Journal of Survey Statistics and Methodology (2020) 8, 4–36

A REVIEW OF CONCEPTUAL APPROACHES AND EMPIRICAL EVIDENCE ON PROBABILITY AND NONPROBABILITY SAMPLE SURVEY RESEARCH

CARINA CORNESSE*
ANNELIES G. BLOM
DAVID DUTWIN
JON A. KROSNICK
EDITH D. DE LEEUW
STÉPHANE LEGLEYE
JOSH PASEK
DARREN PENNAY
BENJAMIN PHILLIPS
JOSEPH W. SAKSHAUG
BELLA STRUMINSKAYA
ALEXANDER WENZ

There is an ongoing debate in the survey research literature about whether and when probability and nonprobability sample surveys produce accurate estimates of a larger population. Statistical theory provides a justification for confidence in probability sampling as a function of the survey design, whereas inferences based on nonprobability sampling are entirely dependent on models for validity. This article reviews the

Carina Cornesse is a postdoctoral researcher; she is with the Collaborative Research Center SFB 884 "Political Economy of Reforns" at the University of Mannheim, 68131 Mannheim, Germany. Annelies G. Blom is Professor for Data Science at the Department of Political Science, School of Social Sciences, and Principal Investigator of the German Internet Panel (GIP) at the Collaborative Research Center (SFB 884) "Political Economy of Reforms" at the University of Mannheim, 68131 Mannheim, Germany. David Dutwin is Senior Vice President; he is with NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, USA. Jon A. Krosnick is Frederic O. Glover Professor in Humanities and Social Sciences, Professor of Communication, and Professor of Political Science at Stanford University, 432 McClatchy Hall, 450 Serra Mall, Stanford, CA 94305, USA. Edith D. de Leeuw is Professor Emerita of Survey Methodology at the Department of Methodology and Statistics at Utrecht University, Sjoerd Groenmangebouw, Padualaan 14, 3584 CH Utrecht, Netherlands. Stéphane Legleye is Head of the Team "Drug Use in the General Population" at CESP, Inserm, faculté de médecine UVSQ, faculté de médecine, Université Paris-Saclay, Université Paris-Sud and Head of the "Household Living Conditions" Unit at Institut National de la Statistique et des Études

doi: 10.1093/jssam/smz041 Advance access publication 9 January 2020

current debate about probability and nonprobability sample surveys. We describe the conditions under which nonprobability sample surveys may provide accurate results in theory and discuss empirical evidence on which types of samples produce the highest accuracy in practice. From these theoretical and empirical considerations, we derive best-practice recommendations and outline paths for future research.

KEYWORDS: Accuracy; Nonprobability sampling; Probability sampling; Survey data quality; Survey inference; Weighting adjustments.

1. INTRODUCTION

In recent years, several cases of mispredicted election outcomes have made the news across the world. Prominent examples include the 2015 Israeli parliamentary election where the majority of polls predicted that Benjamin Netanyahu

Économiques, 88, avenue Verdier, CS 70058, 92541 Montrouge cedex, France. Josh Pasek is Associate Professor, Department of Communication and Media and Faculty Associate, Center for Political Studies, Institute for Social Research at the University of Michigan, 105 S. State Street, 5413 North Quad, Ann Arbor MI 48109. Darren Pennay is Founder, Board Member, and Executive Director Research, Methods & Strategy at the Social Research Centre, Campus visitor at Australian National University Centre for Social Research, Level 9, 277 William St, Melbourne Victoria Australia 3000 and Methods and Adjunct Professor with the Institute for Social Science Research (ISSR), University of Queensland. Benjamin Phillips is Senior Research Director, Survey Methodology at the Social Research Centre, Level 9, 277 William St, Melbourne Victoria Australia 3000 and Campus Visitor at the Centre for Social Research & Methods, Australian National University. Joseph W. Sakshaug is Distinguished Researcher, Head of the Data Collection and Data Integration Unit, and Acting Head of the Statistical Methods Research Department at the Institute for Employment Research (IAB), 104 Regensburgerstr. 90478 Nuremberg, Germany, University Professor of Statistics in the Department of Statistics at the Ludwig Maximilian University of Munich, 80539 Munich, Germany, and Honorary Full Professor in the School of Social Sciences at the University of Mannheim, 68131 Mannheim, Germany. Bella Struminskaya is assistant professor with the Department of Methodology and Statistics at the University of Utrecht, Sjoerd Groenmangebouw Padualaan 14, 3584 CH Utrecht, Netherlands. Alexander Wenz is a postdoctoral researcher with the Collaborative Research Center SFB 884 "Political Economy of Reforms" at the University of Mannheim, 68131 Mannheim, Germany and the Institute for Social and Economic Research at the University of Essex. The authors would like to thank the Collaborative Research Center (SFB) 884 "Political Economy of Reforms" (projects A8 and Z1), funded by the German Research Foundation (DFG), for organizing the Workshop on Probability-based and Nonprobability Survey Research in July 2018 at the University of Mannheim. In addition, the authors would like to thank Melvin John, Sabrina Seidl, and Nourhan Elsayed for their assistance with manuscript preparation. The authors are grateful to Andrew Mercer for valuable input and feedback on earlier versions of this manuscript. The order of authors is alphabetical after the lead author.

*Address correspondence to Carina Cornesse, Collaborative Research Center SFB 884 "Political Economy of Reforms," University of Mannheim, 68131 Mannheim, Germany, E-mail: carina.cornesse@uni-mannheim.de

would lose his presidency,[1] the 2016 Brexit referendum where the majority of polls predicted that Britain would vote to remain in the European Union,[2] and the 2016 US presidential election where the majority of polls predicted that Hillary Clinton would defeat Donald Trump.[3] When investigating potential reasons for these and other polling failures, researchers have pointed toward the fact that election polls based on probability samples usually reached more accurate predictions than election polls based on nonprobability samples (Sohlberg, Gilljam, and Martinsson 2017; Sturgis, Kuha, Baker, Callegaro, and Fisher 2018).

The finding that election polls based on probability samples usually reach better predictions than election polls based on nonprobability samples is not new. Already in the 1920s and 1930s, the scientific community debated which sampling design was better: probability sampling, as initially introduced by Arthur L. Bowley in 1906, or nonprobability sampling, as initially introduced by Anders N. Kiaer in 1895 (Lessler and Kalsbeek 1992; Bethlehem 2009). After several dramatic cases of mispredicted election outcomes in the United States (Literary Digest in 1936, see Crossley 1937; Gallup, Crossley, and Roper in the polling debacle of 1948, see Converse 1987), nonprobability sampling was identified as a principal cause of prediction inaccuracy and was replaced by probability sampling in most high-quality social research.

With the rise of the internet in the late 20th century, however, nonprobability sampling rose to popularity again as a fast and cheap method for recruiting online panels (Göritz, Reinhold, and Batinic 2000). However, nonprobability online panels face a number of challenges, such as noncoverage of people without internet access and selection bias due to the reliance on convenience samples of volunteers who might participate in multiple online panels (Bethlehem 2017). Despite these challenges, a vast amount of opinion polls today are conducted using nonprobability online panels (Callegaro, Villar, Yeager, and Krosnick 2014a; Callegaro, Baker, Bethlehem, Göritz, and Krosnick 2014b). In addition, nonpanel-based recruitment of online respondents, for example through river sampling (American Association for Public Opinion Research 2013) has been on the rise. As a result, the majority of survey data collected online around the world today rely on nonprobability samples (Callegaro et al. 2014a, 2014b).

Regardless of whether they are designed to predict election outcomes or to measure public opinion and regardless of whether they are conducted online or offline, when they are used for research and polling purposes, probability and

1. https://edition.cnn.com/2015/03/18/middleeast/israel-election-polls/index.html, accessed on November 30, 2019.

2. https://www.theguardian.com/politics/2016/jun/24/how-eu-referendum-pollsters-wrong-opinion-predict-close, accessed on September 30, 2019.

3. https://www.forbes.com/sites/startswithabang/2016/11/09/the-science-of-error-how-polling-botched-the-2016-election/#babf86437959, accessed on September 30, 2019.

nonprobability sample surveys often share a common goal: to efficiently estimate the characteristics of a large population based on measurements of a small subset of the population. Therefore, both probability and nonprobability sample surveys require that (i) the sampled units are exchangeable with non-sampled units that share the same measured characteristics, (ii) no parts of the population are systematically excluded entirely from the sample, and (iii) the composition of the sampled units with respect to observed characteristics either matches or can be adjusted to match the composition of the larger population (Mercer, Kreuter, Keeter, and Stuart 2017).

Despite their shared objective of providing accurate insights into a population of interest, probability and nonprobability sample surveys differ in a critical aspect. The key difference between probability and nonprobability sample surveys lies in the type and strength of the justification for why each approach should achieve accuracy. In the case of probability sample surveys, the justification is probability sampling theory, which is based on a set of established mathematical principles (Fisher 1925; Neyman 1934; Kish 1965). This sound theoretical basis makes it possible to compute the accuracy of estimates (e.g., in the form of confidence intervals or margins of error) and gives a universal validity to the estimation method. Furthermore, because providers of probability sample surveys routinely describe the details of the data-generating process, researchers are able to make adjustments that account for potential coverage, sampling, and nonresponse biases (e.g., Harter, Battaglia, Buskirk, Dillman, and English 2016; Blumberg and Luke 2017).

For nonprobability sample surveys, the justification for expecting accurate measurements rests on untested modeling assumptions that are based on a researcher's beliefs about the characteristics that make the sample different from the rest of the population and how those characteristics relate to the research topic. These assumptions can take different forms, such as quasi-randomization or superpopulation modeling (Deville 1991; Elliott and Valliant 2017). However, there is no general statistical theory of nonprobability sampling that justifies when and why accurate inferences can be expected: the validity is topic and survey dependent. Furthermore, online nonprobability sample providers often consider their data collection procedures to be proprietary, thus making it difficult or impossible to know what factors to include in any model aimed at correcting for selection bias in key estimates (Mercer et al. 2017).

This article is intended to move the debate about probability and nonprobability sample surveys forward: We first describe the assumptions that must be made in order to expect nonprobability samples to yield accurate results (section 2). We then summarize the empirical evidence on the accuracy of probability and nonprobability sample surveys to date (section 3). Finally, we conclude our review with practical recommendations and paths for future research (section 4).

2. CONCEPTUALIZATION OF NONPROBABILITY SAMPLING APPROACHES

The Total Survey Error (TSE; Groves and Lyberg 2010) framework that forms the bedrock of quality assessments for probability samples does not cleanly translate to the world of nonprobability sample data collection. Nonprobability samples do not involve a series of well-controlled and well-understood departures from a perfect sampling frame. Instead, most such samples rely on a collection of convenience samples that are aggregated and/or adjusted, with the goal of reducing the final difference between sample and population. Nonprobability samples cannot be evaluated by quantifying and summing the errors that occur at each stage of the sampling process. Instead, nonprobability samples can only be evaluated by assessing how closely the final modeled sample compares to the population in terms of various characteristics.

Although little has been proposed by way of formal statistical theory justifying the use of nonprobability samples for population inferences, the methods adopted for engaging with these samples suggest that a few combinations of assumptions could justify such an approach. In general, justifications can stem from four basic types of claims: (i) that any sample examining a particular question will yield the same inferences, (ii) that the specific design of the sample, as related to the questions at hand, will produce conclusions that mirror the population of interest, (iii) that a series of analytical steps will account for any differences between the sample and the population, and (iv) that the particular combination of sample and/or analytic approaches will produce accurate population estimates. Hence, the suggestion that any particular method, when used for a specific research question, is appropriate depends on underlying claims about the question of interest, the sample, and any adjustment procedures used.

2.1 Design Ignorability Due to the Question of Interest

In some cases, the method of sampling may be unrelated to the phenomena of interest. For instance, researchers may be trying to understand some process that occurs for all individuals, as in most physiological and some psychological studies. Under these circumstances, it may be reasonable to presume that any given group of individuals would behave like any other group of individuals unless there are known confounding factors. Researchers may claim that nonprobability sampling is irrelevant when answering particular questions for a few different reasons. They may believe that the process they are investigating is universal and, thus, that all people would behave similarly. On a more limited scope, they might contend that the particular phenomenon they are studying is appropriately distributed in any broad population sample and that the specific composition of that sample is unlikely to influence their conclusions.

The suggestion that some particular inference is unrelated to the sample being drawn could derive either from theoretical expectations of orthogonality or

prior empirical evidence of orthogonality. Of particular note, there are some theoretical and empirical reasons to believe that certain classes of inference may be more or less susceptible to sample imbalance. Some scholars have argued that trends over time in attitudes and behaviors should be less strongly dependent on sample composition than estimates of the distributions of those attitudes and behaviors (Page and Shapiro 1992). Similarly, a few empirical studies have found that relations between variables were more similar across probability and nonprobability samples than other types of estimates (Berrens, Bohara, Jenkins-Smith, Silva, and Weimer 2003; Pasek 2016). The claim that some kinds of inferences may be made equivalently well regardless of sampling strategy may sometimes be correct. The challenge, however, is determining when this might be the case.

2.2 Fit for Purpose Designs

Researchers do not need to establish that the question they are studying is impervious to sampling strategies to make the case that a nonprobability sample is appropriate for their inferences. Instead, they can assert that the design employed mitigates whatever biases might have emerged in the sampling process. The classic example of this type of argument stems from the use of quota samples. Quota samples are typically designed to ensure that the set of respondents matches the population on certain key demographic parameters. The idea underlying this approach is that the demographic parameters that form the basis for the quotas capture the sources of bias for researchers' inferences. To the extent that this is true, inferences made from quota samples will be accurate because all potential confounds are neutralized by the design. That is, the remaining error induced from the sampling process would be orthogonal to the questions of interest.

Notably, demographic quotas are not the only way to select respondents such that they reflect the population across key confounds. Scholars have proposed techniques ranging from matching individuals in nonprobability samples with individuals from probability samples as a means to recruit respondents to surveys (Rivers 2007) to blending together samples drawn from sources that have known opposing biases (Comer 2019). For any of these processes, if a researcher can be confident that the sample selection strategy eliminates all potential confounds for their particular question of interest, then the use of that sampling strategy is not only justifiable but will yield accurate inferences.

The challenge with these sorts of approaches is that the accuracy of critical assumptions can only really be established empirically. It is also unclear what to make of evidence that a particular conclusion is robust to a particular sampling decision. It may be the case that the nature of the question and/or type of inference renders that conclusion accurate for any similar question on a similarly derived sample or it might be that the particularities of a single analysis

happen to have yielded identical conclusions by mere chance. There is no obvious way to establish which of these is the case, though claims of empirical robustness are strengthened by a clear theoretical rationale.

2.3 Global Adjustment Approaches

In the next two sections, we describe modeling approaches that have been used to improve the accuracy of nonprobability sample data. Researchers have long known that even probability samples are sometimes inaccurate either by chance or due to variations in the likelihood that certain subgroups of the population will participate in a survey. For this reason, many statistical adjustment procedures typically used to adjust for systematic biases in probability samples have been adopted to adjust for selection biases in nonprobability samples. These approaches can be divided into two types: global adjustments and outcome-specific adjustments. *Global adjustments* refer to approaches that use a model to create a single adjustment that can be applied in any subsequent analysis, regardless of the outcome of interest. *Outcome-specific adjustments* tailor the adjustment model to a specific outcome of interest.

Regarding global adjustments, one commonly used approach is calibration weighting (Deville and Särndal 1992; Roshwalb, Lewis, and Petrin 2016; Santoso, Stein, and Stevenson 2016). Calibration weighting involves weighting the respondent pool such that the weighted sample totals of a certain characteristic correspond to known population totals of that same characteristic. The known population totals might come from census data, official statistics, or other data sources assumed to be of high quality. The procedure produces a global adjustment weight that can be applied to the analysis of any outcome variable. Using such weights amounts to making the assumption that once the known sources of deviation are accounted for in the adjustment procedure, the remaining errors will be unrelated to the likelihood that a particular unit in the population participated in the survey. This strategy presumes that the sampled units within specified population subgroups will be roughly equivalent to the nonsampled units within those subgroups with respect to any inferences that will be made with the data.

Although calibration weighting only requires access to population-level benchmark data, alternative global adjustment procedures make use of unit-level reference data to improve the accuracy of nonprobability sample estimates. One approach, known as sample matching, attempts to compose a balanced nonprobability sample by selecting units from a very large frame, such as a list of registered members of an opt-in panel, based on an array of auxiliary characteristics (often demographic) that closely match to the characteristics of units from a reference probability sample (Rivers 2007; Vavreck and Rivers 2008; Bethlehem 2016). The matching procedure, which may be performed before any units are invited to the nonprobability survey, relies on a

distance metric (e.g., Euclidean distance) to identify the closest match between pairs of units based on the set of common auxiliary characteristics.

Another approach that uses unit-level reference data is propensity score weighting. This approach is performed after the survey data have been collected from units in a nonprobability sample. The basic procedure is to vertically concatenate the nonprobability sample survey data with a reference dataset, typically a large probability sample survey. Then a model (e.g., logit or probit) is fitted using variables measured in both datasets to predict the probability that a particular unit belongs to the nonprobability sample (Rosenbaum and Rubin 1983, 1984; Lee 2006; Dever, Rafferty, and Valliant 2008; Valliant and Dever 2011). A weight is then constructed based on the inverse of this estimated inclusion probability and used in any subsequent analysis of the nonprobability survey data. Like calibration weighting, propensity score weighting only works if two conditions are satisfied: (i) the weighting variables and the propensity of response in the sample are correlated; and (ii) the weighting variables are correlated with the outcome variables of interest. A related approach is to use the concatenated dataset to fit a prediction model using variables measured in both datasets, which is then used to impute the values of variables for units in the nonprobability sample that were only observed for units in the reference probability sample (Raghunathan 2015).

All of the previously described adjustment approaches do not require the use of the reference data beyond the weighting, matching, or imputation steps and are discarded during the analysis of the nonprobability survey data. Alternative approaches combine both data sources and analyze them jointly. One such approach is pseudo design-based estimation (Elliott 2009; Elliott and Valliant 2017), where pseudo-inclusion probabilities are estimated for the nonprobability sample units based on a set of variables common to both nonprobability and probability samples. Different techniques may be used to estimate the inclusion probabilities. For example, one could concatenate the nonprobability and probability datasets and predict the probability of participating in the nonprobability survey, similar to the aforementioned propensity score weighting procedure. Alternatively, one could employ sample matching with both surveys and donate a probability sample unit's inclusion probability to the closest recipient match in the nonprobability sample. Once the pseudo-inclusion probabilities have been assigned to all nonprobability sample units, then these units can be treated as if they were selected using the same underlying sampling mechanism as the probability sample units. The datasets may then be combined and analyzed jointly by using the actual and pseudo weights. For variance estimation, Elliott and Valliant (2017) recommend the use of design-based resampling approaches, such as the bootstrap or jackknife to account for variability in both the pseudo weights and the target quantity of interest. For nonprobability samples that have an underlying cluster structure (e.g., different types of persons recruited from different web sites), cluster resampling approaches should be used.

Another approach to combining and analyzing probability and nonprobability samples jointly is blended calibration (DiSogra, Cobb, Chan, and Dennis 2011; Fahimi, Barlas, Thomas, and Buttermore 2015). Blended calibration is a form of calibration weighting that combines a weighted probability sample with an unweighted nonprobability sample and calibrates the combined sample to benchmark values measured on units from the weighted probability sample survey. The combined sample is then analyzed using the calibrated weights.

In summary, each of the previously described approaches produces a single global adjustment that can be applied to any analysis regardless of the outcome variable of interest. These methods entail the assumption that the selection mechanism of the nonprobability sample is ignorable conditional on the variables used in the adjustment method. For example, selection bias is assumed to be negligible or nonexistent within subgroups used in the calibration and propensity scoring procedures. It is further assumed that the adjustment variables in the reference data source (census or probability sample survey) are measured without error and are highly correlated with the target analysis variables and correlated with the probability to participate in the nonprobability sample survey. A very large and diverse reference data source with an extensive set of common variables is therefore needed to maximize the validity of these strong assumptions. One should always keep in mind that global adjustment procedures may improve the situation for some estimates but not others, and there is no guarantee that biases in the nonprobability data will be removed completely. In practice, one never knows with certainty that ignorability assumptions hold, although in some cases it may be possible to place bounds on the potential magnitude of bias through sensitivity analysis (Little, West, Boonstra, and Hu 2019).

2.4 Outcome-Specific Adjustment Approaches

Turning now to outcome-specific adjustment approaches, these approaches utilize adjustment models that are tailored to a specific outcome variable. That is, they adjust for the selection mechanism into a nonprobability sample with respect to a given outcome variable, Y, which is of interest to the researcher. Such approaches attempt to control for variables that govern the selection process and are correlated with the target outcome variable. One example of such a framework is the notion that probability and nonprobability samples both constitute draws from a hypothetical infinite "superpopulation" (Deville 1991). The goal of the analyst is then to model the data-generating process of the underlying superpopulation by accounting for all relevant variables in the analysis model. In practice, this means that a researcher may fit a prediction model for some analysis variable Y based on the sample at hand, which is then used to predict the Y's for the nonsampled units. The sampled and nonsampled units are then combined to estimate the quantity of interest (e.g., mean, total,

regression coefficient) for Y in the total population (Elliott and Valliant 2017). The key assumptions of this approach are that the analysis variable, Y, is explained through a common model for the sampled and nonsampled units and that all parameters governing the superpopulation model are controlled for in the analysis model. The approach also requires the availability of auxiliary data on the population to make predictions for the nonsampled units. Variance estimation for the predictions can be implemented using a variety of frequentist methods, including jackknife and bootstrap replication estimators as described in Valliant, Dorfman, and Royall (2000).

Another model-based approach, which can be applied in a superpopulation framework, is model-assisted calibration. This approach involves constructing calibrated weights by using a model to predict the values for a given analysis variable (Wu and Sitter 2001). The calibrated weights are generated based on constraints placed on the population size and population total of the predicted values. Various model selection approaches (e.g., LASSO) have been proposed for parsimonious modeling of the analysis variable (Chen, Valliant, and Elliott 2018). A key assumption of the method is that the model is correctly specified and capable of making reliable predictions across different samples of the population. It is also assumed that all relevant superpopulation parameters are included in the model if the method is implemented in a superpopulation framework.

Multilevel regression and poststratification is another approach used to estimate a specific outcome of interest from a nonprobability sample (Wang, Rothschild, Goel, and Gelman 2015; Downes, Gurrin, English, Pirkis, Currier 2018). The basic idea is to fit a multilevel regression model predicting an outcome given a set of covariates. The use of a multilevel model makes it possible to incorporate a large number of covariates or high order interactions into the prediction model (Ghitza and Gelman 2013). The model is then used to estimate the mean value for a large number of poststratification cells defined by the cross-classification of all variables used in the regression model. It is necessary that the relative size of each cell in the population is known or can be reliably estimated from external data sources such as a census or population registry. Population quantities are estimated by aggregating these predicted cell means with each cell weighted proportionally to its share of the population. The multilevel regression model allows for cell-level estimates to be generated even when few units exist within the sample cells. The method can also be implemented in a Bayesian hierarchical modeling framework (Park, Gelman, and Bafumi 2004). Like all of the previously mentioned model-based approaches, the model is assumed to control for all variables that affect the probability of inclusion in the nonprobability sample. The method also requires good model fit, and large cell sizes are preferable to generate robust cell-level estimates. For additional hierarchical and Bayesian modeling approaches that have been proposed to estimate outcomes from nonprobability samples, we

refer to Ganesh, Pineau, Chakraborty, and Dennis (2017), Pfeffermann (2017), and Pfeffermann, Eltinge, and Brown (2015).

Collectively, these approaches to using nonprobability sample data to reach population-level conclusions depend on combinations of assumptions about ignorable errors and the availability of information about the sources of nonrandomness in respondent selection that can be used to adjust for any errors that are not ignorable. Although these dependencies are not fundamentally different from the assumptions underlying the use of probability samples, they are more difficult to rely on as we know little about the factors that lead individuals to become members of nonprobability samples.

3. THE ACCURACY OF PROBABILITY AND NONPROBABILITY SAMPLE SURVEYS

Several studies have empirically assessed the accuracy of probability and nonprobability sample surveys by comparing survey outcomes to external population benchmarks (see table 1). The vast majority of these studies concluded that probability sample surveys have a significantly higher accuracy than nonprobability sample surveys. Only a few studies have found that probability sample surveys do not generally have a significantly higher accuracy than nonprobability sample surveys.

Table 1 provides a list of the studies that are included in our overview. A key inclusion requirement is that the studies contain comparisons of probability and nonprobability sample surveys with external population benchmarks. We do not include any studies published in languages other than English or that contain sample accuracy assessments of probability and nonprobability sample surveys only as a minor byproduct or that compare probability sample surveys with nonsurvey convenience samples, such as Amazon MTurk (e.g., Coppock and McClellan, 2019).

3.1 Initial Accuracy Comparisons for Probability and Nonprobability Sample Surveys

As table 1 shows, a number of studies have demonstrated that probability sample surveys have a higher accuracy than nonprobability sample surveys. The higher accuracy of probability sample surveys has been demonstrated across various topics, such as voting behavior (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Sturgis et al. 2018), health behavior (Yeager, Krosnick, Chang, Javitz, and Levendusky 2011), consumption behavior (Szolnoki and Hoffmann 2013), sexual behavior and attitudes (Erens, Burkill, Couper, Conrad, and Clifton 2014; Legleye, Charrance, Razafindratsima, Bajos, and Bohet 2018), and socio-demographics (Malhotra and Krosnick 2007; Chang

Table 1. Studies on the Accuracy of Probability (PS) and Nonprobability (NPS) Sample Surveys

Study	Country	Benchmark	PS modes studied	PS more accurate than NPS?[a]	Did (re)weighting sufficiently reduce NPS bias?[b]
Blom et al. (2018)	Germany	Census data Election data	F2F, web	Yes (univariate)	No (raking)
MacInnis et al. (2018)	USA	High-quality PS	Phone, web	Yes (univariate)	No (poststratification)
Dassonneville et al. (2018)	Belgium	Census data	F2F	Yes (univariate)	No (unspecified approach) – univariate
		Election outcome		No (multivariate)	N/A – multivariate
Legleye et al. (2018)	France	Census data	Phone	Yes (univariate)	N/A
Pennay et al. (2018)	Australia	High-quality PS	Phone	Yes (univariate)	No (poststratification)
Sturgis et al. (2018)	UK	Election outcome	F2F	Yes (univariate)	No (raking, propensity weighting, matching)
Dutwin and Buskirk (2017)	USA	High-quality PS	F2F, phone	Yes (univariate)	No (propensity weighting, matching, raking)
Sohlberg et al. (2017)	Sweden	Election outcome	Phone	Yes (univariate)	N/A
Brüggen et al. (2016)	Netherlands	Population register	F2F, web	Yes (univariate) Yes (bivariate)	No (Generalized Regression Estimation) – univariate, bivariate
Kennedy et al. (2016)	USA	High-quality PS	Web, mail	No (univariate) No (multivariate)	N/A – univariate N/A – multivariate
Pasek (2016)	USA	High-quality PS	Phone	Yes (univariate) No (bivariate) Yes (longitudinal)	No (raking, propensity weighting) – univariate, bivariate, longitudinal

Continued

Table 1. *Continued*

Study	Country	Benchmark	PS modes studied	PS more accurate than NPS?[a]	Did (re)weighting sufficiently reduce NPS bias?[b]
Gittelman et al. (2015)	USA	High-quality PS	Phone	No (univariate)	No (poststratification)
Ansolabehere and Schaffner (2014)	USA	High-quality PS Election outcome	Phone, mail	No (univariate) No (multivariate)	N/A
Erens et al. (2014)	UK	High-quality PS	CASI	Yes (univariate)	N/A
Steinmetz et al. (2014)	Netherlands	High-quality PS	Web	No (univariate) No (bivariate)	Yes (propensity weighting) – univariate, bivariate
Ansolabehere and Rivers (2013)	USA	High-quality PS Election outcome	F2F	No (univariate)	N/A
Szolnoki and Hoffmann (2013)	Germany	High-quality PS	F2F, phone	Yes (univariate)	N/A
Chan and Ambrose (2011)	Canada	Unspecified	Web	No (univariate)	N/A
Scherpenzeel and Bethlehem (2011)	Netherlands	Election outcome Unspecified	F2F, web	Yes (univariate)	N/A
Yeager et al. (2011)	USA	High-quality PS	Phone, web	Yes (univariate)	No (poststratification)
Chang and Krosnick (2009)	USA	High-quality PS	Phone, web	Yes (univariate)	No (raking)
Walker, Pettit, and Rubinson (2009)[c]	USA	Unspecified	Phone, mail	Yes (univariate)	N/A
Loosveldt and Sonck (2008)	Belgium	Census data	F2F	No (univariate)	No (poststratification, propensity weighting)

Malhotra and Krosnick (2007)	USA	High-quality PS	F2F	Yes (univariate)	No (unspecified)
Berrens et al. (2003)	USA	High-quality PS	Web, phone	No (univariate) No (multivariate)	Yes (raking, propensity weighting) – univariate, multivariate

[a]The study results reported here are based on initial comparisons of accuracy as reported by the authors. Some studies use raw (unweighted) data in their initial comparisons, while others use data that are already weighted using either weights that the authors calculated themselves or the weights that a survey vendor delivered with the data.
[b]We report whether (re)weighting has sufficiently reduced bias based on the authors' own judgments as reported in their conclusions. This column includes studies that compare raw (unweighted) data in their initial comparison of accuracy with the data after a weighting procedure was performed and studies that used weighted data in their initial comparison of accuracy and reweighted the data in subsequent comparisons. Studies that only reported raw data or only weighted data, without using any reweighting approaches, are labelled "not applicable" (N/A) in this column.
[c]As reported by Callegaro et al. (2014a, 2014b).

and Krosnick 2009; Yeager et al. 2011; Szolnoki and Hoffmann 2013; Erens et al. 2014; Dutwin and Buskirk 2017; MacInnis, Krosnick, Ho, and Cho 2018). In addition, the higher accuracy of probability sample surveys has been found across a number of countries, such as Australia (Pennay, Neiger, Lavrakas, and Borg 2018), France (Legleye et al. 2018), Germany (Szolnoki and Hoffmann 2013; Blom, Ackermann-Piek, Helmschrott, Cornesse, Bruch, and Sakshaug 2018), the Netherlands (Scherpenzeel and Bethlehem 2011; Brüggen, van den Brakel, and Krosnick 2016), Sweden (Sohlberg et al. 2017), the United Kingdom (Sturgis et al. 2018), and the United States (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Yeager et al. 2011; Dutwin and Buskirk 2017; MacInnis et al. 2018). Furthermore, the higher accuracy of probability sample surveys has been shown over time, with the first study demonstrating higher accuracy of probability sample surveys in 2007 (Malhotra and Krosnick, 2007) to the most recent ones in 2018 (Blom et al. 2018; Legleye et al. 2018; MacInnis et al. 2018; Sturgis et al. 2018). All of these studies from different times and countries that focused on different topics reached the conclusion that probability sample surveys led to more accurate estimates than nonprobability samples.

A prominent study from this line of research is Yeager et al. (2011). In a series of analyses of surveys conducted between 2004 and 2008 in the United States, the authors found that probability sample surveys were consistently more accurate than nonprobability sample surveys across many benchmark variables (primary demographics such as age, gender, and education; secondary demographics such as marital status and homeownership; and nondemographics such as health ratings and possession of a driver's license), even after poststratification weighting.

Another influential study that is based on a particularly rich database was conducted with four probability sample face-to-face surveys, one probability sample online survey, and eighteen nonprobability sample online surveys (the so-called NOPVO [National Dutch Online Panel Comparison Study] project) in 2006 to 2008 in the Netherlands (Brüggen et al. 2016). In line with the findings from Yeager et al. (2011), the authors found that the probability face-to-face and internet surveys were consistently more accurate than the nonprobability internet surveys across a variety of sociodemographic variables and variables on health and life satisfaction.

A recent example that focuses on the current debate about why polls mispredict election outcomes was published by Sturgis et al (2018). Investigating why British polls mispredicted the outcome of the 2015 UK general election, Sturgis et al. (2018) examined data from twelve British pre-election polls and assessed a number of potential reasons for the polling debacle, such as whether voters changed their mind at the last minute (i.e., "late swing"), whether the mode in which the surveys were conducted (telephone or online) played a role, or whether polls failed to correct for the common problem of overestimating voter turnout (i.e., proper turnout weighting). The authors found that sample

inaccuracy due to nonprobability sample selection likely had the biggest impact on the mispredictions.

Although most studies have demonstrated that accuracy is higher in probability sample surveys than in nonprobability sample surveys, two studies have yielded mixed findings depending on the type of estimate examined (Pasek 2016; Dassonneville, Blais, Hooghe, and Deschouwer 2018). Both studies show that accuracy is higher in probability sample surveys for univariate estimates. Pasek (2016) also reports higher accuracy of probability sample surveys for longitudinal analyses. However, these studies found no difference in accuracy regarding bivariate (Pasek 2016) and multivariate (Dassonneville et al. 2018) estimates.

Several studies have found no consistent superiority in accuracy of probability or nonprobability sample surveys over one another. These studies generally yielded mixed findings: a probability sample survey was found to be more accurate than some but not all nonprobability sample surveys examined (Kennedy, Mercer, Keeter, Hatley, McGeeney, and Gimenez 2016); or probability sample surveys were shown to be more accurate than nonprobability sample surveys on some variables while nonprobability sample surveys were more accurate than probability sample surveys on other variables (Loosveldt and Sonck 2008; Chan and Ambrose 2011; Steinmetz, Bianchi, Tijdens, and Biffignandi 2014). In some of the studies, the authors speculated that it might be survey mode rather than the sampling design that led to comparable accuracy (Berrens et al. 2003; Ansolabehere and Schaffner 2014; Gittelman, Thomas, Lavrakas, and Lange 2015).

In general, it should be noted that sample accuracy assessments face some challenges. One common challenge of such studies is to disentangle mode effects (i.e., measurement bias) from sampling effects (i.e., selection bias). This challenge occurs because probability sample surveys are usually conducted offline (e.g., via face-to-face or telephone interviews), whereas nonprobability sample surveys are usually conducted online via nonprobability online panels. However, several studies show that it is possible to disentangle the mode effect from the sampling effect by comparing offline probability sample surveys with online probability sample surveys (mode effect) and comparing online probability sample surveys to online nonprobability sample surveys (sampling effect). The majority of these studies conclude that both offline and online probability sample surveys are more accurate than nonprobability online sample surveys (Chang and Krosnick 2009; Scherpenzeel and Bethlehem 2011; Yeager et al. 2011; Brüggen et al. 2016; Dutwin and Buskirk 2017; Blom et al. 2018; MacInnis et al. 2018).

A related challenge that sample accuracy assessments of probability and nonprobability sample surveys face is the question of how to measure accuracy in a way that accounts for both sampling variability and systematic bias. This challenge occurs because often there is only one probability sample survey and one nonprobability sample survey available for sample accuracy assessment.

However, several large-scale studies that compare a larger number of probability sample surveys with a larger number of nonprobability sample surveys have found that probability sample surveys are consistently more accurate than nonprobability sample surveys (Yeager et al. 2011; Brüggen et al. 2016; Blom et al. 2018; MacInnis et al. 2018; Sturgis et al. 2018). This suggests that although surveys vary on a number of design factors other than their sampling design (e.g., incentive schemes, contact frequency) and might sometimes be more or less accurate by chance, samples are generally more likely to have higher accuracy if they are based on probability sampling procedures rather than nonprobability sampling procedures.

Another common challenge that sample accuracy assessments face is the availability of appropriate gold standard benchmarks. In the current literature, the most commonly used benchmarks are large-scale, high-quality probability sample surveys. A typical example of such a benchmark is the US American Community Survey (Yeager et al. 2011 and MacInnis et al. 2018). Other benchmarks used in the literature are population census data (Legleye et al. 2018), election outcomes (Sohlberg et al. 2017), and population register data (Brüggen et al. 2016).

All of these benchmarks have advantages and disadvantages. A key advantage of using a large-scale, high-quality probability sample survey is that the set of variables available for comparisons usually includes not only sociodemographic variables but also substantive variables on attitudes and behavior. A disadvantage is that large-scale, high-quality probability sample surveys are surveys themselves and might therefore contain typical survey errors, such as coverage, sampling, and nonresponse errors (Groves and Lyberg 2010).

Census data and population register data have the advantage of not suffering from survey errors. However, such data are often not available for the current year and might therefore be outdated at the time of the study. Population register data might also be outdated, for example, if immigration, emigration, births, and deaths are not captured in a timely manner. In addition, census data and population register data are typically limited to a small set of sociodemographic characteristics. With regard to election outcomes, an advantage is that they are key variables of substantive interest to many social scientists. However, if survey data fail to accurately predict election outcomes, there are many potential explanations for this besides the sampling approach; see Sturgis et al. (2018) for a list of reasonable explanations tested after the British polling disaster of 2015.

3.2 Weighting Approaches to Reduce Bias in Nonprobability Sample Surveys

Many studies examining the accuracy of probability and nonprobability sample surveys have attempted to eliminate biases in nonprobability sample surveys.

The majority of these studies found that weighting did not sufficiently reduce bias in nonprobability sample surveys (see table 1). Generally speaking, probability sample surveys were found to be more accurate than nonprobability sample surveys even after (re)weighting.

Although the majority of studies found that weighting did not reduce the bias in nonprobability sample surveys sufficiently (table 1), some studies showed that weighting did reduce the bias somewhat. However, whether researchers considered the bias to be sufficiently reduced by weighting varied from study to study. For example, Berrens et al. (2003) considered the bias in a nonprobability sample survey sufficiently reduced even though an estimate of mean household income deviated from the benchmark by between 4.8 percentage points (after propensity weighting) and 11.9 percentage points (after raking). A study concluding that weighting approaches sufficiently reduced bias in nonprobability sample surveys also reported that weighting increased the variance of the estimates significantly (Steinmetz et al. 2014).

Most of the studies listed in table 1 focus on the success of weighting procedures to reduce bias in nonprobability sample surveys. Only a few studies listed in table 1 have also assessed the success of weighting procedures in reducing bias in probability sample surveys. For example, MacInnis et al. (2018) reported that weighting reliably eliminated the small biases present in unweighted probability sample survey data. This is in line with research by Gittelman et al. (2015), who showed that poststratification weighting successfully reduced biases in a probability sample survey but was less successful across a number of nonprobability survey samples and even increased bias in one instance.

The studies listed in table 1 used one or more common weighting procedures, such as raking, poststratification, and propensity weighting, to improve the accuracy of nonprobability sample survey measurements. Several other studies in the literature have investigated the effectiveness of various weighting procedures in reducing bias in nonprobability sample surveys, without examining the accuracy of probability sample surveys. Table 2 provides an overview of these studies. A key inclusion requirement is that studies assess whether weighting nonprobability sample surveys reduced bias as compared with unweighted estimates and if so, to what extent. We again exclude all studies published in languages other than English and studies that contain assessments of weighting procedures for nonprobability sample surveys only as a minor byproduct or that examine nonsurvey data.

In general, the majority of the studies that investigated the effectiveness of various weighting procedures in reducing bias in nonprobability sample surveys (table 2) reached the same conclusion as the studies that assessed weighting approaches in both probability and nonprobability sample surveys (table 1): weighting does not sufficiently reduce bias in nonprobability sample surveys. Only a few studies found that weighting could sufficiently reduce bias in nonprobability sample surveys.

As table 2 shows, two of the studies that documented a sufficient reduction in the bias of nonprobability sample surveys applied multilevel regression and poststratification weights (Wang et al. 2015; Gelman, Goel, Rothschild, and Wang 2017), one study applied model-based poststratification (Goel, Obeng, and Rothschild 2015), and one study applied propensity weighting and calibration (Lee and Valliant 2009). Two of these studies used weighting to adjust nonprobability sample survey data to accurately predict election outcomes after the actual election outcomes were already known, which reduces confidence in the conclusions (Wang et al. 2015; Gelman et al. 2017).

In sum, the majority of the research on weighting and accuracy finds that the inaccuracy of nonprobability samples cannot be reliably solved by weighting procedures. Some authors conducting such studies also offer explanations as to why their attempts to achieve accurate estimates from weighted nonprobability samples were not successful. Mercer, Lau, and Kennedy (2018), for instance, show that complex weighting procedures outperform basic weighting procedures. Furthermore, the authors show that to get accurate estimates from nonprobability sample surveys by weighting, the availability of variables that predict the outcome of interest are more important than which statistical method is used.

4. CLOSING REMARKS

In this article, we have reviewed conceptual approaches and empirical evidence on probability and nonprobability sample surveys. Probability sampling theory is well established and based on sound mathematical principles, whereas nonprobability sampling is not. Although there are potential justifications for drawing inferences from nonprobability samples, the rationale for many studies remains unarticulated, and the inferences from nonprobability sample surveys generally require stronger modeling assumptions than are necessary for probability samples. The basic problem with these modeling assumptions remains that they cannot be tested. We have therefore proposed a conceptual framework for nonprobability sample surveys to explicate these modeling assumptions, including practical suggestions about when it might be justified to make such assumptions (section 2).

In addition, we have summarized the empirical evidence on the accuracy of probability and nonprobability sample surveys (section 3). Our literature overview shows that, even in the age of declining response rates, accuracy in probability sample surveys is generally higher than in nonprobability sample surveys. There is no empirical support to the claim that switching to nonprobability sample surveys is advisable because the steadily declining response rates across the globe compromise probability sample survey data quality. Based on the accumulated empirical evidence, our key recommendation is to continue to rely on probability sample surveys.

Table 2. Studies Exclusively Investigating Weighting Procedures to Reduce Bias in Nonprobability Sample Surveys

Study	Benchmark	Does weighting sufficiently reduce bias in NPS?[a]
Mercer et al. (2018)	High-quality PS	No (raking, propensity weighting, matching)
Smyk, Tyrowicz, and Van der Velde (2018)	High-quality PS	No (propensity weighting)
Gelman et al. (2017)	High-quality PS	No (raking)
	Election outcome	Yes (multilevel regression and poststratification)
Goel et al. (2015)	High-quality PS	No (raking) Yes (model-based poststratification)
Wang et al. (2015)	Election outcome	Yes (multilevel regression and poststratification)
Lee and Valliant (2009)	High-quality PS	Yes (propensity weighting, calibration)
Schonlau, van Soest, Kapteyn, and Couper (2009)	High-quality PS	No (propensity weighting, matching)
Schonlau, van Soest, and Kapteyn (2007)	PS	No (propensity weighting)
Lee 2006	High-quality PS	No (propensity weighting)
Duffy, Smith, Terhanian, and Bremer (2005)	High-quality PS	No (raking, propensity weighting)
Schonlau, Zapert, Simon, Sanstad, and Marcus (2004)	PS	No (poststratification, propensity weighting)
Taylor (2000)	High-quality PS + Election outcome	Yes (raking, propensity weighting)

[a]We report whether weighting sufficiently reduced bias based on the authors' own judgments as reported in their conclusions.

In the case that only nonprobability sample survey data are available, we recommend carefully choosing among the various modeling approaches based on their underlying assumptions. In order for researchers to be able to justify the modeling approaches used, we recommend obtaining as much information as possible about the data-generating process (see the transparency recommendations in the appendix).

Apart from these key recommendations, this report also shows that there are gaps in the existing literature. To be able to evaluate if and when nonprobability sample surveys can be used as an alternative to probability sample surveys,

we need more insights into the success of nonprobability sample surveys in producing accurate estimates in bivariate and multivariate analyses, longitudinal analyses, and experimental research settings. In addition, we need more research into variance estimation and advanced weighting techniques, in particular with regard to collecting and utilizing weighting variables that are correlated with key survey variables and the data-generating process.

Finally, this report shows that there is great variability across nonprobability sample surveys. Therefore, we would like to end this report with a general call for more transparency in the survey business. For users, researchers and clients, it can be difficult to decide which vendors to trust with data collection. As long as many vendors are unwilling to disclose necessary information about the data collection and processing, researchers will remain unable to make informed decisions about vendors and will lack the information necessary to understand the limitations of their collected data. The availability of research reports that outline the methodology used for data collection and manipulation is therefore of utmost importance (Bethlehem 2017, Chapter 11).

As clients, we can reward vendors who are willing to provide more methodological information to us. As a practical matter, this requires being explicit about our needs prior to contracting and in reaching out to a broad set of nonprobability sample providers. Another form of action we can take as clients is to ensure that when contracting vendors who belong to an organization with relevant standards, such as ESOMAR, or that have ISO 20252 or 26362 certification, these vendors disclose information as required by the respective code or certification (see the appendix and Bethlehem [2017, Chapter 12] for more information on relevant transparency guidelines and standards).

Appendix: Transparency Guidelines

Various codes of ethics and guidelines address the disclosure of methodological information for online panels (e.g., Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute eV 2001; Interactive Marketing Research Organization [IMRO] 2015; International Organization for Standardization [ISO] 2009, 2012; ESOMAR 2012, 2014, 2015a, 2015b; American Association for Public Opinion Research 2015). The lack of information available from some online panel vendors can unfortunately make it impossible for researchers to comply with their own codes or certifications (e.g., ISO 2012, §4.5.1.4; American Association for Public Opinion Research 2015, §III.A.5–8, 12, 14). The unwillingness of some vendors to disclose necessary information is unfortunate for all concerned. Consumers of research are denied information needed to form an opinion as to the quality of the research. Researchers are unable to make informed decisions about vendors and lack information needed to understand the limitations of their data. Further, vendors themselves are unable to benefit from the methodological advances that would follow from greater availability of information on panel operations.

What can be done? Given the abundance of guidelines—several of which are particularly helpful (IMRO 2015; ESOMAR 2015a)—there is little need for another extensive set of recommendations. Therefore, we summarize common reporting recommendations in table A1 and, for the most part, refer the reader to existing guidelines addressing these points (particularly helpful sources are italicized). At times, we make additional recommendations that may go beyond reporting requirements; these contain the words, "We recommend."

In table A2, we also list relevant case-level data that researchers may wish to ensure that the vendor will be able to provide. Such data will likely be most useful to the researcher where they request all data from all cases invited to participate in the survey, not only those that completed the survey or were not removed for data quality reasons.

Table A1. Reporting Recommendations

Construct	Description	References
Universe	What universe does the panel represent? How well does the panel cover that universe?	American Association for Public Opinion Research (2015, §III.A.3); ESOMAR (2014, §5.2); ISO (2009, §4.4.1, §4.7); ISO (2012, §4.5.1.4, §7.2.e, §7.2.f.2)
Sampling frame	What method or methods are used to recruit panel members? We recommend that the percentage of the sample recruited from different sources be included (e.g., online advertisement, email list, aggregator website).	American Association for Public Opinion Research (2015, §III.A.5-8); ESOMAR (2012, §2, §5); ESOMAR (2014, §5.2); ESOMAR (2015a, §3.7); ESOMAR (2015b, §5.2, §6, §6.1); IMRO (2015, §4); ISO (2012, §4.5.1.4, §7.2.f)
Panel size	How large is the panel? What is the percentage of active members? How is "active" defined?	IMRO (2015, §14, §17); ISO (2009, §4.4.2)
Replenishment and retirement	How frequently are new panel members added? Are panel members retired from the panel by the vendor? If so, what criteria are used to determine retirement? What is the panel's annual "churn rate" (the percentage of all panelists who are voluntarily or involuntarily retired each year)?	IMRO (2015, §18, §19); ISO (2009, §4.5.1)
Aggregator sites	Does the panel draw on aggregator sites, which are sites that allow respondents to select from multiple surveys for which they may qualify (see IMRO 2015)	IMRO (2015, §6)
Sample from other panels	Was sample from other providers used in this study?	ESOMAR (2012, §4)
Blending	If there is more than one sample source, how are these blended together?	ESOMAR (2012, §3); ESOMAR (2015b, §6.1)

Sample drawn from frame	How was the sample selected from the panel? Any quotas or other specifications used in selection of the sample should be specified. How many units were drawn?	ESOMAR (2012, §7); ESOMAR (2015a, §3.7); ESOMAR (2015b, §6.2); IMRO (2015, §23); ISO (2009, 4.6.1); ISO (2012, 4.5.1.4)
Routers	Was a router used? If so, how did the vendor determine which surveys are considered for a participant? What is the priority basis for allocating participants to surveys? How is potential bias from the use of a router addressed?	American Association for Public Opinion Research (2015, §III.A.14); ESOMAR (2012, §8-§11); ESOMAR (2015a, §3.7); ESOMAR (2015b, §6.2)
Identity validation	How does the vendor confirm participants' identities? How does the vendor detect fraudulent participants?	American Association for Public Opinion Research (2015, §III.A.17); ESOMAR (2012, §22); ESOMAR (2015a, §3.8); ESOMAR (2015b, §6.1); IMRO (2015, §24); ISO (2009, §4.3.4.2, §4.3.4.3)
Within-panel deduplication	What procedures does the panel employ to ensure that no research participant can complete the survey more than once?	American Association for Public Opinion Research (2015, §III.A.17); ESOMAR (2015a, §3.2, §3.8); IMRO (2015, §25)
Cross-panel deduplication	If multiple sample sources are used, how does the panel ensure that no research participant can complete the survey more than once?	American Association for Public Opinion Research (2015, §III.A.17); ESOMAR (2012, §3); ESOMAR (2015a, §3.8); IMRO (2015, §29)
Category exclusion	How are panel members who have recently completed a survey on the same topic treated? If they are excluded, for how long or using what rules?	IMRO (2015, §15); ISO (2009, §4.6.2)
Participation limits	How often can a panel member be contacted to take part in a survey in a specified period?	ESOMAR (2012, §19, §20); ESOMAR (2015a, §3.8); IMRO (2015, §16); ISO (2009, §4.5.1, §4.6.2)

Continued

Table A1. *Continued*

Construct	Description	References
Other data quality checks	What checks are made for satisficing and fraud? These might include checks for speeding and patterned responses, trap questions, counts of missing or nonsubstantive responses (e.g., don't know), device fingerprinting or logic checks.	American Association for Public Opinion Research (2015, §III.A.17); ESOMAR (2012, §18); ESOMAR (2015a, §3.3, §3.8); ESOMAR (2015b, §6, §6.1, §6.4); IMRO (2015, §30); ISO (2009, §4.6.6, §4.7); ISO (2012, §7.2.k)
Panel profile	How often are profiles updated?	IMRO (2015, §21)
Fieldwork dates	When was the survey fielded? If there were multiple sample sources, did the field period vary between sources?	American Association for Public Opinion Research (2015, §III.A.4); ESOMAR (2014, §5.2); ESOMAR (2015a, §3.8); ESOMAR (2015b, §6.3); ISO (2009, §4.7); ISO (2012, §4.8.3, §7.2.g)
Invitations and reminders	What text is used for invitations and any reminders? We recommend disclosing the maximum number of reminders sent.	American Association for Public Opinion Research (2015, §III.A.16); ESOMAR (2012, §13); ISO (2009, §4.7)
Incentives	What incentives, if any, are offered to panel members?	American Association for Public Opinion Research (2015, §III.A.16); ESOMAR (2012, §14); ESOMAR (2015a, §3.7); ESOMAR (2015b, §6.1-2); IMRO (2015, §20); ISO (2009, §4.5.2); ISO (2012, §7.2.i)
Questionnaire	In the event that the researcher did not create the questionnaire, the questionnaire should be available, including any screening questions. As web surveys are an intensely visual medium (e.g., Couper, Tourangeau, and Kenyon 2004 and Couper, Tourangeau, Conrad, and Crawford 2004) we recommend that screenshots be provided as part of reporting.	American Association for Public Opinion Research (2015, §III.A.2, §III.A.14-15); ESOMAR (2014, §5.2); ESOMAR (2015a, §3.8); ESOMAR (2015b, §5.2, §6.3); ISO (2009, §4.7); ISO (2012, §7.2.l)

Final dispositions and outcome rates	How many panel members were invited to complete the survey? What is the break-off rate? How are these calculated? We recommend that the definitions of outcome rates developed by Callegaro and DiSogra (2008) and DiSogra and Callegaro (2016) be used; these are largely incorporated in American Association for Public Opinion Research (2015).	American Association for Public Opinion Research (2015, §III.A.18); ESOMAR (2015a, §3.7-8); ESOMAR (2015b, §6.1); IMRO (2015, §26); ISO (2009, §4.7); ISO (2012, 7.2.f.4)
Handling of mobile devices / smartphones	We recommend documenting whether the questionnaire layout is adapted/optimized for smartphone completion and, if so, how it is adapted or optimized. There is an extensive literature on methods for formatting web surveys on mobile devices and implications for data quality (e.g., Peytchev and Hill 2010; Chrzan and Saunders 2012; de Bruijne and Wijnant 2013a, 2013b; Link, Murphy, Schober, Buskirk, Hunter Childs 2014; Mavletova and Couper 2014, 2016; Arn, Klug, and Kolodziejski 2015; Struminskaya, Weyandt, and Bosnjak 2015; Lugtig and Toepoel 2016; Revilla, Toninelli, and Ochoa, 2016; Revilla, Toninelli, and Ochoa 2016; Couper, Antoun, and Mavletova 2017; Peterson, Griffin, LaFrance, and Li 2017; Antoun, Katz, Argueta, and Wang 2018). Although not incorporated into other guidelines, it is important that clients are sufficiently informed as handling mobile devices may have consequences for break-off and measurement	Not referenced in any guideline

Table A2. Case-Level Data Recommendations

Construct	Notes
Sample source	Sample source (e.g., list, web advertisement)
Within-survey paradata	Date-time survey started and completed
	Duration between survey started and completed
	Frequency of invitations to surveys: each week, etc.
	Absence of invitations to concurrent surveys during a certain time (e.g., 2 consecutive weeks after the first invitation to participate in the survey)
	Device information provided for each session for surveys completed in multiple sessions:
	User agent string or equivalent information (e.g., operating system and version number, browser and version number)
	Is JavaScript enabled?
	Is AJAX enabled?
	Screen resolution
Cross-survey paradata	Date joined panel (if relevant)
	Number of surveys the panel member has been invited to complete
	Number of surveys the panel member completed
	Individual-level completion rate (number of surveys completed / number of surveys invited)
Profile data	Profile information relevant to any quotas or other selection mechanisms
	Date panel profile data last updated
Panelists recruitment	Recruitment origin of the selected panelists (respondents and nonrespondents): list, telephone, web, ad, etc.
	Membership to other panels (if possible)
History of contacts	Number of reminders (if any)
	Regular or extra incentive
	Invited to participate in other surveys during the research or not
Quality data	If all sample cases (not only completes) are requested:
	The status of any data quality checks (e.g., duplicate responses, speeding, straight-lining, trap questions)
	For each case removed from the data, the reasons why the case was removed
	We recommend that all sample cases be requested from the panel vendor

References

American Association for Public Opinion Research (2013). "Report of the AAPOR Task Force on Nonprobability Sampling," available at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf (accessed September 30, 2019).

American Association for Public Opinion Research (2015). "AAPOR Code of Ethics," available at https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx (accessed September 30, 2019).

Ansolabehere, S., and B. F. Schaffner (2014), "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison," *Political Analysis*, 22, 285–303.

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

Antoun, C., J. Katz, J. Argueta, and L. Wang (2018), "Design Heuristics for Effective Smartphone Questionnaires," *Social Science Computer Review*, 36, 557–574.

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (2001), "Standards zur Qualitätssicherung für Online-Befragungen," available at https://www.adm-ev.de/wp-content/uploads/2018/07/Standards-zur-Qualit%C3%A4tssicherung-bei-Online-Befragungen.pdf (accessed September 30, 2019).

Arn, B., S. Klug, and J. Kolodziejski (2015), "Evaluation of an Adapted Design in a Multi-Device Online Panel: A DemoSCOPE Case Study," *Methods, Data, Analyses*, 9, 185–212.

Berrens, R. P., A. K. Bohara, H. Jenkins-Smith, C. Silva, and D. L. Weimer (2003), "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples," *Political Analysis*, 11, 1–22.

Bethlehem, J. (2016), "Solving the Nonresponse Problem with Sample Matching?," *Social Science Computer Review*, 34, 59–77.

———. (2017), *Understanding Public Opinion Polls*, Boca Raton, FL: Chapman and Hall/CRC.

Bethlehem, J. G. (2009), "The Rise of Survey Sampling," CBS Discussion Paper 09015.

———. (2009), "Weighting Adjustment," in *Applied Survey Methods: A Statistics Perspective*, ed. J. G. Bethlehem, pp. 249–275, Hoboken, NJ: John Wiley and Sons.

Blom, A. G., D. Ackermann-Piek, S. Helmschrott, C. Cornesse, C. Bruch, and J. Sakshaug (2018), "An Evaluation of Sample Accuracy in Probability-Based and Nonprobability Surveys," under review.

Blumberg, S. J., and J. V. Luke (2017), "Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, January–June 2017," available at https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201712.pdf (accessed September 30, 2019).

Bowley, A. L. (1906), "Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science," *Journal of the Royal Statistical Society*, 69, 540–558.

Brüggen, E., J. A. van den Brakel, and J. Krosnick (2016), "Establishing the Accuracy of Online Panels for Survey Research," Statistics Netherlands. Available at https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research (accessed September 30, 2019).

Callegaro, M., and C. DiSogra (2008), "Computing Response Metrics for Online Panels," *Public opinion quarterly*, 72, 1008–1032.

Callegaro, M., A. Villar, D. Yeager, and J. A. Krosnick (2014a), "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*, eds. M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J.A. Krosnick, and P. J. Lavrakas, pp. 23–53, UK: John Wiley and Sons.

Callegaro, M., R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (2014b), "Online Panel Research: History, Concepts, Applications and a Look at the Future," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas, pp. 1–22, UK: John Wiley and Sons.

Chan, P., and D. Ambrose (2011). "Canadian Online Panels: Similar or Different," *Vue*, pp. 16–20.

Chang, L., and J. A. Krosnick (2009), "National Surveys via RDD Telephone Interviewing versus the Internet," *Public Opinion Quarterly*, 73, 641–678.

Chen, J. K. T., R. L. Valliant, and M. R. Elliott (2018), "Calibrating Non-Probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41, 657–681.

Chrzan, K., and T. Saunders (2012), "Improving the Mobile Survey Experience: A Closer Look at Scale Orientation, Grids and Modality Effects," Maritz research whitepaper, available at http://www.websm.org/uploadi/editor/doc/1470813336Chrzan_Saunders_2012_Improving_the_Mobile_Survey_Experience.pdf (accessed September 30, 2019).

Comer, P. (2019), "Sampling in the Digital Age," available at https://luc.id/blog/sampling-in-the-digital-age/ (accessed September 30, 2019).

Converse, J. M. (1987), *Survey Research in the United States: Roots and Emergence, 1890-1960*, Berkeley: University of California Press.

Coppock, A., and O. A. McClellan (2019), "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents," *Research & Politics*, 6, 1–14.

Couper, M. P., R. Tourangeau, and K. Kenyon (2004a), "Picture This! Exploring Visual Effects in Web Surveys," *Public Opinion Quarterly*, 68, 255–266.

Couper, M. P., R. Tourangeau, F. G. Conrad, and S. D. Crawford (2004b), "What They See Is What We Get," *Social Science Computer Review*, 22, 111–127.

Couper, M., C. Antoun, and A. Mavletova (2017). "Mobile Web Surveys: A Total Survey Error Perspective," in *Total Survey Error in Practice: Improving Quality in the Era of Bid Data* (1st ed.), eds. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West, pp. 133–154, Hoboken, NJ: John Wiley and Sons.

Crossley, A. M. (1937), "Straw Polls in 1936," *Public Opinion Quarterly*, 1, 24–35.

Dassonneville, R., A. Blais, M. Hooghe, and K. Deschouwer (2018), "The Effects of Survey Mode and Sampling in Belgian Election Studies: A Comparison of a National Probability Face-to-Face Survey and a Nonprobability Internet Survey," *Acta Politica*, 11, doi: 10.1057/s41269-018-0110-4.

de Bruijne, M., and A. Wijnant (2013a), "Can Mobile Web Surveys Be Taken on Computers? A Discussion on a Multi-Device Survey Design," *Survey Practice*, 6, 1–8.

———. (2013b). "Comparing Survey Results Obtained via Mobile Devices and Computers," *Social Science Computer Review*, 31, 482–504.

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

Deville, J. C., and C. E. Särndal (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Deville, J.-C. (1991), "A Theory of Quota Surveys," *Survey Methodology*, 17, 163–181.

DiSogra, C., and M. Callegaro (2016), "Metrics and Design Tool for Building and Evaluating Probability-Based Online Panels," *Social Science Computer Review*, 34, 26–40.

DiSogra, C., C. Cobb, E. Chan, and J. M. Dennis (2011), "Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics," JSM Proceedings, Survey Research Methods Section, pp. 4501–4515, Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2011/Files/302704_68925.pdf

Downes, M., L. C. Gurrin, D. R. English, J. Pirkis, D. Currier, M. J. Spittal, and J. B. Carlin (2018), "Multilevel Regression and Poststratification: A Modelling Approach to Estimating Population Quantities from Highly Selected Survey Samples," *American Journal of Epidemiology*, 187, 1780–1790.

Duffy, B., K. Smith, G. Terhanian, and J. Bremer (2005), "Comparing Data from Online and Face-to-Face Surveys," *International Journal of Market Research*, 47, 615–639.

Dutwin, D., and T. D. Buskirk (2017), "Apples to Oranges or Gala versus Golden Delicious?," *Public Opinion Quarterly*, 81, 213–239.

Elliott, M. R. (2009). "Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights," *Survey Practice*, 2, 1–7.

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

Erens, B., S. Burkill, M. P. Couper, F. Conrad, S. Clifton, C. Tanton, A. 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, e276.

ESOMAR (2012), "28 Questions to Help Buyers of Online Samples," available at https://www.esomar.org/uploads/public/knowledge-and-standards/documents/ESOMAR-28-Questions-to-Help-Buyers-of-Online-Samples-September-2012.pdf (accessed September 30, 2019).

ESOMAR (2014), "ESOMAR/WAPOR Guideline on Opinion Polls and Published Surveys," available at https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-WAPOR-Guideline-on-Opinion-Polls-and-Published-Surveys-August-2014.pdf (accessed September 30, 2019).

ESOMAR (2015a), "ESOMAR/GRBN Guideline for Online Sample Quality," available at https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-GRBN_Online-Sample-Quality-Guideline_February-2015.pdf (accessed September 30, 2019).

ESOMAR (2015b), "ESOMAR/GRBN Online Research Guideline," available at https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-GRBN-Online-Research-Guideline-October-2015.pdf (accessed September 30, 2019).

Fahimi, M., F. M. Barlas, R. K. Thomas, and N. Buttermore (2015), "Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse," *Survey Practice*, 8, 1–15.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Ganesh, N., V. Pineau, A. Chakraborty, and J. M. Dennis (2017), "Combining Probability and Non-Probability Samples Using Small Area Estimation," JSM Proceedings, Survey Research Methods Section, pp. 1657–1667, Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2017/files/593906.pdf.

Gelman, A., S. Goel, D. Rothschild, and W. Wang (2017), "High-Frequency Polling with Non-Representative Data," in *Routledge Studies in Global Information, Politics and Society: Vol. 12. Political Communication in Real Time: Theoretical and Applied Research Approaches*, eds. D. Schill, R. Kirk, and A. E. Jasperson, pp. 89–105, London: Routledge.

Ghitza, Y., and A. Gelman (2013), "Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups," *American Journal of Political Science*, 57, 762–776.

Gittelman, S. H., R. K. Thomas, P. J. Lavrakas, and V. Lange (2015), "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples," *Journal of Advertising Research*, 55, 368–379.

Goel, S., A. Obeng, and D. Rothschild (2015), "Non-Representative Surveys: Fast, Cheap, and Mostly Accurate," available at http://researchdmr.com/FastCheapAccurate.pdf (accessed September 30, 2019).

Göritz, A. S., N. Reinhold, and B. Batinic (2000), "Online Panels," in *Online Social Sciences*, eds. B. Batinic, U.-D. Reips, M. Bosnjak, and A. Werner, pp. 27–47, Göttingen: Hogrefe Publishing.

Groves, R. M., and L. Lyberg (2010), "Total Survey Error: Past, Present, and Future," *Public Opinion Quarterly*, 74, 849–879.

Harter, R., M. P. Battaglia, T. D. Buskirk, D. A. Dillman, N. English, M. Fahimi, M. R. Frankel, et al. (2016), "Address-Based Sampling," available at https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx (accessed September 30, 2019).

Interactive Marketing Research Organization (IMRO) (2015), "IMRO Guidelines for Best Practices in Online Sample and Panel Management," available at https://www.insightsassociation.org/issues-policies/best-practice/imro-guidelines-best-practices-online-sample-and-panel-management (accessed September 30, 2019).

International Organization for Standardization (2009), *ISO 26362: 2009. Access Panels in Market, Opinion and Social Research - Vocabulary and Service Requirements*, Geneva, Switzerland: ISO.

International Organization for Standardization (2012), *ISO 20252: 2012. Market, Opinion and Social Research - Vocabulary and Service Requirements*, Geneva, Switzerland: ISO.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez (2016), Evaluating Online Nonprobability Surveys: Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics," Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/12/2016/04/Nonprobability-report-May-2016-FINAL.pdf (accessed September 30, 2019).

Kiaer, A. N. (1895), "Observations et Expériences Concernant les Dénombrements Représentatifs," *Bulletin of the International Statistical Institute*, 9, 176–183.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley and Sons.

Klein, R. A., M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1, 443–490.

Lee, S. (2006), "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys," *Journal of Official Statistics*, 22, 329–349.

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

Legleye, S., G. Charrance, N. Razafindratsima, N. Bajos, A. Bohet, and C. Moreau (2018). "The Use of a Nonprobability Internet Panel to Monitor Sexual and Reproductive Health in the General Population," *Sociological Methods and Research*, 47, 314–348.

Lessler, J. T., and W. D. Kalsbeek (1992), *Nonsampling Error in Surveys*, New York: John Wiley and Sons.

Link, M. W., J. Murphy, M. F. Schober, T. D. Buskirk, J. Hunter Childs, and C. Langer Tesfaye (2014), *Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research*, Oakbrook, IL: American Association for Public Opinion Research (AAPOR). Available at https://www.aapor.org/Education-Resources/Reports/Mobile-Technologies-for-Conducting,-Augmenting-and.aspx#4.0%20%20%20%20%20PRIVACY%20CONSIDERATIONS (accessed September 30, 2019).

Little, R. J. A., B. T. West, P. Boonstra, and J. Hu (2019), "Measures of the Degree of Departure from Ignorable Sample Selection," *Journal of Survey Statistics and Methodology*, available at https://academic.oup.com/jssam/advance-article/doi/10.1093/jssam/smz023/5556334.

Loosveldt, G., and N. Sonck (2008), "An Evaluation of the Weighting Procedures for an Online Access Panel Survey," *Survey Research Methods*, 2, 93–105.

Lugtig, P., and V. Toepoel (2016), "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error," *Social Science Computer Review*, 34, 78–94.

MacInnis, B., J. A. Krosnick, A. S. Ho, and M.-J. Cho (2018), "The Accuracy of Measurements with Probability and Nonprobability Survey Samples," *Public Opinion Quarterly*, 16, 307.

Malhotra, N., and J. A. Krosnick (2007), "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples," *Political Analysis*, 15, 286–323.

Mavletova, A., and M. P. Couper (2014), "Mobile Web Survey Design: Scrolling versus Paging, SMS versus e-Mail Invitations," *Journal of Survey Statistics and Methodology*, 2, 498–518.

———. (2016), "Grouping of Items in Mobile Web Questionnaires," *Field Methods*, 28, 170–193.

Mercer, A. W., F. Kreuter, S. Keeter, and E. A. Stuart (2017), "Theory and Practice in Nonprobability Surveys," *Public Opinion Quarterly*, 81, 250–271.

Mercer, A., A. Lau, and C. Kennedy (2018), "For Weighting Online Opt-In Samples, What Matters Most?," available at http://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/ (accessed September 30, 2019).

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.

Page, B. I., and R. Y. Shapiro (1992), *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*, Chicago: University of Chicago Press.

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

Pasek, J. (2016), "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence," *International Journal of Public Opinion Research*, 28, 269–291.

Pennay, D. W., D. Neiger, P. J. Lavrakas, and K. Borg (2018), "The Online Panels Benchmarking Study: A Total Survey Error Comparison of Findings form Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia," available at https://www.srcentre.com.au/our-research/csrm-src-methods-papers/CSRM_MP2_2018_ONLINE_PANELS.pdf (accessed September 30, 2019).

Peterson, G., J. Griffin, J. LaFrance, and J. Li (2017), "Smartphone Participation in Web Surveys: Choosing between the Potential for Coverage, Nonresponse, and Measurement Error," in *Total Survey Error in Practice: Improving Quality in the Era of Bid Data*, (1st ed.), ed. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West, pp. 203–233, Hoboken, NJ: John Wiley and Sons.

Peytchev, A., and C. A. Hill (2010), "Experiments in Mobile Web Survey Design," *Social Science Computer Review*, 28, 319–335.

Pfeffermann, D. (2017), "Bayes-Based Non-Bayesian Inference on Finite Populations from Non-Representative Samples: A Unified Approach," *Calcutta Statistical Association Bulletin*, 69, 35–63.

Pfeffermann, D., J. L. Eltinge, and L. D. Brown (2015), "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture," *Journal of Survey Statistics and Methodology*, 3, 425–483.

Raghunathan, T. (2015), *Missing Data Analysis in Practice*, Boca Raton: Chapman and Hall/CRC.

Revilla, M., C. Ochoa, and D. Toninelli (2016), "PCs versus Smartphones in Answering Web Surveys: Does the Device Make a Difference?," *Survey Practice*, 9, 1–6.

Rivers, D. (2007), "Sampling for web surveys," http://www.websm.org/uploadi/editor/1368187629Rivers_2007_Sampling_for_web_surveys.pdf (accessed September 30, 2019).

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

———. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

Roshwalb, A., Z. Lewis, and R. Petrin (2016), "The Efficacy of Nonprobability Online Samples," JSM Proceedings, Survey Research Methods Section, pp. 3657–3666, Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2016/files/389791.pdf.

Santoso, L. P., R. Stein, and R. Stevenson (2016), "Survey Experiments with Google Consumer Surveys: Promise and Pitfalls for Academic Research in Social Science," *Political Analysis*, 24, 356–373.

Scherpenzeel, A. C., and J. G. Bethlehem (2011), "How Representative Are Online Panels? Problems of Coverage and Selection and Possible Solutions," in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, eds. M. Das, P. Ester, and L. Kaczmirek, pp. 105–132, New York: Routledge.

Schonlau, M., A. van Soest, and A. Kapteyn (2007), "Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?," RAND Working Paper Series No. WR-506, retrieved from http://dx.doi.org/10.2139/ssrn.1006108.

Schonlau, M., A. van Soest, A. Kapteyn, and M. Couper (2009), "Selection Bias in Web Surveys and the Use of Propensity Scores," *Sociological Methods and Research*, 37, 291–318.

Schonlau, M., K. Zapert, L. P. Simon, K. H. Sanstad, S. M. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. H. Berry (2004), "A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey," *Social Science Computer Review*, 22, 128–138.

Smyk, M., J. Tyrowicz, and L. Van der Velde (2018), "A Cautionary Note on the Reliability of the Online Survey Data: The Case of Wage Indicator," *Sociological Methods & Research*,

0049124118782538. Available at https://journals.sagepub.com/doi/pdf/10.1177/00491241187 82538.

Sohlberg, J., M. Gilljam, and J. Martinsson (2017), "Determinants of Polling Accuracy: The Effect of Opt-in Internet Surveys," *Journal of Elections, Public Opinion and Parties*, 27, 433–447.

Steinmetz, S., A. Bianchi, K. Tijdens, and S. Biffignandi (2014), "Improving Web Survey Quality: Potentials and Constraints of Propensity Score Adjustments," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Goritz, J. A. Krosnick, and P. Lavrakas, pp. 273–298, UK: John Wiley and Sons.

Struminskaya, B., K. Weyandt, and M. Bosnjak (2015), "The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-Based General Population Panel," *Methods, Data, Analyses*, 9, 261–292.

Sturgis, P., J. Kuha, N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, B. E. Lauderdale, and P. Smith (2018), "An Assessment of the Causes of the Errors in the 2015 UK General Election Opinion Polls," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 757–781.

Szolnoki, G., and D. Hoffmann (2013), "Online, Face-to-Face and Telephone Surveys—Comparing Different Sampling Methods in Wine Consumer Research," *Wine Economics and Policy*, 2, 57–66.

Taylor, H. (2000), "Does Internet Research "Work"? Comparing on-Line Survey Results with Telephone Surveys," *International Journal of Market Research*, 42, 51–63.

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

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

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

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

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015), "Forecasting Elections with Non-Representative Polls," *International Journal of Forecasting*, 31, 980–991.

Wu, C., and R. R. Sitter (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data," *Journal of the American Statistical Association*, 96, 185–193.

Yeager, D. S., J. A. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang (2011), "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples," *Public Opinion Quarterly*, 75, 709–747.