AAPOR
AMERICAN ASSOCIATION FOR
PUBLIC OPINION RESEARCH

Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis

November, 2022

Cameron McPhee (Chair), SSRS

Frances Barlas, Ipsos

Nancy Brigham, Dynata

Jill Darling, University of Southern California

David Dutwin, NORC

Chris Jackson, Ipsos

Mickey Jackson, SSRS

Ashley Kirzinger, KFF

Roderick Little, University of Michigan

Emily Lorenz, Gallup

Jenny Marlar, Gallup

Andrew Mercer, Pew Research Center

Paul J. Scanlon, National Center for Health Statistics

Steffen Weiss, Morning Consult

Laura Wronski, Momentive

This report was commissioned, reviewed, and accepted by the AAPOR Executive Council as a service to the profession. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the AAPOR Executive Council. The authors, who retain the copyright to this report, grant AAPOR a non-exclusive perpetual license to the version on the AAPOR website and the right to link to any published versions.

The Task Force would like to thank Natalie Jackson, Katrina Sostek, James Wagner, and Brady West for their invaluable feedback and support with writing this report.

Table of Contents

Section 1: Introduction and Scope

Sampling from online panels, recruited either using probability-based sampling techniques or opt-in web-based data collection procedures, has become an increasingly common methodology for survey researchers over the last decade.

In 2016, an AAPOR task force issued a report titled "Evaluating Survey Quality in Today's Complex Environment" which outlined 17 questions that users of survey data should ask to help them make judgements about the survey's results, regardless of the survey methodology. These questions work in tandem with the guidelines outlined by the AAPOR Transparency Initiative for survey disclosure in providing consumers of survey data an excellent framework for better understanding and assessing the possible error associated with a particular survey project. Yet, there is still a gap of guidance surrounding assessing the quality of online panels *prior* to data collection. In addition, the existing AAPOR task force reports on online survey panels and nonprobability sampling are aging and largely do not address probability-based online panels, which have become increasingly common since 2016. This task force examines the characteristics of online survey panels and gives guidelines for evaluating the quality of various online panel methodologies.

The goal of this task force is to provide audiences who have a basic understanding of survey methodology with an overview of the various types of online survey sampling methodologies currently being employed by survey researchers and major survey firms, as of the release of this report. Specifically, the report provides an overview of the landscape of online survey data collection, focusing mainly on probability and nonprobability online panels. We discuss how alternative methodologies for initial recruitment, decisions around panel freshening, respondent attrition, and missing data may impact sampling and data quality. Finally, we outline some ways to assess the quality of online samples, including well-known measures such as cumulative response rates and cooperation rates, as well as newer metrics that can be applied to online samples to evaluate representativeness and inferential reliability. We conclude with some key questions for researchers who are designing research studies based on online panels to potentially ask panel vendors as they develop their research design.

Specific objectives:

1. Develop a clear, concise, updated explanation of survey sample sources for online panels.
2. Describe the representativeness and fitness for use of common sampling strategies that providers use to construct and replenish online panels.
3. In particular, assess coverage error of online panels, including systematic error related to recruitment methods, self-selection, and coverage of internet non-users.
4. Propose alternative metrics of sample quality, beyond completion rates and cumulative response rates that measure the underlying representativeness and utility of online samples.
5. Discuss whether sample quality metrics developed for use with probability-based panels can be applied to samples from panels that do not recruit using probability-based methods.
6. Discuss the application of AAPOR's Code of Ethics' reporting guidelines to studies using online panels and outline important issues regarding methodological transparency.

Section 2: A Short History and the Current Landscape of Online Panels

More than nine in ten adults in the U.S. said they used the internet in 2021, according to data from the Pew Research Center's Internet / Broadband Fact sheet. As internet usage has become ubiquitous in this country, survey research organizations have become increasingly reliant on the internet to collect self-reported data from U.S. households. This section provides a brief history of both probability-based and nonprobability based online panels.

History & Timeline

Online panels have been around since the late 1990s. As phone-based methods have become more expensive due to declining response rates, online household panels have become ubiquitous in modern survey research. The rise of these methods is attributable to the development of new technologies, the adaptation of older technologies, and the economic pressures that accompany the decline of the telephone as a reliable means of reaching U.S. households.

The first internet panels were created in the early 1980s, at about the same time as the rise of random digit dialing (RDD) methods in the United States (Lavrakas, 2008), made possible by the ubiquity of landline telephones, which were a fixture in 93% of U.S. households by 1980. The utility of self-administered research was apparent to many researchers; however the availability of the technology was limited. The first national "online panel" was implemented in 1986 for 1000 households by Dutch Gallup (NIPO) in the Netherlands (Saris, 1998), and was based on a computer-assisted data collection system created by Dutch sociologist Willem E. Saris, founder of the European Survey Research Association (ESRA), and his colleagues in the Sociometric Research Foundation (SRF) (Saris & de Pijper, 1986). As part of SRF's overall interest in improving survey measurement, Saris and his colleagues recognized that conducting surveys without the use of interviewers had the potential to remove interviewer bias, as was true for mail surveys. When dial-up modems that could operate over regular telephone lines became commercially available in 1984, SRF supplied a representative sample of Dutch households with modems and computers, since not many households had a computer at that time. This "telepanel" system and its successors were the foundation for CentERpanel, the world's first academic probability-based panel, which began at Tilburg University (Saris, 1998; Hays, Liu & Kapteyn, 2015).

Modern online probability-based panels, including the U.S.'s oldest panel *Knowledge Networks* (which became GfK and then Ipsos KnowledgePanel), the LISS panel in Holland, and RAND's American Life Panel (ALP) were established as the immediate successors of these groundbreaking early telepanels (de Leeuw, 2013). A discussion of the recruitment for online probability-based panels as well as panel maintenance is found below.

While probability-based panels have become increasingly relied on by researchers as telephone response rates declined, nonprobability panels and other online nonprobability samples proliferated freely once the internet was available to provide a ready source of participants. The first to implement nonprobability surveys were pioneers by necessity as they worked to understand and develop methodologies to best leverage the consumer/customer lists they had access to or worked to identify ways to cut operational costs (e.g., using a brand's customer list to save money on CATI costs and merging email and telephone lists).

The present ubiquity of online samples being used to collect data for marketing, public opinion research, and government and academic studies was predicted by some, as were some potential associated issues. For example, Couper noted quite presciently in 2000 that the rise of the internet was likely to be able to democratize large-scale survey data collection, while also posing risks to the quality of data collected by increasing burden on the public to respond and making it more difficult to determine what was and was not reliable information (Couper, 2000). Even in the case of probability-based panels, where the ability to measure individual change over time was seen as a benefit, methodologists were also concerned about the potential impact of interviewing the same panelists over time, as well as the effects of attrition. The following sections outline the current landscape of recruitment, retention, and weighting of both probability-based and nonprobability-based panels.

Probability-based panels: The Current Landscape

This section outlines the methodologies of several probability-based panels currently operating in the U.S. The methodologies outlined in this section reflect the current landscape of probability-based panels, but these methods are evolving and subject to change. This section includes a discussion of probability-based panels that are either commercial panels and/or make their data products available for public use (or offer opportunities to add questions)[1]. The task force requested information from probability-based panels in the United States. Panels were contacted to complete questions, and a call for responses was also shared on AAPORnet. All major probability-based panels (at the time of this writing) responded and were forthcoming in their responses. The list of questions asked and panels that provided a response are shared in Appendix B.

In that spirit of transparency, many of the organizations and institutions that have built probability-based panels have invested resources into methodological experimentation and sharing these findings with the broader research community. A number of published journal articles[2], book chapters[3], conference presentations[4] and white papers[5] have been dedicated to panel methodology and the design of recruitment materials.

[1] The terms "panel" and "panel data" in this report does not include longitudinal survey panels recruited for specific studies or projects, such the panel design used by the Consumer Expenditure Survey.

[2] Literature on panel data collection appeared in Public Opinion Quarterly in the 1930s. Articles on modern panel data collection have been published since the early 2000s. A few examples of papers include, but are not limited to:

- Callergaro, M., and DiSogra, C. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly*, 72, 1008-1032.
- Rao, K., Kaminska, O., McCutcheon, A. 2010. "Recruiting Probability Samples for a Multi-Mode Research Panel with Internet and Mail Components." *Public Opinion Quarterly*, 74, 68-84.
- Yeager, D., Krosnick, J., Chang, L., Javitz., Levendusky, M., 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.
- Scherpenzeel, A., and Toepoel, V. 2012. "Recruiting a Probability Sample for an Online Panel: Effects of Contact Mode, Incentives, and Information." *Public Opinion Quarterly,* 76, 470-490.
- Special Issue: Recent Advances in Probability-Based and Nonprobability Survey Research. February 2020. Journal of Survey Statistics and Methodology, 8.
- Bretschi, D., Schaurer, I., and DIllman, D., 2021. "An Experimental Comparison of Three Strategies for Converting Mail Respondents in a Probability-Based Mixed-Mode Panel to Internet Respondents." *Journal of Survey Statistics and Methodology.*

[3] Callegaro, M., Baker, R., Bethlehem, J., Goritz, A., Krosnick, J., Lavrakas, P., eds., *Online Panel Research: A Data Quality Perspective (1st ed)* (West Sussex, United Kingdom, Wiley, 2014).

[4] A search of the 2022 AAPOR Conference found over 35 paper presentations related to panel methodology. USC also annually organizes the Current Innovations in Probability-based Household Internet Panel Research (CIPHER) Conference.

[5] Most organizations and institutions listed in the Appendix share methodological information about their panels on their websites, in the form of web pages, white papers and other documents and can be found through a web search.

Panel Recruitment

All probability-based panels reviewed for this report share a common core feature – recruitment samples are selected from an address-based sample (ABS) or RDD frame with good coverage of the population and from which selection probabilities can be calculated. The most common recruitment sample frame for probability-based panels is the USPS delivery sequence file which provides near perfect coverage of U.S. households (Harter, et al., 2016; Iannacchione, 2011). When an address-based frame is used, recruitment materials are mailed to selected households. The contents of the recruitment mailer typically contain information about the panel, instructions on how to complete an enrollment survey, and typically, a pre-paid incentive and/or promised postpaid incentive. In addition to the mailer, a variety of contact attempts may be used, such as pre-notification and reminder postcards. Phone numbers may also be appended to the frame to conduct nonresponse follow-ups via phone. In-person visits may also be used, although these are less common.

Although less common currently than using an ABS frame for recruitment, some panels use (or have used in the past) a dual-frame RDD sample and call randomly selected phone numbers to recruit. Addresses may also be appended to this frame so that mail can be used as a supplementary mode of contact, either in the form of pre-notification mailings or nonresponse follow-ups.

Most panels recruit adults aged 18 and older, while a limited number of panels also recruit teens to the panel. Recruitment efforts may attempt to recruit all eligible members of the household or may randomly select a qualified respondent from the household.

Recruitment response rates, for individual recruitment efforts[6], can be calculated for probability-based recruitment efforts, using the AAPOR Standard Definitions (2022). Recruitment effort response rates, however, can vary considerably, even within the same panel, making overall calculation challenging. Among the panels reviewed for this report (see Appendix A), most recruitment response rates ranged between 5% and 15%, although some were higher, and several organizations declined to provide their exact recruitment rates. One practical challenge in calculating a final panel response rate that considers the recruitment phase and the survey phase (Callegaro and DiSogra 2008) is accounting for the variability of response rates across recruitment efforts over time. These issues are discussed in further detail in Section 4 of this report.

A key distinction between many panels is how they cover the offline population. An important feature of probability-based panels is the ability to reach individuals, using a representative and probabilistic frame, to complete online surveys. Individuals who do not have Internet access are covered by panels using two possible strategies[7]. The first is to offer the offline population access to a web enabled device. The second is to offer the offline population an alternative mode of responding, such as telephone or mail. Offering the offline population a way to complete an online survey eliminates concerns about mixing modes and measurement mode effects, and there are timeline and cost benefits to conducting studies 100% online. The disadvantages of offering a web enabled device include the cost to provide a web enabled device and concerns that individuals provided with a web enabled device may now have

[6] The AAPOR Standard Definitions discusses the inclusion of a profile rate in the recruitment rate calculation for online probability panels. This is discussed in greater detail in Section 4.

[7] As of the publication of this report, the Pew Research Center estimates that approximately 6% of U.S. Adults rarely or never access the internet (https://www.pewresearch.org/methods/fact-sheet/national-public-opinion-reference-survey-npors/)

attitudes and behaviors more similar to the online population, rather than representing the offline population.

Panels must also consider whether to offer languages other than English. Most of the major panels reviewed for this report recruit English and Spanish speaking respondents, and some regional panels recruit in other languages.

The validation of recruited probability-based panel members is generally less thorough than the techniques used by nonprobability panels due to the partial validation provided by the sample frame itself. Potential probability-based panel recruits are reached using random sampling methods and their contact information, such as address, phone number, surname and other household characteristics, is known. For this reason, many of the panels reviewed do not have specific validation procedures. Among those that do validate, a common validation technique is matching the mailing address used during recruitment to the address collected from the respondent during empanelment or requiring double entry of certain attributes (such as asking for a select set of demographics at the time of an individual survey and confirming that they match the panel profile information).

Panel Maintenance: Attrition, Retention, and Replenishment

All panels, probability and nonprobability alike, experience attrition of panelists. Attrition can be passive, where panel members stop participating and are removed from the panel, or it can be active, where panel members request to be removed. Most panels allow members to continue their membership in the panel if they are active participants, although the definition of active participation can vary from panel to panel.

Attrition rates can be calculated and should be made available by a panel if requested. Attrition rates are typically calculated as the percent of panel members that are removed from the panel (either through active or passive attrition) during a specified timeframe. Among the panels that disclosed an attrition rate for this report, the range was, on average, 1 to 2% per month.

Based on information provided by the panels reviewed for this report, panelists who have lower education levels, are younger, and who are Black or Hispanic tend to have higher attrition rates than respondents from other demographic backgrounds. Because of this, attrition rates can also be a function of the demographic groups that are actively being recruited and maintained in the panel. Similarly, recruitment is commonly a function of the individuals who attrite from the panel. For example, a panel may primarily conduct recruitment efforts to replace individuals from demographic groups that attrite from the panel at higher rates (i.e., young, those with low education levels, and Black or Hispanic panelists). These newly recruited members are more likely to leave the panel than other demographic groups, like older people, more highly-educated individuals, or white panelists. Very high attrition rates could indicate a problem with panel member retention, however, it could also indicate that the panel only maintains the most cooperative respondents. Conversely, low attrition rates could be an indication of positive panel engagement or suggest that the vendor does not monitor panelist behavior, instead continuing to sample those who stopped responding to surveys and have for all intents and purposes attritted. There is no known published research on an acceptable level of attrition, and more work is needed in this area. Therefore, the quality of a panel cannot be directly measured by the attrition rate. However, it is one metric that can help explain how the panel is managed.

A variety of methods are used to reduce attrition and improve retention rates (the complement of attrition rates). Incentives are the most prominent method for motivating panel member participation.

Incentives may be paid to respondents in the form of cash, gift cards, or other rewards with monetary value shortly after the completion of a survey. Pre-paid incentives may also be used. Other panels use a point system, where panelists accumulate points and can cash in points for rewards. Panels may also use non-contingent incentives, which are gifts or monetary rewards that are not tied to completing a survey. Other methods of retention include panel member communications or websites where they can view information about the panel or panel findings. Some panels also carefully control panel member participation and limit the number of studies a respondent can be invited to during a specified timeframe (to prevent survey fatigue). To date, there is little evidence of a link between survey fatigue and increased attrition in probability panels. However, as discussed in the Panel Conditioning section below, more research is needed in this area. Panels may also conduct nonresponse follow-ups, such as calling or using other modes of communication to reach chronic nonresponders to encourage their future participation in the panel.

Probability panels conduct recruitment activities to replace members who are lost through attrition. Replenishment recruitment may be an ongoing activity or there may be periodic replenishment efforts. Replenishment efforts typically follow the recruitment procedures described in the previous section. These efforts often focus on replacing members who have left the panel (versus building the size of the panel). As such, replenishment efforts may target subgroups that have a greater propensity to attrite from the panel. Sections 3 and 4 provide further detail on the implications of ongoing recruitment and panel replenishment on the computation of sampling probabilities and response rates for survey samples selected from online probability panels.

Sampling into specific studies

Once recruited into the panel, members are invited to participate in individual surveys. The panel may conduct a census of all panel members, or the panel may draw a subsample from within the larger population. Subsamples may be selected using a simple random sample or more complex designs such as a stratified sample or a probability proportionate to size (PPS) sample using a calculated panel weight as the measure of size (MOS) to select a balanced sample.

The frequency of participation and the types of surveys completed can vary widely across panels. Some panels conduct one survey at a time, at some specified interval (such as a single survey conducted one time per month), while other panels have numerous studies in the field at any given time and members may be invited to several studies per month. Some probability-based panels report trying to control the number of surveys their members are invited to in a given time frame. Other probability-based panels have less stringent caps on participation and reported that increased opportunities for participation can help members remain engaged participants. The number of studies a member is invited to may also depend on other factors such as their demographic attributes and/or the target populations of the studies being fielded. Many researchers using a panel for data collection may be concerned about "professional" respondents. Although there is some limited literature on panel conditioning effects within the context of commercial panels (see the section on Panel Conditioning), there is no known literature on the ideal number of survey invitations to send in a given time frame or the impact on data quality. More work is needed in this area.

One distinct advantage of panels is that they typically maintain extensive demographic and psychographic data for all panel members. This information can be used to target specific populations, including low-incidence populations that may otherwise be challenging to reach in an efficient way.

Participation information is also typically maintained for panel members and response rates can be calculated, either at the respondent level or subgroup level. This participant information can be used to improve sampling and weighting adjustments.

Some panels blend probability and nonprobability samples. This technique is most commonly used to reach very low-incidence populations when adequate sample sizes cannot be achieved by using a probability sample alone. If a panel combines probability and nonprobability samples, this information should be disclosed, including the source of the nonprobability sample and the sample size of each source. The exact methods used for sampling and blending samples are beyond the scope of this report, but it is a method that is becoming more common (for example, see Wiśniowski, et al, 2020).

Weighting

Probability-based panels generally include several weighting phases, each of which include several components. The first is calculation of a base weight, which accounts for the probability of recruitment into the panel as well as selection into the specific survey for which the panelist is selected. The selection probabilities are based on the initial frame or frames from which the panelist is recruited, and the selection method used to sample panelists for specific surveys[8]. The second potential weighting component is a nonresponse adjustment. Often this adjustment will occur separately at both phases. Nonresponse at the recruitment phase may use frame information available for both respondents and nonrespondents. Adjustment for nonresponse to the specific survey can make use of information collected about the panelists once they have been recruited. In this case, there is a great deal of known information that can be used for nonresponse adjustments or nonresponse bias analysis at the individual survey-response level.

Most panels also construct adjustments, which use raking or post-stratification procedures to match the sample to population targets. This commonly includes standard demographics such as age, gender, race, ethnicity and educational attainment. However, some panels include additional adjustments, such as political party identification or measures of civic engagement. Again, this may be done initially for the full recruited panel and/or may only be computed using the final set of survey respondents. Finally, if the selected sample included blending of probability and nonprobability samples, then weighting techniques, such as propensity score matching (PSM), internal calibration, or other estimation techniques may be used to blend the samples and calibrate the nonprobability sample to the probability sample (for example, see Robbins, Ghosh-Dastidar and Ramchand, 2021).

A more complete discussion of weighting of both probability and nonprobability-based panels follows in Section 3.

Nonprobability panels: The Current Landscape

The landscape for nonprobability panels has grown large and varied over the three decades since their introduction. This section attempts to outline the major operational or methodological aspects of the most common nonprobability panels. It is not an exhaustive review of all the unique applications offered by commercial companies, nor can it be with many companies offering intellectual property-protected innovations on the practice. However, the concepts below are fundamental to this type of research and are present in the large majority of nonprobability panels.

[8] Repeated or continuous recruitment to a panel can complicate the computation of selection probabilities because the sample frames are not static. This is discussed in more detail in Section 3.

Nonprobability panels range from “expert panels”, where members are specifically recruited because of specific characteristics—such as working in a certain industry (e.g., medical professionals) or having a particular area of domain authority (IT decision makers in companies)—to mass consumer panels where recruitment is open to all individuals as long as they fulfill basic requirements (e.g., age, country of residence).

Panel Recruitment

Nonprobabilistic sampling is sourced through a number of pathways. This section describes several of the most significant and provides a brief definition of each. More complete discussion of the implications of nonprobabilistic sampling will be addressed later in the report.

Panels are a list or database of people who have agreed to participate in research tasks, usually for some type of reward. Originally, the term “research panel” referred to panels whose respondents only participated in research tasks. However, in order to meet respondents where they are on the internet, research panels have evolved to include actively managed groups who, in addition to taking surveys, also do non-research tasks in exchange for rewards (such as buying products at a certain store). Panels are typically actively managed, panelist PII (personally identifiable information) is known, and panelists are often extensively profiled on various demographic, attitudinal, and behavioral characteristics. Panels are recruited using a wide variety of techniques, ranging from existing probability surveys to online ads. Common methods of panel building include using affiliate networks (i.e., companies who help link panel builders with websites who advertise for panelists), placing banners on websites, as part of rewards programs, and through “co-registration”, when a person uses their email to sign up for membership to an online service, they are also offered other services to sign up.

For some panels, respondents must be specifically invited to join (invitation-only), but most allow respondents to sign up without an invitation (organic). Multiple ways are offered to reach the panel portal (mobile apps, mobile phones, desktop/laptop, smart watch, etc.), so that access is not inadvertently limited for a part of the population. Recruitment is usually focused on the targets and volume that a panel needs to supply to meet their business requirements.

Business to business (B2B) and specialty panels mirror this approach, but are tailored to their target population. For example, Hispanics may be recruited on Spanish-language websites[9], while certain B2B targets may be recruited via Linkedin and similar outlets.

Intercept sampling is another source of nonprobabilistic sample. The primary characteristic of intercept sampling is that respondents are recruited directly into a router or survey, rather than into a pre-existing panel first, and no effort is made by the buyer to maintain a relationship with them after that survey. Intercept sourcing can vary quite a bit due to the nature of how it is being used by a company. For example, some intercept sources are re-contactable by a buyer and have some known characteristics, while others are little more than “river” where they flow in and out for one survey only. Both are discussed in greater detail in the next section on non-panel, nonprobability methods.

[9] The recent analysis by Trejo, et. al. (2022) provides specific recommendations for the recruitment of vulnerable populations into opt-in online panels, in particular for non-English-speaking populations, to encourage greater participation by these types of respondents and suggest population-specific methods may be helpful in maintaining participation and data quality among non-English speaking panelists.

Panel Maintenance: Attrition, Retention, and Replenishment

In the early days of panels, online research was novel and relatively highly incentivized, thus panel sizes and responsiveness were high. As internet access proliferated and many other online activities became available, panel membership started to decline. It is now being challenged even more by the increased competition for people's attention and the corresponding monetization of a respondent's spare time (replacing time to take surveys). In addition, recruiting certain in-demand demographics has become more costly, and some panels no longer work to recruit these demographics into panels, instead getting them through intercept approaches.

Because of these challenges, retention is a major focus of panel owners today. Panel owners are employing many techniques borrowed from other areas, such as direct marketing. This could include, for example, offering incentives tied to the respondent journey (such as a birthday or panelist anniversary), taking an omni-channel approach to engaging with panelists, and establishing a community (interact with other panelists, read blogs, take part in fun games). Incentive programs are being tailored to appeal to the diverse types of respondents on panels (e.g., offering charitable donations, or different gift cards for different demographics). "Resuscitation campaigns" are often initiated if a panelist has not taken a survey in a certain amount of time.

Many panels are regularly "cleaned" by purging nonresponsive or poor-quality respondents from the database. Criteria for panel cleaning most commonly include activity level (i.e., if a panel member does not respond to a survey invitation for a long period) or quality control issues (i.e., the respondent is flagged for disingenuous or careless responding such as failing trap questions, straight-lining, speed, etc.).

A healthy panel will likely contain all tenures, from new panelists to those who have been on the panel for three or more years. With the relatively high rates of turnover that are standard for nonprobability panels, recruiting should take place continuously so there is a constant stream of new panelists who then gain tenure over time. In some cases, the experience of being on the panel could influence how someone responds to a survey – for example, if a brand conducts a survey on the panel, anyone who completed that survey had their awareness of the brand raised and that could affect subsequent measures of brand awareness. There is currently no best practice or ideal mixture of tenures at the panel or survey/study level; more research is needed to understand the optimal blend of panel tenures and under what conditions controlling panel tenure is necessary to support study quality.

Sampling into specific studies

When a specific study is launched, the survey is programmed into a hosting platform (i.e., Qualtrics, Decipher) and a sample is ordered or purchased from a nonprobability panel manager or vendor. In some cases, the panel will direct their panel members to a survey hosted outside their environment. In other cases, the panel may also host the survey. The sample is directed into the specific study using one or more of the following approaches.

Quota sampling. Nonprobability panel samples are typically not random samples from a panel that itself attempts to be representative, but usually a quota sample, or a specified number of respondents that match certain target quotas. Sometimes quotas can be reached by specifying a target population and pulling that population directly from the panel; other times it might be reached by specifying a target population and using a router to hit that number of respondents, before directing router traffic to another survey.

In nonprobability sampling, study quotas should be set on any variables that impact the key metrics, as there is no guarantee that the respondents being sent to a study will conform to the desired profile naturally. This is true whether respondents are being sent to study quotas via a router, or when a balanced sample is pulled directly from a panel. Quotas should be set on completed surveys, at the target population level (e.g., total population representation), or at a lower level (e.g., category purchasers within the population). Basic demographic quotas (e.g., Age/Gender) are always recommended. Some studies set quotas/balancing criteria on "starts" (who clicks into the study and starts the screening process) but this can be inefficient and will waste sample. If the population incidence and demographic distribution is not known a priori (such as Age/Gender distribution of cat owners), then approaches such as a soft launch can be used. With a soft launch, a small sample that is demographically balanced is released and used to then estimate the incidence and demographic distribution of the qualified sample (e.g., the percentage and demographic composition of cat owners in the US); this can then be used to set the quotas for the remainder of the data collection.

Router-based sampling. Routers are a way to manage an online sample. A router is a software system that facilitates the real-time assignment of online respondents from diverse sources into available individual surveys. Routers may contain any type of sample (e.g., panel and intercept), and source from multiple suppliers of each. Routers exist to maximize the utility of available samples by matching respondent criteria with survey quota or qualification criteria.

While the first routers were created at the onset of online research in the mid-1990s and were rudimentary, routers today are much more sophisticated, incorporating sampling principles, methodological considerations (e.g., randomization), and respondent experience enhancements. Routers are very common in the industry today as the need for sampling efficiency continues to increase.

Commonly, an individual survey will be connected to a router where sample will initially be allocated. A single router will have multiple surveys that it is sourcing sample for, often with varying qualification criteria.

In a serial router, respondents first go through some minimal screening (such as age and gender), and that information will be used to identify which of the studies currently active on the router the respondent might qualify for. The respondent then is directed to a specific study and enters the screening process for that survey. If they do not qualify for that study, the respondent is returned to the router and directed to another survey, where they attempt to qualify.

In a parallel router, respondents are shown screening questions from multiple surveys at one time, and that information is used to select a study for which the respondent can attempt to qualify. Sometimes there are two levels of screening before a study is selected. Once the study is selected, they then enter that study's screening process and attempt to qualify.

Router-based sampling can be used in conjunction with quota sampling, or each can be used individually. For example, a sample can be pulled from a panel (bypassing a router) and sent directly to a study with quotas. Alternatively, a router can do initial screening and send only the needed respondents to a study with quotas or a router can send any available respondent to a study that does not have any quotas.

Blended sampling or panels is the practice of sourcing from multiple individual panels for a single survey. Due to the decrease in people joining panels, this practice has become a very common way of executing nonprobabilistic research. Each panel is built or recruited through its own approach and the sample purchaser decides from what panels and in what ratios to source for a specific survey. Sample from each panel will be sent to the router to be allocated to individual surveys. The amount of overlap across panels in terms of multiple memberships is unclear; typically, information like IP address or other panelist or device identification information is used to deduplicate respondents when sample from multiple vendors is utilized. The Advertising Research Foundation conducted some research-on-research in 2009 that showed respondents can be on multiple panels without impact on their survey-taking behavior (Walker et al., 2009). Many routers/panels have restrictions on survey participation, with this goal carried out in various ways depending on the particular system being employed. For example, a router may cap the number of surveys in a session, or the number of surveys in a day. A panel may cap the number of invitations a panelist can receive, or the number of surveys they can take in a week. The intent is to balance letting a respondent take a survey when they want to, and not letting them take so many that they become fatigued. Survey length and complexity play a big role here. Taking six 5-minute surveys may not be as cognitively taxing as taking one 30-minute survey, for example.

Sample companies typically track and monitor the entire survey funnel – from initial invitation to survey completion and tally this information on respondents. As such, they monitor how many email invitations, offers, or app notifications panelists are sent; how many they clicked on or into; how many surveys they completed; how many surveys they qualified or failed to qualify for; how many surveys they broke off or abandoned part way through; and how many notifications or invitations they failed to respond to. As such, most sample providers have extensive paradata on panelists.

Weighting

The main goal of weighting with nonprobability samples is to reduce bias, although it also can improve accuracy if the available information is strongly related to the survey variables of interest. Methods for nonprobability weighting may include:

- Raking (marginal distributions or a mix of marginal and joint distributions)
- Post-stratification (joint distribution of a full cross-classification like age × gender × race × education)
- Model-based weighting such as matching, calibration, or propensity score weighting (often requires a representative respondent-level dataset like the public-use ACS, CPS data, or a synthetic microdata frame)

A more complete discussion of weighting of both probability and nonprobability-based panels follows in Section 3.

Non-panel nonprobability designs

As described in detail above, nonprobability-based panels recruit potential respondents to join, collect basic demographic information from each individual, and then repeatedly target them for future surveys with the goal of retaining them in the panel. Traditional panel membership has been on the decline for at least a decade, with people opting not to make a firm commitment to taking surveys over a longer term. However, they are willing to take surveys now and again, therefore vendors have begun to pivot methodologically to provide a way for respondents to still participate without making a longer-term commitment. In this context, intercept sampling has become more and more common.

In intercept sampling, the goal is to encourage the respondent to complete just one survey rather than to join a panel. This, theoretically, provides access to a larger universe of potential respondents that would not be restricted by their willingness to complete the panel enrollment process. Some practitioners may use the term “river” to describe certain intercept sample methodologies. This term is usually used for intercept sample designs that have minimal to no known respondent information (e.g., may only know that they are visiting a certain website). They are typically for short surveys and may or may not give an incentive/reward for participation.

In general, intercept respondents are in the process of completing another task online and are recruited to participate in a study. Recruitment can be achieved in various ways: the respondent could actively seek out a survey opportunity to gain a desired reward, they may click on a banner invitation, or they may be interrupted via a pop-up ad. The various types of intercept recruitment often stem from different research goals and may have their own associated methodologies to accommodate those varying designs. For example, intercept sampling may be done to assess visitors to a particular website, or based on who has been exposed to a particular advertisement, or may be brought into a router for a general population target.

As compared to panels, sample providers typically have less information on intercept respondents. Depending on how a router operates, a unique identifier or fingerprint can allow an intercept respondent to be followed across multiple surveys. Thus, information on this respondent can be built up over time. A company may also broker partnerships with intercept sample providers that allow more information to be passed over from the provider. For some sources, there may be only minimal information available. PII (personally identifiable information) is usually not available for intercept respondents (although some providers can facilitate re-contact of these respondents).

For some intercept (river) sources, the only information known about respondents may be the self-reported data within the survey itself or that which is inferred through other means. It many cases, information from previously administered surveys or from intake/screener data collected prior to empanelment cannot be appended to the survey data. However, some sample providers are able to maintain identifying information about a river respondent in order to leverage for future data appends. Furthermore, some sample providers use metadata collected through means other than direct survey questions and predictive modeling to infer information about who a respondent is or how they would respond to other survey questions based on their answer to a direct survey question. If possible, asking respondents in the survey to provide the demographic information necessary for weighting or analysis is preferred, rather than relying on imputation, modeling, or other assumptions. One well-known example of this methodology is Google Surveys, in which respondents are recruited through pop-up ads that appear on partners’ webpages. Respondents complete a maximum of 10 survey questions, and their demographic information is inferred from metadata accessed by Google through the respondent’s browser information and web history.

There are many limitations of river intercept sampling, of which coverage error is probably the most obvious, because a person cannot be selected to participate in a survey if they do not encounter a river sample survey administration mode. This methodology then excludes anyone who does not use the internet, who does not visit certain websites, or who has ad-blockers enabled on their web browsers. Researchers should keep in mind the inherent lack of representation due to this methodological structure and ensure that the research question they are trying to answer is possible given the river sample design. A broader discussion of these inferential challenges is provided in Section 3 of this report.

Panel conditioning

Panel conditioning is a cause of survey error where the cumulative effect of participating in multiple surveys over time can change a panelist's survey responses to questions about behaviors, attitudes, or knowledge. Even more than affecting survey responses, cumulative survey participation could in fact change actual behavior, attitudes or knowledge (Neter and Waksberg, 1964). Some literature refers to this as "professional" responding. Professional respondents may enroll in multiple non-probability panels, provide fraudulent responses, or take less care in responding in an effort to maximize their survey taking opportunities. Panel conditioning and professional responding are sources of error with the potential to affect both probability and nonprobability panel responses.

Research into panel conditioning effects has been conducted for over 70 years. While papers have found examples of panel effects, the focus of the majority of this work has been longitudinal research where the respondent is recruited to respond to participate in a single study over time. We refer the reader to the sizable literature on longitudinal conditioning effects for further details. This research has largely excluded an analysis of participation in probability or nonprobability panels where respondents are asked to complete surveys on a variety of topics over time. It is important to note this distinction. In longitudinal panels (the basis of most panel conditioning literature), respondents are exposed to the same survey over time. In most commercial probability and nonprobability panels, members are exposed to a variety of survey topics over time.

Despite the literature, there are few definitive conclusions on whether panel conditioning is consistently produced by repetitive survey participation or whether such an effect has a definitive direction or mean effect size (see Bailar, 1989; Sturgis et al., 2009). Notably, panel conditioning does not necessarily lead to panelists developing only bad behaviors like speeding, satisficing, straightlining, skipping, etc. (Hillygus, Jackson, and Young, 2014; Greszki, Meyer and Schoen, 2014). Rather, conditioning could produce improved respondent behavior, providing more truthful and complete responses (Bailar 1989; Mathiowetz and Lair, 1994; Shields and To, 2005; Toepoel, Das, and van Soest, 2008; van der Zouwen and van Tilburg, 2001; Wang et al., 2000; Warren et al., 2012; Waterton and Lievesley 1989; Yan and Eckman, 2012; Zhang, Antoun, and Conrad, 2020). Nor does research suggest that panel conditioning leads to biased estimates for susceptible attitudinal and behavior metrics such as frequency of news consumption, discussing politics, political partisanship or voting, though, in Pew's 2021 analysis, empanelment was found to correlate slightly with voter registration (Amaya et. al. 2021). Further, in commercial panels, respondents may not remember previous surveys or topics, and each new survey or topic in effect resets the respondent experience (Bach and Eckman, 2018).). Removing respondents deemed as "professional" could also do more harm than good and introduce unintended bias (Hillygus, Jackson, and Young, 2014).

More work is needed on panel conditioning effects, within the context of modern commercial probability and nonprobability panels. We encourage researchers to publish on this topic and explore the effect of panel tenure on survey responses, while accounting for panel retention and attrition and the demographics of those who remain in the panel for longer periods of time.

Section 3: Inference, bias, variance, and risk in the context of online surveys

The statistical properties of random sampling allow us to make valid inferences about the characteristics of a population while having observed only a sample of its members. If every member of a population has a known, non-zero probability of inclusion in a sample, we can use the sample to produce unbiased statistics about the population by simply weighting each sampled unit by the inverse of this probability; with some additional information, measures of sampling error (e.g., confidence intervals) can also be produced (Cochran 1953; Kish 1965; Neyman 1934). That is, as long as selection probabilities are known, inference about finite population quantities (e.g., means or proportions) depends only on the sample design and does not require statistical models or further assumptions about how the data were collected. This is known as the **design-based** approach to statistical analysis.

The assumptions of the design-based approach often do not hold for data collected from online samples. For such samples, deviations from random sampling are the norm and can introduce bias into estimates. To adjust for these deviations and recover unbiased estimates, statistical models are required; these models, in turn, require (usually untestable) assumptions. This is easy to see for nonprobability samples, which make no pretense of being random samples from a well-defined population. But even for panels that rely on probability-based recruitment methods, there may be substantial deviations from random sampling due to low response rates, panel maintenance and curation practices, such as those described in Section 2. Inferences made from both kinds of samples require potentially strong assumptions about how nonrandom features of the data collection process affect survey estimates.

Reliance on statistical modeling to correct for nonrandom aspects of data collection is not a foreign concept to most survey researchers. These kinds of models are employed routinely for non-online probability-based surveys in the form of poststratification and weighting adjustments that correct for undercoverage and nonresponse (Kalton and Flores-Cervantes 2003; Brick 2013; Brick and Montaquila 2009). Thus, even estimation approaches that are nominally "design-based"-- in the sense of creating a unit-level weight that is assumed to reflect the inverse of the unit's probability of being included in the sample—usually incorporate some kind of model to correct for nonrandom undercoverage and nonresponse. There is also a robust literature on "model-based" approaches to survey inference that are not based on probabilities of selection but instead on prediction models and auxiliary data about the population (Ghosh and Meeden 1997; Little 2012; Royall 1970; Valliant, Dorfman, and Royall 2000).

In all of these applications, the use of models involves the assumption that, after adjusting some set of observable characteristics, the units in the sample do not differ from the remaining units in the population with respect to the outcomes that the survey is trying to measure. If this assumption is met, estimates from the sample will be unbiased after adjustment; if not, some bias will remain even after adjustment. Therefore, for nonprobability samples, or probability samples with high nonresponse or other deviations from randomization, the validity of survey estimates depends on how well this assumption reflects reality.

The goal of this section is to clarify what kinds of modeling assumptions are being made when making inferences from online samples, provide practitioners with a framework for making their assumptions explicit, and outline considerations for online samples that can affect the risk that these assumptions are not met. Although online panels are the focus of this report, the concepts in this section apply equally to any survey affected by under-coverage or nonresponse.

Key technical terms used throughout this section and Section 4:

- **Unit**: a member of the population being surveyed. For example, in a survey of the U.S. adult population, each adult living in the U.S. is a unit.
- **Sample**: the set of population units from which survey data are collected and therefore from which estimates for the population will be produced. We use **included units** to describe those units that are in the sample, and **non-included units** to describe the remaining population units that are not in the sample. In our discussion, unless otherwise noted, "sample" refers to the final sample available for analysis, that is, survey respondents. This differs from an alternative usage that is common in the context of probability sampling, in which "sample" refers to all units that are initially invited to complete the survey, regardless of response status. This is because our discussion encompasses both probability and nonprobability samples, and in some nonprobability samples, it is not possible to identify an "invited sample" in any meaningful sense.
- **Selection mechanism**: the full process that determines which units in the population ultimately respond to a survey and therefore are included in the sample. Again, because our discussion encompasses both probability and nonprobability samples, this differs from usage that is common in the context of probability samples, in which "selection" may refer only to the process by which units are *invited* to the survey. In our discussion, we use this term to refer collectively to all of the factors that determine 1) which units in the population are covered and could potentially be invited to take a survey, 2) which of the covered units are actually chosen to receive a survey invitation, and 3) which of the invited units ultimately complete the survey and are included in the final responding sample.
- **Adjustment**: any statistical procedure that corrects for bias in survey estimates caused by differences between the makeup of the sample and the larger population. A number of common adjustment methods, such as weighting, are described below.
- **Outcome variable**: a variable of primary substantive interest that is measured on the survey and used to learn something about the larger population. Examples might include presidential approval or intention to vote for a given candidate in a political survey; willingness to buy a particular product in a market research survey; and so forth. Outcome variables are generally only observed for the units in the sample and are unknown for the remainder of the population.
- **Auxiliary variable**: a variable that is not of primary substantive interest but is meant to be used for adjustment. To be usable for adjustment, an auxiliary variable must have a known population distribution and/or be observed for every unit in the population, both those that are included in the sample and the remainder that are not.

Modeling assumptions and the Rubin framework for missing data

Although there are numerous ways of using models for survey inference, the methods that are most commonly used depend on the assumption that the selection mechanism, that is the process that determines which units in the population are both invited to participate in and ultimately respond to a survey, is *ignorable* for a given outcome variable of interest. The concept of ignorable selection was first formalized by Rubin (1976) in the context of inference in the presence of missing data. Since its introduction, the Rubin framework has been adapted and used extensively in other fields that deal with nonrandom data, including causal inference, survey nonresponse, and nonprobability survey samples (Little and Rubin 2002; Mercer 2018; Rubin 1974, 1987).

In the context of survey research, selection is ignorable if each unit's probability of inclusion in the sample is uncorrelated with the important survey or outcome variable(s) – either across the population as a whole or within subsets defined by auxiliary variables. It means that on average, included and excluded units who share common values on the auxiliary variables will have the same distributions for the outcome variable. This situation is also sometimes described in terms of the included and excluded units being *exchangeable* with one another (Greenland and Robins 1986; Mercer 2018).

In addition to ignorable selection, inferences based on models generally assume an additional condition known as *common support* or *positivity*, which states that all units in the population have a non-zero probability of inclusion in the sample given their auxiliary variables. This means that the combinations of values of auxiliary variables that exist in the population are also represented in the sample.

If the sampling frame covers the entire population, and every unit invited to complete the survey responds, probability-based sampling guarantees ignorable selection and common support for every variable that might be measured on a survey. In this idealized situation, the only auxiliary variables needed to account for systematic differences between the sample and the population are the design variables (e.g., stratum and cluster identifiers) used to determine probabilities of selection, which are known for every unit on the sampling frame.

When selection is not fully random, it is typically not possible to be certain that a chosen set of auxiliary variables is sufficient to fully explain any differences between the sample and the population for a given outcome variable. Instead, the justification for these assumptions is necessarily based on other information about the data collection process and the potential ways in which it might over- or under-represent relevant segments of the population.

Approaches to inference from nonrandom data

If the assumptions described above hold, there are two primary ways of using the auxiliary variables to adjust for potential bias due to nonrandom selection: 1) modeling the probability of inclusion (quasi-randomization) or 2) modeling the outcome variable of interest (superpopulation modeling) (Elliott and Valliant, 2017). These approaches are often used in tandem, a strategy known as "doubly robust" inference.

Modeling the probability of inclusion (quasi-randomization)

Inference approaches based on a model of the probability of inclusion in the sample, called quasi-randomization inference, includes propensity weighting and sample matching. These techniques are conceptually similar to design-based inference, in that they rely on probabilities of selection to assign weights to each sampled unit that link the sample to the larger population. They differ in that selection probabilities are not known from the design; rather, they are estimated using a model.

Propensity weighting is a common quasi-randomization approach. A version of propensity weighting is often used to correct for nonresponse to probability-based samples. This entails estimating a logistic regression model (or similar model for binary outcomes) using the invited sample, where the dependent variable is a response indicator equal to 1 for respondents and 0 for nonrespondents. Predictors are auxiliary variables that are observed for the full invited sample. This model is then used to assign a predicted probability of response to all respondents. The nonresponse-adjusted weight is often calculated as the product of the design weight (the inverse of the known probability of selection for the invited sample) and the inverse of the model-predicted probability of response.

For *nonprobability* samples, propensity weighting relies on a high-quality, probability-based reference survey on which the auxiliary variables have also been measured at the individual level and for which the survey weights are assumed to reflect each respondent's true probability of selection. For example, an external survey such as the Census Bureau's American Community Survey or Current Population Survey could be used as a reference sample.

To implement propensity weighting for a nonprobability sample, the reference survey sample is stacked with the nonprobability sample; a selection indicator is assigned with a value of 0 for records from the reference survey and 1 for records from the nonprobability sample. A weighted logistic regression or a similar model for binary outcomes is estimated with this selection indicator as the dependent variable; predictors can be any auxiliary variables that are measured in both samples. For these models, standard survey weights are used for the reference sample and the weights for the nonprobability sample are all set to the same constant value, typically 1. This model is then used to assign a **propensity score**—the predicted probability of having originated from the nonprobability sample—to all records. This propensity score can then be converted into "pseudo-weights" for units in the nonprobability sample (Hirano, Imbens, and Ridder 2003; Lee 2006; Lee and Valliant 2009; Valliant and Dever 2011). Sample matching is an alternative quasi-randomization technique in which weights from the reference sample are copied over to cases with similar propensity scores in the nonprobability sample (Rivers 2007; Rivers and Bailey 2009).

Modeling the outcome variable of interest (superpopulation modeling)

Inference approaches based on a model for the outcome variable, known as superpopulation modeling, include multilevel regression and poststratification (MRP) as well as calibration weighting techniques such as poststratification and raking. The implementation of these techniques is similar for probability and nonprobability samples. These techniques involve using the survey data together with auxiliary data about the target population to predict values of the outcome variables for all of the nonsampled units in the population. They differ from the quasi-randomization approach in that they rely (implicitly or explicitly) on a model of the outcome, rather than a model of inclusion probabilities. Survey statisticians sometimes simply call this model-based inference, owing to its long history as the main alternative to design-based inference (Ghosh and Meeden 1997; Little 2012; Royall 1970; Valliant, Dorfman, and Royall 2000).

Multilevel regression and poststratification (MRP) is one superpopulation method that is frequently used with online surveys. MRP is informed by the field of small-area estimation in which multilevel regression is used to model the outcome variable using a larger number of auxiliary variables and their interactions than is possible with standard weighting methods. The model is then used to compute cell means for the full cross-classification of the predictor variables used in the model. Although some combinations of values may be very small or nonexistent in the survey data, the multilevel regression borrows information from cells with similar characteristics to reduce the amount of variability in the estimates. The cell means are then poststratified based on external information about the size of each cell in the population (Gelman and Little 1997; Park, Gelman, and Bafumi 2004; Ghitza and Gelman 2013). Variations on the method that use machine learning in place of multilevel regression have also been developed (Breidt and Opsomer 2017; Ornstein 2020). MRP, as with similar **prediction-based methods**, requires estimating a separate prediction model for each outcome of interest; the auxiliary variables included in each model may or may not differ between the outcomes.

Superpopulation modeling also includes **calibration** weighting methods such as poststratification and raking that are well understood by survey researchers and used widely used in practice for both probability and nonprobability samples. Calibration methods create weights such that the weighted sample distributions of the auxiliary variables align with known population distributions for those variables (Kalton and Flores-Cervantes 2003). **Poststratification** matches the population distribution for the full cross-classification of all auxiliary variables. Poststratification is often infeasible due to small cell sizes in the sample and/or the lack of population-level information for the full cross-classification. Therefore, **raking** is a popular alternative that simultaneously matches the marginal distributions, but not the full cross-classification, of multiple auxiliary variables. For probability samples, the input to calibration is the design weight reflecting the inverse of the probability with which each unit was invited to the sample. For nonprobability samples, true design weights do not exist and the input weight to calibration is typically set to 1 for all sampled units.

Calibration weighting methods are convenient because they yield one set of weights that can be used to analyze all outcomes of interest. Poststratification and raking do not involve explicitly fitting a regression model and predicting values for the target population; however, they are implicitly model-based because they assume that matching the sample distributions of the auxiliary variables to known population distributions will be sufficient to generate unbiased estimates of the outcome variables. In fact, the resulting estimates are equivalent to what would be obtained by explicitly fitting a regression model and predicting values for the target population (Deville, Särndal, and Sautory 1993; Valliant, Dorfman, and Royall 2000). However, the use of a single set of poststratified or raked weights assumes that the underlying model specification is appropriate for all outcomes.

Doubly robust inference

A third set of techniques simultaneously employ separate models for both inclusion and the outcome variable. One feature of these so-called doubly-robust methods is that only one of the two models needs to be correctly specified in order to remove selection bias. Often, propensity weighting or sample matching is used to create an initial pseudo-weight, which is subsequently adjusted through calibration or used in a weighted regression model for a specific outcome variable (Valliant 2020; Ansolabehere and Rivers 2013; Kang and Schafer 2007). Other techniques use penalized splines or Gaussian processes to flexibly include the propensity score as a predictor variable in a regression model (Little and An 2004; Si, Pillai, and Gelman 2015; Zhang and Little 2009).

Key considerations for practice

When using a model to adjust for potential selection bias, the first and most important consideration is the choice of auxiliary variables for inclusion in the model. Two criteria determine the utility of including a given auxiliary variable in the adjustment model: (1) the correlation between the auxiliary variable and the probability of selection into the sample; and (2) the correlation between the auxiliary variable and the outcome. Table 1 provides a stylized illustration, based on Little and Vartivarian (2005), of the implications of different scenarios. The ideal scenario is when auxiliary variables are strongly correlated with both selection and the outcome; in this case, adjustment on those variables can reduce both bias and variance in estimates for that outcome. When auxiliary variables are strongly correlated with the outcome, but not selection, adjustment can reduce variance but will not reduce selection bias. Conversely, when auxiliary variables are strongly correlated with selection, but not the outcome, adjustment on those variables will not reduce bias and can *increase* variance, and thus do more harm than good.

Of course, this is a highly stylized illustration, and reality is more complicated: there is no set threshold for differentiating “strongly” vs. “weakly” correlated auxiliary variables; and most surveys have multiple outcomes and multiple auxiliary variables, which are likely to vary in their correlations with each other. The critical point is that the effectiveness of any adjustment method depends on how strongly correlated the available auxiliary variables are with the outcome(s) of interest; adjustment on auxiliary variables that are uncorrelated with outcomes may not improve, and indeed may worsen, the quality of the estimates. For example, calibrating a sample to match the age distribution of the population will reduce bias in an outcome that is highly related to age, but not in one that is independent of age.

Table 1: Effect of adjustment on auxiliary variables

	Auxiliary variables predictive of outcome		
Auxiliary variables predictive of selection		No	Yes
	No	No effect	Reduce variance No effect on bias
	Yes	Increase variance No effect on bias	Reduce variance Reduce bias

The choice of statistical method is secondary to the choice of auxiliary variables, since without appropriate auxiliary variables, none of the methods described previously can be expected to work as intended. Sometimes the choice of method may be determined by the nature of the available auxiliary data. If auxiliary data for the population is only available in the form of summary statistics such as means and totals, methods that require a reference sample or highly granular auxiliary data will not work.

Two recent studies have compared the empirical performance of the different modeling strategies discussed previously. In one, propensity weighting, sample matching, raking (a form of calibration), and doubly robust methods were used to weight three different online nonprobability samples using two sets of auxiliary variables, one with only demographics and the other a combination of demographics and variables associated with political engagement. Estimates were compared to external population benchmarks to evaluate the accuracy of each method. This study found that the doubly-robust methods were slightly more accurate than the alternatives but the differences between methods were generally small. The use of more expansive auxiliary variables did far more to reduce error on benchmark estimates, especially those related to politics, but even so there was still a great deal of uncorrected error in the estimates (Mercer, Lau, and Kennedy 2018).

The other was a simulation study in which a large, probability-based sample was used as a reference population from which subsamples designed to emulate the characteristics of online, nonprobability samples were used to compare a similar set of statistical methods. In this study, doubly-robust estimates showed slightly less bias than raking and propensity-based methods, but none of the methods were entirely successful (Valliant 2020).

Determinants of risk in different types of online samples

Given the above discussion, in order to appropriately correct for selection bias, the adjustment model—whether operationalized via quasi-randomization, superpopulation, or a doubly robust approach—must incorporate all relevant auxiliary variables that influence a unit's likelihood of being included in the sample and are associated with survey variables. If such variables are not measured and hence omitted from the adjustment model, some selection bias will remain even after adjustment.

Of course, it is never possible to know with certainty whether an adjustment model incorporates all relevant auxiliary variables. It is more useful to think in terms of *risk*: for a given type of sample, *how confident* can the user be that the modeling assumptions described above can be met?

The implication of the Rubin framework is that the risk of using a given sample source is inversely proportional to the amount of information that the researcher has (or can obtain) about:

- The *selection mechanism*—how individual population units are selected (or self-selected) for inclusion in the completed dataset, and how this process may interact with auxiliary variables that are related to substantive outcomes.
- The *differences between included and non-included units*—how much the included units differ from non-included units with respect to substantive outcomes and/or auxiliary variables correlated with those outcomes.

Knowledge about the selection mechanism helps to understand what auxiliary variables *should* be included in the adjustment model. The extent of knowledge about the characteristics of non-included units determines what variables *can* be included in the adjustment model.

Therefore, the more information a particular sample affords about each of these, the more confident the researcher can be that the adjustment model will yield reasonably accurate population estimates from that sample.

Knowledge of the selection mechanism

By opting to purchase an online sample rather than design a stand-alone data collection, the researcher foregoes most control over the selection mechanism. The vendor – rather than the researcher – controls the construction of sampling frames (to the extent that they exist), the methods used to select units for invitation to the panel or study, and the data collection procedures used to encourage response. While this reduces costs, it also limits the researcher's window into the selection mechanism, making online samples inherently riskier than carefully designed standalone samples.

However, some types of online samples, by their nature, still afford more information about the selection process than others. The most important distinction in this regard is between samples selected from *probability panels* and those obtained from *nonprobability* (panel and non-panel) sources.

Selection and inference with probability panels

With probability panels, the selection process that yields the final, completed sample for a given study can usually be disaggregated into several phases: a *recruitment* phase at which population units are initially recruited to join the panel; a continuous *maintenance* phase in which panel members choose to either remain on or leave the panel; and a *study* phase at which panel members are chosen for the study-specific sample. Within the recruitment and study phases, the process can be further disaggregated into *invitation* (the designation of units for invitation to join the panel or participate in a study) and *response* (units' decision on whether to join the panel or complete a study questionnaire). Within the intermediate maintenance phase, the decision to remain on the panel (or not) can be considered a form of response. The adjustment model typically takes the form of a weighting procedure that incorporates both *known* invitation probabilities and *modeled* response probabilities across all phases, yielding a final analytic weight for each unit.

The distinguishing feature of online probability panels, relative to their nonprobability counterparts, is that they afford a "baseline" set of information about the *invitation* process at the recruitment and study phases. In principle, therefore, they allow greater confidence that the weighting (or other adjustment model) accounts for all relevant characteristics that determine whether a unit was invited into the sample.

At the recruitment phase, probability panels are distinguished by invitation *from a defined list (frame) of population members*. The ABS and RDD frames most commonly used for probability panel recruitment do not perfectly cover the U.S. population, but their coverage properties and limitations have been well-studied (for ABS, see Iannacchione et al. 2003, English et al. 2009, Shook-Sa et al. 2013, Harter et al. 2016, and Amaya et al. 2021; for RDD, see Blumberg and Luke 2007, Lavrakas et al. 2017, and Blumberg and Luke 2022). This means that users have at least some information about population groups that may be undercovered by the recruitment frame, making it possible, in principle, for adjustment models to include auxiliary variables that can help correct for undercoverage.

Also at the recruitment phase, probability panels are distinguished by the fact that *population members are invited for recruitment with known probabilities based on a defined sample design*. Vendors commonly employ complex design features, such as over- and undersampling of certain strata, that can introduce relationships between invitation probabilities and unit characteristics.

However, for many panels, recruitment is not a once and done process. Often panel replenishment is ongoing or multiple recruitment replenishment waves are implemented over the life of the panel. Additionally, recruitment sample design and methodology may evolve over the life of the panel. For example, several panels that initially recruited using RDD telephone surveys later transitioned to ABS mail surveys when RDD became too costly. Recruitment sample designs may also evolve in order to address panel needs over time – for example, a new stratification approach may be implemented in a given recruitment wave in an attempt to address nonresponse or coverage concerns that the panel is facing. When one considers that many panelists likely had some probability of selection not only for the recruitment survey in which they were empaneled but also other recruitment surveys in which they could have been sampled but were not, it becomes clear that calculating true inclusion probabilities for panel members can be extremely difficult or impossible. Techniques for aggregating multiple samples such as those described by Kish (1999) should be considered to combine different recruitment samples.

Similarly, at the study phase, a probability panel user knows (at minimum) that panel members are invited into studies based on some probability-based design. Many panels adopt practices to reduce panelist burden or improve sample quality that may affect an individual panelist's probability of selection into a given study. Examples at the study phase include efforts to limit overlap between study samples, limit the frequency with which individual panelists are selected, and meet multiple overlapping demographic quotas. It may not always be clear how these practices affect study selection probabilities.

In practice, the complexity of recruitment, panel replenishment, and panel maintenance with its implications for study-level sample selection means that probability panels may depart in some ways from the theoretical ideal of wholly known invitation probabilities. Vendors take varying approaches to addressing these challenges in the calculation of selection probabilities for their starting samples; as such, it may not always be clear as to whether the invitation probabilities incorporated into vendor weights represent exact calculations or approximations.

However, while it is important to understand how panel vendors calculate the selection probabilities that inform the base weights, the most important departure from pure probability selection is likely to be *nonresponse*, which will likely impact the sample weights in a more substantive way. In probability panels, nonresponse occurs at the recruitment phase (invited population members declining to join the panel), the maintenance phase (attrition of panel members), and the study phase (invited panel members declining to complete the questionnaire).

Nonresponse to surveys has been well-studied and is known to be a non-random phenomenon that can introduce biases into otherwise probability-based samples to the extent that it is related to substantive study outcomes or their correlates (Groves 2006, Groves and Peytcheva 2008). This is true for standalone probability samples as well as for panels. However, unique aspects of online panels could be expected to amplify some biases. For example, real or perceived burden can drive nonresponse (Yammarino et al. 1991; Bogen 1996; Galesic and Bosnjak 2009) and may drive nonresponse bias to the extent that it interacts with respondent characteristics or attitudes, for example if respondents with busier lifestyles are more sensitive to burden (Fricker and Tourangeau 2010; Vercruyssen et al. 2014). A request to not only complete a one-off survey, but to join a panel and receive many further survey requests, could add to the perceived burden and thereby amplify any such biases at the recruitment phase.

Given that panels collect data mostly or exclusively over the Web, a particular concern is hard-to-measure nonresponse biases related not only to *whether* someone uses the Internet, but to how frequently and willingly they do so. Observational and experimental research has demonstrated that both expressed and observed preferences between Web and non-Web survey modes correlate with certain demographic characteristics, attitudes, and behaviors (Millar et al. 2009; Smyth et al. 2010; Messer and Dillman 2011; Olson et al. 2012; Smyth et al. 2014; Baumgardner et al. 2014; Nichols et al. 2015; Brick et al. 202; Jackson et al. 2021). This suggests that, even within the universe of persons with access to the Internet, willingness to be recruited to an online panel, or to complete an online survey, may vary. Similarly, even if some effort is made to recruit persons without regular Internet access (for example, by providing a Web-enabled device), those willing to participate in a panel via these channels may differ in degree of resistance from other non-Web-users. To the extent that access to, facility with, and/or willingness to use the Internet are related to substantive study measures, bias may result unless the adjustment model includes auxiliary variables that correct for bias due to these relationships (Dever, Rafferty, and Valliant 2008; Bethlehem 2010).

Abstracting from specific examples, the key overarching point is that by offering multiple “opportunities” for nonresponse, panels amplify *uncertainty* as to whether an adjustment model accounts for the full set of characteristics that determine inclusion in the final sample. The decisions to join a panel, remain on the panel, and respond to a specific study are all voluntary and independent decisions on the part of the respondent, and in principle all could be driven by different characteristics, attitudes, and/or behaviors (cf. Lugtig et al. 2014).

Selection and inference with nonprobability samples

As discussed in Section 2, there are many types of online nonprobability samples. However, with respect to statistical inference, the crucial distinguishing characteristic of all nonprobability samples is that the *entire* selection process—including, in principle, the likelihood that a given unit was *available* for recruitment at all (Valliant and Dever 2011)—must be modeled as part of the adjustment procedure. This contrasts with probability panels, in which it is primarily the *response* component of the selection process that must be modeled. While some types of nonprobability sample sources (particularly panels) allow for meaningful distinction between recruitment and study phases, and/or between invitation and response, others (e.g., some intercept samples) do not. Uncertainty is further compounded when nonprobability samples combine multiple panel and non-panel sources, each of which may use different recruitment, invitation, and data collection procedures. As noted in Section 2, this approach of combining multiple sources has become very common in nonprobability sampling.

With nonprobability samples, there is no "sampling frame" in the traditional sense of a defined list of target population units, with known coverage rates and properties, from which samples are selected. Rather, as described in Section 2, nonprobability vendors typically attempt to identify a "sampling pool"—a set of *potential* sources of sample units, which may include multiple panels, routers, recruitment channels, etc.—that plausibly covers the population of interest to the study. This pool becomes the universe from which units can be invited into the study. For many such sources, neither the user nor the vendor observes the underlying universe—for example, the universe of persons who visited a given website that is used for recruitment, and thus were available to be recruited, during the recruitment timeframe. This precludes any firm understanding of the characteristics—particularly non-demographic characteristics, attitudes, and behaviors—that distinguish between (1) population units who had no chance of recruitment, (2) those who had a chance of recruitment but were not successfully recruited, and (3) those who were successfully recruited (Valliant and Dever 2011).

When it comes to determining which members of the sampling pool are invited into a particular study, how respondents are sampled and where respondents are sampled from affects the calculation of a selection probability. If respondents are obtained from a single panel, with the panel defined as the complete sampling pool, the probability of being invited into a study could theoretically be calculated. However, the far more common selection mechanisms used today for nonprobability sampling—routers—are so complex as to effectively preclude the calculation or even approximation of an invitation probability. And, of course, with nonprobability as much as with probability samples, the decision to complete a given study remains voluntary and can introduce nonresponse bias if it is related to study outcomes or their correlates.

Relative to survey nonresponse, there has been less research into determinants of willingness and ability to participate in online nonprobability samples. Most such research involves descriptive comparisons between estimates from nonprobability and probability samples, with observed differences attributed to the nonprobability selection mechanism. For example, Fahimi et al. (2015) found that even after weighting both samples on demographics, nonprobability respondents differed from probability respondents with respect to political engagement, altruistic behavior, community attachment, shopping habits, and happiness and security, among others. Other studies have included comparisons to external benchmarks, generally finding that nonprobability samples show greater divergences from benchmarks than probability samples (e.g., Yeager et al. 2011, Kennedy et al. 2016, Dutwin and Buskirk 2017, MacInnis et al. 2018; see Cornesse et al. [2020] for a detailed review). In an alternative approach, Boyle et al. (2017) asked respondents to a nationally representative RDD sample to self-report whether they were a member of an online panel; they found that self-identified panelists were younger, more likely to be female, more likely to have a college education, and less likely to be married than non-panelists. In general, additional research into the determinants of participation in online nonprobability samples—and how these interact with the specific type of sample (e.g., panels vs. intercept samples) and other design decisions (e.g., incentivization)—would be worthwhile.

Of particular note is that nonprobability samples (unlike some probability panels) typically do not provide any means of recruiting non-Web-users. This likely amplifies any biases driven by access to, facility with, or willingness to use the Internet.

The specification of quotas is a standard approach to ensuring that the distribution of a nonprobability sample matches defined benchmarks (typically demographic) for the target population. Many users assume that quotas reduce the risk of meaningful selection bias and render nonprobability samples comparable to probability-based samples. However, this cannot be assumed in general. While the quota approach will help ensure that results are not skewed on common auxiliary variables (such as age or gender), it does not guarantee that non-quota'd variables will "fall out naturally" at the study level. As noted above, significant differences have been found between probability and nonprobability samples even after correcting for demographic differences; therefore, even a sample that is balanced on basic demographics may be biased on other unknown characteristics that can drive biases in outcome variables. Relatedly, even if a particular demographic cell accounts for the "correct" percentage of the final sample, there is no guarantee that the included units *within* that cell are representative of the corresponding subpopulation. Kennedy et al. (2016) found that, across several nonprobability samples, biases within demographic cells commonly used for quotas (e.g., race/ethnicity cells and age groups) were sometimes larger than full-sample biases. Indeed, one could speculate that the specification of overly detailed quotas could *increase* risk by requiring vendors to pull from non-standard sources to meet all necessary targets, potentially introducing new biases that would be unobservable to the user.

In general, therefore, the use of quotas cannot be assumed to obviate the risks associated with nonprobability-based selection mechanisms. Without a defined sampling frame and known selection probabilities, there is inherently greater uncertainty as to the characteristics that influence the likelihood of inclusion in the final sample, and therefore that should be included in the adjustment model.

Knowledge about non-included units

Even in an idealized scenario in which the user is aware of all characteristics that influence the likelihood of inclusion in the sample—that is, has a full understanding of the selection mechanism—the ability to measure and correct for selection bias would depend on whether those characteristics were observable as auxiliary variables.

To be usable in an adjustment model, auxiliary variables must be observable in at least one of the following forms:

- Individual-level data for included units, plus population means or distributions that are known (or can be estimated) for the full target population. Auxiliary variables observed in this way can be used for raking and other superpopulation approaches.

- Individual-level data for *both* included and non-included units. This can include scenarios in which a separate "reference" dataset (a census or high-quality external survey dataset) provides individual-level data that is representative of the target population (Elliott and Valliant 2017). Auxiliary variables observed in this way can be used for quasi-randomization approaches.

As discussed above, to be useful for assessing and correcting selection bias in any substantive study outcome, auxiliary variables must be related to the outcome itself; variables unrelated to the outcome, if included in the adjustment model, will not reduce bias and may increase variance (Little and Vartivarian 2005). For this reason, the most appropriate set of auxiliary variables is likely to be study-specific, and potentially outcome-specific. In general, however, the fewer the auxiliary variables that are observed, the fewer that can be included in any adjustment model, and therefore the greater the risk that outcomes will be biased in unknown and uncorrectable ways.

A common practice in survey research is to rely entirely on basic demographics (such as age, gender, educational attainment, and race/ethnicity) as auxiliary variables for analyzing and correcting selection bias. Demographics make for convenient auxiliary variables because known population targets are often readily available. However, there is no guarantee that they will be strongly correlated with outcomes and/or that they will fully account for the selection mechanism. Indeed, a large body of research suggests that demographics alone are often inadequate to account for selection bias, both in probability samples with nonresponse (Peytcheva and Groves 2009) and in nonprobability samples (Yeager et al. 2011, Kennedy et al. 2016, MacInnis et al. 2018, Mercer et al. 2018, Cornesse et al. 2020). The likely inadequacy of primarily demographic-based weighting models is underscored by recent high-profile polling errors, such as those surrounding the 2016 and 2020 U.S. presidential elections, which were hypothesized to have been driven at least in part by partisan nonresponse (Kennedy et al. 2017, Clinton et al. 2021).

Given that traditional demographics may very well be inadequate for reducing selection bias, researchers should consider whether it is possible to collect additional auxiliary variables that are correlated with study outcomes and meets the other requirements described above for quasi-randomization and/or superpopulation modeling

Potential sources of auxiliary variables

The set of potential auxiliary variables can be divided into *collected* variables, which are collected on the study questionnaire; and *appended* variables, which are linked to the sample from some external source. Depending on the sample type, one or both types of auxiliary variables may be available.

Collected variables

Variables included in the questionnaire are only collected from survey respondents. To be usable for bias analysis and adjustment, collected auxiliary variables must either: (1) have distributions that are known, or estimated with reasonable accuracy and precision, for the target population; or (2) be observed at the individual level in an external reference dataset that is at least approximately representative of the target population.

As noted above, basic demographics have readily available population distributions but are often inadequate to account for non-random selection mechanisms. However, in the U.S. context, researchers can choose from an array of relatively large-sample, high-response-rate probability surveys that provide population benchmarks for a broad set of potential auxiliary variables, both demographic and non-demographic, and range in frequency from monthly to annual to every 10 years. These include the large-scale benchmark surveys administered by the Census Bureau, particularly the Decennial Census; the annual American Community Survey (ACS); the monthly Current Population Survey (CPS); and periodic CPS supplements covering subjects such as earnings, school enrollment, civic engagement, and voting (among others). They also include more subject-specialized surveys sponsored by other agencies within the federal statistical system, such as the National Center for Health Statistics, National Center for Education Statistics, the National Center for Science and Engineering Statistics, the Energy Information Administration, the Bureau of Transportation Statistics, etc.[10] Finally, several long-running non-federal collections, such as the General Social Survey (GSS) and the American National Election Studies (ANES), employ similarly rigorous data collection techniques to attain relatively high response rates and therefore may offer a viable source for non-demographic, topic-relevant benchmarks or reference samples.

Therefore, when designing questionnaires for administration to online samples, researchers should consider including topic-relevant items (beyond traditional demographics) from one or more relevant external collections, to broaden the set of auxiliary variables that could be used for bias analysis and modeling. For non-general-population studies, researchers need to confirm that the benchmarking dataset(s) can be appropriately subset to the subpopulation targeted by the sample.

Appended variables

In the context of online panels—whether probability or nonprobability panels—the most common appended auxiliary variables are profile variables maintained by the panel vendor. As discussed in Section 2, vendors typically administer recruitment and/or registration surveys at the time a panelist agrees to join the panel, creating a set of profile variables available for all panelists. These variables are then updated and expanded over time.

Many panel vendors will append profile variables to study samples upon request. Those that have population benchmarks or are observable in an external reference sample can then be used for adjustment in the same way as collected auxiliary variables. In this way, profile variables offer an opportunity to adjust on some variables without having to ask them on the questionnaire (though, if the variable is critical to study analysis, panel vendors sometimes recommend confirming the profile information by also asking it on the questionnaire). Researchers who use online samples drawn from panels should always ask what profile variables, if any, can be appended to the completed dataset.

Profile variables may also be observable for some non-included units—specifically, for units that are in the panel, but were not invited for or did not respond to a specific study. This means that they are usable for adjustment approaches (such as propensity adjustments for study nonresponse) that require microdata for non-included units. The buyer of an online sample would usually not be able to implement this type of adjustment themselves, unless the vendor provided data for panel members who were invited (offered the survey) but did not respond at the study phase (did not accept the offer). However, vendors may incorporate such adjustments into the weights that they provide, either routinely or as an optional service; users should inquire whether this is the case.

For probability panels recruited from the ABS frame, recruitment samples can be matched to a wide array of auxiliary variables at the address level (from voter files and marketing databases) and/or the

[10] A full list of the principal statistical agencies of the U.S. government is available at https://nces.ed.gov/fcsm/agencies.asp.

neighborhood level (from public-use sources such as the Decennial Census and ACS) (Harter et al. 2016). These can potentially be used to model the first (recruitment) phase of the sampling process (West et al. 2015), particularly the propensity to respond to the recruitment survey and agree to join the panel. Again, it is unlikely that the user could operationalize such adjustments themselves, as it would require access to the full recruitment sample; but users who receive vendor-provided weights can inquire whether such adjustments are built into the weights.

For nonprobability samples obtained from non-panel sources, profile variables or other appended auxiliary variables are less likely to be available; therefore, researchers using such samples need to take particular care to collect information on relevant auxiliary variables in the questionnaire.

Choosing auxiliary variables for adjustment

It is only possible to adjust on auxiliary variables that are observed. Therefore, *prior* to collection, some effort should be made to identify variables that are likely to be related to key substantive outcomes and for which external data (either population parameters or microdata for individual non-included units) can be obtained. To be available as auxiliary variables for adjustment, these need to either be asked on the questionnaire or be available as panel profile variables. Of course, it can be difficult to know ahead of time which potential auxiliary variables are actually correlated with outcomes; if feasible, one or more pretest surveys could be conducted to identify a promising set of auxiliary variables.

After collection, some analysis should be done to confirm whether the available auxiliary variables are, in fact, correlated with study outcomes. This could take the form of a regression of study outcomes on the available auxiliary variables. Variables that do not substantially add to the regression's predictive power could then be excluded from the adjustment model.

It should be noted, however, that relationships between variables may themselves be subject to selection bias. Simulation research has shown that, when relationships between an auxiliary variable and outcomes differ between the sample and the population, including the auxiliary variable in the adjustment model can *increase* selection bias (Kreuter and Olson 2011). In practice, since outcomes are usually observed only in the sample, the extent to which this is the case can usually not be known. Therefore, it represents another source of uncertainty as to whether the adjustment model correctly accounts for the drivers of selection into the sample. Ultimately, post-collection analysis can supplement but not replace experience in choosing relevant auxiliary variables for adjustment, which can also be gleaned from looking at high-quality surveys on similar subjects.

Hybrid designs: combining probability and nonprobability samples

A key constraint on the range of available auxiliary variables is the need for these variables to be observable for non-included units, or at least have observable population distributions. This can raise a challenge when, for example, a researcher hypothesizes that a particular characteristic may be related to selection and the outcome, but this characteristic is not measured in any publicly available extant datasets. In this situation, the lack of external data could preclude the inclusion of this characteristic in the adjustment model, even if it is measured on the survey questionnaire.

Hybrid or blended sampling (Fahimi et al. 2015; Dever 2018; Robbins et al. 2021) has been proposed as a means of broadening the range of auxiliary variables that can be used to adjust a nonprobability sample. In a hybrid design, the same questionnaire is administered to side-by-side probability and nonprobability samples, and estimates are produced from the combined sample.

Hybrid designs allow auxiliary variables that lack external data to be incorporated into quasi-randomization and/or superpopulation adjustments. The key assumption is that the internal probability sample is more representative of the target population than the nonprobability sample and therefore provides a plausible source of data about non-included units.

Quasi-randomization methods can be applied to hybrid designs by using the internal probability sample (rather than an external dataset like the ACS) as the reference sample; this allows the propensity or matching model to include auxiliary variables that are measured on the questionnaire but not in any external dataset. When using superpopulation methods with hybrid designs, a common approach is to first weight the internal probability sample on its own, including calibration on demographics and any other auxiliary variables that have external benchmarks; and then use the weighted probability sample to estimate "internal" benchmarks for additional auxiliary variables that lack external benchmarks. The nonprobability sample is then calibrated on both sets of auxiliary variables: those with externally available benchmarks, and those without externally available benchmarks, the latter relying on the benchmarks estimated from the probability sample. Bayesian model-based methods have also been proposed for blending probability and nonprobability samples (Sakshaug et al. 2019; Wiśniowski et al. 2020).

By providing (via the internal probability sample) a means of adjusting on additional auxiliary variables, hybrid designs can reduce the risk of selection bias, relative to relying entirely on a nonprobability sample. As usual, this will be true only if the auxiliary variables used for adjustment are correlated with the outcomes of interest. It also requires that the internal probability sample be unbiased (or at least, less biased than the nonprobability sample) with respect to those auxiliary variables. In practice, therefore, the utility of hybrid designs is likely to depend on the source of the probability sample, the specific auxiliary variables included on the questionnaire, and the outcomes of interest.

Summary: Risk Assessment

This section has laid out a framework for understanding what conditions must be met to make inferences about a population from any type of online sample, whether probability or nonprobability; and what factors, for a given online sample, may increase the risk that these conditions will not be met and that estimates will therefore be biased. Ultimately, the risk that meaningful selection bias will remain after adjustment is a function of the amount of information a researcher has about (1) the selection process that determines which units are included in the sample and (2) the auxiliary variables that are related both to inclusion in the sample and to study outcomes.

When purchasing an online sample, the first of these risk elements is largely outside the researcher's control, except insofar as different types of samples (probability vs. nonprobability) and different vendors offer different information about the selection process. Therefore, the choice between more and less transparent sample sources comes down to the amount of risk that a researcher is willing to tolerate. All else equal, because probability-based designs inherently afford more information about the selection process, the risk of meaningful selection bias can be expected to be lower in a sample obtained from a probability panel. With nonprobability samples, uncertainty about the selection process is compounded by the lack of even the "baseline" knowledge about the frame from which sample members were recruited and the true (or even approximate) probability with which each was invited to participate.

Likely for this reason, empirical research continues to find that probability-based samples overall yield more accurate and less variable population estimates than nonprobability samples (Cornesse et al. 2020), though some nonprobability samples perform better than others (and better than some probability panels) (Kennedy et al. 2016). That said, probability panels are also subject to selection bias because of limitations on the sampling frame and nonresponse. Therefore, while probability panels are *more* likely *a priori* to meet the modeling assumptions described above, the risk of meaningful selection bias is ultimately study- and outcome-specific.

This makes the second risk element—the range of potentially relevant auxiliary variables available for use in adjustment models—important for both probability and nonprobability samples. This element *is* partially under the researcher's control, regardless of the sample source. By including potentially relevant auxiliary variables on the questionnaire—particularly those that go beyond traditional demographics and are expected to be correlated with key study outcomes—the researcher can increase the likelihood that the adjustment model will meaningfully reduce selection bias. If there are potentially relevant auxiliary variables with no available external data, a hybrid design—blending a probability sample with a nonprobability sample—could reduce risk by allowing the nonprobability sample to be adjusted on these variables, though this requires the additional assumption the probability sample provides reasonably unbiased estimates for those additional auxiliary variables.

With this in mind, the assumptions and risks discussed in this section should be considered *prior* to data acquisition and should enter into researchers' decisions as to (1) what type of sample (probability, nonprobability, or hybrid) will be used and (2) what auxiliary variables will be collected on the questionnaire and/or appended from external sources (if available). Section 5 provides specific guidance to researchers about what considerations should be factored into these decisions and what other questions need to be answered prior to designing and fielding a survey.

When designing questionnaires for administration to online samples, some thought must be given to the auxiliary variables that are likely to be related to study outcomes, and that should therefore be collected to enable assessment of and adjustment for selection bias. Indeed, the need for greater effort to identify and collect potentially relevant auxiliary variables can be thought of as a tradeoff for the cost savings associated with online samples. As will be discussed in Section 4, metrics are available to quantify, or at least place plausible bounds on, the risk of selection bias. However, as with the adjustment methods described in this section, all of these metrics require auxiliary variables that are observable for the full population and/or for non-included units, and their utility depends on the relationship between these variables and study outcomes.

Ultimately, it is not possible to adjust on auxiliary variables that are not collected or appended, and it is not useful to adjust on those that are unrelated to study outcomes. Therefore, the selection of effective auxiliary variables for bias analysis and adjustment requires both advance planning and an understanding of the survey topic.

Section 4: Measures of Accuracy, Bias, and Precision

As noted in the 2013 AAPOR Task Force Report on Nonprobability Samples[11], "the task of quantifying the quality of the nonprobability survey estimates is daunting," in large part because so much of the underlying mechanics of sample selection are unknown or unknowable—and vary from one organization to another.

The challenges in estimating data quality for nonprobability surveys that were identified in the 2013 Task Force report are still in play today. As noted in that report, measures of data quality are typically based on three key assumptions that do not hold in nonprobability designs: that a frame exists for all units of the population, that every unit has a positive probability of selection, and that the probability of being selected can be computed for each unit. The guidance from national statistical agencies, including the U.S. Office of Management and Budget, the U.S. Census Bureau, and Statistics Canada, has not been modified in recent years.

As discussed in Section 3, samples from probability panels afford more information about the sample selection process than online nonprobability samples, but they too are subject to some informational gaps and therefore face similar challenges in reporting on data quality.

However, since the 2013 Task Force report, AAPOR has provided new guidance on how to report precision estimates for nonprobability samples[12], including the pros and cons of each approach along with example reporting statements to guide survey researchers. We will not repeat that guidance in this section, but we will build off of those practical recommendations by expanding upon the assumptions upon which those approaches rely.

Bias vs. Precision

In considering metrics that are available for assessing the accuracy of estimates from any sample, it is important to distinguish between measures of *precision* and measures of *bias*. Measures of precision capture the *random* component of error in a survey estimate. Consider a hypothetical scenario in which the survey was repeated many times, with the sample selected and estimates produced in exactly the same way each time, such that any differences between the samples could plausibly be attributed to random chance. This hypothetical scenario is arguably unrealistic for nonprobability samples but helps to illustrate the conceptual differences between precision and bias. Measures of precision attempt to measure how widely the estimate would vary between these repeated samples. The commonly reported *margin of error* is a measure of precision, as are related metrics such as the standard error, confidence or credible intervals, etc. This random component of error shrinks towards zero (i.e., precision improves) as the sample size increases.

In contrast, *bias* captures the *systematic* component of error in a survey estimate. An *unbiased* estimate is one that, if it were averaged across those many repeated samples, would match the corresponding population parameter. *Bias*, then, refers to the difference between the true population parameter and the hypothetical average of an estimate across many repeated samples. Estimates from any sample may be biased if the modeling assumptions described in Section 3 are not met. Critically, *bias is largely independent of the sample size*, so it cannot be reduced simply by increasing the sample size.

[11] https://www.aapor.org/aapor_main/media/mainsitefiles/nps_tf_report_final_7_revised_fnl_6_22_13.pdf

[12] https://www.aapor.org/getattachment/Education-Resources/For-Researchers/AAPOR_Guidance_Nonprob_Precision_042216.pdf.aspx

In practice, bias cannot usually be exactly measured, since true population parameters are unobserved and (as discussed in detail in Section 3) it cannot be known with certainty whether the assumptions of the adjustment model are met. However, metrics are available to quantify the risk that estimates will be biased.

In the past, survey researchers often relied on two standard metrics of accuracy and precision in estimates derived from survey data. One is response rates and the other is sampling variability. These measures may be useful for stand-alone probability-based studies with extremely high response rates. However, for the reasons outlined above, these standards are not met with online samples, or really most surveys these days. Research conducted in the past decade has identified additional metrics that provide a basis for *sensitivity analysis* to place bounds on the amount of bias that may be present in an estimate if modeling assumptions are violated.

In this section, therefore, we begin by discussing response rates and the ways the AAPOR standards for computing them may become complicated by the nuances of how probability-based panels are recruited and maintained.

We then examine the different means that survey researchers have at their disposal when estimating precision for online probability and nonprobability samples. These include resampling, Taylor Series Linearization, and the application of the simple random sample (SRS) formula for margin of error. We provide guidance on the assumptions that are inherent in these methodologies and some limitations that may apply when these measures are calculated for online samples (particularly nonprobability samples). We note that the AAPOR Code of Professional Ethics and Practices allows measures of precision to be reported for nonprobability samples only "if they are defined and accompanied by a detailed description of how the underlying model was specified, its assumptions validated, and the measure(s) calculated."

We then examine metrics that are available for analyzing bias, with an emphasis on the use of sensitivity analysis for evaluating online probability and nonprobability survey samples, including the various methodologies that have been reported in the academic literature from the past decade.

Response rates

The reliance on response rates as a measure of survey data quality, specifically for probability-based samples, is both ubiquitous and contentious among survey researchers. For over twenty years, AAPOR has emphasized the importance of calculating and reporting response rates for all surveys and has provided guidance for the correct calculation across sampling frames and modes of data collection. As the industry has evolved, it has become increasingly clear that nonresponse is only one component within the larger Total Survey Error framework (Biemer, 2010) and that response rates alone are not necessarily indicative of whether or how much nonresponse error exists in a survey. However, as noted in The AAPOR Standard Definitions, "calculating the rates is a critical first step to understanding the presence of this component of potential survey error" (AAPOR, 2016).

As discussed in detail in Section 3 of this report, the risk of bias in data collected from online panels and other online samples is an important consideration for research conducted with these samples. In this section, we discuss the computation of response rates and other quality metrics for studies conducted using online probability-based panels (noting that such calculations cannot be completed in nonprobability samples, where known probabilities of selection do not exist) as well as limitations of these metrics for these samples. While research has shown definitively that *low* response rates *do not* automatically lead to nonresponse bias (Dutwin and Buskirk, 2021; Groves, 2006; Groves and Peytcheva, 2008), *high* response rates *do*, by definition, reduce the risk of nonresponse bias in probability-sample surveys (Peytchev, 2013). For this reason, this report suggests that response rates should be one of several metrics used in the assessment of data quality from probability-based panel studies and nonresponse bias should be considered along with other potential sources of bias in any study for which population inferences are being made. That said, defined by AAPOR as "the number of complete interviews with reporting units divided by the number of eligible reporting units in the sample", the response rate formula is based on assumptions about the eligible population that pose challenges to direct computation for these types of samples. The section below discusses the utility of response rates for online samples and details the challenges and assumptions inherent in the calculations.

In this section, we limit this discussion to probability-based samples, where known probabilities of selection exist and thus response rates as defined by AAPOR are calculable. For probability-based online panels, AAPOR Standards dictate that nonresponse at all stages, including recruitment and study level, should be included in the response rate calculation (AAPOR, 2022). As detailed by Callegaro and DiSogra (2008), this often involves the recruitment response rate (RECR), the profile rate if applicable (PROR), and the study level completion rate (COMR). The formula is as follows:

Cumulative response rate (CUMRR) = RECR x PROR x COMR

This cumulative response rate has been the standard for probability-based internet panels since 2008[13].

With a straightforward panel build in which there is one recruitment wave and all nonrespondents to a given study were from that wave, the logic for including nonresponse at recruitment and profiling is clear. Nonrespondents to the recruitment could not become a respondent to a given study and are therefore relevant to the final response rate calculation. However, the picture is not as clear when we consider some of the ways in which panels have evolved since these standards were set.

As discussed above, there are a number of different modes of recruitment and replenishment for panels, including the use of multiple modes for recruitment either concurrently or sequentially over time. Panels also use different cadences for replenishment, with some replenishing once a year or less often and others continuously recruiting panelists to maintain or increase panel size. Retention efforts have also evolved over time to combat increasing attrition (both overall and among specific important socio-demographic subgroups). With all of these nuances, and the longer history of some of these panels compared to when the response rate methodology was designed, there are new considerations for calculating probability-based panel response rates.

Consider a panel that was initially built in 1999 and has 4 recruitment/replenishment waves per year. That means that by 2022, the panel has had nearly 100 recruitment waves since inception. Thinking about a nonrespondent to wave 1 – there are numerous concerns about factoring them into the cumulative response rate for a study fielded in 2022. The person could no longer live in the household that was sampled for the panel; the household could have split into two households or the household composition could be very different in 2022 than it was at the time the recruitment attempt was made in 1999; the person could have moved to a new household that was subsequently sampled for the panel

[13] For definitions of these rates, see The American Association for Public Opinion Research. 2016. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition. AAPOR.

and perhaps joined the panel at the subsequent time; the person could no longer be alive; minors will have “aged-in” to eligibility for being an panelist, or it is possible the house itself could no longer exist. Under these scenarios, how meaningful is that 1999 recruitment nonresponse to a survey conducted in 2022? These same scenarios (moving, death, etc.) could apply to recruits who never profiled or who subsequently dropped out of the panel and yet that nonresponse is also factored into the response rate calculation in 2022, under the standard AAPOR calculation.

Additionally, it is important to consider how the sampling frame from which panelists are recruited has changed over the life of the panel. For example, an ABS frame such as the Computerized Delivery Sequence File (CDF) represents a snapshot in time such that the probability of selection for a given set of recruited panelists only represents the true probability of selection at that point in time. Similarly, a panel provider may have utilized multiple recruitment modes concurrently, or perhaps the panel switched recruitment methodology at a given point in time. A given study could include a mix of panelists from various recruitment waves. To obtain a recruitment rate, it is typically necessary to attach the recruitment response rate to each panelist who completed the study from the wave in which they were recruited. This could be a blend of recruitment rates from multiple recruitment modes and at multiple points in time. Is it meaningful to treat these modes with different recruitment rates equally in the calculation? Two studies run on the same panel could have very different recruitment rates depending on the blend of panelists, their tenure, and the mode under which they were recruited. If the study includes more longer tenured panelists from when recruitment rates were higher, the recruitment response rate could be much higher than a study run on the same panel with more recently-recruited panelists joining at a time of declining response rates.

Even within the same mode of recruitment, changes to the recruitment methodology are often introduced in an effort to improve response rates and representativeness of the recruited sample. For example, in the case of ABS recruitment – the stratification plan may be adjusted over time. Or with dual frame RDD (DFRDD), the blend of cellphone and landline sample might evolve. While none of these considerations are necessarily problematic to calculating panel response rates, they do raise considerations for the applicability of the cumulative response rate and the need for further refinements of panel response rate calculations.

At the study level, panels may use disproportionate probabilities of selection within their pool of panelists to draw samples from the panel for a specific study that maximize the representativeness of the selected study sample, regardless of whether the full-panel is 100% representative of the target population. Some practitioners would argue that these adjustments negate the importance of calculating the cumulative response rate for such samples.

However, as discussed in Section 3, the mitigation of nonresponse and noncoverage bias is only as good as the auxiliary variables available for both respondents and nonrespondents (or the full population) and the correlation of those variables with study outcomes. Thus, a cumulative response rate can still communicate the *risk* of bias in a particular sample.

The AAPOR standard definition documentation accurately delineates the components of panel response needed to calculate response: a panel’s recruitment rate, profiling rate, and the completion rate for a given survey from the empaneled and profiled members. However, the standards definitions do not address the loss of panelists over time. This can occur via panelists leaving the panel (mostly due to loss of interest, but as well, potentially, from mortality) or from the panel administrators forcibly retiring panelists. Each case presents complications in terms of documenting response at a particular stage of recruitment and study invitation. This is particularly true for the former of these, because it is not always readily apparent that a panelist has attritted out of the panel, versus just declined to participate in a string of invitations they received for specific surveys.

Callegaro and DiSogra (2008) document attrition rates largely from Clinton's (2001) metric of the difference of current panelists over initially recruited and profiled panelists. Whether from forcible retirement, mortality, or extended lack of participation by the panelist, panels should take into account attrition in their response rates. Technically speaking, a panel that does not retire or consider any panelists to have self-retired can accurately portray response by sampling all empaneled members, agnostic of retirement possibilities, to a given study, and attrition will be naturally folded into the cooperation rate of that study. In practice, however, there will be panelists who specifically ask to be retired and to cease from getting survey invitations. As such, presumably, such sample elements will no longer be sampled for specific surveys, and an attrition rate should be incorporated into the final response rate for that survey.

An additional nuance arises when panel companies randomly retire a set of panelists not due to nonresponse, but due to a concerted effort to balance the panel on certain demographic characteristics. For example, if a panel, on whole, overrepresents older white respondents relative to the population, they may randomly retire these respondents from the panel, akin to making them ineligible for all studies. As such, this type of "attrition" should be accounted for differently than attrition due to chronic nonresponse or poor data quality.

The specifics of these calculations, however, are not standardized or explicitly documented in the AAPOR standards.

Furthermore, response rates, including the cumulative response rate described above, are limited to probability panels. As described above in Sections 2 and 3, the characteristics of nonprobability samples (including nonprobability panels) restrict the ability to meaningfully calculate response rates as even an initial stage metric of data quality. Because nonprobability samples do not have a set sampling frame, the proportion of potential panelists who agree to join the sample versus those that do not is unknowable.

While nonprobability samples do not allow for the calculation and use of response rates, there are a number of other common metrics that nonprobability vendors may provide to serve as analogous sample quality metrics. The most common is the cooperation rate[14], or the proportion of people asked to participate in the survey who actually do so. Some nonprobability panels may provide other measures of population representativeness such as participation rate, completion rate, survey offer acceptance rate, etc. However, the calculation of these metrics is not standardized across vendors and there is limited research about the relationship between such metrics to the reliability or representativeness of data collected from participants in studies that use these samples.

For these reasons, the next sections discuss other measures available to researchers to assist in the evaluation of data quality for studies that use online samples. Some of these measures are applicable only to probability-based panels, while others can be extended to the nonprobability sample space.

[14] See The American Association for Public Opinion Research. 2016. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition. AAPOR.

Measures of Precision

Conceptual background: precision and confidence intervals

The main tool for assessing precision is an interval within which the target quantity is thought to lie, with a given probability. In classical statistics, this interval is called a *confidence interval*; in particular, a 95% confidence interval is a random interval that includes the target quantity in at least 95% of repeated samples. The conventionally reported *margin of error* is simply half the width of the 95% confidence interval. The 95% is the *confidence coefficient* and is chosen by convention. Other coefficients are also used – for example the U.S. Census Bureau typically provides 90% intervals.

In Bayesian statistics, the analog of the confidence interval is the posterior credible interval. A 95% credible interval is designed to include the target parameter with 95% probability. There are conceptual differences between confidence intervals and credible intervals – confidence intervals are random intervals whereas credible intervals are fixed intervals representing posterior uncertainty – but the distinctions are subtle, and perhaps of lesser importance for practitioners. We use the more familiar confidence interval terminology in what follows.

A common measure of precision, particularly in the survey field, is the *standard error* (SE) of the estimate. The interpretation of the standard error is based on its relationship with confidence or credible intervals; specifically, in large samples, the estimate plus or minus $z_{(1-\alpha/2)}$ standard errors is a 100(1- α)% confidence interval, where $z_{(1- \alpha /2)}$ is the 100(1- α/2) percentile of the standard normal distribution. In particular, setting $\alpha = 0.05$, the 97.5th percentile of the standard normal distribution is $z_{0.975}$ =1.96, and thus the 95% confidence interval is the estimate plus or minus 1.96 standard errors. The value 1.96 is often rounded up to 2, yielding a 95% interval as the estimate plus or minus 2 standard errors.

The relationship between standard errors and confidence intervals is based on the Central Limit Theorem and assumes large samples. Thus, in small samples the standard error is less easily interpreted, and confidence or credible intervals are more useful.

Just as there are design-based and model-based point estimates from surveys, there are design-based and model-based confidence intervals. Design-based confidence intervals are based on *sampling variability*, that is, random variation in the specific population units that are included in the sample. Model-based confidence intervals are based on the probability distribution of the survey variables assumed in the model.

Confidence interval coverage refers to the percentage of repeated samples in which the confidence bounds would include the true population parameter. If both the point estimate and the standard error are accurately estimated—that is, if both are free of bias—then the *true* confidence interval coverage will equal the *nominal* coverage (for example, the 95% confidence bounds will in fact contain the true population parameter in 95% of repeated samples). If the estimated standard errors are too small—which can occur if the method used to estimate the standard error does not account for some source of random variability in the estimate—true coverage will be below the nominal coverage. When less-than-nominal coverage results from the underestimation of standard errors, confidence intervals are "too narrow", overstating the precision of (and therefore understating the uncertainty in) the estimate.

Common methods of estimating precision

Here, we review common methods of producing confidence intervals in both probability- and nonprobability-based online samples. The purpose of this discussion is not to provide a detailed theoretical derivation of these methods or guidance on their implementation, both of which are available in numerous textbooks (e.g., Wolter 2007, Valliant et al. 2018). Rather, it is to provide an overview of the available methods and context for the ensuing discussion of considerations specific to online samples. The AAPOR Code[15] includes minimal disclosure standards for measures of uncertainty regardless of the sample source, and nonbinding supplemental guidance[16] provides more detailed recommendations specific to nonprobability samples.

In reporting and discussing precision in estimates from sample surveys, most practitioners implicitly assume a design-based inferential framework. That is, estimates are produced by applying weights that are assumed to reflect the inverse of the probability with which each unit was included in the sample—a probability that, as discussed in Section 3, is partially (in the case of probability samples) or wholly (in the case of nonprobability samples) estimated from a statistical model of the inclusion mechanism.

Online samples, both from probability panels and nonprobability sources, can generally be considered "complex" samples in that they depart from the assumption of a simple random sample (SRS) in which the final analytic units are sampled directly and independently from a single frame with equal probabilities of selection. In probability-based designs, common departures from SRS include stratification, clustering, multi-phase sampling, multi-frame designs, and nonresponse. For nonprobability samples, weights are based entirely on the statistical model of the selection process, which typically assumes that the inclusion probability is a complex multivariate function of the auxiliary variables included in the model. For this reason, formulas that assume an SRS will typically not yield correct confidence intervals for either probability or nonprobability samples.

Common methods for estimating confidence intervals in complex weighted probability and nonprobability samples, under a design-based inferential framework, include:

- **SRS formulas adjusted by the unequal weighting effect (UWE)**: the UWE (Kish 1965) is a common approximation of the design effect (DEFF), the factor by which a complex sample design inflates the variance of an estimate. The UWE is equal to 1 plus the squared coefficient of variation (CV) of the analytic weights, where the CV is the standard deviation of the weights divided by their mean. A common approach is to divide the raw sample size by the UWE to obtain an *effective* sample size. SRS formulas are then applied using the (smaller) effective sample size in lieu of the raw sample size. While appealing in its simplicity, this method has several important limitations. It yields the same value for all estimates, whereas the DEFF is in fact estimate-specific. It also assumes a stratified single-stage sample in which weights are independent of the outcome variables and may overestimate the DEFF if the weights are correlated with the outcome. Therefore, it does not account for clustering effects, correlations between the weights and outcomes, or complex weighting adjustments such as calibration and propensity adjustments, all of which can affect the DEFF.
- **Taylor series linearization**: linearization approximates the variances of complex survey estimates, which are nonlinear statistics, using formulas for their linear approximations.

[15] Available at https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx (accessed February 19, 2022).

[16] Available at https://www.aapor.org/getattachment/Education-Resources/For-Researchers/AAPOR_Guidance_Nonprob_Precision_042216.pdf.aspx (accessed February 19, 2022).

Linearization formulas are the default in the complex-sample functions offered by most statistical software packages. However, to accurately estimate variance, the user must specify all relevant aspects of the sample design, including cluster and stratum identifiers. Furthermore, the ability to account for common weighting adjustments (such as propensity adjustments and raking) varies across software implementations, and this may require specifying information that is typically not available to the analyst (Valliant et al. 2018). In practice, therefore, linearization formulas often do not fully account for the impact of such adjustments on sampling variance.

- **Resampling (or replication) approaches**, such as the bootstrap or jackknife. These approaches rely on drawing subsamples ("replicates") from the main sample, repeating all weighting and estimation steps on each subsample, and then applying a formula that estimates the standard error based on the variability of an estimate across the weighted replicates. Replicate weights can be created to allow users to use replication-based formulas for many estimates without repeating the subsampling each time. In practice, resampling methods usually offer the best means of accounting for complex, multi-step weighting adjustments such as propensity and raking adjustments. However, for these methods to yield correct variance estimates, the subsampling must fully replicate the sample design that was used to select the main sample, including any complex design features such as stratification and clustering. Furthermore, all weighting steps must be rerun on each subsample as though it were an independently selected sample—for example, if raking is used, the weights for each individual subsample must be independently re-raked to the applicable control totals.

Several simulation studies (e.g., Lee and Valliant 2009, Valliant 2020, Robbins et al. 2021) have compared linearization to resampling methods when applied to nonprobability samples, generally finding that resampling methods yield the closest-to-nominal confidence interval coverage. Though the SRS formula adjusted by the Kish (1965) UWE is a very common approach in practice, other research (Little and Vartivarian 2005) shows that it is an inadequate approximation when weighting variables are correlated with substantive outcomes–that is, *in precisely the scenario in which the weighting is effective at reducing bias.*

For a detailed discussion of estimating precision under a model-based inferential framework, readers can refer to chapter 5 of Valliant et al. (2000). Elliott and Valliant (2017) and Valliant et al. (2018) discuss the application of model-based variance estimators to nonprobability samples in particular, and the simulation study reported by Valliant (2020) evaluates several such estimators.

Considerations specific to online samples

In a purely mechanical sense, therefore, the *methods* available for estimating precision in online probability and nonprobability samples are the same as those available for any complex probability sample. However, it is important to understand that measures of precision (like point estimates) are based on assumptions. If these assumptions are violated, these measures may substantially overestimate the precision of a statistical estimate—that is, confidence intervals will be "too narrow", and their true coverage will be below the nominal coverage.

Several features of online samples can make it particularly difficult or impossible to apply these methods in a way that makes their underlying assumptions plausible. Therefore, *regardless of the methods used to produce these measures, measures of precision reported for online samples (standard errors, confidence intervals, margins of error, statistical significance tests, etc.) are likely to be over-optimistic,* *and hence should be interpreted with caution*. This applies particularly, but not exclusively, to nonprobability samples. Below, we discuss some considerations that can limit the accuracy and utility of these measures for online samples. We do so not to discourage users of online samples from calculating and reporting measures of precision, but rather to ensure that researchers and their stakeholders are aware of what these measures can and cannot tell them.

First, as noted above, *design-based variance estimation formulas—whether based on a linearization or resampling approach—require information about the sample design and weighting that may not be available to users of online samples*. Stratum and cluster identifiers are needed to apply the proper linearization formulas or to properly replicate the sample design in a resampling approach. In the case of probability panels, these could be provided by the vendor, and users should inquire whether they are available. Note, however, that *all* stages of the sampling process—not only sampling *from* the panel for a specific study, but sampling *for* the panel as part of the recruitment phase—must be specified (for linearization estimators) or replicated (for resampling estimators) to correctly estimate standard errors. Differences in stratification and/or clustering between the recruitment and study phases, as well as panel management practices that complicate the calculation of true selection probabilities, can make it difficult or impossible to calculate standard errors that reflect the full multi-phase sample design inherent in probability panels.

When probability panel vendors provide weights, resampling methods may be further limited by a lack of access to the data needed to rerun all weighting steps on each replicate. For example, if vendor-provided weights include a propensity adjustment for nonresponse at the recruitment phase, replicating this adjustment would require access to the full recruitment sample, which vendors typically would not provide. Vendors could, in principle, implement the resampling and reweighting themselves and provide the user with replicate weights, allowing the use of replication variance formulas. Though common for public-use data files produced by statistical agencies, the provision of replicate weights is not currently a common practice among probability panel vendors, and therefore would likely need to be requested as a special service.

With nonprobability samples, the challenge of correctly estimating standard errors is exacerbated by the fact that there is no replicable “sample design” in the traditional sense, and the methods used to recruit respondents are largely unobservable to the user. Thus, nothing even approximating a stratum or cluster identifier is likely to be available, even though certain aspects of the recruitment structure may approximate these design features. For example, individual recruitment websites may function similarly to clusters (Brick 2015, Elliott and Valliant 2017). Similarly, while the use of quotas in nonprobability sampling is somewhat analogous to stratification in probability sampling, there are important differences. While strata are disjoint (mutually exclusive) by definition, quotas are often specified non-disjointly—for example, a vendor may be told to match marginal population distributions by race and gender, but not the joint distribution of race *by* gender, which would have different implications for sampling variability. Therefore, when applied to nonprobability samples, linearization estimators are unlikely to account for all design features that impact sampling variability. Likewise, when using resampling estimators with nonprobability samples, it is likely impossible to draw subsamples in a way that fully mimics the original recruitment process, requiring the user to resort to approximations (e.g., drawing simple random samples from the obtained nonprobability sample) that are likely to understate sampling variability.

Second, *all measures of precision are conditional on a statistical model of the selection process, and therefore leave out an important source of uncertainty that is particularly salient in nonprobability samples*. As discussed in Section 3, modeling is needed to account for the stages of the selection process at which inclusion probabilities are not known—nonresponse for probability panels, and the entire process for nonprobability samples. When weights are used for estimation, models are "built into" those weights via any weighting steps (such as propensity and raking adjustments) that are not based on known selection probabilities. Measures of precision, just like point estimates, rest on the assumption that the underlying adjustment model is "correct"—for example, that weighting adjustments account for all auxiliary variables that influence the likelihood of inclusion in the sample and are related to the outcome being estimated.

Because confidence intervals are conditional on the adjustment model, they do not incorporate the "meta-uncertainty" as to whether the adjustment model is correct. As discussed in detail in Section 3, various features of online samples—particularly nonprobability samples—tend to amplify uncertainty as to whether the adjustment model incorporates all relevant auxiliary variables. This, in turn, implies that confidence intervals from such samples are likely to be too narrow. The extent of this undercoverage is usually unknown, but is likely greater than in standalone probability samples in which the researcher has more control over the selection process.

Third, and relatedly, *measures of uncertainty capture only the random component of error in an estimate; they do not capture any systematic biases in the estimate*. Even if the uncertainty in the estimate (i.e., the *width* of the confidence interval) is correctly estimated, confidence interval coverage will be below the nominal level if the estimate itself is biased (Valliant 2020). Thus, for online samples—particularly nonprobability samples—measures of precision tell only part of the story. In addition to reporting measures of precision, some effort should be made to also assess and report the risk of bias. We now turn to metrics that are applicable for that purpose.

Bias and sensitivity analysis

A central issue is that, while the variance of an estimate can typically be estimated from the sample, the bias of an estimate is not estimable without some external information on the population. In the case of probability samples with nonresponse, this information takes the form of variables measured for respondents and nonrespondents, or for the whole population. Many measures that were originally defined to assess the impact of nonresponse in probability samples can be adapted to measure selection bias in nonprobability samples. Below we discuss several such measures.

We also provide a brief overview of how these measures are calculated, which necessitates the use of some statistical notation. Throughout the ensuing discussion, suppose Y is an outcome that is observed *only* for units from the sample that respond—for example, a measure collected by the survey instrument for which no external information is available. Thus, bias in Y cannot be directly measured. Suppose Z is a set of auxiliary variables available for the full sample. Unlike with Y, assume the Z variables either (1) are observed for individual non-included units or (2) have available summary statistics for the population as a whole. Therefore, it is possible to directly measure bias in Z.

The R-indicator

Many measures incorporate comparisons of the distribution of Z between included and non-included units – interpreting differences in these distributions are evidence of potential selection bias.

When multiple auxiliary variables are available, a common way of aggregating them into a single measure of bias is to estimate a *propensity score* Pr(S = 1|Z), defined as the estimated probability that a unit was included in the sample. The propensity score is often estimated by a logistic model:

$$logit\,\Pr(S = 1 \mid Z, \gamma_0, \gamma) = \gamma_0 + \gamma Z \quad (1)$$

where S is an indicator of selection (1 = included in the sample, 0 = not included) and the logit of a proportion p is the log odds, $\log(p / (1-p))$.

Note that estimating the propensity score from a logit model requires individual-level data on non-included units. As with the propensity adjustments discussed in Section 3, this could take the form of a separate "reference sample" that is representative of the target population.

The propensity score can be used to estimate the **Representativity or *R* indicator** (Bethlehem, Cobben and Schouten, 2009):

$$R = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p_i - \bar{p})^2}, \quad (2)$$

where p is the propensity score, N is the population size, and the subscript i denotes the ith unit of the population. That is, R is equal to 1 minus twice the population standard deviation of the propensity scores. The values of R range between 0 and 1 and average 0.5, with the maximum value of 1 attained when the probability of selection is unrelated to Z, and the minimum value when the probability of selection is perfectly predicted by Z. Intuitively, when R = 1, all population units are estimated to have the same probability of inclusion in the sample, which implies no risk of selection bias. Therefore, larger values of R are interpreted as indicating a lower risk of selection bias.

The R indicator is relatively popular and easy to estimate, but it has two important limitations. First, because it relies on the propensity score, it requires individual-level data for non-included units. Second, it takes the same value for all survey outcomes Y. That is, it measures the extent to which the auxiliary variables Z, but not the outcomes Y, are related to selection. In reality, different outcomes can show different levels of selection bias, depending on whether Y is strongly or weakly related to selection.

The H_1 indicator

The measure ***H*₁** in Sarndal and Lundstrom (2010) is more tailored to individual outcomes Y, and assesses how the *best predictor of Y* based on Z differs between included and non-included units. Thus, while R requires modeling selection as a function of Z, H_1 requires modeling the outcome Y as a function of Z.

For example, assuming a linear regression of Y on Z, let $X = E(Y \mid Z) = \hat{\beta}_0 + \hat{\beta} Z$, where the regression coefficients β are estimated on the selected sample. Applying these regression coefficients to both the selected and non-selected units yields X, the best predictor of Y based on Z. Let $X^* = cX$, with c being a constant that scales X to have the same variance as Y in the selected sample. H_1 is then calculated as $\bar{x}_n^* - \bar{X}_N^*$ the difference between the sample and population means of X^*. Intuitively, the larger this difference, the greater the risk of selection bias.

Unlike R, H_1 can be estimated without individual-level data for non-included units as long as the outcome is modeled using a linear regression and the population means of the Z variables are known. In this case, the population mean of X^* can be obtained by applying the regression coefficients to the population means of the Z variables.

Sensitivity analysis metrics

Though H_1 is tailored to individual outcomes, both *R* and H_1 are limited in that they only reflect selection bias as it relates to the auxiliary variables *Z*, and do not reflect selection bias related to variables other than *Z*. In other words, they implicitly assume that selection is ignorable, that is, independent of *Y* after conditioning on *Z*. In practice, *Z* is often limited to demographic variables weakly related to the outcomes (*Y*) of interest; therefore, the degree of dependence of selection on *Z* may not provide reliable evidence about the potential selection bias in *Y*.

Furthermore, any auxiliary variables *Z* where there is information on non-included units can be used for the adjustment methods discussed in Section 3 such as regression, poststratification or raking. Adjusting on *Z* using one or more of these methods would eliminate any bias in *Y* attributable to bias in *Z*. Since any bias in *Y* attributable to bias in *Z* can be corrected, the real concern is potential bias due to variables other than *Z*, which is not captured by either *R* or H_1.

Ultimately, in the absence of external information about *Y*, the ignorability assumption cannot be verified. However, measures developed by Little et al. (2020) can be used to analyze the *sensitivity* of estimates to violations of ignorability. Specifically, these measures allow analysts to estimate potential selection bias in the estimated mean of *Y* under varying assumptions about the degree to which selection depends on *Y* after controlling for *Z*.

These measures are based on the proxy pattern-mixture model for nonresponse in Andridge and Little (2011), and can be used to evaluate both probability and nonprobability online panels. For continuous outcomes, the model leads to the **standardized measure of unadjusted bias (SMUB)**:

$$SMUB(\phi) = \frac{[\phi+(1-\phi)r](\bar{x}_n^* - \bar{X}_n^*)}{[\phi r+(1-\phi)]\sqrt{s_x}}, \quad (3)$$

and the **standardized measure of adjusted bias (SMAB)**, which assesses bias from deviations from ignorability:

$$SMAB(\phi) = SMUB(\phi) - SMUB(0) = \frac{[\phi(1-r^2)](\bar{x}_n^* - \bar{X}_n^*)}{[\phi r+(1-\phi)]\sqrt{s_x}}, \quad (4)$$

where $\bar{x}_n^*$ is the sample mean of the best predictor of *Y* based on *Z* (as defined above); $\bar{X}_n^*$ is the population mean of this predictor; s_x is the variance of this predictor estimated from the sample; and *r* is the correlation (estimated from the sample) between *Y* and X^*. The SMUB provides an estimate of selection bias in the unadjusted estimate, and the SMAB provides an estimate of the amount of bias that remains after adjustment.

These measures are based on normality and not assumption-free. They make the key assumption that selection depends on *Y* and X^* through the linear combination $(1-\phi)X^* + \phi Y$, and other variables in *Z* unrelated to *Y*.

The key parameter for sensitivity analysis is ϕ , which must be chosen by the analyst and reflects the assumed degree to which selection depends on *Y* after conditioning on auxiliary variables *Z*. In particular, if $\phi = 0$, then selection does not depend on *Y* after controlling for Z, and$SMUB(0) = \frac{r(\bar{x}_n^* - \bar{X}_n^*)}{\sqrt{s_x}}$. That is, SMUB(0) provides an estimate of selection bias if the ignorability assumption is met. If $\phi = 1$ then selection depends entirely on *Y* and we obtain $SMUB(1) = \frac{(\bar{x}_n^* - \bar{X}_n^*)}{r\sqrt{s_x}}$Values of ϕ between 0 and 1 assume that selection depends on both *Y* and *Z*, with values closer to 1 reflecting a stronger dependence on *Y* (i.e., a greater divergence from ignorability).

There is no information about the parameter ϕ in the sample. Little et al. (2020) suggest two forms of sensitivity analysis to assess potential bias for different choices of this parameter. One is to compute inferences for three choices of ϕ reflecting weak, moderate or strong dependence of selection on *Y,* namely $\phi = 0, \phi = 0.5, \phi = 1$. The second is to do a Bayesian analysis based on a prior distribution for ϕ . Another option is simply to set $\phi = 0.5$, a middle value, which yields $SMUB(0.5) = \frac{(\bar{x}_n^* - \bar{X}_n^*)}{\sqrt{s_x}}$, the bias as estimated by the auxiliary proxy for *Y*. This is equivalent to the H_1 indicator and closely related to the Bias Effect Size proposed by Biemer and Peytchev (2011).

The SMUB (Equation 3) and SMAB (Equation 4) are derived for a continuous *Y*. Extensions are available for a binary *Y* (Andridge et al. 2019) and for regression coefficients (West et al. 2021).

For a continuous *Y*, the SMUB and SMAB can be estimated without individual-level data for non-included units. They require only that the population mean of *X** be known, which in turn, as noted above, requires only that the population mean of each *Z* variable be known. The extensions have additional informational requirements. The extension for a binary *Y* requires knowledge of the population variance of *X**, and the extension for regression coefficients requires knowledge of the population covariance matrix between the auxiliary variables and the regression predictors.

Boonstra et al. (2021) describe a simulation study to compare the SMUB to several other measures, extending previous simulations in Nishimura et al. (2016). They conclude that "Our simulation study showed that the middle value of ϕ = 0.5, which Little et al. (2020) heuristically suggested for default use, resulted in a diagnostic that most consistently estimated the true amount of selection bias."

Note that the sensitivity of inference to different choices of ϕ depends on the correlation *r* between *Y* and the best proxy for *Y*, X^* . Andridge and Little (2011) call X^* a strong proxy when *r* is large and a weak proxy when *r* is small. Inference is much less sensitive to ϕ if X^* is a strong proxy than if it is a weak proxy. This reinforces the point that finding auxiliary variables that are predictive of the survey variables is critical to both measuring and correcting selection bias.

Section 5: Reporting and Transparency

Identifying whether a probability-based or nonprobability-based panel sample is the most appropriate method for your research is contingent on a variety of factors, many of which have been addressed in previous AAPOR task force reports. For example, the intended use of the results, timeline, and budget for the project all play significant roles in a researcher's decision to rely on a probability or nonprobability-based sample. Regardless of the decision of which panel type bests suits the research project, the *quality* of the panel should also be considered. The following section describes questions and considerations that one should think through before deciding which sample type or sample provider to use.

It is important to note that AAPOR makes no judgment about the answers to the questions posed below. Rather, the goal of these questions is to promote transparency and disclosure so that researchers are able to make informed decisions when choosing a panel provider. Additionally, not all questions below are directly applicable to every project and it is important to consider which of these items are most relevant to *your* unique research.

Recruitment & Attrition

How are people recruited to join the panel?

For probability-based panels, panelists are mainly recruited through address-based sampling (ABS; randomly selected addresses) and random digit dialing (RDD; randomly selected landline or cellphone numbers). For nonprobability panels, panelists are recruited using a wider range of methods that include, but are not limited to: placing banners on websites, leveraging affiliate networks, rewards program membership, co-registration when someone signs up to another online service, and more. Across both panel types, it's important to understand the recruitment methodology and whether the panel recruits non-internet users to participate in surveys (either through providing an online way to participate or via the phone). Panels that only recruit through online methods will have underrepresentation of non-internet users in their panel (Pew Research Center estimates that around 6% of U.S. adults do not use the internet as of the publication of this report)[17].

What methods are used to ensure respondents are who they say they are?

There a range of reasons why respondent verification can be important – first, people joining a panel could potentially be dishonest about their demographics (including geography), and would be mis-represented in any analysis that compares demographic or geographic groups. Next, there is some evidence to show that survey-taking "bots" can be programmed to take surveys and provide flagrantly false answers so the creator can earn incentives quickly with minimal work. Lastly, with address-based sampling, the panel invitation usually provides specific instructions for *who* to select within the household (e.g., "person with the next birthday"), and it's possible that a person who was not intended to be invited joins the panel.

Panels take different approaches to identity verification, and there's no single best way – however, it's important that a panel provider be able to openly discuss their approach with you. In addition, panel

[17] https://www.pewresearch.org/methods/fact-sheet/national-public-opinion-reference-survey-npors/

providers should be incorporating data quality checks in their survey administration that may help ensure respondent identity (see question below regarding data quality).

What does the panel attrition look like?

It's typical for all panels to experience some level of attrition – meaning people who have joined the panel, and participated in one or multiple surveys, but are now no longer actively participating in the survey panel. Attrition can occur *passively*, whereby panelists who stop participating are removed from the panel by the panel owner, or *actively*, whereby panelists request to be removed from the panel. This can be done by the panel owner because panelists are no longer participating in surveys, or can be done because the respondent isn't giving quality data. A high panel attrition rate may indicate that survey-takers are having a poor user experience or that the panel provider is removing panelists who are failing data quality checks (mentioned below).

An additional consideration for panel attrition is that the size of a panel may be inflated due to the inclusion of passive respondents that have not been removed by the panelists. For example, if a panel claims to have 50,000 members, but 10,000 of them are inactive, the panel could be advertising a size of 50,000 but only be able to offer 40,000 viable survey takers. Additionally, response rate estimates would be inaccurate due to the number of respondents who are passively in the panel (e.g., you expect a certain response rate, but a large portion of the survey invitations went to panelists who are not actually going to participate, so your response rate will be lower than expected).

What is the panel size and maximum representative feasibility?

Panels vary in size. On average, nonprobability panels typically have more active panelists than probability-based panels. This is, in part, due to the greater ease and lower cost of recruiting respondents through nonprobability methods. Panel size is not a direct indicator of panel *quality*, however it is critical to identify whether the panel of interest has *enough* panelists to fulfill your research goals, especially if you are looking to conduct research with smaller population segments or groups.

A panel's capabilities to reach people of different walks of life is a critical factor in deciding which panel to use. For example, a panel may have several hundred-thousand people ready to take a survey but if these people do not fit the demographic criteria of the group you want to survey then this panel will not be useful to you.

The primary consideration for maximum representative feasibility is to understand what your most limiting demographic group is. Before selecting a panel, we recommend that you discuss exactly how many survey *completes* you are expecting from your sample so the panel can determine whether their service can support your research. This answer might not be obtainable as an exact number, but the key point is transparency between the panel provider and the panel consumer. In some instances, a panel provider may not know how many people in their panel fit a certain criteria if the criteria is unmeasured or not regularly tracked.

Panel Use

Is the researcher looking to conduct time-series or longitudinal research?

Online panels vary widely in size and retention of panelists which can cause sample composition to be significantly different at different points in time. For that reason, for time-series or longitudinal research, make sure to discuss with sample providers the entire program’s needs, including number of waves of research, if unique respondents are needed in each wave, or if the research program would like to sample the same individuals over time.

Generally, panels with low attrition rates (i.e. probability panels) provide an easier path to longitudinal research with the same respondents with most nonprobability panels having significant turnover in the panel over a moderate to long time frame. Many panel vendors also have specific policies about re-contacting panelists that should be considered in the design phase.

What are policies around survey scripting and sample provision?

Sample providers differ in how they prefer clients access their sample. Some offer the sample direct to client-supplied or external survey environments in a “sample only” arrangement. In these, the research client is required to script the survey instrument and the panel vendor provides minimal additional services, often not providing sample quality control, weighting, or data processing. Other sample vendors provide a more “full service” offering where they will program the survey instrument, host it on a company platform, and frequently include data processing and data checks. However, full-service companies are often reluctant to provide their sample to external research platforms. Researchers seeking to do their own survey scripting should discuss this with vendors up-front.

How complicated is the survey instrument?

Most online sample providers have moved to a “device agnostic” mode allowing respondents to complete the survey on a computer or mobile device (smartphone or tablet) with the majority of survey completes for most panels coming from mobile devices. While this approach allows more flexibility for the respondent, it can limit the researcher. Panel providers using device agnostic survey platforms will often have relatively strict limits on the number of characters allowed in question and response options and often other limits on complex questionnaire formulation. Panel companies can limit survey respondents to computer only, but often with feasibility implications. Researchers should discuss any complicated instrument design concepts with a sample provider before commissioning research.

How are panelists contacted to participate in a specific survey?

Panels providers use wide-ranging approaches to include panelists in a given study. Approaches include mail-push-to-web, email invitation, text/SMS, app or program-based, or online river sample among many others. These modes of contact each carry difference pluses and minuses, particularly when it comes to what portions of the population they are most effective with. Email continues to be the most common approach, but researchers should discuss contact methods, particularly if the research program is targeting hard-to-reach populations. In addition, if non-internet individuals are included in the panel, the researcher should know how they are contacted to participate in the survey and whether this will require designing the survey to be completed online and using another mode such as computer-assisted telephone interviewing.

How are panelists selected for specific studies and how many surveys are panelists invited to participate in?

Previous research has demonstrated that having panel members take too many surveys in too short of a time interval can yield results that are not representative of the population. It is important for researchers to understand how panel providers build their sample of panel members and whether the panel limits "over-participating" in surveys, especially with surveys on similar topics during a specific time frame.

Does the researcher want to monitor field progress?

Studies conducted with online panel providers differ widely in the experience once the study launches. Basic nonprobability samples are often executed very quickly, with tens of thousands of survey invitations and thousands of completes possible in as little as 24 hours. If researchers would like to monitor field work, and potentially make changes while in field, these needs should be addressed early so the research design can take them into account. Research vendors have relatively fine control over the how the sample is released (the "outgo") and most can schedule data collection around client needs.

Along the same lines, interest in any sort of "soft launch" or pre-test of the research instrument should also be discussed at the initial stages of research design.

How are panel members incentivized?

Most online panel companies provide some sort of incentive for their panelists to complete individual surveys. These methods range from cash, to gift cards, to point systems. Policies around incentives can have a direct impact on sample quality. Researchers should have an idea of the level of incentivization for a particular study, how respondents qualify for the incentives, and if there are caps on the amount that can be earned in a given time period.

Additionally, incentive policy might or might not be variable at the study level depending on the panel vendor and panel. With panels that allow modification of the incentives, they can be used as an additional factor in sample design.

What secondary information is available about respondents?

Online sample companies who maintain panels often have pre-collected data such as demographic questions or other non-attitudinal measures on panelists. These "profile variables" are often available for the researcher and provides a way to reduce the number of questions that have to be included in the research instrument and can provide additional variables for weighting adjustments. Researchers should discuss with the panel vendor how much coverage (i.e. what percentage of panelists have answered a specific question), how recently the data was collected or re-asked, and whether the panel provider allows researchers to re-ask similar questions as part of the specific survey project.

Data Privacy, Weighting, and Quality

What data privacy and security practices does the panel abide by?

Panels each abide by different standards for respondent privacy. It's important to ask a potential panel provider what practices they have in place regarding data confidentiality and/or privacy agreements, data security practices, and data transfer practices to ensure that respondents' expectations for data security and privacy are met for your research. These privacy and confidentiality arrangements should be transparent—both to you as the researcher and (as applicable) to panelists. The panel should have documentation they can share regarding these arrangements, including the legal authorities they use to promise confidentiality (i.e. a Certificate of Confidentiality). Researchers may also need to know how the panel stores the data they collect from respondents. Some companies use de-centralized or cloud storage, while others store data on servers the panel company controls and maintains itself. Relatedly, researchers may wish to ask panels whether and how they co-mingle data from different clients—i.e.

will respondent data collected from your project be available to the panel's other clients? If so, what are the limitations to this as they affect potential privacy and confidentiality concerns?

What weighting methods, if any, are used by this panel?

The goal of weighting a dataset is to reduce bias and to improve accuracy of results. As such, the weighting methods (if used, at all) implemented by the panel will have implications for your statistical inference. For some panels, weighting the data based on targets is a standard practice and the information needed to weight data is readily available. For other panels, weighting may not be fully possible due to the absence of populations targets or absence of information about respondents that would be required to create and apply such weights.

It is important to note that, in the first instance, weighting is not the same across probability and nonprobability panels. With probability panels, weights are generally based on the inverse probability of selection from a known survey frame; however, since there is no known frame for nonprobability panels, this weighting approach is not possible. It is important for a researcher to discuss whether weighting will be done with your data, and the approaches used to do it can be an important part of determining which panel provider to use.

Researchers should ask panels to provide information regarding the weighting approach they take, and ascertain what information the panel is willing and able to share regarding their approach. For instance, a researcher may wish to have the ability to recreate the weights that the panel calculates—therefore, they need to understand whether the panel is able to share the type of panelist- and respondent-level data needed to replicate the weights, including separating probability of selection weights from other adjustment factors.

What auxiliary variables are available for use in weighting and bias assessment?

As discussed in Section 3, a key factor that must be considered is what auxiliary variables will be available for use in weighting (or other adjustment approaches) and bias assessment. To effectively reduce selection bias, auxiliary variables must be associated with *both* inclusion in the sample and the substantive outcomes measured on the study. Therefore, the selection of effective auxiliary variables for bias analysis and adjustment requires both advance planning and an understanding of the survey topic.

The availability of effective auxiliary variables will depend on the study topic—some outcomes may have strong correlates that have available population-level benchmarks, while others may not. It will also depend on practical factors such as the amount of available questionnaire space and the range of profile variables maintained by the panel vendor.

The key point is that the availability of auxiliary variables for adjustment should be considered ahead of data collection and should factor into your decision on sample sourcing. If effective auxiliary variables are not likely to be available, there may be a stronger case for using a probability-based panel, or a hybrid approach that blends probability and nonprobability data. Probability panels may be more robust to an incomplete adjustment model due to the use of a known sampling frame and known probabilities of invitation for recruitment and individual studies. At the same time, probability panels also depart in important ways from the ideal of purely probability-based selection (primarily due to nonresponse), so the use of a probability panel does not obviate the need to carefully consider what auxiliary variables can and should be used in the adjustment model.

How do they evaluate data quality?

In order to ensure the quality of the survey data collected from online panels where there is not a live interviewer administering the survey, it is important that the panel provider include data quality checks in the survey instrument. There is no one "correct" approach to evaluating data quality, but what is important is whether or not a panel is willing to transparently explain their approach. Researchers should inquire about any data quality procedures a panel uses. Many organizations employ trap questions to ensure the respondent is the individual who was sampled for the survey or that the respondent is paying attention to the questions being asked. In addition, some panels use speeding checks that remove respondents who complete a survey under some standard time, such as completing a survey less than 1/3 of the median completion time. However, as demonstrated by Kennedy et. al. (2021), the traditional trap questions and speeding checks often employed by opt-in sample vendors are insufficient at detecting the majority of insincere respondents from the data. It, therefore, is important for a researcher to understand how panels make these decisions, and whether or not a researcher is able to set their own bespoke limits or data quality checks.

In addition, researchers should also ask panels about how they handle missing data—both for required survey items and for non-required items. Likewise, panels should be able to explain how they handle missing characteristic data (i.e. demographic or household income data that is typically collected during panel recruitment).

What information (if any) can be provided to aid in estimating precision?

As discussed in Section 4, there are standard methods for estimating measures of precision (e.g., margins of error or confidence intervals) for complex probability samples; these include Taylor series linearization and resampling methods such as the jackknife and bootstrap. To yield correct estimates of sampling variability, these methods require the specification of certain aspects of the sample design, such as stratum and cluster identifiers. Vendors may be able to provide such identifiers for probability panels, though likely not for nonprobability samples. Therefore, if you are planning to use one of these methods to produce measures of precision for a sample obtained from a probability panel, it is important to ask the vendor whether such identifiers can be included in the data file. If this information is not available, be aware that measures of sampling variability may themselves be biased in unknown ways. For additional considerations related to the reporting of sampling variability for probability and nonprobability panels, refer to Section 4.

Section 6: Summary and Recommendations

The ubiquity of online samples, both probability-based and opt-in, has led to increased concern that our historical indicators of data quality may not be adequate or meaningful for these types of samples. The goals of this report were to 1) provide audiences with an overview of the various types of online survey sampling methodologies currently being employed by survey researchers and major survey firms; 2) discuss how alternative methodologies for initial recruitment and decisions around panel freshening, respondent attrition, and missing data may impact sampling and data quality; and 3) outline some ways to assess the quality of online samples. In addition, the report provides researchers with some important questions to consider, and potentially ask sample vendors, as they develop their research design.

Online samples can provide researchers with substantial flexibility at reasonable costs. However, it is crucial that researchers are thoughtful about the goals of their study, whether the sample can provide adequate coverage of the full population of interest, and the level of risk of bias that is acceptable for a given project.

Most generally, probability-based panels can be considered an extension of other types of probability-based survey samples, requiring many of the same considerations historically applied to other probability samples. Researchers must consider the recruitment frame as well as the sample selection methodology. They must be sure to understand any undercoverage due to the mode(s) of recruitment and/or data collection. They must be aware of how the data are typically weighted and understand the assumptions inherent in such weighting models. In addition, researchers relying on probability-based online panels need to understand how panelists are retained, the effects of attrition on panel representativeness, and the methods for incorporating attrition and freshening into panel-reported response rates.

Nonprobability online samples present added considerations for researchers that this report details in Section 2. Specifically, recruitment methods cannot be easily mapped on to existing statistical frameworks for probability-based inference. For instance, the use of partner networks for recruitment, intercept or “river” sampling, and the use of routers to direct respondents to particular studies all introduce added complexity into the samples available to researchers. Therefore, added consideration must be given to who is included or excluded from nonprobability samples and how can clients work with sample vendors to obtain the appropriate sample for the study at hand. The diversity of methodologies available to researchers using nonprobability online samples magnifies the importance of understanding the limitations of all imperfect survey samples.

The third section of this report, therefore, discusses directly how deviations from random sampling (inherent in both probability-based and nonprobability online samples) can introduce bias into estimates and how statistical models are often used to “recover” true population parameters. Ultimately, the risk that meaningful selection bias will remain after adjustment (such as weighting) is a function of the amount of information a researcher has about (1) the selection process that determines which units are included in the sample and (2) the auxiliary variables that are related both to inclusion in the sample and to study outcomes. Probability-based designs inherently afford more information about the selection process; and as such the risk of meaningful selection bias can be expected to be lower in a sample obtained from a probability panel. With nonprobability samples, uncertainty about the selection process is compounded by the lack of even the “baseline” knowledge about the frame from which sample members were recruited and the true (or even approximate) probability with which each was invited to participate.

In order to mitigate some risk and increase the likelihood that the adjustment model will meaningfully reduce selection bias, researchers can include potentially relevant auxiliary variables on the questionnaire, particularly those that go beyond traditional demographics and are expected to be correlated with key study outcomes. Since it is not possible to adjust on auxiliary variables that are not collected or appended, and it is not useful to adjust on those that are unrelated to study outcomes, the selection of effective auxiliary variables for bias analysis and adjustment requires both advance planning and an understanding of the survey topic.

Oftentimes, researchers cannot know *a priori* who is missed or omitted in the selection, recruitment, or empanelment for online samples, nor can they know all the auxiliary variables relating to this selection as well as the key survey variables. This report, therefore, also aims to provide specific statistical methods researcher can use to assess the reliability of the collected data and, importantly, transparently report about the limitations of any survey collection.

For over twenty years, AAPOR has emphasized the importance of calculating and reporting response rates for all surveys. As the industry has evolved, however, it has become increasingly clear that response rates alone are not necessarily indicative of whether or how much bias exists in a survey. While research has shown definitively that *low* response rates *do not* automatically lead to nonresponse bias (Dutwin and Buskirk, 2021; Groves, 2006; Groves and Peytcheva, 2008), *high* response rates *do*, by definition, reduce the risk of nonresponse bias in probability-sample surveys (Peytchev, 2013). However, the added layers of complexity inherent in a probability-based panel, such as continuous or repeated recruitment, panel maintenance and attrition, and varying methods for sampling into specific studies, make the standard two-stage response rate computation outlined in the AAPOR *Standard Definitions* (2016) an important but imperfect metric. Additionally, there is no meaningful parallel metric for nonprobability online samples.

This report, therefore, recommends that researchers report additional measures of both precision and representativeness in order to present a comprehensive picture of the inferential capacity of the collected data. Specifically measures of precision (like point estimates) are based on assumptions. If these assumptions are violated, these measures may substantially overestimate the precision of a statistical estimate and certain features of online samples can make it particularly difficult or impossible to apply these methods in a way that makes their underlying assumptions plausible. Therefore, regardless of the methods used to produce these measures, measures of precision reported for online samples (standard errors, confidence intervals, margins of error, statistical significance tests, etc.) are likely to be over-optimistic, and hence should be interpreted with caution. This applies particularly, but not exclusively, to nonprobability samples.

Furthermore, measures of uncertainty capture only the random component of error in an estimate; they do not capture any systematic biases in the estimate. Even if the uncertainty in the estimate is correctly estimated, confidence interval coverage will be below the nominal level if the estimate itself is biased (Valliant 2020). Thus, for online samples—particularly nonprobability samples—measures of precision tell only part of the story. In addition to reporting measures of precision, some effort should be made to also assess and report the risk of bias.

While the variance of an estimate can typically be estimated from the sample, the bias of an estimate is not estimable without some external information on the population. In the case of probability samples with nonresponse, this information takes the form of variables measured for respondents and nonrespondents, or for the whole population. Many measures that were originally defined to assess the impact of nonresponse in probability samples, such as *R-indicators* and *H-indicators*, can be adapted to measure selection bias in nonprobability samples. Other newer measures, such as the *SMUB* and *SMAB*, are available that can help assess the sensitivity of an estimate to the assumptions inherent in the selection models applied. These metrics can and should become more broadly used in the context of surveys that employ online samples.

Finally, in light of the complexity of these issues, researchers are encouraged to consider multiple factors when evaluating which panel, either a probability-based or nonprobability-based sample, is the most appropriate method for their research project. The questions provided in Section 5 of this report are intended to help both researchers and practitioners think through the key issues prior to decision on a study design and also provide possible questions for researchers to ask of the panel providers as they consider the options available.

One of the core tenets of the AAPOR community is the importance of methodological disclosure and transparency. Previously, many aspects of both probability and nonprobability-based panels were seen as proprietary and researchers were left with questions about sampling, recruitment, and adjustments. The goal of the last section of this report is to provide researchers and practitioners with the tools to demystify these types of samples, and in return, legitimize them as sample sources. This methodology is becoming a cornerstone of the industry and it is incumbent upon the AAPOR community to figure out how and when to use it.

AAPOR Online Task Force References

Amaya, A., Hatley, N., and Lau, A. (2021). Measuring the Risks of Panel Conditioning in Survey Research. *Pew Research Center*. https://www.pewresearch.org/methods/2021/06/09/measuring-the-risks-of-panel-conditioning-in-survey-research/

Amaya, A., Zimmer, S., Morton, K., & Harter, R. (2021). Does undercoverage on the US address-based sampling frame translate to coverage bias? *Sociological Methods & Research*, 50, 812-836.

Andridge, R.R. and Little, R.J.A. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, 27 (2), 153 – 180.

Andridge, R. R., West, B. T., Little, R. J., Boonstra, P. S., & Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68, 1465-1483.

The American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* 9th edition. AAPOR

Ansolabehere, S., & Rivers, D. (2013). Cooperative Survey Research. *Annual Review of Political Science*, 16(1), 307–329.

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

Bach, R., and Eckman, S., 2018. Motivated Misreporting in Web Panels. *Journal of Survey Statistics and Methodology*, 6 (3) 418–430.

Bailar, B.A. (1989), Information Needs, Surveys, and Measurement Errors. In: Kasprzyk, Duncan, Kalton Singh, pp. 1–24.

Baumgardner S. K., Griffin D. H., and Raglin D. A. (2014). The Effects of Adding an Internet Response Option to the American Community Survey. American Community Survey Research and Evaluation Report, U.S. Census Bureau, available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Baumgardner_04.pdf. Accessed August 28

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78 (2), 161 – 188.

Bethlehem, J., Cobben, F., and Schouten, B. (2009). Indicators for the Representativeness of Survey Response. Statistics Canada's International Symposium Series: Proceedings. Available at: https://www150.statcan.gc.ca/n1/pub/11-522-x/2008000/article/10976-eng.pdf. Accessed August 28, 2022.

Biemer, P.P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74 (5) 817–848.

Biemer, P. and Peytchev, A. (2011). A Standardized Indicator of Survey Nonresponse Bias Based on Effect Size. Paper presented at the International Workshop on Household Survey Nonresponse, Bilbao, Spain.

Blumberg, S.J. and Luke, J.V. (2022). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July – December 2021. National Health Interview Survey Early Release

Program. Available at: https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless202205.pdf. Accessed August 28, 2022.

Blumberg, S.J. and Luke, J.V. (2007). Coverage Bias in Traditional Telephone Surveys of Low-Income and Young Adults. *Public Opinion Quarterly*, 71 (5): 734 – 749.Bogen, K. (1996). The Effect of Questionnaire Length on Response Rates – A Review of the Literature. Proceedings of the American Statistical Association Survey Research Methods Section. Available at: http://www.asasrms.org/Proceedings/papers/1996_177.pdf. Accessed August 28, 2022.

Boonstra, P. S., Little, R. J., West, B. T., Andridge, R. R., & Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of official statistics*, 37, 751-769.

Boyle, J.M., Fakhouri, T.H., Freedner-Maguire, N., and Iachan, R. (2017). Characteristics of the Population of Internet Panel Members. *Survey Practice*, 10 (4).

Breidt, F. J., and Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques." *Statistical Science*, *32*(2), 190–205.

Bretschi, D., Schaurer, I., & DIllman, D., 2021. An Experimental Comparison of Three Strategies for Converting Mail Respondents in a Probability-Based Mixed-Mode Panel to Internet Respondents. *Journal of Survey Statistics and Methodology.*

Brick, J.M. (2015). Compositional Model Inference. Proceedings of the American Statistical Association Survey Research Methods Section. Available at: http://www.asasrms.org/Proceedings/y2015/files/233896.pdf. Accessed August 28, 2022.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329–353.

Brick, J. M., and Montaquila, J. M. (2009). Nonresponse and Weighting. In *Sample Surveys: Design, Methods and Applications* (Vol. 29A). Elsevier Inc.

Brick, J.M., Kennedy, C., Cervantes-Flores, I., and Mercer, A.W. (2022). An Adaptive Mode Adjustment for Multimode Household Surveys. *Journal of Survey Statistics and Methodology*, 10 (4): 1024 – 1047.

Callegaro, M., Baker, R., Bethlehem, J., Goritz, A., Krosnick, J., Lavrakas, P., eds., *Online Panel Research: A Data Quality Perspective (1st ed)* (West Sussex, United Kingdom, Wiley, 2014).

Callegaro, M., and DiSogra, C. (2008), Computing Response Metrics for Online Panels. *Public Opinion Quarterly*, 72 (5): 1008-1032.

Clinton, J. (2001) Panel bias from attrition and conditioning: A case study of the Knowledge Networks panel. Unpublished Manuscript, Stanford University. Available at: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=6umiavkAAAAJ&cstart=20&pagesize=80&citation_for_view=6umiavkAAAAJ:W7OEmFMy1HYC

Clinton, J., Agiesta, J., Brenan, M., Connelly, M., Edwards-Levy, A., Fraga, B., Guskin, E., Hillygus, D. S., Jackson, C., Jones, J., Keeter, S., Khanna, K., Lapinski, J., Saad, L., Shaw, D., Smith, A., Wilson, D., & Wlezien, C. (2021). *2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls*. American Association for Public Opinion Research.

https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR-Task-Force-on-2020-Pre-Election-Polling_Report-FNL.pdf

Clinton, J. D., Eubank, N., Fresh, A., & Shepherd, M. E. (2021). Polling place changes and political participation: evidence from North Carolina presidential elections, 2008–2016. *Political Science Research and Methods*, 9, 800-817.

Cochran, W. G. (1953). *Sampling Techniques*. John Wiley & Sons.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., ... & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1), 4-36.

Couper, M. P. (2000). Review: Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64(4), 464–494.

de Leeuw, E. D. (2013). Thirty Years of Survey Methodology / Thirty Years of BMS. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 120(1), 47–59.

Dever, 2018. Combining Probability and Nonprobability Samples to Form Efficient Hybrid Estimates: An Evaluation of the Common Support Assumption. Proceedings of the 2018 Federal Committee on Statistical Methodology Research Conference. Available at: https://copafs.org/wp-content/uploads/2020/05/COPAFS-A4_Dever_2018FCSM.pdf. Accessed August 28, 2022.

Dever, J. a, Rafferty, A., & Valliant, R. (2008). Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? *Survey Research Methods*, *2*(2), 47–60.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87, 376-382.

Dutwin, D. and Buskirk, T.D. (2021). Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates. *Journal of Survey Statistics and Methodology*, 9(3), 353–380.

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 (S1): 213 – 239.

Elliott, M. R., and Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2), 249–264.

English, N., O'Muircheartaigh, C., Dekker, K., Latterner, M., and Eckman, S. (2009). Coverage Rates and Coverage Bias in Housing Unit Frames. Proceedings of the American Statistical Association Survey Research Methods Section. Available at: http://www.asasrms.org/Proceedings/y2009/Files/303284.pdf. Accessed August 28, 2022.

Fahimi, M., Barlas, F. M., Thomas, R. K., & Buttermore, N. (2015). Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice*, *8*(6).

Fricker, S. and Tourangeau, R. (2010). Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly*, 74 (5): 934 – 955.

Galesic, M. and Bosnjak, M. (2009). Effect of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, 73 (2): 349 – 360.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, *23*(2), 127–135.

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3), 762–776.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3), 413–419.

Greszki, R. Meyer, M., and Schoen, H. (2014). The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. In Callegaro, M., Baker, R., Bethlehem, J., Goritz, A.S., Krosnick, J.A., and Lavrakas, P.J. (eds.), *Online Panel Research: A Data Quality Perspective*. John Wiley & Sons.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646-675.

Groves, R. M., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, 72, 167-189.

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., ... & Zukerberg, A. L. (2016). Address-based sampling. *Prepared for AAPOR Council by the Task Force on Address-based sampling, Operating Under the Auspices of the AAPOR Standards Committee. Oakbrook Terrace, Il.*

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. *Behavior Research Methods*, 47(3), 685–690.

Hillygus, D. S., Jackson, N. and Young, K. (2014). Professional respondents in nonprobability online panels. In Callegaro, M., Baker, R., Bethlehem, J., Goritz, A.S., Krosnick, J.A., and Lavrakas, P.J. (eds.), *Online Panel Research: A Data Quality Perspective*. John Wiley & Sons.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189.

Iannacchione, V. 2011. The Changing Role of Address-Based Sampling in Survey Research. *Public Opinion Quarterly,* 75(3), 556-575.

Iannacchione, V. G., Staab, J. M., & Redden, D. T. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. *Public Opinion Quarterly*, 67, 202-210.

Jackson, M.T., Medway, R.L., and Megra, M.W. (2021). Can Appended Auxiliary Data Be Used to Tailor the Offered Response Mode in Cross-Sectional Studies? Evidence From An Address-Based Sample. *Journal of Survey Statistics and Methodology* (advance articles).

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4), 523–539.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., Asare-Marfo, D. (2021). Strategies for Detecting Insincere Respondents in Online Polling. *Public Opinion Quarterly,* 85(4), 1050-1075.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., Mcgeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, E., & Wlezien, C. (2017). *An Evaluation of 2016 Election Polls in the United States*. American Association for Public Opinion Research.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., Mcgeeney, K., & Gimenez, A. (2016). Evaluating Online Nonprobability Surveys. *Pew Research Center*. http://www.pewresearch.org/files/2016/04/Nonprobability-report-May-2016-FINAL.pdf

Kennedy, R., Wojcik, S., & Lazer, D. (2017). Improving election prediction internationally. *Science*, *355*, 515-520.

Kish, L. (1965). *Survey sampling*. Wiley.

Kish, L. (1969). Cumulating/Combining population surveys. *Survey Methodology*, 25(2), 129-138.

Kreuter, F., and Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research*, *40*(2), 311–332.

Lavrakas, P.J. (2008) Encyclopedia of survey research methods. Sage Publications, Inc., Thousand Oaks.

Lavrakas, P., Benson, G., Blumberg, S., Buskirk, T., Cervantes, I. F., Christian, L., ... & Shuttles, C. (2017). The future of US general population telephone survey research. *AAPOR Report*.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22, 329.

Lee, S., and Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, *37*(3), 319–343.

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, *72*(1), 6–27.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics*, *28*(3), 309–334.

Little, R., & An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949–968.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley and Sons.

Little, R. J., and Vartivarian, S. L. (2005). Does weighting for nonresponse increase the variance of survey means?” *Survey Methodology*, *31*(2), 4–11.

Little, R. J., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of survey statistics and methodology*, 8, 932-964.

Lugtig, P. Das, M., and Scherpenzeel, A. (2014). Nonresponse and Attrition in a Probability-Based Online Panel for the General Population. In Callegaro, M., Baker, R., Bethlehem, J., Goritz, A.S., Krosnick, J.A., and Lavrakas, P.J. (eds.), *Online Panel Research: A Data Quality Perspective*. John Wiley & Sons.

MacInnis, B., Krosnick, J. A., S. Ho, A., & Cho, M.-J. (2018). The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension. *Public Opinion Quarterly*, 82(4), 707-744.

Mathiowetz, N.A., & Lair, T.J. (1994). Getting Better? Change or Error in the Measurement of Functional Limitations*. Journal of Economic and Social Measurement*, 20, 237-262.

Mercer, A., Lau, A., & Kennedy, C. (2018). For Weighting Online Opt-In Samples, What Matters Most? *Pew Research Center*. http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/

Mercer, A. W. (2018). *Selection Bias in Nonprobability Surveys: A Causal Inference Approach* [Doctoral Dissertation, University of Maryland]. http://drum.lib.umd.edu/handle/1903/20943

Messer, B.L. and Dillman, D.A. (2011). Surveying the General Public Over the Internet Using Address-Based Sampling and Mail Contact Procedures. *Public Opinion Quarterly*, 75 (3): 429 – 457.

Millar, M.M., O'Neill, A.C., and Dillman, D.A. (2009). Are Mode Preferences Real? Technical Report 09-003 of the Social and Economic Sciences Research Center, Washington State University.

Neter, J., & Waksberg, J. (1964). A Study of Response Errors in Expenditures Data from Household Interviews. *Journal of the American Statistical Association*, 59(305), 18–55.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625.

Nichols, E., Horwitz, R., and Tancreto, J.G. (2015). An Examination of Self-Response for Hard-to-Interview Groups When Offered an Internet Reporting Option for the American Community Survey. American Community Survey Research and Evaluation Report, U.S. Census Bureau, available at: https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Nichols_01.pdf. Accessed August 28, 2022.

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *International Statistical Review*, 84(1), 43-62.

Olson, K., Smyth, J.D., and Wood, H.M. (2012). Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Evaluation. *Public Opinion Quarterly*, 76 (4), 611 – 635.

Ornstein, J. T. (2020). Stacked Regression and Poststratification. *Political Analysis*, 28(2), 293–301.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, *12*(4), 375–385.

Peytchev, A. (2013). Consequences of Survey Nonresponse. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 88–111.

Peytcheva, E. and Groves, R. M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, 25, 193.

Rao, K., Kaminska, O., McCutcheon, A. (2010). Recruiting Probability Samples for a Multi-Mode Research Panel with Internet and Mail Components. *Public Opinion Quarterly*, 74, 68-84.

Rivers, D. (2007). Sampling for web surveys. *JSM Proceedings (Survey Research Methods Section)*.

Rivers, D. and Bailey, D. (2009). Inference from matched samples in the 2008 US national elections. *Presented at the 2009 American Association for Public Opinion Research Annual Conference, Hollywood, Florida*.

Robbins, M., Ghosh-Dastidar, B., Ramchand, R. (2021). Blending Probability and Nonprobability Samples with Applications to a Survey of Military Caregivers. *Journal of Survey Statistics and Methodology*, 9, 1114–1145.

Royall, R. M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. *Biometrika*, *57*(2), 377–387.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons.

Sakshaug,J.,Wiśniowski,A.,Ruiz,D. & Blom,A.(2019).Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. *Journal of Official Statistics*, 35(3) 653-681.

Saris W.E. (1998). Ten years of interviewing without interviewers: The telepanel. In Couper M.P., R.P.Baker, J.Bethlehem, C.Clark, J.Martin, W.L.Nicholls II, J.M. O'reilly (Eds.) *Computer assisted survey information collection*. New York Wiley, 409-431.

Saris, W. E., & de Pijper, W. M. (1986). Computer assisted interviewing using home computers. *European Research*, 14, 144–152.

Särndal, C. and Lundstrom, S. (2010). Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias. *Survey Methodology*, 36 (2): 131 – 144.

Scherpenzeel, A., and Toepoel, V. (2012). Recruiting a Probability Sample for an Online Panel: Effects of Contact Mode, Incentives, and Information. *Public Opinion Quarterly,* 76, 470-490.

Shields, J., and To, N., Learning To Say No: Conditioned Underreporting in an Expenditure Survey. In Proceedings of the Survey Research Methods Section of the American Statistical Association, 2005. http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000432.pdf, accessed 2020-11-30.

Shook-Sa, B. E., Currivan, D. B., McMichael, J. P., & Iannacchione, V. G. (2013). Extending the coverage of address-based sampling frames: beyond the USPS computerized delivery sequence file. *Public Opinion Quarterly*, 77, 994-1005.

Si, Y., Pillai, N.S., Gelman, A., (2015). Bayesian Nonparametric Weighted Sampling Inference, *Bayesian Analysis*, 10(3), 605-625.

Smyth, J.D., Dillman, D.A., Christian, L.M., and O'Neill, A. (2010). Using the Internet to Survey Small Towns and Communities: Limitations and Possibilities in the Early 21st Century. *American Behavioral Scientist*, 53 (9): 1423 – 1448.

Smyth, J.D., Olson, K., and Millar, M.M. (2014). Identifying Predictors of Survey Mode Preference. *Social Science Research*, 48: 135 – 144.

Special Issue: Recent Advances in Probability-Based and Nonprobability Survey Research. February 2020. Journal of Survey Statistics and Methodology, 8.

Sturgis, P., Allum, N., and Brunton-Smith, I., (2009). Attitudes Over Time: The Psychology of Panel Conditioning. In Methodology of Longitudinal Surveys (eds R.M. Groves, G. Kalton, J.N.K. Rao, N. Schwarz, C. Skinner and P. Lynn). pp.113 - 126.

Toepoel, V., Das, M., and van Soest, A.H.O., (2008). Design Effects in Web Surveys: Comparing Trained and Fresh Respondents. CentER Discussion Paper Series No. 2008-51, Available at SSRN: https://ssrn.com/abstract=1140603 or http://dx.doi.org/10.2139/ssrn.1140603

Trejo, Yazmín García, Meyers, Mikelyn, Martinez, Mandi, O'Brien, Angela, Goerman, Patricia and Class, Betsarí Otero., (2022). Identifying Data Quality Challenges in Online Opt-In Panels Using Cognitive Interviews in English and Spanish. *Journal of Official Statistics*, 38(3): 793-822.

Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231–263.

Valliant, R., & Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, *40*(1), 105–137.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons.

Valliant, R., Dever, J.A., and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer.

van der Zouwen, J. and van Tilburg, T., (2001). Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses, *Sociological Methods & Research*, 30(1), 35-56.

Vercruyssen, A. Roose, H., Carton, A. and Van De Putte, B. (2014). The Effect of Busyness on Survey Participation: Being Too Busy or Feeling Too Busy to Cooperate? *International Journal of Survey Research Methodology*, 17 (4): 357 – 371.

Walker, R., Pettit, R., & Rubinson, J. (2009). The Foundations of Quality Initiative: A Five-Part Immersion into the Quality of Online Research. *Journal of Advertising Research, 49*(4), 464-485.

Wang, K., Cantor, D., & Safir, A. (2000). Panel conditioning in a random digit dial survey. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 822-827). Alexandria, VA: American Statistical Association.

Warren, J. R., and Halpern-Manners, A. (2012). Panel Conditioning in Longitudinal Social Science Surveys. *Sociological Methods & Research*, 41(4), 491–534

Waterton J., and Lievesley, D. (1989). “Evidence of Conditioning Effects in the British Social Attitudes Panel Survey.” Pp. 319-39 in *Panel Surveys*, edited by Kasprzyk D., Duncan G. J., Kalton G., Singh M. P. New York: Wiley.

West, B.T., Wagner, J., Hubbard, F., and Gu, H. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, 3 (2), 240 – 264.

West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit, A., & Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

Wiśniowski,A., Sakshaug, J., Ruiz, D.A.P., Blom, A. 2020. Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology*, 8, 120–147.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. Springer.Yammarino, F.J., Skinner, S.J., and Childers, T.L. (1991). Understanding Mail Survey Response Behavior: A Meta-Analysis. *Public Opinion Quarterly*, 55 (4): 613 – 639.

Yan, T., and S. Eckman. (2012). Panel Conditioning: Change in True Value versus Change in Self-Report. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, July 28–August 2, 2012, San Diego, California U.S.A. Available at: http://www.asasrms.org/Proceedings/y2012/Files/306203_76099.pdf

Yeager, D. S., Krosnick, J. a., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747.

Zhang, G., & Little, R. (2009). Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. *Biometrics*, 65(3), 911–918.

Appendix A: Probability-based Panels Contributing Information

The task force requested information from probability-based panels in the United States. Individual panels were contacted to complete questions (see Appendix B) and a call for responses was also shared on AAPORnet. All major probability-based panels (at the time of this writing) responded and were forthcoming in their responses. The following panels provided answers to questions:

National Panels (in alpha order by organization):

- Gallup (The Gallup Panel)
- IPSOS (KnowledgePanel)
- NORC (AmeriSpeak)
- Pew Research Center (The American Trends Panel)
- SSRS (SSRS Opinion Panel)
- University of Southern California - Center for Economic and Social Research (CESR) (Understanding America Study (UAS))

Regional/Local Panels:

- New York City Department of Health and Mental Hygiene (Healthy NYC)
- UNL Bureau of Sociological Research (BOSR) (NebrASKa Voices Panel)

Appendix B: Questions asked of U.S. Probability-based Panels

The task force requested the following information from all the panels included in the list above:

1. When was the panel first recruited?
2. How many active members are part of the panel?
3. What population is covered by your panel?
4. Please describe the recruitment procedures, including sample frame(s) and mode(s). (I.e. RDD, ABS, mixed-frame/mode, etc).
5. Have any of your recruitment methods changed over time? If so, please briefly describe.
6. What is a typical response rate at recruitment?
7. Does the panel recruit households or does it recruit individuals within household?
8. Does the panel cover non-English speakers? If yes, please describe.
9. Is there a validation process to confirm new panel members? If yes, please describe.
10. Does the panel calculate selection probabilities? If yes, please briefly describe.
11. Does the panel cover the offline population? If yes, please describe.
12. Does the panel monitor and calculate attrition rates? If yes, please share brief details about panel attrition rates or any groups that have higher than average attrition rates.
13. Does the panel use any special strategies to minimize attrition?
14. Do panel members receive monetary incentives? If yes, please describe.
15. Do panel members receive any other rewards or panel member benefits? If yes, please describe.
16. Please describe any refreshment sampling strategies. Does the panel recruit general population samples or are samples more targeted to certain demographics?
17. Does the panel conduct a census of panel members when fielding a survey, or does the panel more commonly sub-sample panel members? If a sample of panel members is drawn, please describe methods used. (e.g., random selection, stratified, quotas, PPS based on base weight, selection of a single household member).
18. When fielding a study using the panel, what range of response rates can typically be expected (knowing response rates can vary considerably by study design)? What response rate would be expected for a survey of adults age 18+, 15-minute survey, in the field for one week, with typical incentive?
19. Briefly describe demographic or psychographic data maintained on panel members.
20. Does the panel ever combine probability and nonprobability samples? If yes, please briefly describe how and when this strategy is used.
21. Please describe any weighting procedures that are commonly used (i.e. base weights/selection weights, nonresponse adjustments, poststratification, etc).
22. How (if at all) can the panel be leveraged by outside organizations/researchers? For example, does the panel sell space on the panel? Are results/datasets made public?