Public Opinion Quarterly, Vol. 81, Special Issue, 2017, pp. 250–279

THEORY AND PRACTICE IN NONPROBABILITY SURVEYS
PARALLELS BETWEEN CAUSAL INFERENCE AND SURVEY INFERENCE

ANDREW W. MERCER*
FRAUKE KREUTER
SCOTT KEETER
ELIZABETH A. STUART

Abstract Many in the survey research community have expressed concern at the growing popularity of nonprobability surveys. The absence of random selection prompts justified concerns about self-selection producing biased results and means that traditional, design-based estimation is inappropriate. This paper seeks to provide insight into the conditions under which nonprobability surveys can be expected to provide estimates free of selection bias. In fields such as epidemiology and economics that routinely work with observational data, researchers have identified the necessary conditions for unbiased estimation of causal effects when treatments are not assigned randomly. Similar conditions apply to survey estimates when respondents are not randomly selected. Drawing on this body of research, we propose a framework composed

Andrew W. Mercer is a senior research methodologist at the Pew Research Center, Washington, DC, USA, and a PhD candidate in the Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA. Frauke Kreuter is the director of the Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA, a professor of statistics and methodology at the University of Mannheim, Mannheim, Germany, and head of the Statistical Methods Research Department at IAB, Nuremberg, Germany. Scott Keeter is a senior survey advisor at the Pew Research Center, Washington, DC, USA. Elizabeth Stuart is a professor with appointments in the departments of Mental Health, Biostatistics, and Health Policy and Management at the Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. The authors would like to thank J. Michael Brick, Michael Elliott, Mario Callegaro, Peter Miller, and three anonymous reviewers for their comments and suggestions. Stuart's and Kreuter's time was supported by the National Institutes of Health [R01 MH099010-01A1 to E.A.S.]. Stuart's time was also supported by the Institute of Education Sciences [R305D150003 to E.A.S. and Robert Olsen]. This work was also supported by the Mannheim Center for European Social Research. *Address correspondence to Andrew Mercer, Pew Research Center, 1615 L Street NW, Suite 800, Washington, DC 20036, USA; e-mail: amercer@pewresearch.org.

doi:10.1093/poq/nfw060

of three elements that determine the level of selection bias in survey estimates. In this paper, we first provide a general overview of these components and demonstrate the link between causal inference and survey inference in the probability-based setting. Second, we give simplified examples to demonstrate how each of the components can contribute to bias in survey estimates. Finally, we review current practices in the area of nonprobability data collection and estimation, and specify how these methods relate to the elements identified here.

Introduction

The growing use of surveys that do not use traditional probability sampling has provoked both interest and concern from the survey community. Rising data-collection costs coupled with declining response rates have highlighted the appeal of lower-cost nonprobability surveys that can be fielded rapidly online. However, respondent self-selection into these surveys renders design-based methods of survey inference inapplicable, and raises concerns about the potential for biased results.

Selection bias refers to systematic differences between a statistical estimate and the true population parameter caused by problems with the composition of the sample (rather than errors in measurement). Traditionally, survey researchers think of selection bias as resulting from noncoverage—when the sampling frame omits portions of the target population—or nonresponse—when selected units do not complete the survey. These concepts are tied to a process of starting with a complete population and randomly selecting a subset. These categories may prove limiting when applied in a nonprobability context. Many nonprobability surveys do not originate from anything resembling a sampling frame. Even the idea of a sample as a finite set of units, some of which may fail to respond, does not apply to many nonprobability surveys. For nonprobability surveys, the processes that lead to a respondent being included in a sample are numerous, potentially arbitrary, and may not resemble the traditional probability-based survey process at all.

Rather than evaluate nonprobability surveys using concepts designed for a different inferential framework and different data-collection practices, we propose a more general framework that emphasizes the characteristics of the realized sample, regardless of how it was generated. The underpinnings of this framework are not new, but come from research into the estimation of causal effects from experimental and non-experimental data. In fields such as epidemiology, political science, and economics where randomized experiments are frequently not possible and observational studies are commonplace, research has focused on identifying the conditions under which valid statistical inferences about causal effects can be made using observational data. In the causal context, the parameter of interest is a contrast between experimental treatments, whereas surveys measure a broad range of estimates, including means, totals, correlations, and

other measures of association. Despite differences, the conditions that produce selection bias in causal analyses also apply in a survey context.

Others have noted similarities between causal inference and survey inference. Little and Rubin (2002) apply many of these same concepts to experiments, observational studies, survey nonresponse, and imputation. Groves (2006) uses a causal framework to describe when nonresponse will produce bias in survey estimates. Keiding and Louis (2016) reviewed the many objectives and challenges shared by both epidemiological studies and surveys, and suggest that both fields could benefit from sharing methodologies.

Drawing on this work, we identify three components that determine whether or not self-selection could lead to biased results:

- Exchangeability—Are all confounding variables known and measured for all sampled units?
- Positivity—Does the sample include all of the necessary kinds of units in the target population, or are certain groups with distinct characteristics missing?
- Composition—Does the sample distribution match the target population with respect to the confounding variables, or can it be adjusted to match?

In this paper, we first describe how this framework applies in the familiar context of randomized experiments and probability-based surveys before demonstrating how it extends to cover observational studies and nonprobability surveys. Second, we demonstrate the mechanics by which each component can produce bias in survey estimates by way of a simplified example. Finally, through the lens of this framework, we provide a critical review of current practices in online, nonprobability data collection and their implications for selection bias.

Randomization and Unbiased Inference in Experiments and Surveys

Questions about causal effects are usually framed in terms of potential outcomes or counterfactuals (Rubin 1974). A patient's outcome may be different if he is given Treatment A or Treatment B. Prior to choosing a treatment, either outcome is possible, but we observe only the results under the treatment that is actually provided to the patient. We can never observe what would have happened if a different treatment had been applied. The causal effect is the difference between the two potential outcomes. Although we can never observe both outcomes on a single individual, we can compare the average outcome for people who receive Treatment A to that of people who receive Treatment B to make inferences about which treatment is better. When treatments are assigned randomly, we can be reasonably confident that observed differences in the outcomes across treatment conditions are due to the treatments themselves and not to some other difference between the two groups.

When treatments are not assigned randomly, these assessments are more difficult. For instance, if patients who receive Treatment A tend to do worse, but Treatment A is usually given to sicker patients, it is difficult to know whether the difference is due to the treatment or to the fact that the patients who received it were in worse shape to begin with. The baseline level of sickness is known as a *confounder*. Confounders are variables associated with both the choice of treatment and the outcome of interest, and are the primary source of selection bias in causal analyses.

The parallels between causal inference and survey inference are substantial. A probability-based survey is essentially a randomized experiment where the pool of subjects is the set of units on the sampling frame and the treatment is selection into the survey. Unlike experiments where we observe outcomes on both treated and untreated subjects, in surveys we observe outcomes only on the selected units, with the expectation that there should be no difference between selected and non-selected units. The conditions under which causal effects can be estimated without selection bias are analogous to the conditions that produce unbiased estimates in surveys. Before discussing nonprobability surveys, we will first examine how these conditions are met in the context of randomized experiments and probability-based surveys.

STRONG IGNORABILITY—EXCHANGEABILITY AND POSITIVITY

Rosenbaum and Rubin (1983) devised the notion of strong ignorability to describe the conditions under which inferences about causal effects can be estimated without selection bias for a given sample. Strong ignorability consists of two requirements. The first, known as "exchangeability" (Greenland and Robins 1986, 2009), "ignorability," "no unobserved confounding," or "no hidden bias" (Rosenbaum 2002), requires the mechanism by which subjects are assigned a treatment to be independent of the measured outcome either unconditionally or conditional upon observed covariates. Unconditional exchangeability is analogous to the notion of data that is missing completely at random (MCAR), whereas conditional exchangeability corresponds to data missing at random (MAR) (Little and Rubin 2002). When unobserved confounders are present, it is not possible to isolate the effect of the treatment from the effect of the confounder without additional assumptions.

Second, it must be possible for any subject to have received any of the treatments. This requirement is called *positivity* because it requires that all subjects have a positive probability of receiving treatment. If certain types of subjects receive only treatment or control, it is not possible to learn about causal effects for those subjects, and the treatment and control groups will have systematic differences that cannot be resolved. In practice, we generally require not just a positive probability but also enough cases to produce sufficiently precise statistical estimates (Hernán and Robins 2006; Petersen et al. 2012).

In experiments, random treatment assignment guarantees that on average, the exchangeability and positivity conditions will be met. Randomization ensures exchangeability by preventing any relationship between treatment assignment and unobserved variables and ensures positivity because any subject has a chance of receiving any treatment. In probability-based surveys, random selection functions in much the same way. By randomly selecting a sample from the entire population, there can be no unobserved variables systematically associated with selection, and all members of the population have a chance of being included.

COMPOSITION

For experiments, the composition of treatment groups with respect to potential confounders is important in two respects. First, the distribution of potential confounders in the treatment group needs to match the distribution in the control group. Random treatment assignment guarantees that this will occur naturally on average, and this equivalence between treatment groups is implied whenever unconditional exchangeability holds. Second, the composition of the experimental sample affects the degree to which findings can be generalized to an external population.

Strong ignorability guarantees only that the results of an experiment are generalizable to the group of subjects included in an experiment; in other words, it ensures "internal validity" but does not necessarily imply "external validity" (Shadish, Cook, and Campbell 2002). It is rare for samples in randomized trials, which have historically prioritized internal validity, to match a larger population. Because of this, there has been a growing literature on methods to allow the generalization of experimental results to target populations, including reweighting strategies that aim to equate the experimental sample and the population with respect to observed characteristics (Cole and Stuart 2010; Kern et al. 2016; Stuart, Bradshaw, and Leaf 2015). Pearl and Bareinboim (2014) refer to the transportability of empirical findings from one sample to a separate target population. They note that generalization requires one to know the distribution of the outcome conditional upon treatment and any confounders, as well as the joint distribution of the confounding variables in the target population. Put simply, to generalize beyond the experimental sample to a target population, the sample needs to look like (or be made to look like) the target population with respect to the distribution of confounding variables.

The situation for surveys is somewhat less complex than for causal analyses. Whereas experiments must be concerned with the comparability of treatment and control as well as sample and population, surveys need be concerned only about sample and population. It is understood that the composition of a sample will match that of the population when *all* units have an equal probability of selection, implying unconditional exchangeability. When probabilities of

selection are unequal but known for every unit in the frame, the situation is equivalent to conditional exchangeability, and weighting observations by the inverse of the probability of selection yields unbiased population estimates (Horvitz and Thompson 1952). In either case, random selection ensures that on average the sample will match the target population on the distribution of any variables measured on the survey.

EXTENDING THE FRAMEWORK TO NON-RANDOM SAMPLES

For causal analyses and surveys, random treatment assignment and respondent selection provide a powerful mechanism for producing the conditions necessary for unbiased estimation of causal effects and population parameters. However, these conditions are guaranteed only when randomization is 100 percent successful. In practice, this is rarely the case. In experiments, subjects drop out of trials or are lost to follow-up. In surveys, the sampling frames may not perfectly cover the target population, and nonresponse means that some share of sampled units is never observed. When such problems occur, the usual response is to perform statistical adjustments to correct any imbalance. In experiments, methods such as matching or propensity score weighting can be used to adjust for imbalances between experimental treatment groups (see Imbens and Rubin [2015], part 6). In probability surveys, corrections involve nonresponse weighting adjustments for which a variety of techniques exist (see Kalton and Flores-Cervantes [2003]; Valliant, Dever, and Kreuter [2013]).

When we perform these adjustments to randomized experiments or probability surveys, we are no longer relying solely on randomization to produce unbiased estimates. Rather, these adjusted estimates are conditional upon a model that assumes that positivity and exchangeability hold and that the adjustment reconstructs the correct sample composition for the confounding covariates. Even if we perform no adjustment, we are implicitly assuming a model where the correlation between missingness and the outcome of interest is zero—unconditional exchangeability.

In the causal inference world, it is recognized that as long as exchangeability and positivity hold, it is possible to make unbiased inferences about causal effects from non-experimental data (Rubin 1974, 1978; Rosenbaum and Rubin 1983; Greenland and Robins 1986, 2009). Quasi-experimental designs such as regression discontinuity and instrumental variables models are techniques that can be used to identify causal effects from non-experimental data when the appropriate conditions are met (Angrist and Pischke 2009; West et al. 2008). Methods such as matching, marginal structural models, and structural nested models have been developed to estimate causal effects from observational data and have been proven to produce unbiased estimates when their underlying assumptions are met (Robins 1999a, 1999b; Cole et al. 2003; Stuart 2010). However, for all of these techniques, one can never be certain if the exchangeability and positivity conditions have been met. Therefore, the

bar for accepting results from non-experimental data is much higher than for randomized experiments.

The same is true for surveys that do not use probability sampling. When units are not randomly selected from the target population, researchers must rely on statistical models to generalize back to the target population. Probability-based surveys with undercoverage or nonresponse must also specify a model that relates the observed units to the unobserved (Valliant, Dorfman, and Royall 2000; Brick 2013). For probability samples, the initial design performs most of the work in ensuring exchangeability, positivity, and correct sample composition. Statistical models are employed during estimation to correct what are hopefully minor biases. In contrast, nonprobability samples cannot rely on randomization to help meet these requirements, and instead must rely on models at all stages of the survey process from sample selection to estimation. As in causal analyses, researchers can never know with certainty that these requirements have been met.

Mechanics of Selection Bias in Surveys

In this section, we focus specifically on the survey context and demonstrate through a simplified example the mechanics behind each of the components in this framework to show how they can introduce bias into survey estimates.

EXCHANGEABILITY

Suppose we have a sample intended for estimating what share of the population will vote for the Democratic and Republican candidates in an election, and that we have measured each respondent's candidate preference and age. Let us also assume that some feature of the recruitment process over-represents older people, but there are no additional unmeasured confounders. Because older people tend to vote Republican more than young people, an estimate of the overall vote using this sample would be biased in favor of the Republican candidate. However, because inclusion depends only on age, estimates of the vote within the younger and older subgroups would still be correct. In this case, the sampled individuals are exchangeable with non-sampled individuals within the same age group. When sampled observations are conditionally exchangeable, subgroups are internally unbiased with respect to the outcome of interest, even if some groups are over- or underrepresented relative to their share of the target population. Because there are no additional confounders, the overall proportion of the sample voting Democratic would be biased, but measures of the relationship between age and vote preference would be unbiased prior to any adjustment.

However, if inclusion in the sample depends on an unmeasured characteristic related to the survey outcome, the distribution of the outcome variable

within the observed subgroups will no longer match that of the target population. In our voting example, suppose our sample also over-represents respondents who live in big cities but, unlike age, this has not been measured. Because urban dwellers tend to vote Democratic, the Democratic vote share among both young and old respondents will be too high. In this case, young and old respondents in our sample are not exchangeable with their non-sampled counterparts because they are more urban, making urbanicity an unmeasured confounder. The bias in favor of the Democrat due to an excess of city dwellers could actually offset some of the Republican bias produced by having too many older respondents. In this scenario, the estimated vote for the full sample could be close to the true population value while subgroup estimates would be biased. Note that the crucial aspect of exchangeability is not which cases are included in the sample but what characteristics have been measured. If we knew which cases were urban and which were rural, we could adjust by both age and urbanicity to recover the correct sample composition.

In practice, the biases need not cancel out. The unobserved variable could have opposite effects for young and old respondents, or there could be different unobserved variables affecting different subgroups. Because the confounding variables are unobserved, it is impossible to know from the data alone whether or not the exchangeability requirement is met.

The associations that produce bias need not be direct. If we took this same sample but measured something such as eye color, which is not directly related to either age or urbanicity, we might still achieve biased estimates if eye color is associated with race. Over-representing urban respondents likely also means over-representing racial groups that live in urban areas, which could in turn affect the distribution of observed eye colors. The reverse is also true. Variables that are not confounders themselves but are closely correlated with confounders may help reduce bias by serving as proxies during adjustment.

POSITIVITY

The positivity requirement states that even if we know and have measured all potential confounders, all of the subgroups defined by confounding variables must also be represented in the sample (Hernán and Robins 2006). Groups that are underrepresented but present can be weighted up. However, it is not possible to weight up groups that were not surveyed. Returning to our example where inclusion depends on age and urbanicity, suppose that there are no older, urban respondents included in the sample. Even if we were able to record both age and urbanicity, there is no adjustment we can perform that will make up for the absence of older, urban respondents, although subgroup estimates for those groups that were observed would remain unbiased. On the other hand, if older and younger city dwellers are the same with respect to their voting preference, the absence of older urbanites would not introduce bias, because younger urbanites could stand in for them in the sample with

no change to the estimate. When a group is entirely missing from the set of observed units, the researcher requires a theoretical justification for believing that the missing group is not systematically different from other, superficially similar groups that were surveyed.

COMPOSITION

In our example, we have assumed that our sample composition does not match the target population on age and urbanicity. If it can be adjusted to match the distribution in the target population, our estimates of the vote will be unbiased. We have already alluded to the simplest approach, which is to weight each group to be proportional to its share of the target population.

Sample composition can be managed by design as well as through post hoc adjustment. Random selection yields the correct sample composition in expectation, though individual samples will not match exactly. If the confounders are known in advance, purposive methods such as quota sampling, where we predetermine the number of interviews required in each group, can be used to produce an exact sample match (Gittelman et al. 2015).

Managing sample composition through design or adjustment rather than random selection requires the researcher to be confident that all confounders are truly known and measured. When exchangeability or positivity does not hold, bias will not be eliminated and may even be magnified. In our example, if we adjust only for age but not urbanicity, we would eliminate the pro-Republican bias caused by an older sample but not the pro-Democratic bias due to an excess of urban respondents. The biases no longer offset each other, and the adjusted estimate would be *more* Democratic than it was before weighting.

Current Practices for Managing Bias in Online, Nonprobability Surveys

We can use this framework to consider current practices in fielding nonprobability web surveys and producing statistical estimates from the resulting samples. We distinguish between recruitment, whereby an individual becomes eligible for inclusion in one or more surveys (e.g., joining a panel), and sampling, the process by which an individual is selected for a particular survey after recruitment. After reviewing these two features of the data-collection process, we discuss alternative approaches to post-survey adjustment and estimation.

RECRUITMENT

The most common form of recruitment involves inviting individuals to join opt-in panels, which are lists maintained by sample providers of individuals

who have agreed to participate in surveys on an ongoing basis. Individuals can become empaneled in a variety of ways, such as directly through a panel website, clicking on banner advertisements, or when corporations grant panel vendors access to members of their customer loyalty programs. Panels provide an opportunity to collect a large amount of profile information on their members that can be used in both sampling and adjustment. Maintaining respondent profiles across many dimensions can aid in providing exchangeability only if the correct variables are measured. On the other hand, some fear that panel conditioning and attrition may mean that panel members may become less reflective of their non-empaneled counterparts over time, potentially reducing exchangeability (Couper 2000; Callegaro and DiSogra 2008; Callegaro, Manfreda, and Vehovar 2015).

The main alternative to panels is river sampling, in which potential respondents are recruited via similar sources but are directed to a one-off survey rather than asked to join a long-term panel (Callegaro, Baker, et al. 2014). River sampling avoids panel attrition and conditioning but provides no profile data on respondents in advance. Respondent characteristics must be obtained at the time of the survey, limiting the number of characteristics that can be measured. Some online survey providers have begun using a mixture of panel and river respondents (e.g., Lorch, Cavallaro, and van Ossenbruggen 2010; Young et al. 2012).

Both panels and river sampling face an immediate threat to the positivity requirement because individuals who do not use the Internet cannot participate. Studies conducted on the Pew Research Center's American Trends Panel and the Dutch LISS panel, two probability-based panels that take steps to cover individuals without Internet access, found that the exclusion on non-Internet individuals produced only small differences in most survey estimates. However, for outcomes pertaining to technology use, differences in estimates could be large. The Pew Research study also found that indicators of socioeconomic status differed considerably for some subgroups, such as the elderly or racial minorities (Keeter et al. 2015; Eckman 2016).

Obtaining a diverse array of potential respondents is crucial to the success of any recruitment method. Pettit (2015) demonstrated that respondents recruited via different websites can exhibit dramatically different demographic distributions. Respondents recruited from different sources likely vary on other characteristics as well; for instance, individuals recruited via a website dedicated to video games could differ from those recruited from websites devoted to personal finance with respect to variables such as interest in retirement planning or their use of leisure time. Recruiting from a diverse set of sources necessarily improves the probability of meeting the positivity requirement; however, it also increases the complexity of the recruitment process, potentially creating a trade-off between positivity and exchangeability. As the number of sources increases, it may become more difficult to know which characteristics distinguish between individuals recruited from different sources.

To date, the great majority of research into nonprobability surveys has relied on data from online panels. Many of these studies have compared different panels to one another and have found that while some nonprobability surveys compare favorably to probability-based surveys, the same survey fielded on different panels can result in dramatically different results (Yeager et al. 2011; Craig et al. 2013; Callegaro, Villar, et al. 2014; Erens et al. 2014; Schnorf et al. 2014; Kennedy et al. 2016). However, none of these studies were designed to evaluate alternative methods of panel recruitment or isolate the design features that produce such varying results.

Very little research has directly compared panels to river sampling. One such analysis found that after weighting for demographic characteristics, panel respondents were largely similar to river respondents, although panelists were more likely to be registered to vote and more likely to use Twitter. River respondents were closer to the chosen benchmark on both measures (Clark, Young, and Petrin 2015). A study performed as part of the Foundations of Quality 2 (FOQ2) initiative compared the demographic composition of surveys using panels and river sampling. It found that on average, river samples yielded demographic compositions similar to non-river samples, and required somewhat less extreme weighting when adjusted to match demographics not used in the sampling process (Bremer 2013). Unfortunately, there was no evaluation of differences in other non-demographic estimates.

At present, there is not enough research to recommend one recruitment method over the other. The availability of profile data on panels offers flexibility and control for the purposes of sampling and adjustment, but the limited empirical research discussed previously does suggest some possible advantages to river samples. Other practices, such as profiling, sampling, or quota design, may also be more important than the recruitment process.

SAMPLING

Nonprobability surveys generally rely on purposive selection to achieve the desired sample composition while data collection is ongoing. This is commonly achieved through quotas, where the researcher pre-specifies a particular distribution across one or more variables. Usually these are cells defined by a cross-classification of demographic characteristics such as gender by age, with each cell requiring a specified number of completed interviews in that category. The end result is a sample that matches the pre-specified distribution across the chosen variables. The use of quotas relies on the assumption that individuals that comprise each quota cell are exchangeable with non-sampled individuals who share those characteristics. If that assumption is met, the sample will have the correct composition on the confounding variables, allowing for the estimation means and proportions that generalize to the target population.

Most contemporary web surveys that employ quotas define the cells across no more than a handful of demographic variables. However, there is a growing

consensus that basic demographic variables such as age, sex, race, and education are insufficient for achieving exchangeability. A recent study using the FOQ2 data compared three progressively more stringent sets of demographic quotas. Across a range of benchmarks, the application of more stringent quotas did nothing to reduce bias, and post-survey weighting actually increased the average bias for all but five out of 17 sample providers. The study also evaluated three quota schemes that incorporated additional, non-demographic variables; however, their success was mixed. (The details of the methods employed were not specified, to avoid identifying the sample providers [Gittelman et al. 2015]). This finding is consistent with research in causal inference suggesting that demographics alone are generally insufficient for eliminating bias in observational studies (Cook, Shadish, and Wong 2008).

If traditional quota methods are insufficient for producing strong ignorability, sampling methods that allow researchers to control both more and different dimensions may improve the ability to condition on a more appropriate set of potential confounders. The best documented of these methods is implemented by YouGov on surveys conducted using its panel in the United States. YouGov first draws a random sample of anonymized cases from a high-quality data source, such as the American Community Survey (ACS) Public Use Microdata Sample, that is believed to reflect the true joint distribution for a large number of variables in the target population. This subsample is referred to as a synthetic sampling frame (SSF) and serves as a template for the eventual survey sample. Each panelist who completes the YouGov survey is matched to a case in the SSF with similar characteristics using a distance measure such as Euclidian distance. When every record in the SSF has been matched with a suitably similar respondent, the survey is complete (Rivers 2007).

Because a limited number of covariates are available in any single survey such as the ACS, it is possible to impute additional variables onto the SSF using models built with other data sources. This was the approach taken in the 2008 Cooperative Congressional Election Study, which augmented an SSF drawn from the ACS with estimates of voter registration and turnout from the Current Population Survey Voting and Registration Supplement, and of Internet use, religion, and interest in politics from Pew Research Center surveys. The resulting survey sample produced estimates of the presidential vote that closely matched national exit polls and the American National Election Studies (Ansolabehere and Rivers 2013).

This approach is appealing in its capacity to flexibly match the target population on a larger number of covariates than is possible with traditional quota methods. For this approach to succeed, the composition of matching variables in the SSF must accurately match the target population, and any models used to combine datasets must be correctly specified. More importantly, the matching variables must be the correct variables for ensuring conditional exchangeability, and the panel must be able to supply respondents that are close matches to each case in the SSF. If there are remaining confounders that are not accounted

for, resulting survey estimates will be biased. One side benefit of this approach is that problems with positivity should be immediately apparent if there are portions of the SSF for which no clear matching respondents can be found.

Another approach to sampling on a higher number of dimensions is the use of propensity score matching to construct quota cells. Under this approach, a probability survey that is assumed to accurately reflect the target population is fielded in parallel with a nonprobability survey. Using a set of common covariates collected on each survey, a propensity model is estimated by combining the two samples and predicting the probability that each respondent belongs to the probability survey. When subsequent online surveys are fielded, the propensity model is used to calculate a propensity score for each respondent as they are screened for the new survey. Quotas are not set on particular respondent characteristics, but are based on quintiles of the propensity score distribution (Terhanian and Bremer 2012).

As with the SSF used in sample matching, much hinges on how well the parallel reference survey matches the target population. If the reference survey suffers from its own nonresponse or coverage bias, those biases will be transferred into the nonprobability survey. On the other hand, the researcher could tailor the contents of the baseline surveys to include any variables believed to be necessary to ensure conditional exchangeability. Under other approaches, researchers are limited to covariates that are available from preexisting data sources. This method performed well in a simulation; however, the data used to construct the propensity model was the same data used to generate the simulated survey. The evaluation also generated only a single simulated dataset (Terhanian and Bremer 2012). As such, it is difficult to know how this technique performs on new samples and over repeated applications. Dividing the propensity score into quintiles will result in a loss of information contained in the full distribution of propensity scores, though it is also possible that quintiles provide a sufficient foundation of balance and positivity that can be further refined through post-survey adjustment. Additional research comparing this approach with the matching approach described above would be valuable, particularly if the same survey and set of covariates can be used.

Another, less understood component of the sampling process for many nonprobability surveys is the use of routers. Most nonprobability survey vendors have many surveys fielding simultaneously. When a router is employed, rather than drawing separate samples for each survey, respondents are invited to participate in an unspecified survey. The actual survey taken is determined dynamically based on the characteristics of the respondent and the needs of active surveys with respect to quotas or screening criteria. This makes for a more efficient use of sample, but means the sample for any one survey depends on what other surveys are in the field simultaneously. If there are ample respondents and few competing surveys, routers may pose little threat of bias. On the other hand, the presence of surveys focused on rare groups may

mean that individuals belonging to those groups are not routed to other surveys. In such an event, the routing process becomes a confounder that would be difficult to observe and account for.

The only empirical study evaluating routers compared the effects of three different routing methods against a non-routed control and found that all four conditions produced similar estimates. In a set of simulations, the authors did find that routing could produce bias for questions that are highly correlated with the selection criteria for other surveys in the field. This study evaluated routing under a narrow set of conditions that the authors recognize may not generalize to many circumstances observed in practice (Brigham, Fallig, and Miller 2014). Additional experiments and simulations testing alternative algorithms and scenarios, or observational studies comparing router performance over time for different vendors, would be of substantial benefit.

POST-SURVEY ADJUSTMENT

Because it may not be feasible to achieve the desired sample composition through sampling alone, post-survey adjustment is still needed. Most of the research on adjusting nonprobability samples has focused on adapting the methods used to perform non-response adjustment with probability samples. Calibration and propensity score weighting are the two most common approaches to weighting.

Calibration methods directly adjust the composition of the sample to match a known distribution of variables in the target population. The simplest form of calibration is poststratification, in which the sample is divided into mutually exclusive cells that are weighted up or down such that the proportion of each cell in the sample matches the corresponding proportion in the target population. Whereas poststratification requires knowledge of the joint distribution of the stratification variables in the target population, other calibration methods such as raking and generalized regression estimation require only knowledge of the marginal distribution of any adjustment variables (Deville and Sarndal 1992; Kalton and Flores-Cervantes 2003). Calibration methods generally require that the outcome be a linear function of the calibration variables, and may not perform well in the presence of nonlinear relationships between the outcome and adjustment variables or unmodeled interactions (Valliant, Dorfman, and Royall 2000).

Propensity score weighting involves combining a nonprobability sample with a parallel probability or gold-standard data source as a reference sample. A model predicting sample membership is fitted to these combined data, and observations in the nonprobability sample are weighted by the inverse of their probability of appearing in the nonprobability sample (Taylor 2000; Terhanian and Bremer 2000; Lee 2006; Valliant and Dever 2011). Valliant and Dever (2011) demonstrated that for propensity score adjustment to be effective, the propensity model must incorporate any nonresponse adjustment and

bias correction that has been applied to the reference sample. Otherwise, those biases will be transferred to the nonprobability sample.

Given the same set of covariates, generalized regression estimation (GREG) has been found to perform comparably to propensity score weighting, suggesting that a parallel reference survey may be unnecessary when the requisite population totals are available (Valliant and Dever 2011). Propensity score weighting can more easily accommodate nonlinear associations and interactions between confounding variables. If there are a large number of confounders or it is unknown which of the observed covariates are confounders, machine learning methods such as boosting or random forests can fit high dimensional propensity models if a suitable reference sample with common covariates is available (Buskirk and Kolenikov 2015; Lee, Lessler, and Stuart 2010).

Some have explored matching as an alternative to weighting for post-survey adjustment of nonprobability surveys. Traditionally, matching is used in causal inference in order to adjust for differences in composition between treatment groups (see Stuart [2010] for a review of their use in causal inference). With matching, the idea is to create groups containing one or more observations from both a reference sample and a nonprobability sample that are similar on a set of auxiliary variables believed to be associated with selection. Groups in the nonprobability sample are then weighted so that their distribution matches the distribution in the reference sample. For example, a reference sample might be divided into cells based on a set of covariates or a propensity score, while cases in the nonprobability sample in matching cells would be weighted so that the proportion in each cell matches the proportion in the reference sample. In this sense, matching is very similar to propensity score weighting or poststratification, with one important exception. In many applications, observations for which there is no acceptable match are removed from the final dataset. When this happens, information is lost, and inference is only possible for those portions of the samples that overlap. On the other hand, identifying a lack of overlap forces researchers to evaluate the validity of the positivity assumption in ways that other methods may not. Unlike standard weighting methods that will generally produce a weight for every observation (even if some are quite large), matching software often automatically identifies those observations in a reference sample for which no counterparts exist in the nonprobability sample (e.g., the MatchIt package for the R statistical software platform [Ho et al. 2011]). Buskirk and Dutwin (2015) found that raking to basic demographics was more effective at reducing bias than matching on a more extensive set of demographics; however, a two-stage process of matching followed by raking reduced bias more than raking alone.

A final approach to post-survey estimation is multilevel regression and poststratification (MRP). In traditional poststratification, a sample is divided up into mutually exclusive cells, each of which is weighted to be proportional to their representation in the target population. As the number of cells becomes large,

the number of observations in each cell becomes small and estimates become unstable. MRP enables poststratification using a large number of cells by fitting a multilevel model that pools information about cells sharing similar characteristics and allows for the estimation of cell means even when cells are sparse. A weighted mean is then constructed using the estimated cell means (Park, Gelman, and Bafumi 2004; Lax and Phillips 2009; Ghitza and Gelman 2013).

This approach performed well when used to predict 2012 presidential election results using a survey conducted via the Microsoft Xbox platform whose sample composition differed radically from the population of voters and for which unadjusted estimates were wildly inaccurate (Wang et al. 2014). Unweighted, the sample was 93 percent male, only 1 percent 65 years old or older, and showed Barack Obama losing badly to Mitt Romney. On the surface, it seems unlikely that such a survey could produce accurate estimates. However, the Xbox study enjoyed two benefits not available to many other studies. The first is a very large sample size (345,858 unique respondents), which means that even groups that are dramatically underrepresented in the sample in relative terms still have enough observations in absolute terms to avoid problems with positivity. The 1 percent of the sample 65 years old or older yields 3,400 observations—more than enough cases to produce stable estimates for that subgroup. The second is that the authors had a very powerful set of covariates, including party identification and ideology, making it much more likely that the exchangeability requirement was satisfied for the purpose of predicting partisan voting behavior.

Another study using only demographic covariates met with less success. It compared MRP-based estimates of presidential approval and country direction to estimates from the Pew Research Center's probability-based telephone surveys over the same time period. For the share of the population that thinks the country is on the right track, the MRP estimates were not different from the estimates obtained using a simple poststratification adjustment, and were lower than the telephone-based estimates. On the other hand, presidential approval changed dramatically, moving from an underestimate to an overestimate relative to the comparison telephone survey (Petrin and El-Dash 2015). Although the telephone survey benchmarks are themselves estimates and have their own biases, if the goal of adjustment was to match that particular benchmark, neither MRP nor traditional poststratification were successful.

Each of these approaches to estimation comes with advantages and disadvantages. When control totals are available for the confounders and their relationship with the survey outcome is linear, calibration methods are quite powerful and easy to apply. Propensity score methods provide a great deal of flexibility at the cost of requiring an auxiliary dataset with a shared set of covariates. It is less clear whether matching offers substantial benefits over propensity score weighting or calibration. For approaches that produce weights, there is some indication that methods applied in combination may offer an improvement over the use of a single method (Lee and Valliant 2009; Brick 2015;

Buskirk and Dutwin 2015). MRP may be most efficient at extracting information from smaller datasets, but at the cost of computational complexity and the fact that a separate model is required for each outcome variable. Additional research directly comparing adjustment methods to one another would be valuable in helping researchers choose the most appropriate tool.

All of these methods will fail if the exchangeability and positivity requirements are not met, or if the model specification does not correctly replicate the target composition on the confounding variables. If exchangeability and positivity are met, the best method is the one that can most closely mirror the correct sample composition using the available data and information. If exchangeability and positivity are not met, there is no a priori reason to believe that any of these methods will perform better than any other.

VARIABLE SELECTION

Given the centrality of exchangeability and positivity in achieving unbiased estimates from nonprobability surveys, what variables should practitioners measure and utilize in sampling and adjustment? A number of researchers have attempted to find sets of variables that can reliably serve to achieve at least partial exchangeability for a broad range of survey topics. These include so-called "webographics," early adopter characteristics, and other behavioral and attitudinal factors intended to differentiate between survey participants and the broader population (Schonlau et al. 2004; Schonlau, van Soest, and Kapteyn 2007; DiSogra et al. 2011; Fahimi et al. 2015). While such general-purpose variables may fill a need, their effect will be limited unless they are correlated with the outcome to be measured.

Researchers will be best served if they can identify a likely set of theoretically grounded confounders prior to data collection, and use these as the starting point for a research design. For example, in studies of US politics, many outcome variables of interest will be related to respondents' underlying political engagement and partisanship. These may be effective confounders to use in sampling and adjustment. In the absence of strong theory regarding the survey topic, achieving exchangeability will prove extremely challenging. Researchers must also be confident that the variables they have identified can account for any indirect confounding resulting from idiosyncrasies associated with recruitment or sampling. Although some vendors consider sampling practices proprietary, vendors must be fully transparent about any variables used in the selection process to ensure that researchers are aware of any potential for confounding.

Discussion

Whereas the emphasis in probability-based surveys has traditionally been to develop processes that minimize confounding, the emphasis suggested here is to first identify likely confounders and design the data collection and analysis

so that they are measured and actively accounted for. To be clear, this is more a shift in emphasis than a full-scale departure. Probability-based surveys generally seek to measure and account for specific characteristics that are associated with bias, and we have discussed how data-collection practices may introduce or mitigate confounding in nonprobability surveys.

Grounding this framework in causal inference suggests that there may be other techniques from that field that can be applied in a survey context. Testing the sensitivity of findings to unmeasured confounding is another common practice in causal inference whose adoption would likely benefit the survey field (Rosenbaum 2005). Unlike probability surveys where the maximum range of bias is bounded by the size of the nonresponding sample, selection bias is unbounded and non-identifiable in nonprobability surveys. Although some methods, such as pattern mixture models, have been developed to evaluate selection bias under such constraints, they not widely used in practice (Andridge and Little 2011). Other techniques that do not rely on assumptions about the probability of selection may also prove useful for nonprobability surveys (e.g., Robins, Rotnitzky, and Scharfstein 1999; Manski 2007). Additionally, the use of causal diagrams and other methods of identifying confounders represents another worthwhile area for future research (e.g., Pearl 2009; Steiner et al. 2010; Myers et al. 2011).

Finally, it is one thing to know in principle that exchangeability, positivity, and composition must be achieved in order to avoid selection bias in nonprobability survey estimates. It is another thing to achieve them successfully in practice. Even when the subject matter is well known and many likely confounders are identified, it may prove difficult to have complete confidence that there is not some yet unknown factor quietly introducing bias into survey estimates. Nevertheless, by making explicit a set of assumptions that to date have been largely implicit, the notions of exchangeability, positivity, and composition provide a framework by which to evaluate and critique specific research findings and improve methodological practice.

References

Andridge, Rebecca R., and Roderick J. A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27(2):153–80.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ansolabehere, Stephen, and Douglas Rivers. 2013. "Cooperative Survey Research." *Annual Review of Political Science* 16(1):307–29.

Bremer, John. 2013. "Research Quality: The Interaction of Sampling and Weighting in Producing a Representative Sample Online: An Excerpt from the ARF's 'Foundations of Quality 2' initiative." *Journal of Advertising Research* 53(4):363–71.

Brick, J. Michael. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29(3):329–53.

———. 2015. "Compositional Model Inference." In *JSM Proceedings (Survey Research Methods Section)*, 299–307. Alexandria, VA: American Statistical Association.

Brigham, Nancy, Michael Fallig, and Chuck Miller. 2014. "The Impact of Survey Routers on Sampling and Surveys: Unraveling the Mysteries of Survey-Router Design and Deployment." *Journal of Advertising Research* 54(4):381–88.

Buskirk, Trent D., and David J. Dutwin. 2015. "Selected or Self-Selected? Part 2: Exploring Non-Probability and Probability Samples from Response Propensities to Participant Profiles to Outcome Distributions." Paper presented at the Conference of the American Association for Public Opinion Research, Hollywood, FL, USA.

Buskirk, Trent D., and Stanislav Kolenikov. 2015. "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification." *Survey Methods: Insights from the Field, Weighting: Practical Issues, and "How To" Approach*. Available at http://surveyinsights.org/?p=5108.

Callegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas. 2014. "Online Panel Research: History, Concepts, Applications and a Look at the Future." In *Online Panel Research: A Data Quality Perspective*, edited by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas, 1–22. New York: John Wiley and Sons.

Callegaro, Mario, and Charles DiSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72(5):1008–32.

Callegaro, Mario, Katja Lozar Manfreda, and Vasja Vehovar. 2015. *Web Survey Methodology*. Los Angeles: Sage.

Callegaro, Mario, Ana Villar, David Yeager, and Jon A. Krosnick. 2014. "A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels Based on Probability and Nonprobability Samples." In *Online Panel Research: A Data Quality Perspective*, edited by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Krosnick, and Paul J. Lavrakas, 23–53. West Sussex: Wiley and Sons.

Clark, Julia, Clifford Young, and Robert Petrin. 2015. "Meta-Analysis of Online Panel and Non-Panel Sampling: Electoral and Non-Electoral Behavior Metrics." Paper presented at the Conference of the American Association for Public Opinion Research, Hollywood, FL, USA.

Cole, Stephen R., Miguel A. Hernán, James M. Robins, Kathryn Anastos, Joan Chmiel, Roger Detels, Carolyn Ervin, et al. 2003. "Effect of Highly Active Antiretroviral Therapy on Time to Acquired Immunodeficiency Syndrome or Death Using Marginal Structural Models." *American Journal of Epidemiology* 158(7):687–94.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial." *American Journal of Epidemiology* 172(1):107–15.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27(4):724–50.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64(4):464–94.

Craig, Benjamin M., Ron D. Hays, A. Simon Pickard, David Cella, Dennis A. Revicki, and Bryce B. Reeve. 2013. "Comparison of US Panel Vendors for Online Surveys." *Journal of Medical Internet Research* 15(11):e260.

Deville, Jean-Claude, and Carl-Erik Sarndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–82.

DiSogra, Charles, Curtiss Cobb, Elisa Chan, and J. Michael Dennis. 2011. "Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics." In *JSM Proceedings (Survey Methods Section)*, 4501–15. Alexandria, VA: American Statistical Association.

Eckman, Stephanie. 2016. "Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?" *Social Science Computer Review* 34(1):41–58.

Erens, Bob, Sarah Burkill, Mick P Couper, Frederick Conrad, Soazig Clifton, Clare Tanton, Andrew Phelps, et al. 2014. "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey." *Journal of Medical Internet Research* 16(12):e276.

Fahimi, Mansour, Frances M. Barlas, Randall K Thomas, and Nicole Buttermore. 2015. "Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse." *Survey Practice* 8(6). Available at http://www.surveypractice.org/index.php/SurveyPractice/article/view/324.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups." *American Journal of Political Science* 57(3):762–76.

Gittelman, Steven H., Randall K. Thomas, Paul J. Lavrakas, and Victor Lange. 2015. "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples." *Journal of Advertising Research* 55(4):368–79.

Greenland, Sander, and James M. Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology* 15(3):413–19.

———. 2009. "Identifiability, Exchangeability and Confounding Revisited." *Epidemiologic Perspectives & Innovations* 6(4). Available at http://epi-perspectives.biomedcentral.com/articles/10.1186/1742-5573-6-4.

Groves, Robert. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5):646–75.

Hernán, Miguel A., and James M. Robins. 2006. "Estimating Causal Effects from Epidemiological Data." *Journal of Epidemiology and Community Health* 60(7):578–86.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42(8):1–28.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47(260):663–85.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81–97.

Keeter, Scott, Kyley McGeeney, Andrew Mercer, Nick Hatley, Eileen Patten, and Andrew Perrin. 2015. *"Coverage Error in Internet Surveys: Who Web-Only Surveys Miss and How That Affects Results."* Pew Research Center. Available at http://www.pewresearch.org/files/2015/09/2015-09-22_coverage-error-in-internet-surveys.pdf.

Keiding, Niels, and Thomas A. Louis. 2016. "Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys." *Journal of the Royal Statistical Society, Series A: Statistics in Society* 179(2):319–76.

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

Kern, Holger L., Elizabeth A. Stuart, Jennifer Hill, and Donald P. Green. 2016. "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations." *Journal of Research on Educational Effectiveness* 9(1):103–27.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103(3):367–86.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine* 29(3):337–46.

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2):329–49.

Lee, Sunghee, and Richard Valliant. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research* 37(3):319–43.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: John Wiley and Sons.

Lorch, Jackie, Kristin Cavallaro, and Robert van Ossenbruggen. 2010. "Sample Blending: 1 + 1 > 2." *Survey Sampling International*. Available at https://www.surveysampling.com/site/assets/files/1584/sample-blending-1-1-2.pdf.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Myers, Jessica A., Jeremy A. Rassen, Joshua J. Gagne, Krista F. Huybrechts, Sebastian Schneeweiss, Kenneth J. Rothman, Marshall M. Joffe, and Robert J. Glynn. 2011. "Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates." *American Journal of Epidemiology* 174(11):1213–22.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–85.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability across Populations." *Statistical Science* 29(4):579–95.

Petersen, Maya L., Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. 2012. "Diagnosing and Responding to Violations in the Positivity Assumption." *Statistical Methods in Medical Research* 21(1):31–54.

Petrin, Robert A., and Neale El-Dash. 2015. "Reaching Wider, Going Deeper: Incorporating Sample Source Variation and Other Considerations into MRP Adjustments of Polling Estimates from Blended River Samples." Paper presented at the Conference of the American Association for Public Opinion Research, Hollywood, FL, USA.

Pettit, Annie. 2015. "Building a Quality Nonprobability Panel: Methods, Problems, and Being Innovative." Paper presented at the 2015 Conference of the American Association for Public Opinion Research, Hollywood, FL, USA.

Rivers, Douglas. 2007. "Sampling for Web Surveys." Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA.

Robins, James M. 1999a. "Association, Causation, and Marginal Structural Models." *Synthese: An International Journal for Epistemology, Methodology, and Philosophy of Science* 121(1–2):151–79.

———. 1999b. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, edited by M. Elizabeth Halloran and Donald Berry, 116:95–134. New York: Springer-Verlag.

Robins, James M., Andrea Rotnitzky, and Daniel O. Scharfstein. 1999. "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models." In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, edited by M. Elizabeth Halloran and Donald Berry, 1–92. New York: Springer-Verlag.

Rosenbaum, Paul R. 2002. *Observational Studies*. Springer Series in Statistics. New York: Springer.

———. 2005. "Sensitivity Analysis in Observational Studies." In *Encyclopedia of Statistics in Behavioral Science*, edited by Brian S Everitt and David C Howell, 4:1809–14. Chichester: John Wiley & Sons.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.

———. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1):34–58.

Schnorf, Sebastian, Aaron Sedley, Martin Ortlieb, and Allison Woodruff. 2014. *"A Comparison of Six Sample Providers Regarding Online Privacy Benchmarks."* Symposium on Usable Privacy and Security Workshop on Privacy Personas and Segmentation. Available at http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42558.pdf.

Schonlau, Matthias, Arthur van Soest, and Arie Kapteyn. 2007. "Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?" *Survey Research Methods* 1(3):155–63.

Schonlau, Matthias, Kinga Zapert, Lisa P. Simon, Katherine Haynes Sanstad, Sue M. Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra H. Berry. 2004. "A Comparison Between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review* 22(1):128–38.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont: Wadsworth Cengage Learning.

Steiner, Peter M., Thomas D. Cook, William R. Shadish, and M. H. Clark. 2010. "The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies." *Psychological Methods* 15(3):250–67.

Stuart, Elizabeth A. 2010. "The Use of Propensity Scores to Assess Generalizability." *Journal of the Royal Statistical Society, Series A: Statistics in Society* 174(2):369–86.

Stuart, Elizabeth A., Catherine P. Bradshaw, and Philip J. Leaf. 2015. "Assessing the Generalizability of Randomized Trial Results to Target Populations." *Prevention Science* 16(3):475–85.

Taylor, Humphrey. 2000. "Does Internet Research Work?" *International Journal of Market Research* 42(1):51–63.

Terhanian, George, and John Bremer. 2000. "Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research." Rochester, NY: Harris Interactive.

———. 2012. "A Smarter Way to Select Respondents for Surveys?" *International Journal of Market Research* 54(6):751–80.

Valliant, Richard, and Jill A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociological Methods & Research* 40(1):105–37.

Valliant, Richard, Jill A. Dever, and Frauke Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, Richard, Alan H. Dorfman, and Richard M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31(3):980–91.

West, Stephen G., Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, José Szapocznik, et al. 2008. "Alternatives to the Randomized Controlled Trial." *American Journal of Public Health* 98(8):1359–66.

Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." *Public Opinion Quarterly* 75(4):709–47.

Young, Clifford, John Vidmar, Julia Clark, and Neale El-Dash. 2012. "Our Brave New World: Blended Online Samples and the Performance of Nonprobablity Approaches." *Ipsos Public Affairs*. Available at http://www.ipsos-na.com/knowledge-ideas/public-affairs/points-of-view/?q=brave-new-world.