Statistical Science
2017, Vol. 32, No. 2, 249–264
DOI: 10.1214/16-STS598

Inference for Nonprobability Samples

Michael R. Elliott and Richard Valliant

Abstract. Although selecting a probability sample has been the standard for decades when making inferences from a sample to a finite population, incentives are increasing to use nonprobability samples. In a world of "big data", large amounts of data are available that are faster and easier to collect than are probability samples. Design-based inference, in which the distribution for inference is generated by the random mechanism used by the sampler, cannot be used for nonprobability samples. One alternative is quasi-randomization in which pseudo-inclusion probabilities are estimated based on covariates available for samples and nonsample units. Another is superpopulation modeling for the analytic variables collected on the sample units in which the model is used to predict values for the nonsample units. We discuss the pros and cons of each approach.

Key words and phrases: Coverage error, hierarchical regression, quasi-randomization, reference sample, selection bias, superpopulation model.

1. INTRODUCTION

Probability sampling became the touchstone for good survey practice some decades ago after Neyman (1934) presented the theory for stratified and cluster sampling based on the randomization distribution. Neyman also showed that a type of nonrandom quota sample of Italian census records drawn by Gini and Galvani had failed to provide satisfactory estimates for many variables in the census. Quoting Smith (1976), "This combined attack was overwhelming and since that day random sampling has reigned supreme." Another early nail in the coffin of nonrandom sampling was the notable failure of a one enormous, but nonprobability, sample to correctly forecast the 1936 US presidential election result. In pre-election polls, the Literary Digest magazine collected 2.3 million mail surveys from mostly middle-to-upper income respondents. Although this sample size was huge, the poll incorrectly predicted that Alf Landon would win by a landslide over the incumbent, Franklin Roosevelt. In fact, Roosevelt won the election in a landslide, carrying every state except for Maine and Vermont (Squire, 1988). As Squire noted, the magazine's respondents consisted mostly of automobile and telephone owners plus the magazine's own subscribers. This pool underrepresented Roosevelt's core of lower-income supporters. In the same election, several pollsters (Gallup, Crossley and Roper) using much smaller but more representative quota samples correctly predicted the outcome (Gosnell, 1937). However, it is worth noting that in the 1948 US presidential elections, Gallup and Roper erroneously forecasted that Dewey would win using quota sampling methods similar to those from 1936. Quota samples are themselves nonprobability samples but are controlled to be distributed more like a random sample from a population would be.

More recent examples of polls that failed to correctly predict election outcomes are the 2015 British parliamentary election (Cowling, 2015), the 2015 Israeli Knesset election (Liebermann, 2015) and the 2014 governor's race in the US state of Maryland (Enten, 2014). The widespread failure of the British 2015 polls led to an extensive evaluation by two professional societies (Sturgis et al., 2016). There were various potential reasons for the misfires, including samples with low contact and response rates, samples based on unrepresentative volunteer panels, inability to predict which

Michael R. Elliott is Professor, Biostatistics Department & Research Professor, Institute for Social Research, University of Michigan, ISR Rm 4068, 426 Thompson St., Ann Arbor, Michigan 48109, USA (e-mail: mrelliott@umich.edu). Richard Valliant is Research Professor, Institute for Social Research, University of Michigan & Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742, USA (e-mail: rvallian@umd.edu).

respondents would actually vote, question wording and framing, deliberate misreporting, and volatility in voters' opinions about candidates. The samples for the 2015 British polls were online or telephone polls that could not be considered probability samples of all registered voters. Demographic population totals for characteristics like age, sex, region, social grade and working status were used to set quota sample and weighting targets. After evaluating eight putative explanations, Sturgis et al. (2016) concluded that the British polls were wrong because of their unrepresentative samples. The statistical adjustment procedures that were used did not correct this basic problem.

On the other hand, selecting a probability sample does not guarantee that the cooperating units will provide a good basis for inference to a population. In many types of surveys response rates have declined dramatically, casting doubt on how well these samples represent the population. Pew Research reported that their response rates (RRs) in typical telephone surveys dropped from 36% in 1997 to 9% in 2012 (Kohut et al., 2012). With such low response rates, a sample initially selected randomly can hardly be called a probability sample from the desired population. Low RRs raise the question of whether probability sampling is a viable methodology for general population surveys without expensive face-to-face data collection methods which usually have higher response.

For some purposes, convenience samples or other types of nonprobability samples have long been acceptable. For example, using convenience samples in experimental studies is standard practice, even when the conclusions are intended to apply to some larger population. The inferences are model-based and come from assuming that the experimental effects are homogeneous among all units in the relevant population. Models are also used for inference in observational studies where, in contrast to designed experiments, assignments of interventions or treatments are not controlled by an experimenter. However, the lack of randomization in those studies may threaten their validity (Madigan et al., 2014). Inferences from nonprobability samples must also rely on models, rather than the distribution generated by random sampling, to project a sample to a larger finite population.

Obtaining data without exercising much control over the set of units for which it is collected is often cheaper and quicker than probability sampling where efforts are made to use a frame that covers most or all of the population, and units are randomly selected from the frame. Repeatedly attempting to get nonrespondents to cooperate, which is standard procedure in probability samples, can be expensive and time-consuming. Eliminating nonresponse followup is also an expedient way of cutting costs. In telephone-only surveys, no amount of nonresponse followup is likely to boost response to the rates that were considered minimally acceptable 10 to 15 years ago. For these reasons, nonprobability sampling is currently staging a kind of renascence (e.g., see Berzofsky, Williams and Biemer, 2009, Dever and Valliant, 2014).

There are also other data sources that are currently receiving attention and might be considered for finite population estimation (Couper, 2013). Social media and other data that can be scraped from the web might be used for gauging public opinion (Murphy et al., 2015) or measuring changes in consumer prices (Cavallo and Rigobon, 2016). Although the inferential issues raised subsequently apply to these "big data," we mainly concern ourselves with nonprobability samples that were directly collected for the purposes of making finite population estimates.

1.1 Types of Nonprobability Samples

There are a number of types of nonprobability samples that are summarized briefly below. Regardless of type, there is quite a bit of controversy about the use of nonprobability surveys for making inferences. Section 2 describes the potential problems with nonprobability samples that can bias inferences. However, these concerns are not limited to finite population inference. Keiding and Louis (2016) is a recent discussion of problems with self-selected entry to epidemiological studies and surveys. Stuart et al. (2011) considers the use of propensity cores to generalize results from randomized trials to populations. Kaizar (2015) reviews approaches that have been proposed for combining randomized and nonrandomized studies in the estimation of treatment efficacy. O'Muircheartaigh and Hedges (2014) describe the use of stratified propensity scores for analyzing a nonrandomized social experiment.

For finite population sampling, the American Association of Public Opinion Research (AAPOR) has issued two task force reports on the use of nonprobability samples—neither of which favored their use. Baker et al. (2010) studied the use of online Internet panels; Baker et al. (2013a, 2013b) cover nonprobability sampling generally. Baker et al. (2010) recommended on several grounds that researchers not use online panels if the objective is to accurately estimate population values. Among other reasons, they noted that (i) some comparative studies showed that nonprobability samples were less accurate than probability samples; (ii) the demographic composition of different panels can affect estimates; and (iii) not all panel vendors fully disclose their methods. Baker et al. (2013a) took a more nuanced view that inferences to a population from nonprobability samples can be valid but that the modeling assumptions needed are difficult to check.

Nonprobability surveys capture participants through various methods. The AAPOR task force on nonprobability sampling (Baker et al., 2013a) characterized these samples into three broad types:

1. Convenience sampling.
2. Sample matching.
3. Network sampling.

Baker et al. (2013a) describe these in some detail; we briefly summarize them here. Convenience sampling is a form of nonprobability sampling in which easily locating and recruiting participants is the primary consideration. No formal sample design is used. Some types of convenience samples are *mall intercepts*, *volunteer samples*, *river samples*, *observational studies* and *snowball samples*. In a *mall intercept sample*, interviewers try to recruit shoppers to take part in some study. Usually, neither the malls nor the people are probability samples.

Volunteer samples are common in social science, medicine and market research. Volunteers may participate in a single study or become part of a panel whose members may be recruited for different studies over the course of time. A recent development is the opt-in web panel in which volunteers are recruited when they visit particular web sites (Schonlau and Couper, 2017). After becoming part of a panel, the members may participate in many different surveys, often for some type of incentive. *River samples* are a version of opt-in web sampling in which volunteers are recruited at a number of websites. Some thought may be given to the set of websites used for recruitment with an eye toward obtaining a cross-section of demographic groups.

In *sample matching*, the members of a nonprobability sample are selected to match a set of important population characteristics. For example, a sample of persons may be constructed so that its distribution by age, race-ethnicity and sex closely matches the distribution of the inference population. Quota sampling is an example of sample matching. The matching is intended to reduce selection biases as long as the covariates that predict survey responses can be used in matching. Rubin (1979) presents the theory for matching in observational studies. A variation of matching in survey sampling is to match the units in a nonprobability sample with those in a probability sample. Each unit in the nonprobability sample is then assigned the weight of its match in the probability sample. Rivers (2007) describes this type of sampling matching in the context of web survey panels. Other techniques developed by Rosenbaum and Rubin (1983) and others for analyzing observational data have also been applied when attempting to develop weights for some volunteer samples.

In *network sampling*, members of some target population (usually a rare one like intravenous drug users or men who have sex with men) are asked to identify other members of the population with whom they are somehow connected. Members of the population that are identified in this way are then asked to join the sample. This method of recruitment may proceed for several rounds. *Snowball sampling* (also called chain sampling, chain-referral sampling or referral sampling) is an example of network sampling in which existing study subjects recruit additional subjects from among their acquaintances. These samples typically do not represent any well-defined target population, although they are a way to accumulate a sizeable collection of units from a rare population.

Sirken (1970) is one of the earliest examples of network or multiplicity sampling in which the network that respondents report about is clearly defined (e.g., members of a person's extended family). Properly done, a multiplicity sample is a probability sample because a person's network of recruits is well-defined. Heckathorn (1997) proposed an extension to this called *respondent driven sampling* (RDS) in which persons would report how many people they knew in a rare population and recruit other members of the rare population. RDS has been used in many applications. For example, Frost et al. (2006) used RDS to locate intravenous drug users; Schonlau, Weidmer and Kapteyn (2014) used it in an attempt to recruit an internet panel. If some restrictive assumptions on how the recruiting is done are satisfied, probabilities of being included in a sample can be computed and used for inferences to a full rare population, but these assumptions can easily be violated (e.g., see Gile and Handcock, 2010). Because the network applications are extremely specialized, we will not address them further.

1.2 General Framework for Inference

Smith (1983) discusses the general problem of making inferences from nonrandom samples. His formulation is to consider the joint density of the population vector of an analysis variable, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)$ and the population vector of 0–1 indicator variables, $\boldsymbol{\delta}_s = (\delta_1, \delta_2, \dots, \delta_N)$ for a sample s. The presentations of Rubin (1976) and Little (1982) on selection mechanisms and survey nonresponse are closely related. Suppose that $\mathbf{X}$ is an $N \times p$ matrix of covariates that can be used in designing a sample or in constructing estimators. The conditional density of $\mathbf{Y}$ given $\mathbf{X}$ and a parameter vector $\boldsymbol{\Theta}$ is $f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta})$. The density of $\boldsymbol{\delta}_s$ given $\mathbf{Y}$, $\mathbf{X}$, and another unknown parameter $\boldsymbol{\Phi}$ is $f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \Phi)$. The joint model for $\mathbf{Y}$ and $\boldsymbol{\delta}_s$ is

$$(1) \quad f(\mathbf{Y}, \boldsymbol{\delta}_s|\mathbf{X}; \boldsymbol{\Theta}, \boldsymbol{\Phi}) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta}) f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi}).$$

Note that this allows the possibility that being in the sample depends on $\mathbf{Y}$, that is, to be not missing at random (NMAR). In a probability sample (without nonresponse or other missingness that is out of control of the sampler), $f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \Phi) = f(\boldsymbol{\delta}_s|\mathbf{X})$. The density $f(\boldsymbol{\delta}_s|\mathbf{X})$ is the randomization distribution and is the basis for design-based inference. However, in a nonprobability sample, the distribution of $\boldsymbol{\delta}_s$ can depend on both $\mathbf{Y}$ and an unknown parameter $\boldsymbol{\Phi}$. Depending on the application, inference can be based on either $f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta})$ or $f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \Phi)$ or on a combination of both.

We term two general approaches to making inferences from nonprobability samples as *quasi-randomization* and *superpopulation*. Quasi-randomization is described in Section 3 and requires modeling $f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \Phi)$. Ideally, the probability of being in the sample is not NMAR and a model can be found for $f(\boldsymbol{\delta}_s|\mathbf{X}; \Phi)$. The superpopulation approach is covered in Section 4 and involves modeling $f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta})$. Both of these approaches involve models, but the approaches are fundamentally different. In the quasi-randomization approach the probability of a unit's being included in the sample is modeled. In the superpopulation approach, the analytic variables (y's) collected in the sample are modeled. Deville (1991) also covers these approaches in the context of quota sampling.

Descriptive statistics, like means and totals, and analytic statistics, like model parameters, are common estimands in finite population estimation. Detailed discussion of the latter is given in Lumley and Scott (2017). Finite population totals are the simplest target to discuss. A total of some quantity Y can be written as the sum of the values over the set of sample units, s, and the sum over the nonsample units $\bar{s}$:

$$t_U = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} Y_i \equiv t_s + t_{\bar{s}}.$$

Since the sample values are observed, we use lower case y for them; upper case is used for the unobserved, nonsample values. In this simple case, the nonsample sum, $t_{\bar{s}}$, is often estimated (or predicted) by a weighted sum of the sample observations, that is, $\hat{t}_{\bar{s}} = \sum_{i \in s} w_i y_i$ where w_i is a weight that may be dependent on the units in the sample. [Alternative ways of calculating weights in probability samples are discussed in Haziza and Beaumont (2017)]. Typically, the estimator can also be written as $\hat{t}_{\bar{s}} = \sum_{i \in \bar{s}} \hat{y}_i$ where $\hat{y}_i$ is a prediction for nonsample unit i. Thus, for totals the estimation problem is one of prediction.

Estimation of model parameters often requires solving a set of estimating equations for the parameter estimates. The estimating equations can be linear in the parameters, as for linear regression or nonlinear, as for generalized linear models. In design-based finite population estimation, the estimating equations include survey weights and are estimators of types of finite population totals (Binder and Roberts, 2009). If weights are constructed for a nonprobability sample that are appropriate for estimating totals, then those weights can also be used in the estimating equations. Consequently, weight construction for nonprobability samples can play the same role in estimation as in probability sampling.

Baker et al. (2013a) discuss the methods that have been proposed for weighting nonprobability samples. Such samples lack many of the features that guide weighting in probability samples. A nonprobability sample is not selected randomly from an explicit sampling frame. Consequently, selection probabilities cannot be computed, and the usual method of computing base weights (inverses of selection probabilities) does not apply. Weights can, however, be computed using the quasi-randomization or superpopulation approaches noted above.

2. POTENTIAL PROBLEMS WITH NONPROBABILITY SAMPLES

Since nonprobability samples are often obtained in a poorly controlled or uncontrolled way, they can be subject to a number of biases when the goal is inference to a specific finite population. Several issues are listed here in the context of voluntary Internet panels, but other types of nonprobability samples can suffer from similar problems.

Selection bias occurs if the seen part of the population (the sample) differs from the unseen (the nonsample) in such a way that the sample cannot be projected

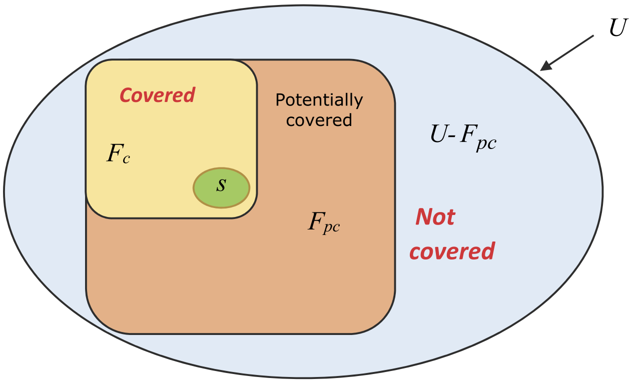

FIG. 1. *Illustration of potential and actual coverage of a target population.*

to the full population. Whether a nonprobability sample covers the desired population is a major concern. For example, in a volunteer web panel only persons with access to the Internet can join a panel. To describe three components of coverage survey bias, Valliant and Dever (2011) defined three populations, illustrated in Figure 1: (1) the target population of interest for the study U; (2) the potentially covered population given the way that data are collected, F_{pc}; and (3) the actual covered population, F_c, the portion of the target population that is recruited for the study through the essential survey conditions. For example, consider an opt-in web survey for a smoking cessation study. The target population U may be defined as adults aged 18–29 who currently use cigarettes. The potentially covered population F_{pc} would be those study-eligible individuals with Internet access who visit the sites where study recruitment occurs; those actually covered F_c would be the subset of the potential covered population who participate in the study. Selecting a sample only from F_c results in selection bias. The sample s are those persons who are invited to participate in the survey and who actually do. The $U - F_{pc}$ area in the figure are the many persons who have Internet access but never visit the recruiting websites or who do not have Internet access at all. In many situations, $U - F_{pc}$ is vastly larger than either F_c or F_{pc}.

To illustrate a case that is rife with coverage problems, we further consider surveys done using panels of persons recruited via the Internet. Table 1 lists percentages of households in the US in 2013 estimated from the American Community Survey (ACS) that have some type of Internet subscription (File and Ryan, 2014). The ACS estimates are based on a sample of about 3.5 million households. About 25% of households had no Internet subscription, which in itself is a substantial amount of undercoverage of the full population. The coverage varies considerably by demographic group. Only 58.3% of households where the head is 65 or older have the Internet. Black non-Hispanic and Hispanic households are less likely to have access than other race-ethnicities. Households in metropolitan areas are more likely to have access. There is also a clear dependence on income and education. As income and education increase, so does the percentage of households with access. As illustrated in Dever, Rafferty and Valliant (2008), these coverage errors can lead to biased estimates for many items.

Selection bias occurs when some groups are also more likely to volunteer for a panel. Bethlehem (2010) reviews this issue for web surveys. Vonk, van Ossenbruggen and Willems (2006) report that ethnic minorities and immigrant groups were systematically underrepresented in Dutch panels. They also found that, relative to the general population, the Dutch online panels contained disproportionately more voters, more Socialist Party supporters, more heavy Internet users and fewer churchgoers.

Nonresponse of several kinds affects web panels. Many panel vendors have a “double opt-in” procedure for joining for a panel. First, a person registers his/her name, email and some demographics. Then the vendor sends the person an email that must be responded to in order to officially join the panel. This eliminates people who give bogus emails but also introduces the possibility of *registration nonresponse* since some people do not respond to the vendor’s email. People may also click on a banner ad advertising the panel but never complete all registration steps. Alvarez, Sherman and Van Beselaere (2003) report that, during the recruitment of one panel, just over 6% of those who clicked through a banner ad to the panel registration page eventually completed all the steps required to become a panel member. Finally, a panel member asked to participate in a survey may not respond.

Attrition is another problem—persons may lose interest and drop out of a panel. Many surveys are targeted at specific groups, for example, young Black females. A panelist that is in one of these “interesting” groups may be peppered with survey requests and drop out for that reason. Another reason that some groups, like the elderly, are over-burdened is that they may be oversampled to make up for anticipated nonresponse.

Measurement error is also a worry in nonprobability surveys as they are in any survey. The types of error that have been demonstrated in some studies are effects

TABLE 1
Percentages of US households with Internet subscriptions; 2013 *American Community Survey*

	Percent of households with some Internet subscription
Total households	74.4
Age of householder	
15–34 years	77.7
35–44 years	82.5
45–64 years	78.7
65 years and older	58.3
Race and Hispanic origin of householder	
White alone, non-Hispanic	77.4
Black alone, non-Hispanic	61.3
Asian alone, non-Hispanic	86.6
Hispanic (of any race)	66.7
Limited English-speaking household	
No	75.5
Yes	51.4
Metropolitan status	
Metropolitan area	76.1
Nonmetropolitan area	64.8
Household income	
Less than $25,000	48.4
$25,000–$49,999	69.0
$50,000–$99,999	84.9
$100,000–$149,999	92.7
$150,000 and more	94.9
Educational attainment of householder	
Less than high school graduate	43.8
High school graduate	62.9
Some college or associate's degree	79.2
Bachelor's degree or higher	90.1

due to questionnaire design, mode and peculiarities of respondents. For example, the persons who participate in panels tend to have higher education levels. The motivation for participating may be a sense of altruism for some but may be just to collect an incentive for others. Participants are often paid per survey completed. Some respondents speed through surveys, answering as quickly as possible to collect the incentive. This is a form of "satisficing" where respondents do just enough to get the job done (Simon, 1956). On the other hand, self-administered online surveys do tend to elicit more reports of socially undesirable behaviors, like drug use, than do face-to-face surveys. Higher reports are usually taken to be more nearly correct. But, it may be that the people taking those surveys just behave undesirably more often than the general population.

Baker et al. (2010, page 739) list 19 studies where the same questionnaire was administered by interviewers to probability samples and online to nonprobability samples. As they noted, "Only one of these studies yielded consistently equivalent findings across methods, and many found differences in the distributions of answers to both demographic and substantive questions. Further, these differences generally were not substantially reduced by weighting."

Despite all of these actual and potential problems, online panels are now widely used. For example, the Washington Post newspaper and the company, SurveyMonkey, have recently mounted a nonprobability, online poll of over 75,000 registered voters that covers all 50 states in the US (Clement, 2016). Baker et al. (2010) quotes the market research newsletter, *Inside Research* as estimating the total spent on online research in 2009 at about $2 billion USD, the vast majority of which is supported by online panels.

3. QUASI-RANDOMIZATION APPROACH

In the quasi-randomization approach, pseudo-inclusion probabilities are estimated and used to correct for selection bias. Given estimates of the pseudo-probabilities, design-based formulas are used for point estimates and variances. Using the earlier notation, the goal is to estimate $f(\boldsymbol{\delta}_s|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi})$ or $f(\boldsymbol{\delta}_s|\mathbf{X}; \boldsymbol{\Phi})$. Having a situation where the sample inclusion probabilities do not depend on the Y's is ideal since the nonsample Y's are unknown, but verifying that this is the case is impossible in most applications. There is some literature on estimation when nonsample data are NMAR (e.g., see Little, 2003), but the methods generally require information on nonsample units that is available only in specialized applications. Thus, the practical approach is to estimate $f(\boldsymbol{\delta}_s|\mathbf{X}; \boldsymbol{\Phi})$.

To illustrate how involved estimating these probabilities may be, consider a case in which a volunteer panel of persons is recruited to provide a pool from which a sample of persons is selected. To respond to a survey, a person must have Internet access, volunteer for the panel, be selected for the particular survey and then respond. Considering all of these, the probability of person i participating in that Web survey [using a simpler notation than $f(\boldsymbol{\delta}_s|\mathbf{X}; \boldsymbol{\Phi})$ above] can be decomposed as

$$\begin{aligned} &P(\mathbf{x}_i) \\ (2)\quad &= P(i \in I|\mathbf{x}_i)P(i \in V|I, \mathbf{x}_i) \\ &\quad \cdot P(i \in s_V|V, I, \mathbf{x}_i)P(i \in s_{VR}|s_V, V, I, \mathbf{x}_i), \end{aligned}$$

where

$\mathbf{x}_i$ = a vector of covariates for person i that are predictive of participation;

I = set of persons with Internet access, that is, F_{pc} in Figure 1; V = set of persons who volunteer;

$P(i \in I|\mathbf{x}_i)$ = probability of having access to the Internet;

$P(i \in V|I, \mathbf{x}_i)$ = probability of volunteering for an opt-in panel given that person i has access to the Internet;

$P(i \in s_V|V, I, \mathbf{x}_i)$ = probability that person i was subsampled from the panel and asked to participate with s_V denoting the subsample from the panel;

$P(i \in s_{VR}|s_V, V, I, \mathbf{x}_i)$ = probability that person i responds given selection for the subsample with s_{VR} denoting the set of survey respondents.

Standard methods (e.g., see Valliant, Dever and Kreuter, 2013) can be used to compute the last two terms in (2). The first two probabilities—having Internet access and volunteering for the panel—are more difficult. Both are likely to depend on the $\mathbf{x}_i$ covariates and, in a worse case, upon the Y's. For example, persons with higher socioeconomic status are more likely to have access; younger people are more likely to join a panel than older ones. In some countries, probability samples that represent the full population may include questions on Internet access. The US National Health Interview Survey routinely includes such questions. The probability of volunteering (given Internet access) is harder to estimate.

Reference survey. One approach is to use a *reference survey* in parallel to the nonprobability survey. The reference survey can be a probability survey selected from either (i) the population of persons who have Internet access or (ii) the full population including persons that do not have the Internet. The reference sample might also be a census that covers the entire population. The statistical approach is to combine the reference sample and the sample of volunteers and fit a model to predict the probability of being in the nonprobability sample, as described in Section 3.1.

A key requirement of the reference survey is that it include the same covariates $\mathbf{x}_i$ as the volunteer survey so that a binary regression can be fitted to permit estimation of inclusion probabilities for the volunteers. One possibility for a reference survey is to use a publicly available dataset collected in a well designed and executed probability survey (like one done by a central government agency). Another possibility is for the survey organization to conduct its own reference survey. In the latter case, some specialized questions, beyond the usual age/race/sex/education types of demographics, can be added that are felt to be predictive of volunteering and of the analysis variables for the volunteer survey. Schonlau, van Soest and Kapteyn (2007) refer to these extra covariates as *webographics*. However, identifying webographics that are useful beyond the standard demographics (age, race-ethnicity, sex, income and education) is difficult (Lee and Valliant, 2009). Of course, another problem with conducting your own reference survey is that doing a high quality survey with good coverage of the target population is expensive and may be beyond the means of many organizations.

Sample matching is another approach to attempting to reduce selection biases in a nonprobability sample. As noted in Baker et al. (2013a), the matching can be done on an individual or aggregate level. If, for each case in a volunteer sample, a matching case is found in a probability, reference sample, this would be individual-level matching. The matches would be found based on covariates available in each dataset. This may be done based on individual covariate values or on propensity scores as described in Rosenbaum and Rubin (1983). This is an example of predictive mean matching in which an imputation of an inclusion probability is made for each nonprobability unit.

Matching at the aggregate level consists on making the frequency distribution of the nonprobability sample the same as that of the population. Quota sampling is an example of this. For example, the age × race distribution of the sample might be controlled to be the same as that in the population. If we start with a large panel of volunteers, a subsample might be selected to achieve this kind of distributional balance. Each person would receive the same weight, which is the same way that a proportionally allocated probability sample would be treated. Considered in this way, quota sampling falls into the quasi-randomization framework.

A probability sample used as a reference survey or in sample matching ideally must not be subject to coverage or other types of bias. As noted in Section 1, many probability samples are now subject to high nonresponse rates and are tantamount to nonprobability samples themselves. Poor quality reference or matching samples can lead to biased estimators of the inclusion probabilities in (2) and, consequently, biased estimators from the nonprobability sample. This is an argument for using large, well-controlled samples conducted by central governments for reference or matching samples if at all possible. For example, in a household survey in the US, the American Community Survey (https://www.census.gov/programs-surveys/acs/) would be a good choice.

3.1 Estimation Using Pseudo-Weights

This approach assumes that the nonprobability sample actually does have a probability sampling mechanism, albeit one with probabilities that have to be estimated under identifying assumptions. The goal is to estimate this unknown probability of selection relying on a true probability sample or a census with common variables that explain the unknown sampling mechanism (Elliott, 2009, Elliott et al., 2010). Let S_i denote the sampling indicator for the probability sample, S_i^* denote the indicator for the nonprobability sample, and $\mathbf{x}_i$ be the set of common covariates available to both samples that are assumed to fully govern the sampling mechanism for both. Applying Bayes rule, we have (Elliott and Davis, 2005):

$$
\begin{aligned}
&P(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \\
&\quad = \frac{P(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)P(S_i^* = 1)}{P(\mathbf{x}_i = \mathbf{x}_o)} \\
(3)\quad &\quad = \frac{P(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)P(S_i^* = 1)P(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{P(S_i = 1)P(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)} \\
&\quad \propto \frac{P(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)P(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{P(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)},
\end{aligned}
$$

where $P(S_i = 1)/P(S_i^* = 1)$ can be treated as a normalizing constant.

Estimating $P(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)$ and $P(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)$ can be difficult for a general joint distribution of covariates $\mathbf{x}$, but extensions of discriminant analysis (without making a normality assumption) provide a way around this problem. Combine the probability and nonprobability samples and let $Z_i = 1$ for nonprobability cases (i.e., $S_i^* = 1, S_i = 0$) and $Z_i = 0$ for the probability cases (i.e., $S_i^* = 0, S_i = 1$) conditional on being in the combined probability-nonprobability sample (i.e., $S_i^* + S_i = 1$). Then

$$
\begin{aligned}
&\frac{P(\mathbf{x}_i = \mathbf{x}_o|Z_i = 1)}{P(\mathbf{x}_i = \mathbf{x}_o|Z_i = 0)} \\
(4)\quad &\quad = \frac{P(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)P(\mathbf{x}_i = \mathbf{x}_o)/P(Z_i = 1)}{P(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)P(\mathbf{x}_i = \mathbf{x}_o)/P(Z_i = 0)} \\
&\quad \propto \frac{P(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{P(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)}.
\end{aligned}
$$

As long as sampling fractions are small, $P(S_i = 1, S_i^* = 0) \approx P(S_i = 1)$ and $P(S_i = 0, S_i^* = 1) \approx P(S_i^* = 1)$, so $P(\mathbf{x}_i|Z_i = 0) = P(\mathbf{x}_i|S_i = 1, S_i^* = 0) \approx P(\mathbf{x}_i|S_i = 1)$ and $P(\mathbf{x}_i|Z_i = 1) = P(\mathbf{x}_i|S_i = 0, S_i^* = 1) \approx P(\mathbf{x}_i|S_i^* = 1)$. Thus,

$$
\begin{aligned}
&P(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \\
&\quad \dot{\propto}\, P(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)\frac{P(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{P(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)}.
\end{aligned}
$$

The resulting "pseudo-weight" is given by

$$
\begin{aligned}
&w_i = 1/\hat{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \\
(5)\quad &\quad \propto 1/\hat{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)\frac{\hat{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)}.
\end{aligned}
$$

If the covariates $\mathbf{x}_i$ that are available in both the nonprobability and probability sample match those used to design the probabilities of selection/inclusion in the probability sample, (5) can be written as

$$w_i = 1/\hat{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \propto \tilde{w}_i \frac{\hat{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)}, \tag{6}$$

where $\tilde{w}_i$ is the inverse of the probability of selection for the nonprobability unit in the probability sampling frame. Otherwise, in the more likely setting where $\mathbf{x}_i$ does not correspond precisely to the probability sample design variables, $\hat{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)$ can be estimated by regressing $\mathbf{x}_i$ on $\tilde{w}_i^{-1}$ via beta regression (Ferrari and Cribari-Neto, 2004) in the probability sample, and predicting $P(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)$ for the nonprobability sample elements.

The term $\hat{P}(Z_i = z|\mathbf{x}_i = \mathbf{x}_o)$ can be obtained via logistic regression, or, to reduce model misspecification if $\mathbf{x}_i$ is of high dimensionality, via least absolute shrinkage and regression operator (LASSO) (Tibshirani, 1996, LeBlanc and Tibshirani, 1998), Bayesian additive regression trees (BART) (Chipman, George and McCulloch, 2010), or super learner algorithms that combine estimators from numerous model fitting methods (Van der Laan, Polley and Hubbard, 2007). In some settings, the nonprobability sample will represent only a portion of population; for example, in a setting with a binary outcome Y (e.g., injured/uninjured) only positive outcomes $Y = 1$ (e.g., injuries) might be represented in the nonprobability dataset; in this case (5) is updated as

$$\begin{aligned} w_i &= 1/\hat{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \\ &\propto 1/\hat{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o, Y_i = 1) \\ &\quad \cdot \frac{\hat{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o)}. \end{aligned} \tag{7}$$

An alternative to estimating the probability of unit i's being in the nonprobability sample is used by some panel vendors. The probability (reference) and nonprobability samples are combined, but a logistic regression is run to estimate $P(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)$, not conditioned on being in the combined probability and nonprobability sample (e.g., see Valliant and Dever, 2011). This is done by assigning a weight of 1 to the nonprobability cases, the probability sampling weight to the probability cases, and running a weighted logistic regression. The model predictions, thus, refer to the unconditional probability, $P(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)$, not the probability conditional on being in the combined sample. Whether this method is better or worse than (5) has not been studied, although, as noted above, (5) can be adapted to cases where the nonprobability sample represents only a portion of the population.

If analysis of the nonprobability sample only is required, the pseudo-weight construction is complete. If the nonprobability and probability samples are to be combined, the nonprobability sample pseudo-weights and probability sample weights are normalized so that the weighted fraction of the nonprobability sample is equal to the unweighted fraction of the nonprobability sample cases in the combined dataset, and similarly the weighted fraction of the probability sample is equal to the unweighted fraction of the probability sample cases in the combined dataset (Korn and Graubard, 1999, pages 278–284). This ensures that the sum of the combined weights continues to approximate the population size, and that each sample will contribute in proportion to their unweighted sample size. This is accomplished by setting $\hat{w}_i = C_{S^*} \times w_i$ for $C_{S^*} = n_{S^*}/(n_S + n_{S^*}) \times \sum_i I(Z_i = 0)\tilde{w}_i / \sum_i I(Z_i = 1)w_i$ for the nonprobability sample cases and $\hat{w}_i = C_S \times \tilde{w}_i$ for $C_S = n_S/(n_S + n_{S^*})$.

To obtain inference, the pseudo-weights or the normalized pseudo-weights and probability sample weights in the combined dataset can be used to obtain weighted point estimates. For variance estimation, a bootstrap or jackknife estimator should be used to incorporate both sampling variability in the estimation of the pseudo weights and in the estimation of the main quantity of interest. In the absence of true design information in the nonprobability sample, resampling at the subject level for the bootstrap or leave-one-out computation of the pseudo-estimate for the jackknife can be applied. However, some thought must be given to the structure of the convenience sample. For example, the websites used to recruit a volunteer web panel might properly be considered as clusters if different types of persons visit the different sites (Brick, 2015). For the probability sample, resampling clusters within strata and use of the Rao–Wu bootstrap (Rao and Wu, 1988, Rao, Wu and Yue, 1992) to accommodate weights can be used. For the jackknife, clusters within strata should be dropped, with standard weighting up by the number of clusters divided by the number of clusters retained to maintain the stratum size should be used. For each bootstrap or jackknife iteration, the pseudo-weights should be recomputed as well as the point estimator using the dropped-out or resampled data.

4. SUPERPOPULATION MODEL APPROACH

In the superpopulation modeling approach, a statistical model is fitted for a Y analysis variable from the sample and used to project the sample to the full population. That is, inferences are based on $f(\mathbf{Y}|\mathbf{X};\mathbf{\Theta})$. This approach could, of course, also be used with a probability sample. The difference here is that design-based inference, where the randomization distribution is under the control of the sampler, is not an option for a nonprobability sample. As noted in Smith (1983), the sample selection mechanism can be ignored for model-based inferences about the distribution of $\mathbf{Y}$ if

$$f(\boldsymbol{\delta}_s|\mathbf{Y},\mathbf{X};\mathbf{\Phi}) = f(\boldsymbol{\delta}_s|\mathbf{X};\mathbf{\Phi}), \tag{8}$$

which would be the formal justification for using only $f(\mathbf{Y}|\mathbf{X};\mathbf{\Theta})$. There are purposive, nonprobability samples that satisfy (8). For example, selecting the n units with the largest x values as is done by US Energy Information Administration (2016), or sampling balanced on population moments of covariates (Royall, 1970, 1971) are ignorable, nonprobability plans. However, in nonprobability samples where the selection of sample units is not well-controlled, (8) may not hold and the quasi-randomization and superpopulation approaches could be combined.

Note that $\mathbf{Y}$ can be partitioned between the sample and nonsample units as $\mathbf{Y} = (\mathbf{Y}_s, \mathbf{Y}_{\bar{s}})$. Thus, $f(\mathbf{Y}|\mathbf{X};\mathbf{\Theta}) = f(\mathbf{Y}_s|\mathbf{Y}_{\bar{s}},\mathbf{X};\mathbf{\Theta})f(\mathbf{Y}_{\bar{s}}|\mathbf{X};\mathbf{\Theta})$. If $f(\mathbf{Y}_s|\mathbf{Y}_{\bar{s}},\mathbf{X};\mathbf{\Theta}) = f(\mathbf{Y}_s|\mathbf{X};\mathbf{\Theta})$, then $\mathbf{Y}_s$ and $\mathbf{Y}_{\bar{s}}$ are independent conditional on the covariates, $\mathbf{X}$. If model-based inferences are desired for $\mathbf{\Theta}$, these can be done based only on $f(\mathbf{Y}_s|\mathbf{X};\mathbf{\Theta})$. However, if descriptive inferences are required for the full population $\mathbf{Y}$, then $f(\mathbf{Y}_{\bar{s}}|\mathbf{X};\mathbf{\Theta})$ must be estimated. If this model has the same form as $f(\mathbf{Y}_s|\mathbf{X};\mathbf{\Theta})$, then the model fitted from the sample can be used to predict values for the nonsample. If this is not the case, inference to the full population may be difficult or impossible.

To introduce the superpopulation approach, consider the simple case of estimating a finite population total. The general idea in model-based estimation when estimating a total is to sum the responses for the sample cases and add to them the sum of predictions for nonsample cases. The key to forming unbiased estimates is that the variables to be analyzed for the sample and nonsample follow a common model and that this model can be discovered by analyzing the sample responses. When both the sample and nonsample units follow the same model, model parameters can be estimated from the sample and used to make predictions for the nonsample cases. An appropriate model usually includes covariates, as in $f(\mathbf{Y}_s|\mathbf{X};\mathbf{\Theta})$ above, which are known for each individual sample case. The covariates may or may not be known for individual nonsample cases. For some common estimation methods like poststratification, only population totals of the covariates are required to construct the estimator, so that individual nonsample $\mathbf{X}$ values are unnecessary. Suppose that the mean of a variable y_i follows a linear model:

$$E_M(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\boldsymbol{\beta},$$

where the subscript M means that the expectation is with respect to the model, $\mathbf{x}_i$ is a vector of p covariates for unit i and $\boldsymbol{\beta}$ is a parameter vector. Given a sample s, an estimator of the slope parameter is $\hat{\boldsymbol{\beta}} = \mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{y}_s$ where $\mathbf{A}_s = \mathbf{X}_s^T\mathbf{X}_s$, $\mathbf{X}_s$ is the $n \times p$ matrix of covariates for the sample units, and $\mathbf{y}_s$ is the n-vector of sample y's. (Weighted least squares might also be used if there were evidence of nonhomogeneous model variances.) A prediction of the value of a unit in the set of nonsample units, denoted by r, is $\hat{y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. A predictor of the population total is

$$\begin{aligned}\hat{t}_1 &= \sum_{i\in s} y_i + \sum_{i\in\bar{s}} \hat{y}_i \\ &= \sum_{i\in s} y_i + (\mathbf{t}_{Ux} - \mathbf{t}_{sx})^T\hat{\boldsymbol{\beta}},\end{aligned} \tag{9}$$

where $\mathbf{t}_{Ux}$ is the total of the $x's$ in the population and t_{sx} is the sample sum of the $x's$. This estimator is also equal to the general regression estimator (GREG) of Särndal, Swensson and Wretman (1992) if the inverse selection probabilities in that estimator are all set to 1. The theory for this *prediction approach* is extensively covered in Valliant, Dorfman and Royall (2000). If the sample is a small fraction of the population, as would be the case for most volunteer web surveys, the prediction estimator is approximately the same as predicting the value for every unit in the population and adding the predictions:

$$\hat{t}_2 = \sum_{i\in U} \hat{y}_i = \mathbf{t}_{Ux}^T\hat{\boldsymbol{\beta}}. \tag{10}$$

The population mean of y can be estimated by $\hat{\bar{Y}} = \bar{X}_U^T\hat{\boldsymbol{\beta}}$ where $\bar{X}_U = \mathbf{t}_{Ux}/N$, the population vector of covariate means.

The estimators in (9) or (10) are quite flexible in what covariates can be included. For example, we might predict the amount that people have saved for retirement based on their occupation, years of education, marital status, age, number of children they have and region of the country in which they live. Constructing the estimator would require that census counts be available for each of those covariates. Another possibility is to use estimates from some other larger or more accurate survey (e.g., Dever and Valliant, 2010, 2016). The reference surveys mentioned earlier could be a source of estimated control totals in which webographic covariates might be used.

Both (9) and (10) can be written so that they are weighted sums of y's. If (9) is used, the weight for unit i is $w_{1i} = 1 + \mathbf{t}_{rx}^T \mathbf{A}_s^{-1} \mathbf{x}_i$ where $\mathbf{t}_{rx} = \mathbf{t}_{Ux} - \mathbf{t}_{sx}$. In (10), the weight is $w_{2i} = \mathbf{t}_{Ux}^T \mathbf{A}_s^{-1} \mathbf{x}_i$. The estimated total for an analysis variable can be written as $\hat{t} = \sum_s w_i y_i$ where w_i is either w_{1i} or w_{2i}. Notice that these weights depend only on the x's not on y. As a result, the same set of weights could be used for all estimates. It is true that a single set of weights will not be equally efficient for every y, but this situation is also true for design-based weights.

In the superpopulation (y-model) approach, statistical properties, like bias and variance, are computed conditional on the set of sample units that is observed. This contrasts to the quasi-randomization approach where the pseudo design-based calculations average over the random appearance in the sample of units that have the same configuration of covariates observed in the sample. A quasi-randomization estimator that only uses inverse estimated inclusion probabilities as weights will be biased under a y-model where $E_M(y|\mathbf{x})$ depends on covariates. Consequently, the y-model approach to constructing estimators can produce more precise estimators than the quasi-randomization approach alone. Chen (2015) gives some numerical illustrations of this approach applied to a nonprobability sample.

4.1 Variance Estimation for Prediction Estimators

For the frequentist methods, estimating the variance of an estimator is the usual step toward making inferences about population values. There are several choices for variance estimators when model-based weighting is used. These are described in Valliant, Dorfman and Royall (2000, Chapter 5). To fully define the model, we need to add a variance specification. The ones we summarize here are appropriate for models in which units are mutually independent. Although model-based estimators have been extended to cases where units are correlated within clusters (Valliant, Dorfman and Royall, 2000, Chapter 9), these clustered structures are often unnecessary for the web surveys and similar cases that we cover here. Suppose that the full model is

$$\begin{aligned} E_M(y_i|\mathbf{x}_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\ V_M(y_i|\mathbf{x}_i) &= v_i, \end{aligned} \tag{11}$$

where v_i is a variance parameter that does not have to be specifically defined. The variance estimators below will work regardless of the form of v_i (as long as it is finite).

For use below, define a_i to be $w_i - 1$ where w_i is either w_{1i} or w_{2i}. The variance estimators below then apply for either of the w_{1i} or w_{2i} weights. The *prediction variance* of an estimator of a total, $\hat{t}$, is defined as

$$V_M(\hat{t} - t_U) = \sum_{i \in s} a_i^2 v_i + \sum_{i \in r} v_i. \tag{12}$$

The population total of y, t_U, is subtracted on the left-hand side because the sum is random under the model. As long as the fraction of the population that is sampled is very small, the second term on the right-hand side above is inconsequential compared to the first. The variance estimators are built from the model residuals, $r_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$. An estimator of the dominant, first term is

$$\sum_s a_i^2 \hat{v}_i, \tag{13}$$

where $\hat{v}_i$ can be any of three choices: (i) r_i^2, (ii) $r_i^2/(1 - h_{ii})$, or (iii) $[r_i/(1 - h_{ii})]^2$ where h_{ii} is the leverage for unit i, defined as the diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}_s^T \mathbf{A}_s^{-1} \mathbf{X}_s$. As the sample size increases and if no x is extreme, each leverage will converge to zero.

The estimators of the first term are robust in the sense that they are approximately model-unbiased regardless of the form of v_i (which is unknown) as long as the sampling fraction is small. The first choice, $\hat{v}_i = r_i^2$, when used in (13), gives an example of a sandwich estimator. The second choice adjusts for the fact that r_i^2 is slightly biased for v_i. The third choice is very similar to the jackknife in which one sample unit at a time is deleted, a new estimate of the total computed, and the variance among those delete-one estimates is used. Since the second term in (12) is usually negligible compared to the first, misspecifying its form is likely to be unimportant. Valliant, Dorfman and Royall (2000) provide some options for estimating that term.

The bootstrap is another replication estimator that should be equally robust, although, to our knowledge, finite population, model-based theory has not been worked-out for the bootstrap. The bootstrap should also be consistent for estimating the variance of estimated quantiles, unlike the jackknife. If the population totals for some of the covariates are estimated from an independent survey, then the variance in (12) should be modified by adding a term to reflect that additional uncertainty (e.g., see Dever and Valliant, 2010, 2016).

4.2 Hierarchical Regression Modeling

This approach can be explained by viewing calibration approaches such as poststratification and raking as flowing from special cases of model (11). In the case of poststratification, this can be viewed as regression on all of the (discrete) calibration variables and their interactions. Assume that the calibration variable $\mathbf{x}_i$ consists of p binary indicators, $x_{i1}, \dots, x_{ip}$:

$$\begin{aligned}\mu_{yi} &= E_M(y_i|\mathbf{x}_i)\\ &= \beta_0 + \sum_{k_1=1}^{p} \beta_{k_1} I(x_{ik_1}=1)\\ &\quad + \sum_{k_1=1}^{p}\sum_{k_2=2}^{p} \beta_{k_1,k_2} I(x_{ik_1}=1) I(x_{ik_2}=1) + \cdots\\ &\quad + \sum_{k_{p-1}=p-1}^{p} \beta_{k_1,k_2} I(x_{ik_{p-1}}=1) I(x_{ik_p}=1)\\ &\quad + \cdots + \beta_{k_1,k_2,\dots,k_p} \prod_{l=1}^{p} I(x_{ik_l}=1),\end{aligned} \tag{14}$$

where $I(\cdot)$ is a binary indicator variable. Raking assumes main effects only:

$$\mu_{yi} = E_M(y_i|\mathbf{x}_i) = \beta_0 + \sum_{k=1}^{p} \beta_k I(x_{ik}=1). \tag{15}$$

Denote the 2^p possible combinations of values of $x_1, \dots, x_p$ by $h = 1, \dots, 2^p$. The resulting estimates of a population mean are given by

$$\hat{\bar{Y}} = \sum_{h=1}^{2^p} P_h \hat{\mu}_h, \tag{16}$$

where P_h is the proportion of the population whose combination of binary indicator variables is equal to h. That is, the P_h are special cases of $\bar{X}_U$ at the beginning of this section.

The estimated mean, $\hat{\mu}_h$, of the hth combination is found by replacing each β with an estimator, $\hat{\beta}$, in (14) for the poststratification estimator and in (15) for the raking estimator. (Note that $\hat{\mu}_h$ is an estimator of μ_{yi} for each unit in combination h.) These correspond to the weighted estimates obtained from poststratification or raking. Both of these models can be extended to generalized linear regression by replacing μ_{yi} with the appropriate link function $g(\mu_{yi})$ (logistic link for logistic regression of a binary outcome, log link for a count outcome, etc.). Intermediate models between poststratification (14) and raking (15) can be fit by incorporating some but not all possible interaction terms.

To deal with instabilities in the estimation of β, a number of authors have considered adding hierarchical models to the mean regression model. Holt and Smith (1979) first suggested a model for unit i in combination h of the form:

$$\begin{aligned} y_{ih}|\mu_h &\sim N(\mu_h, \sigma^2)\\ \mu_h &\sim N(\mu, \tau^2).\end{aligned} \tag{17}$$

The mean estimator is again given by (16), where $\hat{\mu}_h = E(\mu_h|y) = \frac{\tau^2}{\sigma^2/n_h+\tau^2}\bar{y}_h + \frac{\sigma^2/n_h}{\sigma^2/n_h+\tau^2}\bar{y}$ for known σ^2 and τ^2 and sample sizes n_h within the hth combination of x's, and $n = \sum_h n_h$; $\bar{y}_h$ is the sample mean for units in the hth combination and $\bar{y}$ is the mean for all units. In practice, σ^2 and τ^2 are replaced, for example, with empirical Bayes estimators. Simulation studies in Elliott and Little (2000) showed that exchangeable priors of the form (17) were somewhat fragile, tending to oversmooth when σ^2 and τ^2 were approximately equal. Alternative priors that ordered the strata or poststrata h by sampling weights $w_h = N_h/n_h$ for population size N_h and included information about this structuring in either the prior mean or the variance (e.g., having the mean be a function of w_h, or the variance an autoregressive structure as a function of $|h - h'|$) had much better performance with respect to coverage and mean square error.

Wang et al. (2015) used an extension of this hierarchical model approach, termed multilevel regression and stratification (MRP), to obtain estimates of voting behavior in the 2012 US Presidential election from a highly nonrepresentative convenience sample of nearly 350,000 Xbox users, empaneled 45 days prior to the election. This large sample, combined with highly predictive covariates about voting behavior, including information about party identification and 2008 Presidential election voting behavior, allowed for a refined prediction model that incorporated numerous interactions and used priors on the βs to stabilize parameter estimates and resulting values of μ_h. The values of P_h were estimated via probability sample exit polls from the 2008 US Presidential election, themselves of very large size (over 100,000). Wang et al. (2015) showed that, despite the fact that the raw Xbox estimates were severely biased in favor of Romney, reflecting its largely male and white sample composition, accurate estimates of voting behavior were obtained, based on comparisons with aggregated probability sampling polls as well as the final election result. This accuracy was due to the large sample size that allowed prediction of voting behavior among decidedly under-represented elements of the population (e.g., older minority females), combined with the hierarchical regression modeling to stabilize predictions.

4.2.1 *Multilevel regression and stratification via Bayesian finite population inference.* Wang. et al.'s implementation of MRP ignored uncertainty in the estimation of the P_h from the probability sample. While this may have been warranted due to its large size, in general failure to account for this variance will lead to anti-conservative inference (too narrow confidence intervals). An alternative approach would be utilize a Bayesian finite population inference approach that treats the unsampled elements in the population as missing data, together with the variable Y that is missing in the probability sample data but available in the nonprobability sample data.

Let X be the variables available in the probability and nonprobability sample for prediction of Y, Z be the probability sample design variables, and let (X_{ns}, Z_{ns}) and (X_p, Z_p) represent the nonsampled and probability-sampled elements of the population, respectively. Dong, Elliott and Raghunathan (2014) obtain nonparametric draws from the posterior predictive distribution of the nonsampled elements $(X_{ns}|X_s, Z_p)$

$$\begin{aligned} &p(X_{ns}|X_s, Z_p) \\ &\quad\propto \int p(X_{ns}, Z_{ns}|X_p, Z_p) p(X_p, Z_p)\, dZ_{ns} \end{aligned} \tag{18}$$

under the assumption of ignorable sampling (X is independent of the sampling indicator I conditional on Z) by making draws of $p(X_p, Z_p)$ from a Bayesian bootstrap (Rubin, 1981) and draws from $p(X_{ns}, Z_{ns}|X_p, Z_p)$ via a finite population Bayesian bootstrap (FPBB) procedure that accounts for probabilities of selection, clustering and weighting. Treating the nonprobability sample (Y_{np}, X_{np}) as a certainty sample and concatenating it with the probability sample to obtain $Y_s = Y_{np}$ and $X_s = (X_p, X_{np})$, we have (Zhou, Elliott and Raghunathan, 2016c)

$$\begin{aligned} &p(X_{ns}|Y_s, X_s, Z_p) \\ &\quad\propto \int p(X_{ns}, Y_{ns}|Y_s, X_s, Z_p)\, dY_{ns} \\ &\quad\propto \int\int p(Y_{ns}|X, Y_s, Z_p, \theta) p(X_{ns}|Y_s, X_s, Z_p, \theta) \\ &\qquad \cdot p(Y_s, X_s, Z_p, |\theta) p(\theta)\, d\theta\, dY_{ns} \end{aligned}$$

under the assumption that $p(Y|X, \theta) = p(Y_s|X_{np}, \theta)$, that is, the model for Y given X holds in both the probability and nonprobability samples. Draws of $p(X_{ns}|Y_s, X_s Z_p)$ can be made under (18) and imputations of Y_{ns} made by alternating between draws of $p(\theta|Y, X)$ and $p(Y_{ns}|Y_s, X, \theta)$. Full implementation is made by obtaining L Bayesian draws of Y_s, X_s, Z_p, S draws of X_{ns} via a weighted FPBB, and finally M draws of Y_{ns} via standard multiple imputation methods (including, possibly, MRP models of the form used in Wang et al., 2015). Inference about Y, or, more typically, functions $Q \equiv Q(Y)$ can then be made via the approximate posterior distribution of Q given by $t_{L-1}(\overline{Q}_L, (1 + L^{-1})V_L)$ where $\overline{Q}_L = \frac{1}{LMS} \sum_l \sum_m \sum_s q^{(lms)}$ and $V_L = \frac{1}{L-1} \sum_l (\tilde{Q}^{(l)} - \overline{Q}_L)^2$ for $\tilde{Q}^{(l)} = \frac{1}{MS} \sum_m \sum_s q^{(lms)}$ and $q^{(lms)}$ is $Q(Y^{(lms)})$ where $Y^{(lms)} = (Y_s, Y_{ns}^{lms})$ for Y_{ns}^{lms} obtained from the sth imputation of the mth weighted FPBB of the lth BB. Details are available in Zhou, Elliott and Raghunathan (2016c, 2016a, 2016b), where empirical results are also presented.

5. CONCLUSION

Although selection of probability samples has been the standard for inference in finite populations for over 60 years, there are now many other sources of data that seem useful. Data obtained from convenient sources like internal business records or the internet are plentiful and tempting to use in estimation. Another mitigating factor is that selecting and maintaining probability samples becomes more difficult all the time, particularly when surveying households and persons. Because of these considerations, methods of statistical inference other than the design-based, repeated sampling approach are required.

Two alternatives are quasi-randomization and superpopulation modeling. In the former, probabilities of being included in a sample are estimated based on covariates. Unit-level covariates must be available for both the nonprobability sample and either a census of the population or a well-controlled, reference dataset that represents the nonsample units. The reference sample may or may not be a probability sample. But, in any case, the reference sample must permit inclusion probabilities to be estimated for the nonprobability units when the two covariate sources are combined. The superpopulation approach constructs models for y variables and uses them to predict finite population quantities like means or totals. The quasi-randomization and superpopulation approaches can also be combined to create estimators.

There are pros and cons to the two. In quasi-randomization, general inclusion probabilities can be estimated that are not specific to particular analytic y variables. Thus, they can apply to estimation for any y. An estimator generated for a particular y using the superpopulation approach may use a model specific to the y. Such an estimator can have a lower model-variance than a quasi-randomization estimator because it accounts for the population structure of the y. On the other hand, it can be model-biased if the superpopulation model is misspecified by, say, omitting important covariates. Although a quasi-randomization estimator may be unbiased with respect to repeated "pseudo-sampling," it can also be model-biased with respect to the superpopulation y model. Which of these two approaches is the most useful and statistically efficient appears to be an open question. Comparing these two approaches in the context of the many different types of data now available should be fertile ground for research.

Finally, a broader issue is whether there are certain situations in which nonprobability samples should be avoided altogether if some sort of probability sample is available. This consideration can be viewed through the "fit for purpose" framework (Baker et al., 2013b). Though perhaps most commonly used in defense of nonprobability samples by adding factors such as timeliness, accessibility and cost to the assessment of survey design, fit for purpose suggests that, when critical estimates of descriptive quantities such as means, quantiles or cell probabilities are required, nonprobability designs should be avoided or utilized only when it is reasonably certain that there are available covariates in both datasets related to the nonprobability selection mechanism that can be used to appropriately incorporate information from the nonprobability sample. If a sufficiently large probability sample is available for estimating descriptive statistics, methods to incorporate nonprobability data are likely not warranted. (We focus on descriptive statistics because there may be a smaller impact of nonprobability samples on model estimators. The effect on model estimators results from interactions between the probability of selection and the model effects, which we might presume to be less prevalent to the degree that models are correctly specified. However, the possibility of nonignorable selection related to model residuals remains present in nonprobability samples.) Developments of methods to assess the sensitivity of results to failures of the observed covariates to fully capture the selection mechanism for the nonprobability sample is, thus, yet another avenue for future research.

ACKNOWLEDGMENTS

The authors thank the Editor, Associate Editor and referees for their comments which led to considerable improvements in the paper.

REFERENCES

ALVAREZ, R., SHERMAN, R. and VAN BESELAERE, C. (2003). Subject acquisition for web-based surveys. *Polit. Anal.* **11** 23–43.

BAKER, R., BRICK, J., BATES, N., COUPER, M., COURTRIGHT, M., DENNIS, J., DILLMAN, D., FRANKEL, M., GARLAND, P., GROVES, R., KENNEDY, C., KROSNICK, J., LAVRAKAS, P., LEE, S., LINK, M., PIEKARSKI, L., RAO, K., THOMAS, R. and ZAHS, D. (2010). AAPOR report on online panels. *Public Opin. Q.* **74** 711–781.

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. and TOURANGEAU, R. (2013a). Report of the AAPOR Task Force on Non-probability Sampling. Technical report, American Association for Public Opinion Research, Deerfield, IL.

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. and TOURANGEAU, R. (2013b). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

BERZOFSKY, M., WILLIAMS, R. and BIEMER, P. (2009). Combining probability and non-probability sampling methods: Model-aided sampling and the O*NET data collection program. *Survey Practice*.

BETHLEHEM, J. (2010). Selection bias in web surveys. *Int. Stat. Rev.* **78** 161–188.

BINDER, D. and ROBERTS, G. (2009). Imputation of business survey data. In *Handbook of Statistics*, *Sample Surveys*: *Inference and Analysis*, *Volume* 29*B* (D. Pfeffermann and C. Rao, eds.). Elsevier, Amsterdam.

BRICK, J. (2015). Compositional model inference. In *Proceedings of the Section on Survey Research Methods* 299–307. Amer. Statist. Assoc., Alexandria, VA.

CAVALLO, A. and RIGOBON, R. (2016). The billion prices project: Using online prices for measurement and research. *The Journal of Economic Perspectives* 151–178.

CHEN, J. K.-T. (2015). Using LASSO to Calibrate Non-probability Samples using Probability Samples. Ph.D. thesis, Univ. Michigan, Ann Arbor, MI.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

CLEMENT, S. (2016). How the Washington Post-SurveyMonkey 50-state poll was conducted. Available at https://www.washingtonpost.com/news/post-politics/wp/2016/09/06/how-the-washington-post-surveymonkey-50-state-poll-was-conducted/.

COUPER, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods* **7** 145–156.

COWLING, D. (2015). Election 2015: How the opinion polls got it wrong. Available at http://www.bbc.com/news/

uk-politics-32751993. BBC News online; accessed 06-November-2016.

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

Dever, J. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Surv. Methodol.* **36** 45–56.

Dever, J. and Valliant, R. (2014). Estimation with nonprobability surveys and the question of external validity. In *Proceedings of Statistics Canada Symposium* 2014. Statistics Canada, Ottawa, ON.

Dever, J. and Valliant, R. (2016). GREG estimation with undercoverage and estimated controls. *Journal of Survey Statistics and Methodology* **4** 289–318.

Deville, J. (1991). A theory of quota surveys. *Surv. Methodol.* **17** 163–181.

Dong, Q., Elliott, M. and Raghunathan, T. (2014). A nonparametric method to generate synthetic populations to adjust for complex sample designs. *Surv. Methodol.* **40** 29–46.

Elliott, M. (2009). Combining data from probability and nonprobability samples using pseudo-weights. *Survey Practice*.

Elliott, M. R. and Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 595–609. MR2137256

Elliott, M. and Little, R. J. A. (2000). Model averaging methods for weight trimming. *J. Off. Stat.* **16** 191–209.

Elliott, M., Resler, A., Flannagan, C. and Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention* **42** 530–539.

Enten, H. (2014). Flying Blind Toward Hogan's Upset Win In Maryland. Available at http://fivethirtyeight.com/datalab/governor-maryland-surprise-brown-hogan/. FiveThirtyEight online; accessed 06-November-2016.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31** 799–815. MR2095753

File, T. and Ryan, C. (2014). Computer and internet use in the United States: 2013. Available at http://www.census.gov/content/dam/Census/library/publications/2014/acs/acs-28.pdf. US Census Bureau; accessed 06-November-2016.

Frost, S., Brouwer, K., Firestone-Cruz, M., Ramos, R., Ramos, M., Lozada, R., Magis-Rodriguez, C. and Strathdee, S. (2006). Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: Recruitment dynamics and impact on estimates of hiv and syphilis prevalence. *Journal of Urban Health* **83** 83–97.

Gile, K. J. and Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.

Gosnell, H. F. (1937). How accurate were the polls? *Public Opin. Q.* **1** 97–105.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statist. Sci.* **32** 206–226.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.

Holt, D. and Smith, T. M. F. (1979). Poststratification. *J. R. Stat. Soc., A* **142** 33–46.

Kaizar, E. (2015). Incorporating both randomized and observational data into a single analysis. *Annual Review of Statistics and Its Application* **2** 49–72.

Keiding, N. and Louis, T. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. R. Stat. Soc., A* **179** 319–376. MR3461587

Kohut, A., Keeter, S., Doherty, C., Dimock, M. and Christian, L. (2012). Assessing the representativeness of public opinion surveys. Available at http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/. Pew Research Center; accessed 06-November-2016.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys*. Wiley, New York.

LeBlanc, M. and Tibshirani, R. (1998). Monotone shrinkage of trees. *J. Comput. Graph. Statist.* **7** 417–433.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys uing propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **37** 319–343.

Liebermann, O. (2015). Why were the Israeli election polls so wrong? Available at http://www.cnn.com/2015/03/18/middleeast/israel-election-polls/. CNN online; accessed 06-November-2016.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **77** 237–250.

Little, R. J. A. (2003). Bayesian methods for unit and item nonresponse. In *Analysis of Survey Data* (R. Chambers and C. Skinner, eds.). Wiley, Chichester.

Lumley, T. and Scott, A. (2017). Fitting regression models to survey data. *Statist. Sci.* **32** 265–278.

Madigan, D., Stang, P., Berlin, J., Schuemie, M., Overhage, J., Suchard, M., Dumouchel, W., Hartzema, W. and Ryan, P. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application* **1** 11–39.

Murphy, J., Link, M., Childs, J., Tesfaye, C., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M. and Harwood, P. (2015). Social media in public opinion research. *Public Opin. Q.* **78** 788–794.

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** 558–625. MR0121942

O'Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 195–210. MR3234340

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241. MR0941020

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Surv. Methodol.* **18** 209–217.

Rivers, D. (2007). Sampling for web surveys. Amazon Web Services. Available at https://s3.amazonaws.com/yg-public/Scientific/Sample+Matching_JSM.pdf.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROYALL, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57** 377–387.

ROYALL, R. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference* (V. Godambe and D. Sprott, eds.). Holt, Rinehart, and Winston, Toronto.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SCHONLAU, M. and COUPER, M. (2017). Options for conducting web surveys. *Statist. Sci.* **32** 279–292.

SCHONLAU, M., VAN SOEST, A. and KAPTEYN, A. (2007). Are "Webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods* **1** 155–163.

SCHONLAU, M., WEIDMER, B. and KAPTEYN, A. (2014). Recruiting an Internet panel using respondent-driven sampling. *J. Off. Stat.* **30** 291–310.

SIMON, H. (1956). Rational choice and the structure of the environment. *Psychological Review* **63** 129–138.

SIRKEN, M. (1970). Household surveys with multiplicity. *J. Amer. Statist. Assoc.* **65** 257–266.

SMITH, T. M. F. (1976). The foundations of survey sampling: A review. *J. Roy. Statist. Soc. Ser. A* **139** 183–204. MR0445669

SMITH, T. M. F. (1983). On the validity of inferences from nonrandom samples. *J. R. Stat. Soc., A* **146** 394–403. MR0769995

SQUIRE, P. (1988). Why the 1936 literary digest poll failed. *Public Opin. Q.* **52** 125–133.

STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc., A* **174** 369–386. MR2898850

STURGIS, P., BAKER, N., CALLEGARO, M., FISHER, S., GREEN, J., JENNINGS, W., KUHA, J., LAUDERDALE, B. and SMITH, P. (2016). Report of the Inquiry into the 2015 British general election opinion polls. Available at http://eprints.ncrm.ac.uk/3789/1/Report_final_revised.pdf. accessed 06-November-2016.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

US ENERGY INFORMATION ADMINISTRATION (2016). Weekly petroleum status report. Available at https://www.eia.gov/petroleum/supply/weekly/pdf/appendixb.pdf. US Department of Energy online: accessed 06-November-2016.

VALLIANT, R. and DEVER, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **40** 105–137. MR2758301

VALLIANT, R., DEVER, J. A. and KREUTER, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York. MR3088726

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**.

VONK, T. W. E., VAN OSSENBRUGGEN, R. and WILLEMS, P. (2006). The effects of panel recruitment and management on research results. Available at https://www.esomar.org/web/research_papers/Web-Panel_1476_The-effects-of-panel-recruitment-and-management-on-research-results.php. ESOMAR; accessed 06-November-2016.

WANG, W., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* **31** 980–991.

ZHOU, H., ELLIOTT, M. and RAGHUNATHAN, T. (2016a). Multiple imputation in two-stage cluster samples using the weighted finite population Bayesian bootstrap. *Journal of Survey Statistics and Methodology* **4** 139–170.

ZHOU, H., ELLIOTT, M. and RAGHUNATHAN, T. (2016b). Synthetic multiple imputation procedure for multi-stage complex samples. *J. Off. Stat.* **32** 251–256.

ZHOU, H., ELLIOTT, M. and RAGHUNATHAN, T. (2016c). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics* **72** 242–252. MR3500593