Jackson, Michael T, Arifah Hasanbasri, Cameron McPhee, and Jordon Peugh. 2022. "The Utility of a Random Forest Propensity Adjustment in Recurring Hybrid Probability-Nonprobability Samples: Evidence from a Tracking Poll." *Survey Practice* 15 (1). https://doi.org/10.29115/SP-2022-0004.

ARTICLES

The Utility of a Random Forest Propensity Adjustment in Recurring Hybrid Probability-Nonprobability Samples: Evidence from a Tracking Poll

Michael T Jackson[1] [a], Arifah Hasanbasri[1], Cameron McPhee[1], Jordon Peugh[1]

[1] SSRS

Keywords: survey sampling, survey weighting, nonprobability samples, hybrid samples

https://doi.org/10.29115/SP-2022-0004

Survey Practice

Vol. 15, Issue 1, 2022

The optimal approach to weighting samples that rely wholly or partially on nonprobability sources (such as opt-in Web panels) remains an active area of research. Aside from traditional raking, a wide array of advanced statistical techniques, including propensity adjustments using advanced predictive modeling methods, can potentially be applied to adjust for selection bias in nonprobability samples. However, prior research has shown that the choice of adjustment variables matters more than the choice of techniques—in particular, for a given set of adjustment variables, the addition of a propensity adjustment usually leads to minimal additional bias reduction, relative to traditional raking alone. In this paper, we expand on this prior research to consider the more complex scenario of a recurring (specifically, repeated cross-sectional) study. As a case study, we analyze a tracking poll that measures media consumption in a major metropolitan area using a "hybrid" sample that blends probability and nonprobability sources. We demonstrate that in studies that rely on recurring nonprobability samples, the set of characteristics that best explains selection into the sample can change dramatically over time, diminishing the effectiveness of raking adjustments at controlling selection bias. We further demonstrate that in such scenarios, propensity adjustments offer a flexible method of controlling for these changes and thereby reducing selection-driven disruptions in trends. Tradeoffs in terms of the impact on the precision of estimates are discussed. Overall, this research suggests that in recurring nonprobability studies, propensity adjustments can be a valuable addition to traditional raking by providing "insurance" against unexpected changes in the selection mechanism.

INTRODUCTION

Empirical research consistently finds that population estimates from nonprobability sample sources—such as opt-in Web panels—are less accurate than those from traditional probability-based sources, despite declining response rates to the latter (for a detailed review, see Cornesse et al. 2020). Hybrid sampling—surveying parallel probability-based and nonprobability samples and then blending the resulting completes—has been proposed to balance the lower cost of nonprobability sampling with the statistical rigor of probability sampling (DiSogra et al. 2011; Fahimi et al. 2015; Robbins, Ghosh-Dastidar, and Ramchand 2021; Wiśniowski et al. 2020).

[a] Corresponding author:
Michael T. Jackson
Director of Data Science Innovation
SSRS
mjackson@ssrs.com
1-484-840-4326

Hybrid approaches leverage the parallel probability sample to allow adjustment on characteristics that are related to the nonprobability selection mechanism (and therefore must be weighted on to control selection bias) but lack external population benchmarks. Two common weighting-based approaches to doing so, which can be used in tandem, are

- **Propensity adjustment** (Valliant, Dever, and Kreuter 2018): this entails estimating a statistical model predicting presence in the nonprobability versus the probability sample, in which the predictors are characteristics collected from both samples expected to be related to the nonprobability selection mechanism. This model is then used to assign an estimated selection probability to each nonprobability complete, which is transformed into a weight.
- **Calibration (raking) to internal benchmarks** (DiSogra et al. 2011; Fahimi et al. 2015): calibration creates weights such that the weighted distribution of specified characteristics among respondents matches known benchmarks. In hybrids, this entails first calibrating the probability sample to external benchmarks (typically demographic) obtained from high-quality external surveys such as the American Community Survey (ACS). The weighted probability sample is then used to produce "internal benchmarks" for non-demographic characteristics that lack external benchmarks. The joint sample is calibrated to these internal benchmarks plus the available external benchmarks. A widely used calibration algorithm is raking (Deming and Stephan 1940), which adjusts the weights to match the marginal distributions of multiple characteristics.

Prior research has found that the choice of adjustment variables matters more than the choice of weighting methods: in particular, for a given set of adjustment variables, the addition of a propensity adjustment does not yield greater bias reduction than raking alone (Mercer, Lau, and Kennedy 2018).

We extend this prior research to the special case of a recurring (repeated cross-sectional) study in which the estimation of trend is of interest. We address the following questions:

1. **Is there evidence that nonprobability selection mechanisms change over time in recurring studies**? Changes in the selection mechanism imply that selection bias may vary across waves. In this case, a raking model that is effective at removing selection bias in one wave may later become ineffective, potentially disrupting trend estimates.
2. If so, **can the addition of a propensity adjustment help correct for changes in the selection mechanism** and thereby recover an accurate trend?

Table 1. Number of completes and percent nonprobability, by survey year.

Survey year	Sample size			Percent nonprobability
	Nonprobability	Probability	Total	
2016	1,503	1,000	2,503	60.0%
2017	1,658	1,001	2,659	62.4%
2018	1,771	800	2,571	68.9%
2019	1,871	700	2,571	72.8%
2020	1,929	583	2,512	76.8%

METHODS

As a case study, we use data from an annual tracking poll of media consumption among adults ages 18+ in a large U.S. metropolitan area, which was sponsored by a major media outlet for market research purposes. Key outcomes were the percentage of adults reading the Sunday and daily editions of two major U.S. newspapers; to maintain client confidentiality, these are referred to here as "Newspaper A" and "Newspaper B," respectively.

From 2016 through 2020, SSRS administered the annual study using a repeated cross-sectional hybrid design, with an annual sample size of approximately 2,500 completes. Table 1 shows the evolution of the probability and nonprobability sample sizes from 2016 through 2020. From 2016 through 2019, the probability completes were collected by phone via dual-frame (landline and cell) random digit dialing (RDD). In 2020, a mixed-mode approach was employed, combining dual-frame RDD phone completes with Web completes from the SSRS Opinion Panel, a probability panel recruited via address-based sampling. The nonprobability completes were obtained from several Web panel vendors.[1] Nonprobability sample quotas were specified to match the population distribution of the target metropolitan area with respect to age, sex, and geographic subregion.

The original weighting approach (the "raking-only weight") relied solely on raking. The probability sample completes were assigned base weights reflecting their original selection probabilities. The probability sample was then raked to eight ACS demographic benchmarks: sex, age, Hispanic ethnicity, race, educational attainment, phone use, region, and household income. Next, the weighted probability sample was used to obtain internal benchmarks for two nondemographic characteristics: whether the respondent accessed the Internet in the past 30 days, and when the respondent last visited the website of Newspaper A using a laptop. Finally, the hybrid sample (blending the probability and nonprobability completes) was raked to the eight ACS demographic benchmarks plus the two internal non-demographic benchmarks.

1 The same three vendors were used in 2019 and 2020. The authors do not have information on the identity of the vendors used prior to 2019.

For the 2019 and 2020 samples, we calculated an alternative weight (the "propensity + raking weight"), which incorporated a propensity adjustment prior to raking. The "nonprobability propensity" score was generated as a case's predicted probability of having been sourced from a nonprobability vendor, using a random forest (Breiman 2001) in which the dependent variable was presence in the nonprobability sample (1 = nonprobability complete, 0 = probability complete).

To control weighting variability, we used a propensity stratification approach (Valliant, Dever, and Kreuter 2018) to transform this probability into a "pseudo-base weight" for nonprobability completes. We divided the combined sample into deciles based on the propensity score and calculated the following adjustment factor for each decile d:

$$NPA_d = \frac{1 - \left(\frac{N_{n,d}}{N_{n,d} + N_{p,d}}\right)}{\left(\frac{N_{n,d}}{N_{n,d} + N_{p,d}}\right)}$$

where $N_{n,d}$ is the unweighted count of nonprobability completes in decile d, and $N_{p,d}$ is the sum of the base weights of probability completes in decile d. NPA_d was assigned as the pseudo-base weight to all nonprobability completes in decile d. This adjustment is designed to make the pseudo-base-weighted nonprobability sample resemble the base-weighted probability sample, with respect to the random forest predictors, prior to raking. Raking then proceeded as with the original weight.

The random forest predictors included the demographics and non-demographic "background" characteristics used in raking, as well as all other variables that were asked of at least half the sample and were consistent across waves—including the key study outcomes (newspaper readership). By including outcomes in the propensity model, we can account for the possibility that, even after controlling for other characteristics, selection into the nonprobability sample may be directly related to outcomes. A propensity model that includes outcomes (in lieu of direct calibration on outcomes) provides a means of adjusting for such relationships without relying entirely on the probability sample to produce estimates of outcomes, which would render the nonprobability completes superfluous.

RESULTS

Do nonprobability selection mechanisms change over time?

Figure 1 shows the estimated trend in the four outcomes from 2016 through 2020 using the raking-only weights. The trend using only the probability sample is compared to the trend using the hybrid sample. In 2020, the hybrid

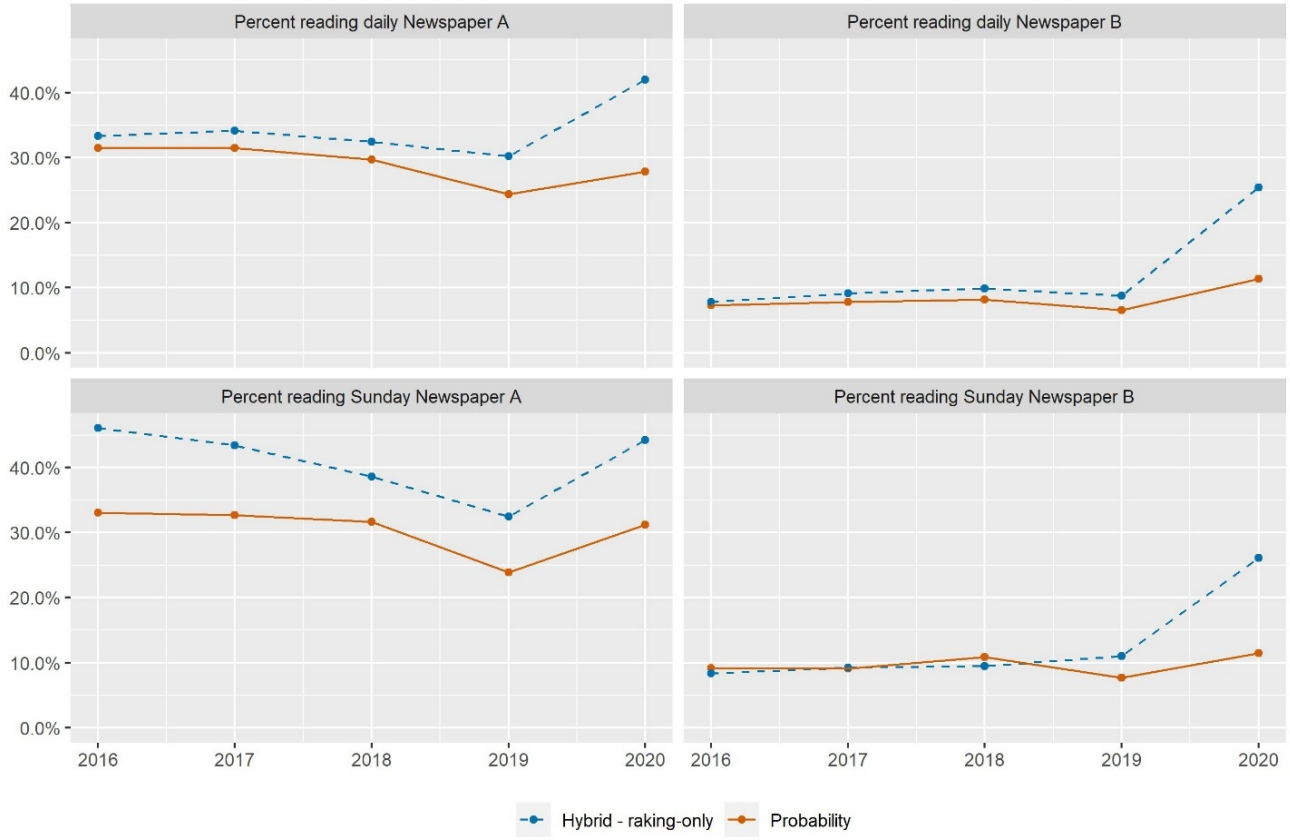

Figure 1. Estimated trends in key outcomes, hybrid sample vs. probability sample.

NOTE: The probability sample is raked to external demographic benchmarks obtained from the American Community Survey (ACS). The hybrid sample is raked to ACS benchmarks plus internal benchmarks obtained from the weighted probability sample.

sample estimates sharp increases[2] in the outcomes, while the probability sample generally estimates more modest increases. If the probability sample is assumed to be approximately unbiased, this suggests increased selection bias in the hybrid estimates. This, in turn, suggests a change in the nonprobability sampling mechanism that was not corrected by the raking alone.

To disentangle the drivers of this pattern, Table 2 shows coefficients from logistic regressions predicting presence in the nonprobability sample (relative to the probability sample) in the 2019 and 2020 studies. In 2019, controlling for demographics, one of the four outcomes was a marginally statistically significant predictor of having been sampled from a nonprobability source. In 2020, two of the four outcomes became strongly significant predictors of presence in the nonprobability sample, though the sampling specifications provided to the nonprobability vendors did not change. This provides additional evidence that the nonprobability selection mechanism changed over the life of this study—specifically, that it became more independently associated with substantive outcomes.

2 A separate analysis, not shown here, found that similar increases were observed within the nonprobability sample alone, ruling out the possibility that the change in the hybrid estimate was driven by the increase in the nonprobability sample share shown in Table 1.

Table 2. Logistic regression coefficients predicting presence in nonprobability sample.

	2019			2020		
Predictor	Coefficient	Standard error		Coefficient	Standard error	
Intercept	2.68	0.44	***	1.89	0.55	***
Sex: Female	0.69	0.10	***	0.52	0.13	***
Sex: Missing	13.80	0.48	***	14.08	0.50	***
Age: 25 - 34	-0.36	0.21	*	0.07	0.25	
Age: 35 - 44	-0.33	0.22		-0.09	0.25	
Age: 45 - 54	-0.94	0.22	***	-0.86	0.25	***
Age: 55 - 64	-1.01	0.21	***	-0.56	0.26	**
Age: 65+	-0.79	0.21	***	-0.35	0.25	
Race: Black	-0.58	0.15	***	0.08	0.16	
Race: Asian	0.17	0.22		0.12	0.22	
Race: Other	-0.65	0.20	***	-0.05	0.22	
Race: Missing	-0.21	0.28		-1.03	0.69	
Income: $50K – Less than $100K	-0.48	0.16	***	-0.58	0.18	***
Income: $100K – Less than $150K	-0.70	0.19	***	-0.71	0.21	***
Income: $150K+	-0.90	0.18	***	-0.93	0.20	***
Income: Missing	-1.32	0.18	***	-4.72	0.63	***
Ethnicity: Not Hispanic	0.10	0.19		0.09	0.22	
Ethnicity: Missing	-0.63	0.46		4.94	1.72	***
Education: High school graduate	-0.46	0.36		-0.77	0.47	
Education: Some college	-0.22	0.35		-0.70	0.46	
Education: Bachelor's degree or higher	-0.41	0.35		-1.21	0.46	***
Education: Missing	0.28	0.62		14.02	1.30	***
Did not access Internet in past 30 days	-0.02	0.92		1.02	0.69	
Last visited Newspaper A website on laptop: 2 days+	-0.13	0.12		-0.22	0.15	
Last visited Newspaper A website on laptop: Never	-0.76	0.15	***	-0.12	0.20	
Last visited Newspaper A website on laptop: Don't know/Refused	-1.68	0.90	*	-2.31	0.62	***
Reads Sunday Newspaper A	0.21	0.11	*	0.25	0.18	
Reads Sunday Newspaper B	-0.08	0.16		0.58	0.21	***
Reads daily Newspaper A	0.14	0.11		0.26	0.18	
Reads daily Newspaper B	0.23	0.18		0.52	0.20	**

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

NOTE: The dependent variable was coded as 0 for probability completes and 1 for nonprobability completes. Probability completes were base-weighted by their original probability of selection; nonprobability completes were unweighted. Reference category is Male for sex; 18 – 24 for age; White for race; Under $50K for income; Hispanic for ethnicity; Less than high school for education; and Today/yesterday for last visited Newspaper A website on laptop. All other predictors are dichotomous.

Does the addition of a propensity adjustment help correct the trend?

Figure 2 shows the same estimates as Figure 1; but for 2019 and 2020, adds the hybrid estimates using propensity + raking weights. The propensity adjustment brings the 2020 hybrid estimates much closer to the probability-only estimates. Thus, the propensity adjustment partially corrects for the change in the nonprobability selection mechanism, reducing the distortion in the estimated trends.

Figure 2. Estimated trends in key outcomes, hybrid sample (with propensity adjustment) vs. probability sample.

NOTE: The probability sample is raked to external demographic benchmarks obtained from the American Community Survey (ACS). The hybrid sample is raked to ACS benchmarks plus internal benchmarks obtained from the weighted probability sample.

Table 3. Measures of precision.

	2019			2020		
Design	Sample size	Unequal weighting effect	Effective sample size	Sample size	Unequal weighting effect	Effective sample size
Probability	700	1.67	419	583	1.95	299
Hybrid - raking-only	2,571	1.51	1,703	2,512	1.64	1,528
Hybrid - raking + propensity	2,571	1.85	1,391	2,512	2.21	1,134

NOTE: The unequal weighting effect is calculated using the Kish (1965) formula. Due to rounding, the effective sample size may not exactly equal the sample size divided by the unequal weighting effect.

Is there a tradeoff to the addition of a propensity adjustment?

Table 3 reports the Kish (1965) unequal weighting effect (UWE) and effective sample size (ESS) for the hybrid sample with both sets of weights, compared to the probability sample alone, for 2019 and 2020. Particularly in 2020, the addition of the propensity adjustment sharply increases the UWE and therefore decreases the hybrid ESS.

Table 3 captures the tradeoff to the addition of a propensity adjustment. Effectively, the propensity adjustment transforms the increase in bias into a reduction in precision, via an increased UWE. The more accurate trend

therefore comes at the expense of a smaller ESS. However, the hybrid ESS remains higher than the probability-only ESS, suggesting that the nonprobability completes continue to add value despite the additional weighting required to correct for the changing selection mechanism.

DISCUSSION

Though prior research has found that propensity adjustments (relative to raking alone) add little value in weighting nonprobability or hybrid samples, these results add an important caveat.

Specifically, in the case of recurring studies, it is important to consider that patterns of selection bias in the nonprobability sample may change over time. As demonstrated here, propensity adjustments using random forests can help correct for such changes and thus obtain more accurate trends than raking alone. Therefore, in recurring studies that use nonprobability or hybrid samples, the incorporation of a pre-raking propensity adjustment provides "insurance" against selection-driven distortions of estimated trends.

In this case study, the utility of the propensity adjustment was driven by the fact that the random forest included a wider range of predictors than the raking. The increase in bias in the raking-only estimates in 2020 (Figure 1) reflects the fact that selection became more strongly related to outcomes that were not included in the raking model (Table 2). The propensity adjustment provided an alternative vehicle by which such characteristics could be accounted for in the hybrid weighting and thus helped mitigate this increase in selection bias (at the cost of a higher UWE).

Two alternatives to a propensity adjustment would be to (1) reevaluate the raking model at each wave of the study to ensure that any newly relevant characteristics are included or (2) include all potentially relevant characteristics (including study outcomes) in the raking model from the first wave.

Relative to the first alternative, the operational advantage of a propensity adjustment is that it operates largely automatically—particularly if, as in this case study, it is operationalized using nonparametric machine-learning techniques such as random forests. The number of potential raking margins in a typical study is considerable, particularly if interactions between characteristics are allowed. Therefore, choosing a new raking model may require extensive time and effort to remodel the selection mechanism at each wave. In contrast, given a large set of potentially relevant predictors, random forests can automatically identify those characteristics (and interactions between characteristics) that are most relevant in differentiating between probability and nonprobability completes. These potentially complex relationships are then built into the resulting propensity scores. This flexibility implies that a random forest propensity adjustment, once built into the

weighting workflow, can pick up changes in the nonprobability selection mechanism with little intervention by the user, providing an efficient form of insurance against such changes.

Similar considerations favor the propensity approach over the second alternative. Since changes in the selection mechanism are unpredictable, the incorporation of all potentially relevant raking margins from the start is likely to be impractical. A raking model that is extensive enough to control for all characteristics that *might* become relevant may lead to an excessively high UWE and/or convergence problems, particularly when many margins are correlated with each other (Brick, Montaquila, and Roth 2003). In contrast, nonparametric methods such as random forests are designed to deal with a large set of candidate predictors whose relative importance is unknown in advance. Therefore, a pre-raking propensity adjustment allows the weighting procedure to account for a larger number of potentially relevant characteristics, while maintaining a parsimonious raking model.

In particular, a propensity adjustment provides a means of adjusting on substantive outcomes (if necessary) without raking on them. As demonstrated in Table 2, substantive outcomes may well be relevant to the nonprobability selection mechanism even after controlling for other characteristics, in which case adjustment on outcomes is needed to avoid selection bias. In most cases, however, raking on outcomes is likely to be undesirable. Raking on an outcome implies that the estimate from the hybrid sample would be forced to match the estimate obtained from the probability completes alone. This implies that sampling variability in the estimate would be driven by the size of the probability sample, not the (larger) hybrid sample. This, in turn, means that measures of sampling variability calculated using the hybrid sample size may overstate the precision of the estimates. At the extreme, raking on all outcomes of substantive interest would render the nonprobability completes superfluous (Robbins, Ghosh-Dastidar, and Ramchand 2021). In contrast, including outcomes in a propensity model allows the weighting to account for any independent influence that they may have on the nonprobability selection mechanism, without directly controlling them to the probability-based estimates and thus continuing to allow the nonprobability completes to contribute to the final estimates. The ability to control for a potential "missing not at random" selection mechanism (Rubin 1976) provides a general argument for a "doubly robust" approach combining propensity adjustment and calibration (Valliant, Dever, and Kreuter 2018), even in nonrecurring studies.

It is important to emphasize that this case study used a hybrid sample in which, at each wave, the full survey instrument was administered to side-by-side probability and nonprobability samples. This allowed the propensity model to include study outcomes as well as non-demographic covariates that were collected on the questionnaire. A similar propensity adjustment could be

applied to a nonprobability-only study by using an external public-use dataset (such as the ACS) as the "reference" probability sample (Elliott and Valliant 2017). In this case, however, the predictors would be limited to variables that are available in both the study and the external dataset. These may not be well-tailored to the study's outcomes and in practice are likely to be limited to demographics—which, as demonstrated in Table 2, are often not sufficient to account for nonprobability selection mechanisms. Therefore, the utility of a propensity adjustment could be more limited in a study that relied solely on a nonprobability sample.

Submitted: February 09, 2022 EDT, Accepted: May 24, 2022 EDT

REFERENCES

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/a:1010933404324.

Brick, Michael, Jill Montaquila, and Shelley Roth. 2003. "Identifying Problems with Raking Estimators." In *Proceedings of the American Statistical Association Section on Survey Research Methods*. http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000472.pdf.

Cornesse, Carina, Annelies G. Blom, David Dutwin, Jon A. Krosnick, Edith D. de Leeuw, Stéphane Legleye, Josh Pasek, et al. 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.

Deming, W. Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *The Annals of Mathematical Statistics* 11 (4): 427–44. https://doi.org/10.1214/aoms/1177731829.

DiSogra, Charles, Curtiss Cobb, Elisa Chan, and J.Michael Dennis. 2011. "Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics." In *Proceedings of the American Statistical Association Section on Survey Research Methods*. http://www.asasrms.org/Proceedings/y2011/Files/302704_68925.pdf.

Elliott, Michael R., and Richard Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2): 249–64. https://doi.org/10.1214/16-sts598.

Fahimi, Mansour, Frances M. Barlas, Randall K. Thomas, and Nicole Buttermore. 2015. "Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse." *Survey Practice* 8 (6). https://doi.org/10.29115/sp-2015-0031.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. "For Weighting Online Opt-In Samples, What Matters Most?" Pew Research Center. January 26, 2018. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most.

Robbins, Michael W., Bonnie Ghosh-Dastidar, and Rajeev Ramchand. 2021. "Blending Probability and Nonprobability Samples with Applications to a Survey of Military Caregivers." *Journal of Survey Statistics and Methodology* 9 (5): 1114–45. https://doi.org/10.1093/jssam/smaa037.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. https://doi.org/10.1093/biomet/63.3.581.

Valliant, Richard, Jill A. Dever, and Frauke Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer. https://doi.org/10.1007/978-3-319-93632-1.

Wiśniowski, Arkadiusz, Joseph W Sakshaug, Diego Andres Perez Ruiz, and Annelies G Blom. 2020. "Integrating Probability and Nonprobability Samples for Survey Inference." *Journal of Survey Statistics and Methodology* 8 (1): 120–47. https://doi.org/10.1093/jssam/smz051.