methods, data, analyses | Vol. 17(2), 2023, pp. 171-206 DOI: 10.12758/mda.2023.04

Comparing and Improving the Accuracy of Nonprobability Samples: Profiling Australian Surveys

Sebastian Kocar[1] *& Bernard Baffour*[2]

[1] *Institute for Social Change, University of Tasmania*

[2] *School of Demography, The Australian National University*

Abstract

There has been a great deal of debate in the survey research community about the accuracy of nonprobability sample surveys. This work aims to provide empirical evidence about the accuracy of nonprobability samples and to investigate the performance of a range of post-survey adjustment approaches (calibration or matching methods) to reduce bias, and lead to enhanced inference. We use data from five nonprobability online panel surveys and compare their accuracy (pre- and post-survey adjustment) to four probability surveys, including data from a probability online panel. This article adds value to the existing research by assessing methods for causal inference not previously applied for this purpose and demonstrates the value of various types of covariates in mitigation of bias in nonprobability online panels. Investigating different post-survey adjustment scenarios based on the availability of auxiliary data, we demonstrated how carefully designed post-survey adjustment can reduce some bias in survey research using nonprobability samples. The results show that the quality of post-survey adjustments is, first and foremost, dependent on the availability of relevant high-quality covariates which come from a representative large-scale probability-based survey data and match those in nonprobability data. Second, we found little difference in the efficiency of different post-survey adjustment methods, and inconsistent evidence on the suitability of 'webographics' and other internet-associated covariates for mitigating bias in nonprobability samples.

Keywords: nonprobability sampling, volunteer online panels, post-survey adjustment, calibration, matching methods, benchmarking

It has become increasingly evident that traditional surveys face challenges in measuring and understanding emerging and complex social issues, since they often fail to accurately measure individual behavior, attitudes and perceptions on various issues (Baker et al. 2010; Malhotra & Krosnick 2007; Tourangeau et al. 2014). Recent notable failures of polls to predict the outcomes of referenda and elections have shown that the way in which data are collected from the population must be responsive to people's dynamic lifestyles, choices, and attitudes (e.g., Goot 2021; Kennedy et al. 2018; Wang et al. 2015). Further, the widespread availability of and access to the internet and social media leads to a quick diffusion of ideas that may rapidly shift social attitudes and behaviors (e.g., Wang et al. 2021).

Compared to traditional (probability-based) survey methods which usually include offline data collection (mail, telephone, face to face (f2f)) and have been proven to be inadequate to capturing new to emerge, quick to change events, web-based surveys are advantageous given their convenience, quick turn-around times, and relatively low respondent costs (Baker et al. 2013). Additionally, nonprobability online panel surveys allow tests for consistency and reliability to be performed in a timelier manner than telephone and interviewer administered surveys. While there are web-based surveys that are probability-based (for instance push-to-web surveys with a 'population' frame of emails (Cornesse et al. 2020)), the majority of online surveys rely on being quick and efficient through reaching potentially millions of internet users which comes at the expense of being representative of the population (Bethlehem & Biffignandi 2012; Baker et al. 2013). We focus on nonprobability online (web-based) panel[1] surveys in this research, although the findings can be applied to other types of nonprobability surveys.

There are four main issues associated with nonprobability online panel surveys, which are related to type of sampling, sampling frame, nonresponse, and coverage. First, respondents are not selected based on probability sampling. Even though they may be 'randomly' selected, it is often not possible to work out their chance of being selected into the survey. Consequently, it is unknown what respondents with a non-zero chance of being selected comprise the population that the sample is selected from, and so the reliability of those sample survey estimates can-

[1] Nonprobability online panels are also known as volunteer, opt-in or access online panels.

Acknowledgements

We would like to acknowledge the contributions of the Social Research Centre for collecting and providing access to their survey data (Online Panels Benchmarking Study and Life in Australia™) and the Australian Government Department of Education, Skills and Employment for PhD program funding.

Direct correspondence to

Sebastian Kocar, Institute for Social Change, University of Tasmania
E-mail: sebastian.kocar@utas.edu.au

not be assessed with confidence (Callegaro & DiSogra 2008b). This is associated with the second issues, which is that there is no general population online sampling frame for the internet (e.g., Couper 2000). While virtually all internet users have an email address, there is no list comprising all of these email addresses that could be used to draw a random sample. Third, online survey respondents have been found to have different characteristics and behaviors to respondents from more traditional surveys. Online surveys generally have higher levels of item and unit nonresponse (e.g., Couper 2000; Daikeler et al. 2020), which has potential to introduce more nonresponse bias. Four, the internet does not have universal coverage: in Australia in early-2022, it was estimated that 9% of people did not use the internet (DataReportal 2022)[2]. This can introduce (under)coverage error, since this lack of access is concentrated amongst those of older age, rural location, Indigenous ethnicity, and lower education levels. Those are groups which are increasingly important to policymakers, hence limiting the utility of internet-based surveys. Collectively, these limitations mean that the data collected online with nonprobability panels are less reliable than those gathered by traditional survey methods, since they are generally more prone to the above-mentioned sampling, nonresponse, and undercoverage bias (which is, at the same time, challenging to estimate). Hence, the existing evidence suggests that we cannot be confident that the results from nonprobability panels accurately represent trends in the general population.

Our research aims are three-fold. First, we quantify the differences in survey estimates obtained from the same survey administered through a probabilistic sampling framework in contrast with those collected from a non-probabilistic framework. Second, we compare and contrast the performance of different post-survey adjustment methods on reducing bias in nonprobability-based online panel surveys. Third, we compare how the inclusion of different external data sources (such as Census) and covariates (such as non-demographics[3]) in post-survey adjustment affect the accuracy of survey estimates. This investigation adds value to the existing literature on approaches to mitigate bias in nonprobability surveys. As such, it provides valuable evidence to survey practitioners using samples from nonprobability online panels of better quality (e.g., those with ESOMAR or ISO accreditation), as well as survey researchers interested in implementing other types of nonprobability surveys.

[2] The last official statistics estimate for Australian households with no internet access at home was from 2016-17, i.e., 14.0%. The same estimate for households without children under 15 was even higher, i.e., 18.1% (Australian Bureau of Statistics 2018a).

[3] We define non-demographics as attitudinal, behavioral, knowledge and factual questions that do not ask about person's socio-demographic characteristics (see Yeager et al. 2011).

Background and Literature Review

With probability sampling we ensure that every unit in the population has a known, and non-zero, chance of being selected into the sample. This randomisation is a key design attribute of probability sampling, and enables the calculation of standard errors, confidence intervals, and making generalized inferences regarding the target population of interest from the sample (Hade & Lemeshow 2011). However, while most (probability) surveys have known selection probabilities, whether people respond cannot be controlled for, in spite of all the best efforts of survey practitioners. Rivers (2013) argues that it is the probability of sample inclusion not selection that matters, since whether people cooperate in probability surveys cannot be controlled for, and low response rates introduce skews similar to those in volunteer panels. Trends of high nonresponse rates with a large proportion of probability-based surveys reporting response rates of under 10% (Kennedy & Hartig 2019), and the associated nonresponse biases may lead to flawed results and problems in statistical inference (Baker et al. 2010; Baker et al. 2013). However, the fact that the selection probabilities for a sample are unknown does not imply that they cannot be estimated or adjusted for in a nonprobability sample, just as adjustments are used in probability-based surveys to compensate for issues around coverage and response (Rivers 2013).

Opportunities to Improve Accuracy of Nonprobability Samples

There is a whole gamut of online nonprobability-based surveys, from the opt-in click-through unsolicited surveys which are advertised on websites, to more structured recruitment of a panel of respondents; as a result of these idiosyncratic designs which make it difficult to work out the rates of contact, response, and (non)coverage, it is almost impossible to make reasonable statistical inferences from data obtained with nonprobability-based surveys (Rivers 2013). However, the characteristics of the nonprobability online panel sample may closely resemble the population being studied and identifying the conditions under which valid statistical inferences can be made using the realized sample is important (Mercer et al. 2017). This selection bias - which leads to the sample misrepresenting the population - can be controlled for using several different approaches, underpinned by an existing framework based on causal inference used in numerous fields such as epidemiology, political science and economics (Heckman 1979; Hug 2003; Rothman et al. 2008).

Valliant (2020) and Elliott and Valliant (2017) showed that it is not necessary and sufficient that (i) every unit in the population has some probability of being included in the sample, and that (ii) there is a structural model based on the observed sample which can be used to describe the variables we are interested in measuring, meaning that you do not need both conditions to hold. This implies that reweighting or matching schemes can be used to (a) estimate the probability of response and (b) calibrate to known benchmark population totals, to correct for any selection biases in the estimates derived from nonprobability sample surveys (Matei 2018). We distinguish between matching approaches and reweighting approaches, and the key goal of both approaches is to ensure that there is no (or little) bias in the observed data, meaning that the empirical distribution of the observed data is similar to the population (Baker et al. 2013; Elliott & Valliant 2017; Mercer et al. 2017; Mercer et al. 2018; Valliant 2020).

Post-Survey Adjustments in Nonprobability Samples

Post-survey adjustments correct for the unequal probabilities of selection and are common in both nonprobability and most probability surveys: virtually no probability sample uses simple random sampling. As such, in both probability and nonprobability samples, the objective for inference is to ensure that the composition of the sampled units with respect to the observed characteristics either matches or can be adjusted to match the population of interest. Post-survey adjustments have the dual purpose of reducing the bias and producing more accurate population estimates (Elliott & Valliant 2017; Mercer et al. 2017).

There are several approaches which have been proposed to improve accuracy and inference for data collected under a nonprobability sample. These approaches are predicated from the issues facing probability samples caused by differences in response and coverage of surveys. To cope with these issues, statistical adjustments typically correct for any systematic biases, including in nonprobability samples (Cornesse et al. 2020; Elliott 2009; Lehdonvirta et al. 2021; Rivers 2007).

This study compares six primary methods of reweighting and matching survey data: raking, generalized regression estimation (GREG), propensity score weighting (PSW), multilevel regression and poststratification (MRP), Mahalanobis distance matching (MDM) and coarsened exact matching (CEM). Reweighting methods directly adjust the sample distribution to the target population distribution, to achieve the desired sample composition in the presence of nonresponse and/or other factors. Matching methods attempt to create a balanced nonprobability sample which closely resembles the characteristics of a probability sample from the ‘true’ population (when compared with a selected array of auxiliary, often non-demographic, characteristics) (Bethlehem 2016; Cornesse et al. 2020). Assessing performance of different post-survey adjustment methodology is important as all methods come with certain limitations – for example, raking was reported to be less effective to mitigate bias in nonprobability online panel samples than in probability samples (Mercer et al. 2018), the GREG estimator becomes less precise the larger the number of benchmarks (Deville et al. 1993), MRP requires knowledge of the joint distribution of the poststratification variables in the target population (Deville & Särndal 1992), and matching methods cannot be used with all types of data.

In the next paragraphs, we provide more information about each of the post-survey adjustment methods investigated in this study.

Raking

Raking, also known as iterative proportional fitting, is the most common weighting method and is simple to implement as it relies on knowing the marginal distribution of population covariates. As part of the procedure, the weights for each individual are repeatedly adjusted until the sample distribution is perfectly aligned with the population distribution for the selected set of variables. As the utility of a large set of weighting covariates diminishes, using key socio-demographic variables is often sufficient to reduce the selection bias in probability samples (Kalton & Flores-Cervantes 2003).

Generalized Regression Estimation (GREG)

Generalized regression estimation (GREG) is a calibration[4] approach where the sampling weights are adjusted to make certain the survey estimators match to the set of known population totals (benchmarks). In contrast to raking which repeatedly reweights the sample to the marginal distributions of the known population totals, the GREG estimator is based on the minimizing the distance measure between the sample and the benchmark information and it is supposedly more efficient and provides more accurate population estimates (Deville & Särndal 1992).

Propensity Score Weighting (PSW)

In the simplest version of probability-based sampling, survey respondents are assumed to have a non-zero chance of being included in the sample and weighting each sample individual by the inverse of its sample selection probability removes any selection bias (Cochran 1977). When data are collected through a nonprobability-based sample, we can use the same ideas, and although selection probabilities from a nonprobability sample are unknown, it does not mean that they cannot be estimated (Rivers 2013). In PSW, a synthetic population assumed to "represent" the full target population is created by using external high-quality data representative of the population. Then pseudo-inclusion probabilities are estimated using binary (i.e., probit or logistic) regression modeling, which leads to a probability-based (or

[4] Calibration is a general framework for weighting in which the following conditions for adjustment weights have to be satisfied: (1) the weights have to be as close to 1 as possible, (2) after calibration, the sample distribution of the auxiliary variables should match the population distribution (Bethlehem 2008). Deville et al. (1993) distinguish between complete post-stratification, generalized raking, and GREG as calibration methods.

synthetic) reference sample which is combined with the nonprobability sample (Schonlau & Couper 2017; Valliant 2020). Like in calibration, PSW is efficient in bias reduction if the weighting variables and the propensity of response in the nonprobability sample are (strongly) associated with outcome variables (Rosenbaum & Rubin 1983; Valliant & Dever 2011).

Multilevel Regression and Poststratification (MRP)

The MRP approach (Gelman 2007; Gelman & Little 1997) is based on assuming the existence of a super-population model which can be fitted to the analytic survey variables and can be used to project the observed sample to the full population. The key assumption here is that sampled and non-sampled data are driven by an underlying model (for the analysis variables) and this model can be revealed by analyzing the sample responses. In the presence of nonresponse, this model also specifies the relationship between the observed units and the unobserved data (Brick 2013). Poststratification, which includes creating a set of post-strata and estimating the mean value by fitting mixed effects (multilevel) model in the case of MRP, requires knowledge of the joint distribution of the poststratification variables in the target population unlike other reweighting methods (Deville & Särndal 1992), except for interactions between covariates. In political science, this approach is useful in obtaining state-level predictions based on relatively small national samples (for example, Bon et al. 2019; Park et al. 2004; Park et al. 2006; Wang et al. 2015).

Mahalanobis Distance Matching (MDM)

MDM is a distance matching method which creates groups containing one or more observations from both the reference sample and the nonprobability sample that are similar on a set of auxiliary variables believed to be associated with the probability of selection. In MDM, we measure the distance between a pair of observations, y_i and y_j, with the Mahalonobis distance calculated as presented in Equation 1:

$$M(y_i, y_j) = \sqrt{(y_i - y_j)^T S^{-1} (y_i - y_j)} \tag{1}$$

where S is the sample covariance matrix of y. Two observations are matched if they have the minimum distance out of a set of pairs, e.g., through nearest neighbour matching. Since the population of possible match-pairs exponentially increases as the nonprobability sample size increases, usually some procedure is used to remove pairs that are unreasonably distant through defining calipers which are chosen cutoffs for which the maximum distance is allowed (Stuart & Rubin 2008).

Coarsened Exact Matching (CEM)

Coarsened exact matching (CEM) is a matching method like the MDM, but the key difference is that it is a stratification-based method (Sizemore & Alkurdi 2019), and calipers are not required to remove unreasonably bad matches (Iacus et al. 2011). In CEM, units with the same values of the selected covariates (in contract to exact matching, they can be coarsened, i.e., recategorized into fewer groups) are placed in a single stratum. Within each stratum, the units in the nonprobability sample are weighted to be equal to the number of units in the reference sample. Strata without at least a single nonprobability sample or reference sample unit, are given a zero weight which effectively prunes them from the dataset. By removing unmatched units, the inference is generally improved because it achieves a better balance between the empirical distributions of reference sample and the nonprobability sample (Iacus et al. 2009; Stuart 2010).

Scope of this Study

Following from Mercer et al. (2017), we use the general framework which emphasizes the characteristics of the realized sample (regardless of how it was generated), and therefore correct for any self-selection bias in survey inference (Groves 2006; Keiding & Louis 2016; Little & Rubin 2002). The authors identify three components that determine whether the presence of self-selection ultimately leads to biased survey estimates: exchangeability, positivity, and composition (Mercer et al. 2017). These components of self-selection bias are not fundamentally different for nonprobability samples, but what differs between probability and nonprobability samples are the underlying assumptions which lead to individuals becoming members of nonprobability samples (Kennedy et al. 2016; MacInnis et al. 2018; Pfeffermann et al. 2015).

Notwithstanding, this can be useful in investigating if there is (a) improved inference of sample data from a nonprobability survey, and (b) through comparing different post-survey adjustment methods under different external data sources scenarios we can ascertain their suitability/performance under various conditions. There have been a number of authors – for instance, DiSogra et al. (2011), Baker et al. (2013), Mercer et al. (2017), Mercer et al. (2018), and Valliant (2020) – who have undertaken similar research into the performance of different methods, and also discussed the requirements with respect to the external data sources for the various approaches.

Therefore, we will examine a range of survey estimates against two categories of population benchmarks: secondary demographics (such as citizenship and employment status), and non-demographics (such as alcohol consumption and life satisfaction), as well as against both categories combined. First, we compare the accuracy of probability and nonprobability samples from two Australian survey projects, by presenting updated evidence. Second, we investigate the performance of different post-survey adjustments to improve accuracy of nonprobability samples. We do that under four realistic scenarios which differ in terms of the nature of the auxiliary data that is available for use in post-survey adjustment for nonprobability surveys. The scenarios under which we are assessing performance of adjustment methods are the following:

- *Scenario 1 – availability of census aggregated statistics utilized to improve accuracy in nonprobability samples*
  Under this scenario, aggregated[5] population census data matching to *primary demographics*[6] from a nonprobability sample are used to adjust the sample distribution for those key auxiliary variables to match the population distribution (e.g., for *sex, age*, and *education*).
- *Scenario 2 – availability of additional census aggregated statistics utilized to improve accuracy in nonprobability samples*
  Under this scenario, aggregated population census data matching to *primary* and, additionally, *secondary demographics*[7] from a nonprobability sample are used to adjust the sample distribution for those selected auxiliary variables to match the population distribution (e.g., besides for *sex*, *age*, and *education*, *employment status* covariate can be included in the post-survey adjustment).
- *Scenario 3 – availability of census aggregated statistics and a representative source of non-demographic benchmarks (i.e., a large national survey) utilized to improve accuracy in nonprobability samples*
  Under this scenario, besides the aggregated population census data from Scenario 1, we can use *secondary demographics* and *non-demographics* from a large probability-based national survey (e.g., *household composition* and *health status* from a government survey on health) that are matching to those covariates in the nonprobability sample. This time, microdata[8] are a source of *secondary demographics* and *non-demographics*.

[5] Aggregated or tabular data are produced by grouping information into categories. Within these categories, values are combined (e.g., a count of respondents of particular age). They are also known as macrodata (Australian Bureau of Statistics n.d.-b).

[6] Primary demographics as defined by Pennay et al. (2018) are socio-demographic variables which were used in post-stratification weighting.

[7] In contrast to primary demographics, secondary demographics as defined and used by Pennay et al. (2018) were additional socio-demographic variables which were not included in post-stratification weighting but rather in accuracy calculations only (such as Indigenous status or voluntary work).

[8] Microdata, also known as unit record files, are a type of data including unit records containing detailed information about analytical units such as persons or organizations. They often include individual responses to survey questions or from administrative forms (Australian Bureau of Statistics n.d.-b).

- *Scenario 4 – availability of census aggregated statistics, and a smaller scale probability-based survey data utilized to improve accuracy in nonprobability samples*
  Under this scenario, besides the aggregated population census data from Scenario 1, we can use *non-demographics* from a smaller-scale non-government survey that are matching to selected covariates in the nonprobability sample. While we apply a less representative external data source to improve accuracy, there are additional non-demographic covariates which could be used to balance the samples as noted in the literature. An example of those non-demographics is 'webographic' variables, which are available in a microdata form. Webographic variables are attitudinal or lifestyle variables accounting the difference between web survey participants and those who do not do surveys online (Baker et al. 2013). Different authors considered different questions as 'webographic' questions, such as: feeling alone, eagerness to learn new things, willingness to take chances, lifestyle questions (on travelling, participation in sports, reading a book), opinions on what is a violation of privacy, knowing a 'lesbian, gay, bisexual, transgender, and queer or questioning' (LGBTQ) person (Schonlau et al. 2007), early-adopter items (DiSogra et al. 2011; Dutwin & Buskirk 2017) or media use (Baker et al. 2013). On the other hand, Mercer et al. (2018) used political attitude variables in post-survey adjustments. In our study, besides early-adopter items, we also consider internet connection, access and use, and number of surveys completed as 'webographic' variables or, simpler, 'webographics' (see Table 10 in the Appendix for more information).

The difference between Scenarios 3 and 4 is the type and the source of auxiliary survey data available for post-survey adjustment. Under Scenario 3, we have access to a large-scale nationally representative survey (large sample, e.g., 20,000+, with higher accuracy), such as the National Drug Strategy Household Survey. Under Scenario 4, we can use a smaller probability-based sample (e.g., about n=600), but with an ability to collect tailor-made data including key covariates which could help mitigate bias after matching or propensity scoring weighting (e.g., 'webographics'); data collectors attempting to improve the accuracy of their nonprobability samples could conduct a smaller-scale probability-based survey, e.g., a probability-based sample from Online Panels Benchmarking Study, to improve inference in opt-in panel samples.

This study will address the following research question: *How accurate are nonprobability online samples in comparison to probability samples and to what extent can inference be improved by using post-survey adjustment methods under different scenarios?*

Methods

Data

Original Online Panel Benchmarking Study (2015 OPBS)

The 2015 Online Panels Benchmarking Study (OPBS, Pennay et al. 2016[9]) was conducted in June 2015 and administered the same questionnaire to eight samples, made up of three probability samples and five nonprobability online panel samples. Each sample aimed to achieve approximately six hundred completed interviews; in the end, the smallest sample comprised of 538 respondents (Pennay et al. 2018), as presented in Table 1. The design was similar to the US study by Yeager et al. (2011) which compared the accuracy of seven online samples and two probability samples. The main objective of OPBS was to inform the debate in Australia on the issues pertaining to inference from nonprobability online panel surveys.

Life in Australia™ – Probability-Based Online Panel: OPBS Replication (2017 OPBS)

Life in Australia™ is a probability-based internet panel for the Australian general adult population, and in January-February 2017, all active Life in Australia™ panellists were asked to participate in the replication of the OPBS. Social Research Centre administered the same questionnaire used for the original 2015 OPBS to determine the accuracy of their probability-based online panel (Kaczmirek et al. 2019). This was the second wave of Life in Australia, referred to as the Online Panel Benchmarking Study Replication or 2017 OPBS (Pennay & Neiger 2020[10]).

Life in Australia™ panellists were recruited in 2016 via their landline or mobile phones to take part in incentivized monthly surveys, and the final sample of registered panellists was 3,322 individuals (overall recruitment rate, AAPOR RR3: 15.5%). Since the recruitment of panellists was through probability-based dual-frame sampling, the results from the surveys are generalizable to the Australian population. Life in Australia™ is a mixed-mode probability online panel, and to take into account the population with no access to the internet, the study also contacted panel members who happened to be offline via phone (representing 13.6% of Wave 2 sample) (Kaczmirek et al. 2019).

Population, Sampling and Samples

Both the 2015 and 2017 OPBS surveys collected information from an in-scope population of all Australians aged 18 years and over. The studies were carefully designed to assess accuracy of nonprobability online panel samples relative to prob-

[9] Data DOI: 10.4225/87/FSOYQI

[10] Data DOI: 10.26193/YF8AF1

ability-based surveys using different probabilistic sampling methodology through applying the same data collection instrument to provide data on the demographic, social characteristics and wellbeing of people in Australia (Kaczmirek et al. 2019; Pennay et al. 2018).

As previously explained, the OPBS 2015 study data comprised of eight samples, three of which were probability-based samples: (i) an address-based sampling (A-BS) survey with Geocoded National Address File (G-NAF) as a sampling frame (survey mode: hard copy/mail, online, telephone), (ii) a standalone dual-frame Random Digit Dialing (RDD) survey sample (survey mode: telephone), and (iii) a RDD end-of-survey recruitment sample (survey mode: telephone, online, hard/copy) (Pennay et al. 2018), also known as ‘piggybacking’ survey sample (Tourangeau & Smith 1985). For the purpose of the 2015 OPBS study, five Australian nonprobability online panels collected data from about 600 of their panellists each. Four of five nonprobability online panels complied with all ESOMAR’s questions to help online research buyers[11] and three of the five were with ISO 26362 accreditation[12] (Pennay et al. 2018). We will analyze accuracy of the whole nonprobability sample combined[13] (n=3,058) and for two purposely selected nonprobability samples, the most and the least accurate.

The OPBS Replication 2017 survey comprised of one probability-based mixed-mode (online and telephone) sample. The cumulative response rate (CUMRR1), which is a product of overall recruitment (RECR x PROR) and survey completion rates (COMR)[14], was 12.2% (AAPOR RR3). A total of 2,580 Life in Australia™ panellists completed Wave 2 questionnaire (Kaczmirek et al. 2019).

[11] ESOMAR’s *Questions to help buyers of online samples* include questions on company profile (such as *What experience does your company have in providing online samples for market research?*), sample sources and recruitment (such as *Is the recruitment process ‘open to all’ or by invitation only?*), sampling and project management (such as *Do you employ a survey router or any yield management techniques?*), data quality and validation (such as *How often can the same individual participate in a survey?*), policies and compliance (such as *How can participants provide, manage and revise consent for the processing of their personal data?*) and metrics (*Which of the following [metrics] are you able to provide to buyers, in aggregate and by country and source?*). For more information, see ESOMAR (2021).

[12] ISO 26362:2009 developed criteria and specified terms, definitions and service requirements for organisations managing online panels, including on sampling, fieldwork, and data management. It has since been revised by ISO 20259:2019 standard (International Organisation for Standardisation 2022).

[13] Combining data from several volunteer panels can increase their overall accuracy (Cornesse et al. 2020), can be thus considered a solution to mitigate representation bias in nonprobability surveys, and is as such a subject of this study. We were particularly interested in the effectiveness of post-survey adjustment on combined data from different nonprobability sources, in comparison to individual volunteer panel samples.

[14] Recruitment rate, completion rate, and cumulative response rate were introduced by Callegaro and DiSogra (2008a) for calculation of response rates in online panels.

Table 1 Studies and subsamples analyzed

Study	Subsample	Response rate (AAPOR RR3)	n[a]
Online Panels Benchmarking Study (2015 OPBS)	Address-based sampling	26.2%	538
	Standalone RDD (dual-frame)	14.7%	600
	RDD “piggybacking” (dual-frame)	9.8%	560
	5 volunteer panel samples[b]	2.6%-15.4%[c]	3,058
Online Panels Benchmarking Study Replication (2017 OPBS)	Life in Australia™ Wave 2	recruitment rate: 15.5%, Wave 2 survey completion rate: 78.6%, cumulative response rate: 12.2%	2,580

[a] We have to acknowledge the fact that with relatively small sample (n=about 600), sampling variance as a component of sampling error is larger. In practice this means that estimates from surveys with smaller samples can be less accurate in benchmarking studies by chance in comparison to those from larger surveys.

[b] Besides the combined nonprobability sample, we will analyze data separately for the most accurate panel (Panel 3, n=601) and the least accurate panel (Panel 1, n=601) (based on the results from Kaczmirek et al. 2019, p. 25). We will not analyze data for all 5 nonprobability panels separately due to space constraints. However, through comparing the best and worst performing nonprobability panel, we can get an indication of the variation in the bias and accuracy of different panel providers.

[c] For nonprobability samples, response rates cannot be calculated and some authors (e.g., Pennay et al. 2018) report sample yields instead.

Generally speaking, there were notable differences in response between the subsamples listed in Table 1, which might result in different levels of nonresponse error. The hope is that we can mitigate against this in our analysis through effective post-survey adjustment procedures applied to nonprobability data.

Benchmarks

Assessing quality in surveys requires an objective standard to which the survey estimates can be compared, such as population benchmarks. Differences between estimates from survey response and population benchmarks can occur through bias or variance, where the bias term captures the systematic (selection) errors that are shared by nonprobability samples. The variance term captures the sampling variation and accounts for the variation due to the differences in survey protocols, statistical modeling or weighting adjustments.

To replicate benchmarking analysis from Pennay et al. (2018)[15] and Kaczmirek et al. (2019), we use the same benchmarks but from updated data sources collected closer in time to 2015 OPBS and 2017 OPBS studies. We primarily use information from the Australian quinquennial Census (Australian Bureau of Statistics 2016) as benchmarks since censuses offer universal coverage of the population by definition. For some instances we use administrative record data and information drawn from large government surveys as benchmarks. Those are electoral registration information from the Australian Electoral Commission, and social and health characteristics from the government funded surveys which are considered as the best quality sources of nationally representative benchmarks in Australia with the highest validity (e.g., Australian Bureau of Statistics 2018b).

Benchmarks will be divided into primary (for post-survey adjustment only), secondary demographics, and substantive items (see Pennay et al. 2018). Table 2 provides a description of the benchmarks used in the study.

[15] The findings presented in Pennay et al. (2018) were further explored and published by Lavrakas et al. (2022).

Table 2 Benchmarking data sources and nationally representative benchmarks

Study	Data collection mode	Sample size	Benchmarks *(modal response category in brackets)* [a]*primary demographics,* [b]*secondary demographics,* [c]*substantive items/non-demographics*
Australian Census 2016 (Australian Bureau of Statistics 2016)	self-administered online, F2F	n=23,401,892 persons	Age, in categories [a] Gender[a] State[a] Residence in state capital city[a] Country of birth[a] Australian citizenship[b] *(Australian citizen)* Employment status[b] *(currently employed)* Home ownership[b] *(with a mortgage)* Indigenous status[b] *(not Indigenous)* Language other than English[b] *(speak only English)* Living at last address 5 years ago[b] Most disadvantaged quintile for area-based socio-economic score[b] Resident of a major city[b] Voluntary work[b]
National Drug Strategy Household Survey (NDSHS) 2016 (Hewitt 2017)	self-administered paper-based or online, CATI	n=23,749 persons	Household status[b] *(couple with dependent children)* Smoking status[c] *(daily smoker)* Alcoholic drink of any kind in the past 12 months[c] (y*es, consumed alcohol)*
National Health Survey 2014-15	F2F	n=19,259 persons	Psychological distress (Kessler 6)[c] *(low distress)* General health[c] *(very good)* Private health insurance[c] *(yes, has insurance)* Wage and salary income[b] *(income $1000-1249 pw)*
General Social Survey 2014	F2F	n= 12,932 persons	Life satisfaction[c] *(8 out of 10)*
Australian Electoral Commission (2015)	administrative data	n=16,405,465 persons	Enrolled to vote[b] *(yes, enrolled)*

F2F – face-to-face; CATI – Computer-assisted telephone interviewing

Data Analysis

Benchmarking Analysis

To carry out our benchmark analysis, we need to balance against variance and bias in the final estimates. There are a wide variety of measures estimating the bias, such as the number of statistically significant differences from the benchmarks, the average absolute error (AAE) (including measures of uncertainty of the AAE, such as the standard deviation of the AAE or the range and ranking) (see Dutwin & Buskirk 2017; MacInnis et al. 2018; Yeager et al. 2011). To provide a measure of the variance, we compute the mean squared error which is a function of both the bias and the variance, and as such it is a good measure of the overall accuracy of the different approaches; it is usual practice to take the square root of the mean square error (RMSE) which is more sensitive to large errors than AAE. The aim of the study is to find the approach which is robust under the different scenarios. As such we present results using the AAE and RMSE to give an absolute measure of the error and the variability measure of the error, respectively.[16]

The AAE was used by Yeager et al. (2011) to compare impact of different weighting approaches for probability and nonprobability surveys in the US. The same measure was used by Pennay et al. (2018) and Kaczmirek et al. (2019), who replicated the study design in Yeager et al. (2011) for Australia.

Our study follows all three of these previous studies, and the AAE is calculated as presented in Equation 2:

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \tag{2}$$

where $\widehat{y_j}$ is the j-th estimate (of a survey item) and y_j is the value for a corresponding (population) benchmark. And similarly, the RMSE is computed as presented in Equation 3:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{k} (\widehat{y_j} - y_j)^2}{k}} \tag{3}$$

where k is the number of benchmarks, $\widehat{y_j}$ is again the j-th estimate from either OPBS surveys, and y_j is the value for a corresponding benchmark. In our study, the estimates ($\widehat{y_j}$) represented proportion estimates for modal response for items with corresponding benchmarks; this is consistent with the approach from the Australian benchmarking studies (Kaczmirek et al. 2019; Pennay et al. 2018) and the US studies described in the literature (e.g., Yeager et al. 2011).

[16] When we computed Relative Absolute Bias (see Dutwin & Buskirk 2021) as a relative measure, we reached the same conclusions about the accuracy of probability and nonprobability samples as when computing AAE as an absolute measure.

To explore the generalizability of these findings, we calculate AAE and RMSE for 12 secondary demographics, 6 substantive items, and all 18 survey items with corresponding benchmarks combined. Most probability and nonprobability surveys apply adjustment for primary benchmarks as a standard approach, and for the majority of surveys the differences between the sample and population for primary benchmarks is expected to be minimal (Cornesse et al. 2020; Mercer et al. 2017). The analysis was facilitated by the statistical coding environment and language R (R Core Team 2020) to carry out all data processing, post-survey adjustments, imputation of missing values[17] and benchmarking analyses. Besides R base or stats packages, the following packages were used: *Hmisc* (for data processing, Harrell et al. 2020), *missForest* (for imputation of missing values, Stekhoven 2013), *fastDummies* (to create dummy variables for MDM, Kaplan 2020), *anesrake* (to perform raking, Pasek 2018), *sjstats* (for data processing, Lüdecke 2020), *questionr* (for data processing, Barnier et al. 2020), *MatchingFrontier* (to perform MDM, King et al. 2015), *cem* (to perform CEM, Iacus et al. 2020), and *rstanarm* (to conduct dominance analysis, Goodrich et al. 2020).

Post-Survey Adjustment Approaches and Parameters

Methods. To improve inference in nonprobability samples, we will test a number of post-survey adjustment methods and techniques:

- raking[18]
- generalized regression estimation (GREG)
- multilevel regression and poststratification (MRP)
- coarsened exact matching (CEM)
- Mahalanobis distance matching (MDM)
- propensity score weighting (PSW).

PSW, MDM and CEM selection/weighting will be later adjusted to match primary demographic benchmarks from Australian Census 2016. This means that those methods will be combined with raking not to introduce bias due to any socio-demographic sample imbalance after the initial adjustment. For more information about each of these methods, see Subsection 2.2, and for post-survey adjustment details from this study, see Table 3.

[17] We imputed missing values using random forest imputation algorithm, which is suitable for both continuous and categorical variables. Missing values were imputed for calibration and matching purposes only, and not for estimation, which means that only valid values of items with corresponding benchmarks were used in calculations of estimates.

[18] In probability samples, a two-stage process can be used for weighting, first calculating a design weight (for the unequal probability of sample members being selected) and second raking (to reduce possible nonresponse). As the same process cannot be used for weighting nonprobability samples, and as the findings on the accuracy of nonprobability samples would not change (see Kaczmirek et al. 2019), we used a consistent one-stage raking approach across all samples (and calibration methods).

Table 3 Post-survey methods, covariates, and parameters

Method	Scenario	Type of covariates	Covariate selection mechanism	Source of covariates	Other post-survey adjustment characteristics
Raking	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	We applied weight trimming to ensure that the maximum weight after post-survey adjustment was 5.
	Scenario 2	Primary demographics(1) Secondary demographics	Weighting variables from the original benchmarking studies Covariates with the largest absolute error relative to Census benchmarks	Australian Census 2016	
	Scenario 3	Primary demographics(1) Non-demographics	Weighting variables from the original benchmarking studies All matching additional covariates from a large-scale survey	Australian Census 2016 NDSHS 2016	
GREG	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	
	Scenario 2	Primary demographics(1) Secondary demographics	Weighting variables from the original benchmarking studies Covariates with the largest absolute error relative to Census benchmarks	Australian Census 2016	
	Scenario 3	Primary demographics(1) Secondary demographics & non-demographics	Weighting variables from the original benchmarking studies All matching additional covariates from a large-scale survey	Australian Census 2016 NDSHS 2016	
MRP	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	

Method	Scenario	Type of covariates	Covariate selection mechanism	Source of covariates	Other post-survey adjustment characteristics
CEM	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	Pruning(3) of maximum 50% of all nonprobability sample units; adjusted to match primary demographic(1) benchmarks
	Scenario 4	Non-demographics, including 'webographics')	Selected with dominance analysis(2) out of all available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	
MDM	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	The same matched sample sizes as for CEM; adjusted to match primary demographic(1) benchmarks
	Scenario 4	Non-demographics, including 'webographics')	All available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	
PSW	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	Adjusted to match primary demographic(1) benchmarks (the whole approach could also be called “doubly-robust”)
	Scenario 4	Non-demographics, including 'webographics')	All available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	

(1) Primary demographics from Australian Census 2016 (and used in the original benchmarking studies by Pennay et al. 2018 and Kaczmirek et al. 2019) were gender, country of birth, and interaction effects between age and education, and state and capital city in state.
(2) Dominance analysis is used to compare the relative importance of predictors in regression models by comparing R2 or Pseudo R2 coefficient with different ranges of selected predictors (Budescu 1993). In practice, with dominance analysis we can select the covariates which distinguish probability and nonprobability samples the most in a multivariate setting (logistic regression was used in our case).
(3) Removing those units from nonprobability data which cannot be matched with any unit from a probability sample.

Based on the literature review from Subsection 2.2, we selected all post-survey adjustment methods appropriate for use with particular types of data. While raking, GREG and MRP can be used with tabular data and estimates from survey microdata (or both at the same time), weighting schemes should include estimates from nationally representative data sources producing known population totals (Kalton & Flores-Cervantes 2003, p. 82); hence, raking, GREG and MRP are not analyzed under Scenario 4, i.e., with a smaller scale probability survey data producing rough estimates of population totals. Also, a disadvantage of MRP is the requirement of the joint distribution of the poststratification variables, and CEM, MDM and PSW can only be used with microdata, i.e., under Scenarios 3 and 4.

Covariates. The theory explains that the selection of covariates for post-survey adjustment should be based on the relationship with nonresponse and non-coverage (e.g., Battaglia et al. 2009). Kalton and Flores-Cervantes (2003) pointed out that the precision of estimates can be increased by benchmarking to external sources with covariates that are closely related to key survey variables. The literature on post-survey adjustment in nonprobability samples (e.g., Dutwin & Buskirk 2017) suggests using covariates that are associated with participation in nonprobability samples/online panels, in attempt to primarily reduce errors associated with coverage, and adjust for inherent selection bias. We will follow these general recommendations/principles by:

- selecting secondary demographic covariates with the largest absolute error relative to Census benchmarks under Scenario 2 – our assumption is that those socio-demographic differences are directly associated with undercoverage (and nonresponse) bias in nonprobability online panels;
- selecting all matching health-related items (besides a secondary demographic item) to reduce error of other health-related items under Scenario 3 – if adjustment covariates are closely related to the target outcome variables, bias could be mitigated
- selecting non-demographic covariates which were previously discussed in the literature as effective in reducing coverage error in non-probability samples, so-called ‘webographic’ variables, under Scenario 4;
- identifying a limited number of ‘webographic’ covariates which distinguish nonprobability and probability samples the most, to be used with CEM under Scenario 4.

At the same time, validity of the sample has to be preserved by including core demographics like age and gender; in the case of calibration, we also have to have in mind that selecting too many covariates can lead to significant variance inflation and inability for raking algorithm to converge (Battaglia et al. 2009).

For details on the final selection of covariates, applied with different methods and under different scenarios, please see *Final selection of post-survey adjustment covariates* section and Table 10 in the Appendix.

Results

Accuracy of Nonprobability Online Panels

The results in this section provide updated evidence regarding the accuracy of nonprobability online samples in comparison to probability samples (with more recent benchmarks, for original results see Kaczmirek et al. 2019). We will use the identified gap in accuracy as a reference for assessment of effectiveness of post-survey adjustments (see Section 4.2).

Table 5 presents the results on the accuracy of OPBS 2015 and OPBS 2017 Replication surveys. The results confirm the findings from Pennay et al. (2018) and Kaczmirek et al. (2019) on the accuracy of nonprobability-based online panels in comparison to probability samples, as well as that raking as a post-survey adjustment method improves the quality of estimates from probability surveys more effectively than for nonprobability-based online panels. While nonprobability panel samples are similarly accurate in measuring secondary demographics as probability samples (AAE: nonprobability samples 4.7-5.4, probability samples 4.2-5.3, all raked), they are less accurate in measuring non-demographics than probability surveys (AAE: nonprobability samples 6.6-9.9, probability samples 3.7-5.4, all raked), which is also confirmed by RMSE measures. We would particularly like to reduce the non-demographic bias with various post-survey adjustments.

Assessment of Effectiveness of Post-Survey Adjustment Methods for Improving Inference in Nonprobability Samples

In this section, we will show if the difference in accuracy between probability and nonprobability samples, i.e., representation bias, can be reduced using different post-survey adjustment methods. The results will be presented by scenarios based on the availability of external data and, as previously explained, not all methods can be used with all data types. Importantly, we will use Life in Australia™ Wave 2 sample as a reference sample for post-survey adjustment efficiency. This sample has been selected as it is similarly accurate as the OPBS 2015 probability samples (see online Appendix Table 5), yet with a much larger sample size (smaller sampling variance) and greater comparability with nonprobability samples in terms of the survey mode (online: 86.4% in Life in Australia™ Wave 2, 100% in volunteer samples). We will use AAE for the raked[19] Life in Australia™ sample, and no further post-survey adjustment will be carried out with this probability sample. Fundamentally, we will assess (i) the efficiency of post-survey adjustments with nonprobability samples relative to (ii) the accuracy of probability-based online panel estimates normally reported in practice (i.e., calibrated using primary demographics).

Scenario 1: Availability of Census Aggregated Statistics, and Only Primary Demographics were Collected from the Nonprobability Sample

To illustrate the effectiveness of post-survey adjustments (i.e., raking, GREG and MRP) using primary demographics (i.e., performing 'basic calibration'), we are presenting results for unweighted and weighted data for the nonprobability online samples in Figure 1 (see Table 6 from the Appendix for more detailed results). The presented evidence shows how basic weighting post-survey adjustments improve the quality of estimates, but the improvement is only slight on average (AAE combined reduction between 0.4 [GREG, Panel 3] and 0.7 [MRP, Panel 1]). We can confirm our previous finding on how raking improves the accuracy of nonprobability samples to a lesser extent than those from probability samples. We can also extend this finding to other calibration methods studied in this article – GREG and MRP.

The improvement in accuracy is more apparent for all 18 survey items combined than for six substantive items combined, which indicates that calibration using primary demographic more consistently improves the quality of secondary demographic estimates than non-demographic estimates. Moreover, the results from Figure 1 show how calibration can deteriorate substantive item estimates from nonprobability samples, especially the least accurate one, but also the combined volunteer panel sample. This is consistent across all calibration methods, with MRP performing just slightly better than GREG and raking. On the other hand, weighting improved accuracy of the most accurate nonprobability panel in a similar fashion for both secondary demographics and non-demographics.

We have to note that the differences in item-level results (not only at the AAE level, see Table 6 in the Appendix) are almost non-existent for raking and GREG and very little between the first two calibration methods and MRP. Based on this finding, as well as due to the limitations of MRP (i.e., requiring a joint distribution), we will only assess the efficiency of the first two calibration methods under Scenario 2. Also, the results for basic raking from Scenario 1 are included as a reference method for Scenarios 2, 3 and 4 (see Figures 2, 3 and 4).

[19] By gender, age group*education (interaction), country of birth, state*capital city in state (interaction)

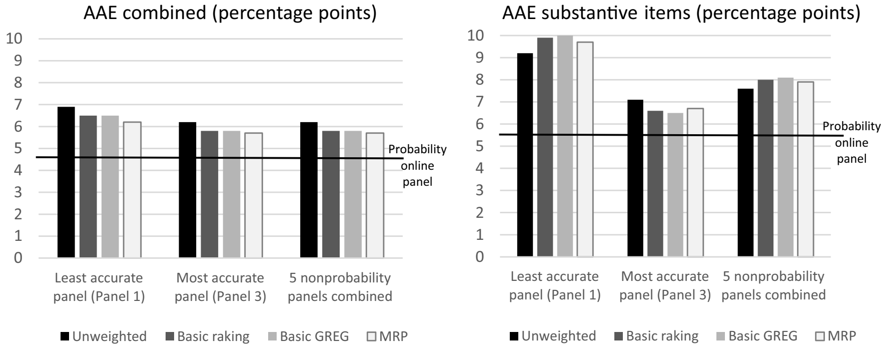

Figure 1 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 1 - average absolute error (AAE) for all sample estimates (see Table 6), un- and weighted (raking, GREG, MRP)*

* AAE for secondary demographics and all RMSE calculations (combined, secondary demographics, and substantive items) are presented in the tables in the Appendix.

Scenario 2: Availability of Census Aggregated Statistics, Both Primary and Secondary Demographics were Collected from the Nonprobability Sample

To illustrate how including new covariates in calibration further improves the accuracy of nonprobability samples, additional socio-demographic items with corresponding census benchmarks were added[20] and 'expanded' calibration 1 (e.g., expanded raking 1) was performed (see Figure 2). The presented evidence suggests that expanded raking and GREG predominantly improved secondary demographic estimates and, in some cases, estimates from substantive items (see Table 7 from the Appendix for more detailed results). For the most and the least accurate online panel, as well as all panels combined, we can see a slight improvement in the combined AAE and RMSE. Generally speaking, we can again report almost negligible differences between estimates adjusted with expanded raking and expanded GREG. We also did not notice a significant increase of design effect compared to basic raking.

Moreover, this time calibration did not increase AAE for substantive items for the least accurate panel and five panels combined. Including three secondary demographic covariates seemed to eliminate the negative effect of raking with primary demographics only. Moreover, we can notice a notable improvement in accuracy of substantive items after using an expanded raking scheme for the most

[20] For more information on selection of additional covariates under Scenarios 2, 3 and 4 (e.g., employment status, language other than English, and voluntary work under Scenario 2), see *Post-survey adjustment approaches and parameters* section (Methods) and Table 10 (Appendix).

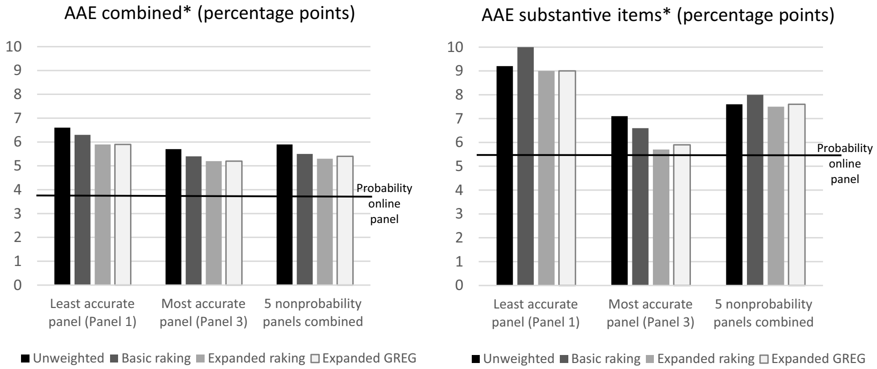

Figure 2 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 2 - average absolute error (AAE) for all sample estimates (see Table 7), unweighted and weighted (raking, GREG)

*AAE were calculated for all items excluding the secondary demographics included in an expanded calibration scheme (employment status, language other than English (LOTE), and voluntary work, see Table 7 in the Appendix for more information)

accurate nonprobability panel (AAE: unweighted 7.1, raking 5.7, GREG 5.9). The selected secondary demographic items seem to be more associated with representation bias in the most accurate nonprobability online panel than our core/primary demographics.

The evidence from Figures 1 and 2 suggests that the highest-quality nonprobability online panels are not only the most accurate for unweighted estimates, but they also respond better to various calibration adjustments.

Scenario 3: Availability of Census Aggregated Statistics and One Other Representative Source of Benchmarks

To illustrate potential added value of having access to an additional external high-quality data source with non-demographic matching covariates, we are presenting results for 'expanded' calibration 2, CEM, MDM, and PSW in Figure 3. The presented evidence shows how including new non-demographic covariates in post-survey adjustment improves the accuracy of nonprobability samples fairly similarly to including new secondary demographic covariates. However, the improvement seems to be more substantial under Scenario 3 – an increase in accuracy measured with AAE combined ranges from 0.4 (Panel 1, MDM) to 1.8 (Panel 1, CEM).

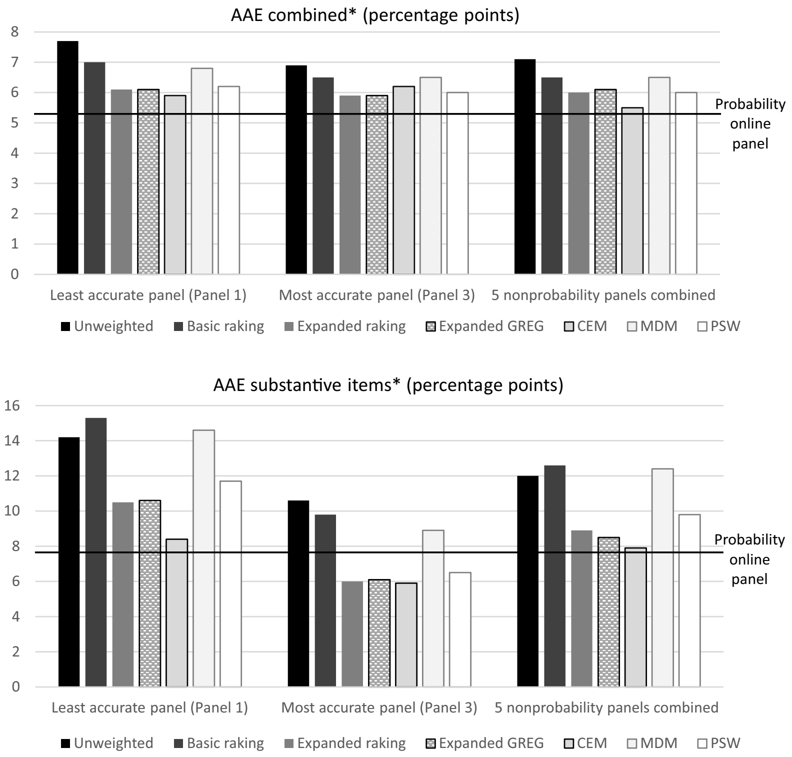

Figure 3 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 3 - average absolute error (AAE) for all sample estimates (see Table 8), unweighted and adjusted post-survey (raking, GREG, CEM, MDM, PSW)

*AAE were calculated for all items excluding the covariates in an expanded post-survey adjustment scheme (household status, frequency of smoking, and drinking alcohol, see Table 8 in the Appendix for more information)

In comparison to the efficiency of calibration under Scenario 2, including non-demographic covariates improved the accuracy of substantive items[21] to a greater extent. The decrease in that AAE (substantive items) was as high as 5.8 (Panel 3, CEM). Generally speaking, post-survey adjustment with a limited number of covariates was more efficient with calibration (raking, GREG) and CEM than distance-based models, i.e., PSW and especially MDM. While CEM seems to com-

[21] The remaining three substantive items for benchmarking were from National Health Survey 2014-15 and General Social Survey 2014 (see Table 2).

pare favourably to other methods using covariates from a large-scale survey, we noticed a larger design effect than for expanded raking 2.

All in all, post-survey adjustment with expanded raking, GREG and CEM under Scenario 3 made nonprobability online panels almost as accurate as a probability-based online panel overall (AAE combined). For the three remaining substantive items, the most accurate nonprobability online panel (Panel 3) was even more accurate after advanced adjustments than the probability online panel after basic raking.

Scenario 4: Availability of Census Aggregated Statistics and a Smaller-Scale Probability-Based Survey Data with Matching Variables from Nonprobability-Based Survey Data

To illustrate potential added value of having access to a smaller-scale external survey data source (i.e., OPBS 2017 replication sample from a probability online panel) with non-demographic matching covariates, we are presenting results for CEM, MDM, and PSW[22] in Figure 4.

The results present mixed evidence on the efficiency of post-survey adjustment methods using smaller-scale external survey data with no demographics or health-associated items. First, there was a fairly moderate and inconsistent effect of post-survey adjustments on the total accuracy of nonprobability samples. In most cases, the decrease of AAE combined was less than 0.5, and no method seemed to have a clear advantage. The only exception to the rule was MDM with the data from five nonprobability-based panels combined (AAE: unweighted 6.2, MDM 5.2, probability panel 4.6). Overall, basic raking with primary demographics from Australian Census seems to be a more reliable method than any other method for improving the combined accuracy of secondary demographic and non-demographic estimates with webographics.

Comparing AAE for substantive items, we can observe as many instances of post-survey adjustment deteriorating estimates as instances of improving estimates. The least accurate nonprobability-based panel stands out as the sample with no decrease in AAE before or after adjustment, and CEM as the method with limited efficiency for only one sample (the most accurate). The best result overall can again be attributed to MDM (AAE: unweighted 7.6, MDM 6.5, probability panel 5.4), and we can also see a positive effect of PSW on the accuracy of Panel 3 (AAE: unweighted 7.1, PSW 6.0, probability panel 5.4).

[22] A variety of other methods and their combinations would be possible under this scenario with auxiliary microdata, including calibration such as raking, GREG and MRP. However, calibration is normally carried out with benchmarks from the highest-quality censuses or large-scale surveys, and smaller-scale probability-based survey tend to introduce more error (see Table 5, probability samples).

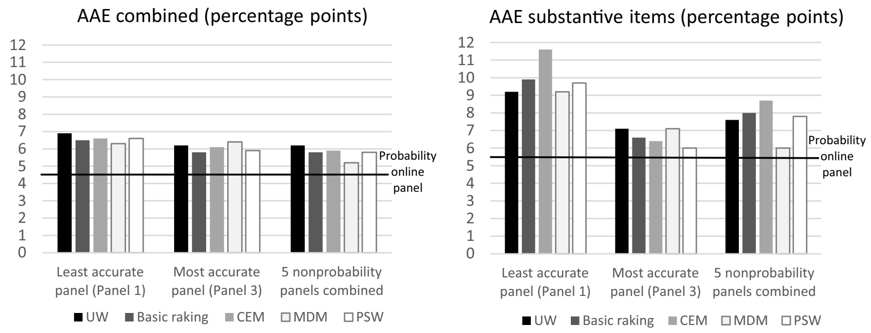

Figure 4 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 4 - average absolute error (AAE) for all sample estimates (see Table 9), unweighted and adjusted (raking and matching methods)

Summary of Post-Survey Adjustment Efficiency

To sum up, we are presenting a review of all post-survey adjustment results by four data availability scenarios. All AAE combined values from Scenarios 1-4 and associated AAE reduction % (as a proportion of unadjusted/unweighted AAE) are now combined.

Based on the results from Table 4 (as well as Figures 1-4), we are offering the following main findings of our study:

- the best post-survey adjustment results can be expected under Scenario 3, i.e., by using a combination of primary, secondary, and non-demographic covariates from nationally representative data sources;
- expanded calibration with additional secondary demographic covariates further improves accuracy, in comparison to basic calibration with primary demographic covariates (and to a similar extent);
- secondary demographic covariates seem to have a better potential to improve the accuracy of secondary demographic estimates, and non-demographic covariates seem to have a better potential to improve the accuracy of non-demographic estimates (in this particular study, those were health-related items);
- webographics from probability-based online panel survey data did not consistently improve the accuracy of nonprobability samples (see Scenario 4 results);
- there are some observable differences between the analyzed methods, albeit they are little in this study, and MDM was the method with the least consistent results;
- while we could not reduce error by more than 23% no matter the chosen auxiliary data, covariates or methods, we have to note that the probability samples

Table 4 Post-survey adjustments efficiency for all methods and samples under four scenarios

Scenario	Post-survey adjustment method	Least accurate nonprobability panel (1), AAE combined			Most accurate nonprobability panel (3), AAE combined			5 nonprobability panels combined, AAE combined		
		AAE (UW)	AAE (adjusted)	AAE reduction (%)	AAE (UW)	AAE (adjusted)	AAE reduction (%)	AAE (UW)	AAE (adjusted)	AAE reduction (%)
Scenario 1	Basic raking	6.9	6.5	6%	6.2	5.8	6%	6.2	5.8	6%
	Basic GREG		6.5	6%		5.8	6%		5.8	6%
	MRP		6.2	10%		5.7	8%		5.7	8%
Scenario 2*	Basic raking	6.6	6.3	4%	5.7	5.4	5%	5.9	5.5	7%
	Expanded raking 1		5.9	11%		5.2	9%		5.3	10%
	Expanded GREG 1		5.9	11%		5.2	9%		5.4	8%
Scenario 3*	Basic raking	7.7	7.0	9%	6.9	6.9	0%	7.1	6.5	8%
	Expanded raking 2		6.1	21%		5.9	14%		6.0	15%
	Expanded GREG 2		6.1	21%		5.9	14%		6.1	14%
	CEM		5.9	23%		6.2	10%		5.5	23%
	MDM		6.8	12%		6.5	6%		6.5	8%
	PSW		6.2	19%		6.0	13%		6.0	15%
Scenario 4	Basic raking	6.9	6.5	6%	6.2	5.8	6%	6.2	5.8	6%
	CEM		6.6	4%		6.1	2%		5.9	5%
	MDM		6.3	9%		6.4	-3%		5.2	16%
	PSW		6.6	4%		5.9	5%		5.8	6%

*AAE were calculated for all items excluding the covariates in an expanded post-survey adjustment scheme, UW – unweighted estimates from OPBS 2015 and the OPBS 2017 Replication sample were about 20-30% more accurate than the studied nonprobability samples[23].

Discussion and Conclusion

This investigation into improving inference in nonprobability sample surveys supports the conclusion that the issue of improving inference in nonprobability sample surveys is a three-dimensional problem. First, the quality of post-survey adjustments is dependent on the availability of relevant high-quality covariates which are associated with either representation bias in nonprobability samples or outcome variables. Second, as the covariates in nonprobability samples should have matching covariates in external representative data sources, the availability and ability to access auxiliary data is a key aspect in mitigating bias. Third, the efficiency of post-survey adjustments is also dependent on the selection and combination of post-survey adjustment methods, albeit to a lesser extent.

In this study, we presented evidence that post-survey adjustment can reduce representation bias in nonprobability online samples to some extent, but cannot consistently eliminate it. These findings are in line with evidence from Tourangeau et al. (2014) and Kalton and Flores-Cervantes (2003). However, we demonstrated a greater potential to mitigate representation bias in nonprobability panels if having access to more external data sources and more covariates matching in nonprobability samples and auxiliary data. Ideally, we would have access to large-scale survey microdata, since smaller-scale surveys come with some nonignorable error. While those probability surveys mostly remain more accurate than nonprobability surveys even after post-survey adjustments, they are more susceptible to coverage, sampling, and nonresponse error (or even measurement mode effect) than most high-quality government surveys, and the total representation error can be carried over to post-survey adjustment results (e.g., after matching or PSW). For that reason, improving inference in nonprobability samples should be planned in the survey design stage, and relevant external data sources reviewed before data collection, if possible.

Moreover, identification of covariates from external data sources which are associated with representation bias or target outcome variables can lead to a more efficient mitigation of bias. While post-survey adjustments using primary demographics have little positive effect on the quality of nonprobability estimates, we have shown how including secondary demographics can improve the quality of other demographics and including non-demographics can decrease the error from

[23] This research did not take into account that the accuracy of probability samples could be further improved with the same post-survey adjustment methods including secondary demographic and non-demographic items.

associated non-demographics. This is consistent with findings from Bethlehem (2002). Similarly, Mercer et al. (2018) reported that including political attitude covariates in adjustment improved the quality of political engagement estimates. However, we found inconsistent evidence on the suitability of ‘webographics’ and other internet-associated covariates for mitigating bias in nonprobability samples. Unfortunately, we could not distinguish between the effect of those covariates and the effect of the data source on the post-survey adjustment efficiency. While auxiliary variables like early adopter items (traditionally used to mitigate bias in nonprobability samples, e.g., DiSogra et al. 2011) did not distinguish our probability online sample and nonprobability online panel samples well, we identified new covariates for post-survey adjustment that could be considered as ‘webographics’, such as the number of surveys participated in. Therefore, we believe it is crucial to carry out more investigation into ‘good’ webographic variables for post-survey adjustment, as previously suggested by Dutwin and Buskirk (2017). Our study also highlights the importance of selection bias and representativeness, and how this varies between different nonprobability samples (Lehdonvirta et al. 2021).

The investigation into the suitability of post-survey adjustment methods did not highlight a particular method or a combination of them which consistently performed better parameter estimates. This supports the finding from Mercer et al. (2018). While a detailed technical investigation into calibration methods was not the focus of this study, we found little differences in efficiency between the investigated methods: raking and the model-based methods (such as GREG or MRP), which was consistent with findings from Kalton and Flores-Cervantes (2003). Therefore, we suggest the selection of calibration methods to be instead based on the availability of joint distributions of covariates weighed against the computational intensity of methods. While matching methods and PSW under limited scenarios might have a better potential for efficient post-survey adjustment, we observed less consistency in bias reduction between different samples and scenarios. We also observed an increase of design effect for CEM and, consequently, confidence intervals for estimates (see Kolenikov 2014).

This study has several limitations, including the availability of external data and covariates both in nonprobability surveys and high-quality government surveys. Having access to additional data sources could improve post-survey adjustments and help distinguish better between the efficiency of covariates, the effect of quality of external data sources, and the efficiency of methods. Moreover, since estimates for only 18 items were compared to benchmarks and the majority of substantive items were more or less associated with one topic (i.e., health status), the findings would be more robust if survey items with corresponding benchmarks would be associated with other aspects of respondent’s lives, not only health. In addition, the total survey error framework (Biemer 2010; Groves et al. 2009; Groves & Lyberg 2010) has been proposed to provide a comprehensive overview of all possible sources of sampling and non-sampling errors and give a systematic measure of survey quality that encompasses not just accuracy but also bias. The framework attempts to account for, and assess, many sources of error that arise through the survey process (which we could not study separately, e.g., measurement mode effects versus representation bias). This framework lends itself to the Bayesian paradigm through incorporating prior information (Shirani-Mehr et al. 2018) or using expert opinion (Toepoel & Emerson 2017) in assessing the survey quality of surveys not based on probability schemes. We would suggest future research on improving inference in nonprobability samples to be more targeted, planned and properly designed in advance. Nonetheless, the approaches discussed in this chapter have distinct long-term benefits in improving the inferences from surveys conducted using nonprobability samples.

References

Australian Bureau of Statistics. (n.d.-a). *TableBuilder.* Retrieved November 1, 2020, from https://www.abs.gov.au/websitedbs/d3310114.nsf/home/about+tablebuilder

Australian Bureau of Statistics. (n.d.-b). *Compare data services.* Retrieved January 16, 2021, from https://www.abs.gov.au/websitedbs/D3310114.nsf/4a256353001af3ed4b2562bb00121564/c00ee824af1f033bca257208007c3bd5!OpenDocument

Australian Bureau of Statistics. (2016). *2016 Census of Population and Housing* [Census TableBuilder], accessed 1 November, 2020.

Australian Bureau of Statistics. (2018a, March 28). *Household use of information technology.* https://www.abs.gov.au/statistics/industry/technology-and-innovation/household-use-information-technology/latest-release

Australian Bureau of Statistics. (2018b, December 12). *National Health Survey: First results.* https://www.abs.gov.au/statistics/health/health-conditions-and-risks/national-health-survey-first-results/latest-release

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly, 74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology, 1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Barnier, J., Briatte, F., & Larmarange, J. (2020). *Questionr: Functions to make surveys processing easier* (R package version 0.7.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=questionr

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2009). Practical Considerations in Raking Survey Data. *Survey Practice, 2* (5). https://doi.org/10.29115/SP-2009-0019.

Bethlehem, J. G. (2002). *Weighting nonresponse adjustments based on auxiliary information.* Wiley.

Bethlehem, J. G. (2008). *Weighting*. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 957-960). Sage.

Bethlehem, J. G. (2016). Solving the nonresponse problem with sample matching?. *Social Science Computer Review, 34*, 59-77.

Bethlehem, J. G., & Biffignandi, S. (2012). *Handbook of Web Surveys*. Wiley & Sons.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817-848.

Bon, J. J., Ballard, T., & Baffour, B. (2019). Polling bias and undecided voter allocations: US presidential elections, 2004–2016. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 467-493.

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

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, *114*(3), 542.

Callegaro, M., & DiSogra, C. (2008a). Computing response metrics for online panels. *Public opinion quarterly*, *72*(5), 1008-1032.

Callegaro, M., & DiSogra, C. (2008b). Probability of Selection. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 617-618). Sage.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, 8(1), 4–36. https://doi.org/10.1093/jssam/smz041

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology,* 8(3), 513-539.

DataReportal. (2022, February 9). *Digital 2022: Australia*. https://datareportal.com/reports/digital-2022-australia

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, *88*(423), 1013-1020.

DiSogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, Survey Research Methods*, 4501-4515.

Dutwin, D., & Buskirk, T. D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, *81*(S1), 213-239.

Dutwin, D., & Buskirk, T. D. (2021). Telephone sample surveys: dearly beloved or nearly departed? Trends in survey errors in the era of declining response rates. *Journal of Survey Statistics and Methodology*, *9*(3), 353-380.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice, 2,* 1-7.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.

ESOMAR (2021, March). *Questions to help buyers of online samples.* https://esomar.org/uploads/attachments/ckqqecpst00gw9dtrl32xetli-questions-to-help-buyers-of-online-samples-2021.pdf

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153-164.

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

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.21.1, https://mc-stan.org/rstanarm.

Goot, M. (2021). How good are the polls? Australian election predictions, 1993–2019. *Australian Journal of Political Science*, 56(1), 35-55.

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology.* John Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Hade, E. N., & Lemeshow, S. (2011). In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 621-623). Sage.

Harrell, F. E. Jr., Dupont, C., & others (2020). *Hmisc: Harrell miscellaneous* (R package version 4.4-1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=Hmisc.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–62.

Hewitt, M. (2017). *National Drug Strategy Household Survey 2016.* (ADA Dataverse, Version 7) [Data set]. ADA. https://doi.org/10.4225/87/JUDY2Y

Hug, S. (2003). Selection Bias in Comparative Research: The Case of Incomplete Data Sets. *Political Analysis, 11*(3), 255-274. https://doi.org/10.1093/pan/mpg014

Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened Exact Matching Software. *Journal of Statistical Software, 30.* http://gking.harvard.edu/cem

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.

Iacus, S. M., King, G., & Porro, G. (2020). *Cem: Coarsened exact matching (R package version 1.1.20)* [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=cem

International Organisation for Standardisation (2022). *ISO 26362:2009.* https://www.iso.org/standard/43521.html

Kaczmirek, L., Phillips, B., Pennay, D. W., Lavrakas, P. J., & Neiger, D. (2019). Building a probability-based online panel: Life in Australia™. *CSRM and SRC Methods Paper, 2019* (2).

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

Kaplan, J. (2020). *fastDummies: Fast creation of dummy (binary) columns and rows from Categorical Variables* (R package version 1.6.1) [Computer software]. The

Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=fastDummies

Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 319-376.

Kennedy, C., & Hartig, H. (2019). *Response rates in telephone surveys have resumed their decline*. Pew Research Center.

Kennedy, C., M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, L. Miringoff, K. Olson, D. Rivers, L. Saad, G. E. Witt, & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1-33.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys.* Pew Research Center.

King, G., Lucas, C., & Nielsen, R. A. (2015). *{MatchingFrontier}: {R} Package for Computing the Matching Frontier ()* [Computer software]. The Comprehensive R Archive Network. Available from http://projects.iq.harvard.edu/frontier

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, *14*(1), 22-59.

Lavrakas, P. J., Pennay, D., Neiger, D., & Phillips, B. (2022). Comparing Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia: A Total Survey Error Perspective. *Survey Research Methods*, *16*(2), 241-266.

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social media, web, and panel surveys: using non-probability samples in social and policy research. *Policy & internet*, *13*(1), 134-155.

Little, R. J. A., & Rubin, D. B. (2002). Single imputation methods. In R. J. A. Little, & D. B. Rubin (Eds.), *Statistical analysis with missing data* (pp. 59-74). Wiley.

Lüdecke, D. (2020). *Sjstats: Statistical functions for regression models* (version 0.18.0) () [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=sjstats

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The accuracy of measurements with probability and nonprobability survey samples: replication and extension. *Public Opinion Quarterly*, *82*(4), 707-744.

Malhotra, N., & Krosnick, J. A. (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.

Matei, A. (2018). On Some Reweighting Schemes for Nonignorable Unit Nonresponse. *Survey Statistician, 77*, 21–33.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most.* Pew Research Center.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250-271.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, *12*(4), 375-385.

Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In J. E. Cohen (Ed.), *Public opinion in state politics* (pp. 209-228). Stanford University Press.

Pasek, J. (2018). *Anesrake: Anes raking implementation* (R package version 0.80) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=anesrake

Pennay, D. W., Borg, K., Neiger, D., Misson, S., Honey, N., & Lavrakas, P. (2016). *Online Panels Benchmarking Study, 2015* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.4225/87/FSOYQI

Pennay, D. W., & Neiger, D. (2020). *Health, Wellbeing and Technology Survey (OPBS replication), 2017* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.26193/YF8AF1

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). The Online Panels Benchmarking Study: a Total Survey Error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia. *CSRM and SRC Methods Paper, 2018* (2).

Pfeffermann, D., Eltinge, J. L., & Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. *Journal of Survey Statistics and Methodology, 3*(4), 425-483.R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, D. (2007, July 29–August 2). *Sampling for Web Surveys* [Conference presentation]. 2007 Joint Statistical Meetings, Salt Lake City, United States of America.

Rivers, D. (2013). Comment on task force report. *Journal of Survey Statistics and Methodology, 1*(2), 111-117.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Validity in epidemiologic studies. In K. J. Rothman, S. Greenland, T. L. Lash (Eds.), *Modern epidemiology* (3rd ed.) (pp. 128-147). Lippincott Williams & Wilkins.

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279-292.

Schonlau, M., Soest, V. A., & Kapteyn, A. (2007). Are "Webographic" or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?. *Survey Research Methods*, *1*, 155–163.

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607-614.

Sizemore, S., & Alkurdi, R. (2019). *Matching Methods for Causal Inference: A Machine Learning Update*. Available from https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/.

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest* (R package version 1.4) [Computer software]. The Comprehensive R Archive Network. Available from https://github.com/stekhoven/missForest

Stuart, E. A. (2010). The Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 174*(2), 369–386.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155-176). Sage.

Toepoel, V., & Emerson, H. (2017). Using experts' consensus (the Delphi method) to evaluate weighting techniques in web surveys not based on probability schemes. *Mathematical Population Studies*, *24*(3), 161-171, DOI: 10.1080/08898480.2017.1330012.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations.* Cambridge University Press.

Tourangeau, R., & Smith, W. (1985). Finding subgroups for surveys. *Public Opinion Quarterly, 49*(3), 351-365.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231-263.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105-137.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980-991.

Wang, Y., Dai, Y., Li, H., & Song, L. (2021). Social Media and Attitude Change: Information Booming Promote or Resist Persuasion?. *Frontiers in Psychology*, 12, 2433.

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