation on a NPS from a commercial vendor. For example, Clancy (2022) describes the process by which Pew Research Center tested items for a political knowledge scale using data from a nonprobability vendor (SurveyMonkey).[15]

Collecting Data on Unique or Hard-to-Reach Subpopulations

Sometimes a researcher wants to target a group that is rare (i.e., low incidence in the general population) or not easily identified from typical administrative data (i.e., the group is "hidden," Heckathorn 1997). In these situations, purposive sampling can be used to identify the sample of interest. Respondents are "chosen subjectively," but that choice involves "expert judgement about the suitability of the sample and a well-reasoned argument for why a particular set of respondents provides an adequate basis for survey-experimental inference" (Klar and Leeper 2019, p. 422). As an illustration, Bergersen, Klar, and Schmitt (2018) worked with community organizations in Pima County, Arizona, to identify lesbian, gay, bisexual, transgender, and queer (LGBTQ) individuals who were nonwhite. The researchers contacted area LGBTQ+ organizations and asked permission to invite their members to participate in an anonymous online survey that included an experiment about government representation of in- versus outgroups. In other situations, a particular subgroup may be of theoretical interest but not easily reached through traditional sampling methods. Cassese et al. (2013) describe a version of purposive sampling—the Socially Mediated Internet Survey (SMIS)—that leverages social networks organized around Web 2.0 platforms. Using the SMIS method, Huddy, Mason, and Aarøe (2015) recruited respondents from political blogs to participate in an online survey about campaign involvement (also see Gutting 2020). In these

15. One could use an NPS to assess measurement properties (e.g., reliability, dimensionality, different forms of validity). However, such analyses are based on models that assume unbiased estimates of population-level correlations (see earlier section, "Are Nonprobability Surveys Fit for Purpose?").

examples, researchers draw up their own NPSs and data collection is tailored to the research question at hand.[16]

In theory, one could contract with a commercial survey firm to obtain samples of unusual or hard-to-reach subpopulations. For example, Lohr observes that NPSs "can increase the sample size and visibility of rare population subgroups," or potentially reach those who are "underrepresented in the probability survey because they are out of scope, undercovered in the sampling frame, or prone to nonresponse" (2022, p. 336). However, and like NPSs with a broader scope, the selection and inclusion of units is proprietary—and essentially a black box. One cautionary tale comes from a study whose authors contracted with a "nationally recognized market research firm" to sample people with army experience (Bell and Gift 2022, p. 149). The authors reported that approximately 80 percent of respondents misrepresented their background and credentials to gain access to the survey. If a researcher uses a commercial firm to collect data about a distinctive subpopulation, they must learn as much as they can about the recruitment and screening procedures employed by the company, and then develop methods for identifying fraudulent respondents.

Combining Probability and Nonprobability Samples

A third use case involves the combination of data that are collected via probability and nonprobability sampling methods (Sakshaug et al. 2019; Wiśniowski et al. 2020; also see Elliott and Haviland 2007). This usage may seem similar to techniques already employed by researchers (e.g., sample matching, propensity score adjustment, weighting), but there is a crucial difference. With existing methods, the researcher adjusts the composition of a NPS in reference to a probability sample or population figure but uses *only* the NPS in the analysis. This strategy is problematic because: (1) the researcher must assume the matching/adjustment variables fully explain the selection mechanism that leads to inclusion in the NPS; and (2) there is no formal way to measure the uncertainty (sampling error) of the resulting estimates (Wiśniowski et al. 2020).

In response to these challenges, researchers have developed estimation techniques that use Bayesian inference to combine a NPS and a smaller-sized probability sample. In the words of Sakshaug et al. (2019, p. 655), this approach integrates "sparse scientific data" (i.e., from probability-based samples) with "less scientific and less reliable but potentially abundant and cheap information" (i.e., from nonprobability sources). This use case effectively *reverses* the logic of adjustment techniques:

16. Both examples also involve experimental research (i.e., identifying group-specific causal effects, not descriptive claims about a subpopulation).

> In contrast to sample matching and post-survey adjustment, which takes an error prone nonprobability sample and skews it towards a presumably less error-prone probability reference sample, the Bayesian approach that we describe does the opposite. That is, the method takes a probability sample and deliberately skews it towards a nonprobability sample reflected in the prior. (Sakshaug et al. 2019, p. 656)

Using data from two high-quality probability surveys and eight nonprobability web surveys in Germany, Sakshaug et al. (2019) show that the method reduces the variance and mean-squared error (MSE) of coefficient estimates and model-based predictions relative to probability-only samples. Using actual and estimated cost data, they also demonstrate that the combined approach yields substantial cost savings (relative to a probability-only sample for a given MSE). The central risk lies with the quality of the NPS: if it contains "large biases" when utilized as the prior distribution, there may be larger MSEs compared to probability-only samples (Sakshaug et al. 2019, p. 676). Wiśniowski et al. (2020) also note that the technique is limited to continuous outcome variables. Nevertheless, the integration of probability and nonprobability samples is an emerging area of research (e.g., Yang, Kim, and Song 2020; Yang and Kim 2020; Wu 2022).

Using NPSs to Study Selection

In the absence of random selection, judgements about the data from an NPS rest on "observed properties of realized samples," as opposed to "intrinsic properties of the survey process" (Mercer et al. 2017, p. 278). This puts a premium on understanding the factors that *shape participation at the relevant stages* (e.g., joining a panel, agreeing to a study). We know a great deal about the correlates of response in the context of probability samples; "what is unclear is whether these variables or types of explanatory models can adequately account for or describe self-selection mechanisms [in NPSs]" (Dutwin and Buskirk 2017, p. 235). Researchers may assume the equivalence between "'opting in' by volunteering and 'opting out' by not responding" (Gotway-Crawford 2013, p. 119). However, that equivalence remains to be established: "The postsurvey adjustment methods applied to non-probability sampling have largely mirrored efforts in probability samples. Although this may be appropriate and effective to some extent, further consideration of selection bias mechanisms may be needed" (Baker et al. 2013, p. 103).

Studies have demonstrated the influence of "webographic" or lifestyle variables (e.g., regarding privacy, use of new products, travel) in predicting selection into NPSs (e.g., Schonlau, Van Soest, and Kapteyn 2007; Terhanian et al. 2016). Yet even when scholars can incorporate these characteristics (e.g., via propensity weights), the composition of NPSs differs from benchmark data (Dutwin and Buskirk 2017). Understanding the self-selection

process of NPSs thus remains a fundamental challenge. Automated methods represent a promising way to identify the relevant predictor variables in an inductive manner (Terhanian et al. 2016). Unangst et al. (2020) use qualitative methods to link practices of nonprobability firms (e.g., panel refreshment) to sources of bias.

NPSs may be of potential value in many situations. But to use them effectively in any of these contexts, researchers must know more about the provenance of these data—what we refer to below as "production transparency."

NPSs and Transparent Research

Across the social sciences, researchers are adopting standards for sharing data and documenting research practices. In their discussion of this development, Elman, Kapiszewski, and Lupia (2018) identify three types of transparency: data access, production transparency, and analytic transparency. According to the authors, "Data access refers to making available to others the data on which empirical claims in published research rest; production transparency implies clearly explicating the most relevant aspects of the data generation process; and analytic transparency entails conveying the processes through which data were analyzed to produce claims and conclusions" (Elman, Kapiszewski, and Lupia 2018, p. 32).

The recruitment of individuals into a sample relates to production transparency, especially the process by which survey data are selected, collected, and rendered "usable" for analysis (Elman, Kapiszewski, and Lupia 2018, p. 33). Yet it is often unclear how firms identify and screen potential respondents. For example, even though a researcher may enter into an agreement with a specific survey organization, data collection can be outsourced to other entities without the researcher's knowledge (Enns and Rothschild 2022). Additionally, there are companies whose sole purpose is to aggregate respondents from multiple online survey firms and to market the aggregated data to researchers (e.g., PureSpectrum). In both situations, the ultimate source of the data is unclear, making it difficult to obtain details about recruitment, sampling, and the de-duplication of data across platforms. These practices also violate the spirit of an oft-cited AAPOR report that asks "survey companies . . . [to share] more about their methods and data, describing outcomes at the recruitment, enrollment, and survey-specific stages" (Baker et al. 2010, p. 759). The fast-changing nature of the modern polling landscape—characterized by one author as "the Wild West" (Berinsky 2017, p. 315)—has made compliance with the AAPOR report exceedingly difficult.[17]

17. From the researcher's standpoint, it can be hard to select a vendor. The ESOMAR organization has created a guide entitled "Questions for Users and Buyers of Online Sample" (https://esomar.org/code-and-guidelines/questions-for-users-and-buyers-of-online-sample). Cornesse et al.

What Are the Stakes?

Elman, Kapiszewski, and Lupia (2018) state that the motivation for transparency is intellectual. "All social science disciplines seek to produce valid knowledge," and transparent research contributes to that goal because researchers can show that "they have complied with their particular tradition's standards and practices for producing valid knowledge" (2018, p. 30). It is worth asking whether the online survey industry's lack of transparency prevents its academic clients from producing valid knowledge.

One answer, in the form of an illustration, comes from the large opt-in surveys conducted by Delphi-Facebook and Census Household Pulse during the Covid-19 pandemic (Bradley et al. 2021). Compared to behavioral data from the US Centers for Disease Control and a probability-based survey by Axios-Ipsos, the two opt-in surveys significantly overestimated Covid–19 vaccine uptake in the United States. Bradley et al. (2021) used data from the first half of 2021, when people would have been eligible for their first dose of the vaccine. Not only was the "snapshot" of vaccine uptake from big surveys biased, but the errors increased over time, from just a few percentage points to *double-digit* discrepancies by May 2021. At that point in the pandemic, with herd immunity thresholds hovering around 70 percent, Bradley et al. write that "a discrepancy of 10 percentage points in vaccination rates could be the difference between containment and uncontrolled exponential growth in new SARS-CoV-2 infections" (2021, p. 2; also see Meng 2018).

The lack of transparency surrounding NPSs also makes it difficult for the *consumers* of polling data, such as journalists and news audiences. "Many surveys are not conducted or reported in such a way that their results can be replicated or validated . . . this stance makes it more difficult to understand why polls vary from one another or miss systematically, such as when they collectively underpredicted conservative turnout in two presidential election cycles in the U.S." (Radford et al. 2022, p. 44). In this context, political reporters, editors, and newsroom staff all struggle to differentiate the quality of various polls (Toff 2019). Among members of the news-consuming public, the perceived "failure" of polls to predict election outcomes can lead to a declining public confidence in survey methodology (Narea 2016; Geraci 2022). Indeed, a 2017 PBS News Hour/National Public Radio Poll found that nearly *two-thirds* of the public is distrustful of public opinion polls.[18] This lack of trust could lead to an ironic spiral in which opinion surveys

(2020) provide a list of case-level data that researchers can request from vendors. However, it can be difficult for researchers to obtain information deemed proprietary.

18. The poll sampled 1,205 adults using probability-based methods. The question read "How much do you trust each of the following: A great deal, a good amount, not very much, or not at all?" In response to the prompt "public opinion polls," 40 percent said "Not very much" and 21 percent said "Not at all." By way of comparison, about two-thirds of respondents were distrustful

become less and less valued—to the point where leaders dismiss the public preferences expressed through surveys (Toff 2019, p. 886).

A Path Forward

Current practices of online survey companies prevent scholars from meeting the aims of transparent research. We believe it is possible to have an honest reckoning with current practices and chart a fruitful path forward. Toward that end, we offer three suggestions. First, continued investment in probability samples is essential to increase the accessibility of these data and to maintain their quality. Second, the standards for production transparency should be similar for all forms of survey data, regardless of sampling method. Third, the channels through which opinion data become public—namely, the mass media and the scientific community—can promote adherence to quality and disclosure goals.

Investment

If probability surveys represent the gold standard, disciplinary structures must make it easier for researchers at all ranks and types of institutions to have access to these data. Such an effort could involve resources to expand competitions at the Time-Sharing Experiments in Social Sciences (TESS) platform and continued funding for the highest-quality probability surveys in our field (e.g., the ANES) that periodically invite ideas for questionnaire content from the larger research community. Additionally, governing organizations (e.g., APSA, AAPOR) and foundations (e.g., National Science Foundation) could redouble their efforts to support research groups that develop these public goods on their own.[19] As the cost disadvantage of probability surveys declines relative to NPS—effectively becoming zero in the case of competitions offering free data—a greater share of survey research will feature probability samples.[20] Because a probability sample is more expensive than a NPS (e.g., given differences in sample construction), there must be an ongoing conversation about the value added of these data. Druckman points out that probability samples are especially useful for

of President Trump (61 percent across the two categories), Congress (68 percent), and the media (68 percent) (PBS NewsHour 2017).

19. For example, the Vanderbilt University Poll conducts probability sampling from a state list of registered voters. Yet it can be cost prohibitive for any given department or research group to undertake these efforts.

20. There needs to be a similar effort in comparative public opinion research where the comparability of samples across countries is a fundamental component of the research design.

investigating heterogeneous treatment effects, which "can be done to isolate a priori heterogeneous predictions or post hoc to build theory" (2022, p. 80).

For others, the continued existence of high-quality probability surveys is essential for the development (and improvement) of NPSs. Scholars who seek to combine probability and nonprobability samples recognize the value of design-based inference, as indicated by Elliott's claim: "Probability sampling in the 21st Century: Now more than ever" (2022, p. 325). That author warns: "The absence of probability samples unmoors the non-probability sample from the possibility of even partial calibration or other adjustment approaches" (2022, p. 326). On this view, the probability sample is an "analytic partner" to the NPS industry, not merely a stand-alone product.

Disclosure

Elman, Kapiszewski, and Lupia (2018) observe that transparent social science will not come into being without the disciplinary structures to incentivize and support those efforts. Their discussion highlights the incentives put in place to eliminate the "file drawer" problem (Franco, Malhotra, and Simonovits 2017), but a parallel point can be made with respect to NPSs and production transparency. Considering the growing use of NPSs in political science, communications, psychology, and related fields, disclosure standards should be augmented to be on par with those for probability samples (e.g., AAPOR's Code of Professional Ethics and Practices).[21] According to AAPOR, "good professional practice imposes the obligation upon all public opinion and survey researchers to disclose sufficient information about how the research was conducted to allow for independent review and verification of research claims, regardless of the methodology used in the research" (2021). That organization's website lists 11 categories of disclosure, two of which are especially relevant for NPSs: "Methods Used to Generate and Recruit the Sample" and "How the Data Were Processed and Procedures to Ensure Data Quality."[22] At present, information regarding sample recruitment and the preprocessing of data in NPSs can be difficult to obtain for all but the most sophisticated clients. Disclosure varies dramatically across survey organizations and even within the same organization over time.

21. Given the multiplicity of decisions that are made when using NPSs, some believe that "there is a *higher* burden than that carried by probability samples to describe the methods used to draw the sample, collect the data, and make inferences" (Baker et al. 2013, p. 100, emphasis added).

22. The list includes: (1) Data Collection, (2) Research Sponsor, (3) Instruments, (4) Population Under Study, (5) Sample Recruitment, (6) Dates of Data Collection, (7) Sample Size, (8) Data Weighting, (9) Data Processing, and (10) Limitations of Design and Data Collection. Jamieson et al. (2023) build upon this list and have other useful recommendations.

Accountability

Members of the mass media and scientific organizations can promote the conversation about transparency and, where possible, enforce agreed-upon standards. For example, professional pollster John Geraci suggests that the media should only report polls that meet AAPOR standards (2022, p. 264; see Terhanian and Bremer 2012, p. 754, for a related discussion). Yet some outlets, like the *New York Times*, are moving in the opposite direction and have relaxed reporting standards over time (Toff 2019, note 2). Some in the academy have urged scientific journals to take a stronger stand when a study uses data from nonrepresentative samples (Bradley et al. 2021).[23] We suspect that if academic gatekeepers (e.g., funding organizations, disciplinary associations, data repositories, editors) articulated the importance of production transparency and maintained common reporting standards for all types of opinion data, practices at online survey organizations would likely change as researchers gravitate to companies with better disclosure practices. Finally, the dichotomy between "probability" and "nonprobability" may be too simplistic given the heterogeneity within each type. But that variation only underlines the importance of disclosure and accountability.

Many of the world's most pressing problems require an understanding of the public's opinions and behavioral intentions. Despite having access to vast amounts of data, opinion researchers may not be better equipped to contribute meaningful solutions to these problems. Indeed, one pollster proclaimed: "Survey research and polling is the only field I can think of where advances in technology over the past 20 years have reduced quality" (Geraci 2022, p. 17). This characterization may be unduly pessimistic. Yet we believe that survey researchers can and must do better, starting with a clear-eyed assessment of the strengths and weaknesses of survey practices in relation to inferential goals.

Acknowledgements

The authors thank Vin Arceneaux, Yuki Atsusaka, Bert Bakker, Ethan Busby, Scott Clifford, Charles Crabtree, Jamie Druckman, Sunshine Hillygus, Bernhard Clemm von Hohenberg, Cindy Kam, Brendan Nyhan, the anonymous reviewers, and the editors of *POQ* for comments on previous versions of this paper.

23. More specifically, Bradley et al. write: "Scientific journals that publish studies based on surveys that may be unrepresentative [e.g., those with large sizes such as Delphi–Facebook] . . . need to ask for reasonable effort from the authors to address the unrepresentativeness" (2021, p. 699). It remains to be specified what constitutes a reasonable effort, however.

References

American Association for Public Opinion Research (AAPOR). 2021. "AAPOR Code of Professional Ethics and Practices." April 2021. https://aapor.org/standards-and-ethics/#aapor-code-of-professional-ethics-and-practices.

Ansolabehere, S., and D. Rivers. 2013. "Cooperative Survey Research." *Annual Review of Political Science* 16:307–29.

Ansolabehere, Stephen, and Brian F. Schaffner. 2014. "Does Survey Model 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 Dennis, 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; Prepared for the AAPOR Executive Council by a Task Force Operating under the Auspices of the AAPOR Standards Committee. 2010. "Research Synthesis: 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. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1:90–143.

Bell, Andrew, and Thomas Gift. 2022. "Fraud in Online Surveys: Evidence from a Nonprobability Subpopulation Sample." *Journal of Experimental Political Science* 10:148–53.

Bergersen, Meghan, Samara Klar, and Elizabeth Schmitt. 2018. "Intersectionality and Engagement among the LGBTQ+ Community." *Journal of Women, Politics & Policy* 39: 196–219.

Berinsky, Adam J. 2017. "Measuring Public Opinion with Surveys." *Annual Review of Political Science* 20:309–29.

Bethlehem, Jelke. 2016. "Solving the Nonresponse Problem with Sample Matching." *Social Science Computer Review* 34:59–77.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74:817–48.

Bradley, Valerie C., Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. "Unrepresentative Big Surveys Significantly Overestimated U.S. Vaccine Uptake." *Nature* 600:695–700.

Brüggen, Elisabeth, and Uptal M. Dholakia. 2010. "Determinants of Participation and Response Effort in Web Panel Surveys." *Journal of Interactive Marketing* 24:239–50.

Callegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas. 2014. *Online Panel Research: A Data Quality Perspective*, 1–22. West Sussex, UK: John Wiley & Sons.

Casey, Logan S., Jesse Chandler, Adam S. Levine, Andrew Proctor, and Dara Strolovitch. 2017. "Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection." *SAGE Open* 7:215824401771277. https://doi.org/10.1177/21582440177127.

Cassese, Erin C., Leonie Huddy, Todd K. Hartman, Lilliana Mason, and Christopher R. Weber. 2013. "Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments." *PS: Political Science & Politics* 46:775–84.

Cavari, Amnon, and Guy Freedman. 2022. "Survey Nonresponse and Mass Polarization: The Consequences of Declining Contact and Cooperation Rates." *American Political Science Review* 117:332–39.

Chandler, Jesse, and Gabriele Paolacci. 2017. "Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Imposters." *Social Psychological and Personality Science* 8:500–508.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaive among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46:112–30.

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.

Clancy, Laura. 2022. "How We Designed a Scale to Measure American's Knowledge of International Affairs." *Medium*, June 30. https://medium.com/pew-research-center-decoded/how-we-designed-a-scale-to-measure-americans-knowledge-of-international-affairs-f2c54acdd0e7.

Clinton, Josh, Jennifer Agiesta, Megan Brenan, Camille Burge, Marjorie Connelly, Ariel Edwards-Levey, Bernard Fraga, Emily Guskin, D. Sunshine Hillygus, Chris Jackson, Jeff Jones, Scott Keeter, Kabir Khanna, John Lapinski, Lydia Saad, Daron Shaw, Andrew Smith, David Wilson, and Christopher Wlezien. 2020. "Task Force Report on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls." American Association for Public Opinion Research Report. https://www.aapor.org/Education-Resources/Reports/2020-Pre-Election-Polling-An-Evaluation-of-the-202.aspx.

Clinton, Joshua, John S. Lapinski, and Marc J. Trussler. 2022. "Reluctant Republicans, Eager Democrats? Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-Election Telephone Polls." *Public Opinion Quarterly* 86:247–69.

Converse, Philip E. 1987. "Changing Conceptions of Public Opinion in the Political Process." *Public Opinion Quarterly* 51:S12–S24.

Coppock, Alexander, Thomas Leeper, and Kevin Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences of the United States of America* 115:12441–46.

Cornesse, Carina, Annelies G. Blom, David Dutwin, Jon A. Krosnick, Edith D. De Leeuw, Stéphane Legleye, Josh Pasek, Darren Pennay, Benjamin Phillips, Joseph W. Sakshaug, Bella Struminskaya, and Alexander Wenz. 2020. "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." *Journal of Survey Statistics and Methodology* 8:4–36.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464–94.

——— 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75:889–908.

Druckman, James N. 2022. *Experimental Thinking: A Primer on Social Science Experiments*. New York: Cambridge University Press.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base.'" In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James Kuklinski, and Arthur Lupia, 41–57. New York: Cambridge.

Dutwin, David, and Trent D. 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.

Elliott, Marc N., and Amelia Haviland. 2007. "Use of a Web-Based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33:211–15.

Elliott, Michael R. 2022. "Comments on 'Statistical Inference with Non-Probability Survey Samples.'" *Survey Methodology* 48:319–29.

Elman, Colin, Diana Kapiszewski, and Arthur Lupia. 2018. "Transparent Social Inquiry: Implications for Political Science." *Annual Review of Political Science* 21:29–47.

Enns, Peter K., and Jake Rothschild. 2022. "Do You Know Where your Survey Data Come From?" *Medium*, May 3. https://medium.com/3streams/surveys-3ec95995dde2.

Franco, Annie, Neil Malhotra, Gabor Simonovits, and L. J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science* 4:161–72.

Fowler, Floyd J. Jr. 2014. *Survey Research Methods*, 5th ed. Los Angeles: Sage.

Gelman, Andrew, Sharad Goel, David Rothschild, and Wei Wang. 2016. "High-Frequency Polling with Non-Representative Data." In *Political Communication in Real Time*, edited by Dan Schill, Rita Kirk, and Amy E. Jasperson, 89–105. New York: Routledge.

Geraci, John. 2022. *POLL-ARIZED: Why Americans Don't Trust the Polls and How to Fix Them before It's Too Late*. Fayetteville, AR: Houndstooth Press.

Gharpure, Radhika, Candis M. Hunter, Amy H. Schnall, Catherine E. Barrett, Amy E. Kirby, Jasen Kunz, Kirsten Berling, Jeffrey W. Mercante, Jennifer L. Murphy, and Amanda G. Garcia-Williams. 2020. "Knowledge and Practices Regarding Safe Household Cleaning and Disinfection for COVID-19 Prevention—United States, May 2020." *American Journal of Transplantation* 20:2946–50.

Goel, Sharad, Adam Obeng, and David Rothschild. 2021. "Non-Representative Surveys: Fast, Cheap, and Mostly Accurate." Unpublished manuscript.

Gotway-Crawford, Carol A. 2013. "Comment." *Journal of Survey Statistics and Methodology* 1: 118–24.

Gutting, Raynee Sarah. 2020. "Contentious Activities, Disrespectful Protesters: Effect of Protest Context on Protest Support and Mobilization Across Ideology and Authoritarianism." *Political Behavior* 42:865–90.

Hargittai, Eszter, and Gokce Karaoglu. 2018. "Biases of Online Political Polls: Who Participates?" *Socius: Sociological Research for a Dynamic World* 4:1–7.

Heckathorn, Douglas D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44:174–99.

Herbst, Susan. 1993. *Numbered Voices: How Public Opinion Polling Has Shaped American Politics*. Chicago: University of Chicago.

Hillygus, D. Sunshine. 2020. "The Practice of Survey Research: Changes and Challenges." In *New Directions in Public Opinion*, edited by Adam Berinsky, 21–40. New York: Routledge.

Hillygus, D. Sunshine, Natalie Jackson, and McKenzie Young. 2014. "Professional Respondents in Nonprobability Online Panels." In *Online Panel Research: A Data Quality Perspective*, edited by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas, 219–37. West Sussex, UK: John Wiley & Sons Ltd.

Huddy, Leonie, Lilliana Mason, and Leone Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109:1–17.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society Series A: Statistics in Society* 171:481–502.

Jamieson, Kathleen Hall, Arthur Lupia, Ashley Amaya, Henry E. Brady, René Bautista, Joshua D. Clinton, Jill A. Dever, David Dutwin, Daniel L. Goroff, D. Sunshine Hillygus, Courtney Kennedy, Gary Langer, John S. Lapinski, Michael Link, Tasha Philpot, Ken Prewitt, Doug Rivers, Lynn Vavreck, David C. Wilson, and Marcia K. McNutt. 2023. "The Integrity of Survey Research." *PNAS Nexus* 2:1–10.

Jhangiani, Rajiv S., I-Chant A. Chiang, Carrie Cuttler, and Dana C. Leighton. 2019. *Research Methods in Psychology*, 4th ed. Vancouver, BC: Open Text Kwantlen Polytechnic University. https://kpu.pressbooks.pub/psychmethods4e/.

Kalmoe, Nathan P., and Lilliana Mason. 2022. *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. Chicago: University of Chicago Press.

Kam, Cindy D., Jennifer R. Wilking, and Elizabeth J. Zechmeister. 2007. "Beyond the 'Narrow Database': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415–40.

Karp, Jeffrey A., and Maarja Lühiste. 2016. "Explaining Political Engagement with Online Panels: Comparing the British and American Election Studies." *Public Opinion Quarterly* 80: 666–93.

Kennedy, Courtney. 2020. "Key Things to Know about Election Polling in the United States." Pew Research Center, August 5. https://www.pewresearch.org/fact-tank/2020/08/05/key-things-to-know-about-election-polling-in-the-united-states/.

Kennedy, Courtney, Nick Hatley, Arnold Lau, Andrew Mercer, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. 2020. "Assessing the Risks to Online Polls from Bogus Respondents." Pew Research Center, February 18. https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/.

Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. 2016. "Evaluating Online Nonprobability Surveys: Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics." Pew Research Center, May 2. https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/.

Keusch, Florian, Bernad Batinic, and Wolfgang Mayerhofer. 2014. "Motives for Joining Nonprobability Online Panels and Their Association with Survey Participation Behavior." In *Online Panel Research: A Data Quality Perspective*, edited by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas, 170–91. West Sussex, UK: John Wiley & Sons.

Klar, Samara, and Thomas J. Leeper. 2019. "Identities and Intersectionality: A Case for Purposive Sampling in Survey Experimental Research." In *Experimental Methods in Survey Research: Techniques That Combine Random Sampling and Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 419–33. Hoboken, NJ: John Wiley & Sons.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1:59–80.

Krupnikov, Yanna, Hannah Nam, and Hilary Style. 2021. "Convenience Samples in Political Science Experiments." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, 165–83. New York: Cambridge University Press.

Langer, Gary. 2013. "Comment." *Journal of Survey Statistics and Methodology* 1:130–36.

Lavrakas, Paul J., Michael W. Traugott, Courtney Kennedy, Allyson L. Hollbrook, Edith D. de Leeuw, and Brady T. West. 2019. *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*. West Sussex, UK: John Wiley & Sons.

Litman, Leib, Zohn Rosen, Chsekie Rosezweig, Sarah L. Winberger-Litman, Aaron J. Moss, and Jonathan Robinson. 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." medRixiv 20246694. https://doi.org/10.1101/2020.12.11.20246694. 2 January 2021, preprint: not peer reviewed.

Lohr, Sharon. 2022. "Comments on 'Statistical Inference with Non-Probability Survey Samples.'" *Survey Methodology* 48:331–38.

Lopez, Jesse, and D. Sunshine Hillygus. 2018. "Why So Serious? Survey Trolls and Misinformation." https://ssrn.com/abstract=3131087.

MacInnis, Bo, Jon A. Krosnick, Annabell S. Ho, and Mu-Jung Cho. 2018. "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension." *Public Opinion Quarterly* 82:707–44.

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* 15:286–323.

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *Annals of Applied Statistics* 12:685–726.

Mercer, Andrew W., Frauke Kreuter, Scott Keeter, and Elizabeth A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." *Public Opinion Quarterly* 81:250–71.

Motta, Matt, Timothy Callaghan, Steven Sylvester, and Kristin Lunz-Trujillo. 2021. "Identifying the Prevalence, Correlates, and Policy Consequences of Anti-Vaccine Social Identity."*Politics, Groups, and Identities* 11:108–22. https://doi.org/10.1080/21565503.2021.1932528.

Mullinix, Kevin, Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2:109–38.

Munger, Kevin, Ishita Gopal, Jonathan Nagler, and Joshua A. Tucker. 2021. "Accessibility and Generalizability: Are Social Media Effects Moderated by Age or Digital Literacy?" *Research & Politics* 8:205316802110169. https://doi.org/10.1177/205316802110169.

Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

Narea, Nicole. 2016. "After 2016, Can We Ever Trust the Polls Again?" *New Republic*, December 14. https://newrepublic.com/article/139158/2016-can-ever-trust-polls-again.

Oliver, Eric J., and Thomas J. Wood. 2014. "Conspiracy Theories and the Paranoid Style(s) of Mass Opinion." *American Journal of Political Science* 58:952–66.

Panagopoulos, Costas. 2021. "Accuracy and Bias in the 2020 U.S. General Election Polls." *Presidential Studies Quarterly* 51:214–27.

Pasek, Josh. 2016. "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28:269–91.

Pasek, Josh, and Jon A. Krosnick. 2020. "Relations Between Variables and Trends over Time in RDD Telephone and Nonprobability Sample Internet Surveys." *Journal of Survey Statistics and Methodology* 8:37–61.

PBS NewsHour/National Public Radio. 2017. *Marist College Institute for Public Opinion*. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research. Question 25. USMARIST.070317ANP.R06

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. "Data Quality of Platforms and Panels for Online Behavioral Research." *Behavior Research Methods* 54:1643–62.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68:109–30.

Prosser, Christopher, and Jonathan Mellon. 2018. "The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses." *Government and Opposition* 53:757–90.

Radford, Jason, Jon Green, Alexi Quintana, Alauna Safarpour, Matthew Simonson, Matthew Baum, David Lazer, Katya Ognyanova, James Druckman, Roy Perlis, Mauricio Santillana, and John Della Volpe. 2022. "Evaluating the Generalizability of the COVID States Survey—A Large-Scale, Non-Probability Survey." https://doi.org/10.31219/osf.io/cwkg7.

Reinhart, Alex, Esther Kim, Andy Garcia, and Sarah LaRocca. 2021. "Using the COVID-19 Symptom Survey to Track Vaccination Uptake and Sentiment in the United States." Carnegie Mellon University Delphi Group. https://delphi.cmu.edu/blog/2021/01/28/using-the-covid-19-symptom-survey-to-track-vaccination-uptake-and-sentiment-in-the-united-states/.

Sakshaug, Joseph W., Arkadiusz Wiśniowski, Diego Andres Perez Ruiz, and Annelies G. Blom. 2019. "Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach." *Journal of Official Statistics* 35:653–81.

Sanders, David, Harold D. Clarke, Marianne C. Stewart, and Paul Whiteley. 2007. "Does Mode Matter for Modeling Political Choices? Evidence from the 2005 British Election Study." *Political Analysis* 15:257–85.

Schaurer, Ines, and Bernd Weib. 2020. "Investigating Selection Bias of Online Surveys on Coronavirus-Related Behavioral Outcomes." *Survey Research Methods* 14:103–8.

Schonlau, Matthias, Arthur Van Soest, and Arie Kapteyn. 2007. "Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?" RAND Center. Unpublished manuscript.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel, and Andrew Gelman. 2018. "Disentangling Bias and Variance in Election Polls." *Journal of the American Statistical Association* 113:607–14.

Silver, Nate. 2019. "The State of the Polls, 2019." FiveThirtyEight, November 5. https://fivethirtyeight.com/features/the-state-of-the-polls-2019/.

Stephensen, Laura B., and Jean Crête. 2010. "Studying Political Behavior: A Comparison of Internet and Telephone Surveys." *International Journal of Public Opinion Research* 23:24–55.

Sturgis, Patrick, Jouni Kuha, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Benjamin E. Lauderdale, and Patten Smith. 2018. "An Assessment of the Causes of the Errors in the 2015 UK General Election Opinion Polls." *Journal of the Royal Statistical Society Series A: Statistics in Society* 181:757–81.

Sudman, Seymour, Norman N. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Terhanian, George, and John Bremer. 2012. "A Smarter Way to Select Respondents for Surveys?" *International Journal of Market Research* 54:751–80.

Terhanian, George, John Bremer, Jonathan Olmsted, and Jiqiang Guo. 2016. "A Process for Developing an Optimal Model for Reducing Bias in Nonprobability Samples: The Quest for Accuracy Continues in Online Survey Research." *Journal of Advertising Research* 56:14–24.

Ternovski, John, and Lilla Orr. 2022. "A Note on Increases in Inattentive Online Survey Takers Since 2020." *Journal of Quantitative Description: Digital Media* 2:1–35. https://doi.org/10.51685/jqd.2022.002.

Toff, Benjamin. 2019. "The 'Nate Silver Effect' on Political Journalism: Gatecrashers, Gatekeepers, and Changing Newsroom Practices around Coverage of Public Opinion Polls." *Journalism* 20:873–89.

Tourangeau, Roger. 2004. "Survey Research and Societal Change." *Annual Review of Psychology* 55:775–801.

Unangst, Jennifer, Ashely A. Amaya, Herschel L Sanders, Jennifer Howard, Abigail Ferrell, Sarita Karon, and Jill A. Dever. 2020. "A Process for Decomposing Total Survey Error in Probability and Nonprobability Surveys: A Case Study Comparing Health Statistics and US Internet Panels." *Journal of Survey Statistics and Methodology* 8:62–88.

Uscinski, Joseph E., Adam M. Enders, Michelle I. Seelig, Casey A. Klofstad, John R. Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar N. Murthi. 2021. "American Politics in Two Dimensions: Partisan and Ideological Identities versus Anti-Establishment Orientations." *American Journal of Political Science* 65:877–95.

Valentino, Nicholas, Kirill Zhirkov, D. Sunshine Hillygus, and Brian Guay. 2020. "The Consequences of Personality Biases in Online Panels for Measuring Public Opinion." *Public Opinion Quarterly* 84:446–68.

Walker, Robert, Raymond Pettit, and Joel Rubinson. 2009. "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 Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31:980–91.

Weinberg, Jill D., Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1:292–310.

Westwood, Sean, Justin Grimmer, Matthew Tyler, and Clayton Nall. 2022. "Current Research Overstates American Support for Political Violence." *Proceedings of the National Academy of Sciences of the United States of America* 119:e2116870119. https://doi.org/10.1073/pnas.211687011.

Wiśniowski, Arkadiusz, Joseph W. Sakshaug, Diego Andres Perez Ruiz, and Annelies G. Blom. 2020. "Integrating Probability and Nonprobability Samples for Survey Inference." *Journal of Survey Statistics and Methodology* 8:120–47.

Wu, Changbao. 2022. "Statistical Inference with Non-Probability Survey Samples." *Survey Methodology* 48:283–311.

Yang, Shu, and Jae Kwang Kim. 2020. "Statistical Data Integration in Survey Sampling: A Review." *Japanese Journal of Statistics and Data Science* 3:625–50.

Yang, Shu, Jae Kwang Kim, and Rui Song. 2020. "Doubly Robust Inference When Combining Probability and Nonprobability Samples with High Dimension Data." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 82:445–65.

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.