sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 653–681, http://dx.doi.org/10.2478/JOS-2019-0027

Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach

Joseph W. Sakshaug[1], *Arkadiusz Wiśniowski*[2], *Diego Andres Perez Ruiz*[2], *and Annelies G. Blom*[3]

Carefully designed probability-based sample surveys can be prohibitively expensive to conduct. As such, many survey organizations have shifted away from using expensive probability samples in favor of less expensive, but possibly less accurate, nonprobability web samples. However, their lower costs and abundant availability make them a potentially useful supplement to traditional probability-based samples. We examine this notion by proposing a method of supplementing small probability samples with nonprobability samples using Bayesian inference. We consider two semi-conjugate informative prior distributions for linear regression coefficients based on nonprobability samples, one accounting for the distance between maximum likelihood coefficients derived from parallel probability and nonprobability samples, and the second depending on the variability and size of the nonprobability sample. The method is evaluated in comparison with a reference prior through simulations and a real-data application involving multiple probability and nonprobability surveys fielded simultaneously using the same questionnaire. We show that the method reduces the variance and mean-squared error (MSE) of coefficient estimates and model-based predictions relative to probability-only samples. Using actual and assumed cost data we also show that the method can yield substantial cost savings (up to 55%) for a fixed MSE.

Key words: Bayesian inference; quota sampling; German Internet Panel; GESIS Panel; web surveys.

1. Introduction

1.1. Background

Scientific surveys based on random, probability-based samples are ubiquitously used in the social sciences to study and describe large populations. They provide a critical source of quantifiable information used by governments and policy-makers to make informed decisions. However, probability-based surveys are increasingly expensive to carry out due to declining response rates and costly intervention strategies (Tourangeau and Plewes 2013). Consequently, many survey organizations have shifted away from probability-based samples in favor of cheaper nonprobability samples usually drawn from volunteer web panels. This shift in practice has prompted significant controversy and skepticism

[1] University of Mannheim and Institute for Employment Research, Nuremberg, 90478 Germany. Email: joe.sakshaug@iab.de
[2] University of Manchester, Manchester, M13 9PL United Kingdom. Emails: a.wisniowski@manchester.ac.uk, and diego.perezruiz@manchester.ac.uk
[3] School of Social Sciences, University of Mannheim, Mannheim, 68131 Germany. Email: blom@uni-mannheim.de

over the representativeness and overall utility of nonprobability samples (Baker et al. 2013). While probability-based surveys have their own concerns regarding representativeness (Gelman et al. 2016; Wang et al. 2015), comparison studies generally show (with same exceptions: see, for example Kennedy et al. 2016) that they produce more accurate population estimates than nonprobability surveys when evaluated against benchmark data (Yeager et al. 2011; Blom et al. 2017; Malhotra and Krosnick 2007; Chang and Krosnick 2009; Dutwin and Buskirk 2017; Pennay et al. 2018; Erens et al. 2014; Callegaro et al. 2014; MacInnis et al. 2018). Hence, the field of survey research is in a situation where probability-based samples are preferred from an error perspective, while nonprobability samples are preferred from a cost perspective.

Given the advantages of both sampling schemes, it makes sense to devise a strategy to combine them in a way that is beneficial from both a cost and error perspective. In some ways, survey organizations already attempt to make use of both sample types, either by drawing a nonprobability sample whose units closely match units from a reference probability sample prior to data collection (Rivers 2007; Rivers and Bailey 2009; Ansolabehere and Rivers 2013), or by devising post-survey weights that adjust the composition of a nonprobability sample survey towards that of a reference probability survey (Lee 2006; Lee and Valliant 2009; Valliant and Dever 2011). While both approaches are cost-effective and have been shown to increase the accuracy of estimates derived from nonprobability surveys, they have some important limitations. Firstly, they assume that the matching/adjustment variables fully explain the underlying selection mechanism that leads to inclusion in the nonprobability sample – a questionable and usually untestable assumption in practice (Mercer et al. 2017). Secondly, the target variable of interest is usually not present in the reference probability survey data, and therefore, these data are usually discarded after the matching/adjustment procedure. The intended analysis is then based solely on the nonprobability survey data, which lacks important properties of randomization theory, including the ability to measure the uncertainty of sample-based estimates.

Instead of forgoing probability-based survey data collection entirely, an alternative approach is to field the same questionnaire in a parallel probability and nonprobability sample and analyze the collected data jointly. For example, Elliott and Haviland (2007) describe a methodology that supplements a traditional probability sample with a web-based convenience sample. They evaluate a composite estimator influenced by Rao (2003) that is a linear combination of a probability and convenience sample, with each sample weighted according to a bias function. The estimator, under certain conditions, yields a smaller mean-squared error (MSE) compared to the probability-only sample. In related work, Elliott (2013) proposes a method of devising pseudo-weights for a nonprobability sample based on probabilities of selection estimated using a parallel probability sample. Both samples can then be combined and analyzed with case weights as if the units were drawn from the same population frame. The method is shown to reduce bias and MSE relative to a probability-only sample.

DiSogra et al. (2012) introduce an idea referred to as "blended calibration" in which available probability sample cases are supplemented with parallel nonprobability opt-in panel cases. The two-step procedure relies on, firstly, weighting the probability sample to known population benchmarks using a raking or poststratification procedure. In the second

step, the weighted probability and unweighted opt-in cases are combined and the combined sample is calibrated to the probability-only sample on a selection of survey variables common to both samples. The method yields smaller bias and MSE compared to more traditional approaches of analyzing probability and nonprobability samples separately and jointly. Fahimi et al. (2014) extend the approach by considering a more effective range of differentiator variables to use in the calibration step.

A practical limitation of the above studies is that they require relatively large probability sample sizes. Elliott and Haviland (2007) recommend a probability sample size of at least 1,000–10,000 cases alongside a convenience sample size in the thousands, and Elliott (2013) uses a probability sample size of 50,000 in the simulation study. Blended calibration also requires a relatively large probability sample size in order to minimize the variability in the probability-based survey benchmarks.

Any data integration strategy that requires fielding a large probability sample is likely to be met with opposition, as such sample sizes are prohibitively expensive for most survey budgets. An alternative, and more budget-friendly, strategy is to draw and field a small probability sample and combine it with a parallel nonprobability sample. On the face of it, the usefulness of deliberately fielding a small probability sample is not intuitively clear. Estimates derived from small probability samples, while inferentially valid, are subject to large variability and are insufficient as a standalone source of population information. Furthermore, a small probability sample is too sparse to be used as a reference sample for sample matching and post-survey adjustment procedures. A natural question, therefore, is whether there exists any scenario in which combining a small probability sample with a nonprobability sample could be beneficial from both a cost and error perspective.

1.2. Bayesian Inference

We address this question from a Bayesian inferential viewpoint. Bayesian inference offers an attractive system of estimation that allows combining sparse scientific data, such as those from probability-based samples, with less scientific and less reliable but potentially abundant and cheap information, such as those derived from nonprobability sources (Gelman et al. 2013). There are several advantages of using Bayesian inference in the context of combining small probability samples with nonprobability samples. First, the Bayesian framework allows for estimating complex models and quantifying measures of uncertainty, which can be problematic when analyzing nonprobability data under traditional estimation frameworks. Second, unlike sample matching and post-survey adjustment procedures, the Bayesian framework allows for the analysis of probability-based sample units through the likelihood function and is principally structured to give priority to these units in the posterior estimations as the probability sample size increases. Put differently, as additional probability sample units are observed, the "prior" information brought in through the nonprobability data becomes less relevant in the estimations, and increasing weight is given to the probability units. And third, because the probability-based likelihood *borrows information* from the informative nonprobability-based prior, the resulting posterior estimates are expected to be more efficient, that is, have less uncertainty, compared to estimates derived from small probability-only samples. This result could yield potential cost savings if large reductions in uncertainty are achieved and

the marginal cost of interviewing a nonprobability sample unit is lower than that of a probability sample unit – a plausible scenario in practice.

However, a disadvantage of applying the Bayesian framework in the aforementioned context is the deliberate incorporation of (potentially) biased data into the estimation process. In contrast to sample matching and post-survey adjustment, which takes an error-prone nonprobability sample and skews it towards a presumably less error-prone probability reference sample, the Bayesian approach that we describe does the opposite. That is, the method takes a probability sample and deliberately skews it towards a nonprobability sample reflected in the prior. The posterior estimates are therefore likely to have more bias than corresponding probability-only estimates. This effect is likely to be most pronounced for small probability samples where the prior will have peak influence on the posterior estimates. On the other hand, the expected reduction in variability due to the supplementary use of nonprobability data may offset any increase in bias, resulting in an estimator that yields a smaller mean-squared error.

1.3. Research Aims

In this article, we investigate whether supplementing a probability sample with nonprobability sample priors can produce more efficient survey estimates under varying probability sample sizes. We consider three specifications of the prior distribution for a target analysis of regression coefficients and model-based predictions: (i) a reference prior that allows for the probability sample to dominate the posterior, (ii) an informative prior that decreases the weight of the nonprobability sample with increasing distance between the maximum likelihood coefficient estimates derived from the probability and nonprobability samples, thus, "protecting" against bias in the latter, and (iii) an informative prior whose weight depends on the variability and size of the nonprobability sample and is able to dominate the posterior. Further, we examine the extent to which varying levels of bias in the nonprobability sample affect the mean-squared error (MSE) of the posterior estimates. To achieve these aims, we carry out a simulation study and real-data application involving two nationally-representative, probability-based surveys and eight nonprobability web surveys fielded in parallel using the same questionnaire. Through the application, we also assess whether the method is likely to yield cost savings for a fixed MSE.

The balance of this article is organized into five sections. Section 2 describes the proposed methodology for combining probability and nonprobability samples under a Bayesian framework. Section 3 presents the simulation study examining the bias-variance tradeoff of the method for various bias and sample size parameters. Sections 4 and 5 describe the real-data application and evaluation. Lastly, Section 6 provides a general discussion of the results, their implications for survey practice, and possible research extensions.

2. Methodology

2.1. Modeling Approach

As introduced in Subsection 1.2, in Bayesian inference (for details, see Gelman et al. 2013), the likelihood distribution is multiplied by a prior distribution, and inferences are

typically summarized by random draws from this product, that is, the posterior distribution. On the one hand, Bayesian inference can utilize prior distributions that "allow data to speak for themselves," that is, to have a negligible influence on the posterior draws. These priors are known as noninformative or weakly informative. On the other hand, informative priors can be used to add information about model parameters. This may be desirable in situations where parameters cannot be identified, or due to a small number of available observations. In this section we present three models, of which *two* use informative prior distributions constructed from a *single* nonprobability sample.

Consider an $n \times 1$ response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ of observations collected from a probability-based survey. The parameter of interest is the expectation of $\mathbf{y}$, denoted by μ. We assume that y is continuous and normally distributed:

$$\mathbf{y} \sim N(\mu, \sigma^2),$$

where σ^2 is the unknown variance of $\mathbf{y}$. The simple model can be expanded to account for covariates if the researcher's substantive interest lies in interpreting their effect on the outcome variable, or in making model-based predictions of the outcome. We focus on these two scenarios. The covariates can be incorporated by using a linear regression with an $n \times p$ design matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, which leads to

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a column vector of length p and I is the $n \times n$ identity matrix. We note that this model does not explicitly reflect the survey design. In our forthcoming application, we include survey weights as a covariate in the proposed modeling approach. Adapting the proposed approach to include additional survey design features (e.g., stratification, clustering) is a topic we leave for future work.

A semi-conjugate prior distribution for a single regression coefficient, β_j, for $j = 1, \ldots, p$ is

$$\beta_j \sim N\left(\beta_{j0}, \sigma^2_{\beta_j 0}\right), \tag{1}$$

with fixed location and variance hyperparameters, β_{j0} and $\sigma^2_{\beta_j 0}$, respectively. The semi-conjugacy (or conditional conjugacy) results from the fact that the variance in (1) does not depend on σ^2 (Gelman et al. 2013, 130). We consider three specifications of these hyperparameters.

In **Model 1** we assume a weakly informative parameterization of the priors, that is;

$$\beta_{j0} = 0, \quad \sigma^2_{\beta_j 0} = 10^6.$$

This specification allows the model parameters to be estimated directly from the probability data. Therefore, we treat this model as a reference to compare the two other specifications in which we introduce information from the nonprobability samples.

In **Model 2** we introduce an informative prior by utilizing information from a single nonprobability sample. First, we define $\hat{\beta}_j^P$ and $\hat{\beta}_j^{NP}$ to be the maximum likelihood (ML) estimators of the regression coefficients using the probability (*P*) and nonprobability (*NP*) survey data, respectively. These ML estimates are equivalent to the means of the posterior distributions of these parameters under the linear regression model using

noninformative Jeffrey's priors. We implicitly assume a simple random sampling design for the nonprobability data. Next, we set the hyperparameter β_{j0} equal to the estimated regression coefficient derived from the nonprobability survey, $\hat{\beta}_j^{NP}$. For the variance hyperparameter $\sigma_{\beta_j 0}^2$ we consider the Euclidean distance between the ML regression coefficients estimated in the probability survey and in the nonprobability survey. Specifically, we consider the square of the distance as the variance hyperparameter in (1):

$$\sigma_{\beta_j 0}^2 = d^2\left(\hat{\beta}_j^P, \hat{\beta}_j^{NP}\right) = \left(\hat{\beta}_j^P - \hat{\beta}_j^{NP}\right)^2, \quad \forall j.$$

Therefore, the prior for the regression coefficient in Model 2 can be written as:

$$\beta_j \sim N\left(\hat{\beta}_j^{NP}, d^2\left(\hat{\beta}_j^P, \hat{\beta}_j^{NP}\right)\right) \tag{2}$$

This method of setting the hyperparameter for the regression coefficient implies that the standard deviation, $\sigma_{\beta_j 0}$, is equal to the difference between the probability- and nonprobability-based ML estimates and does not depend on the size of the nonprobability sample. This, on the one hand, ensures some variability around the mean while keeping the uncertainty relatively small. If the distance d is large, the prior is wider and allows the small probability sample to influence the posterior. The smaller the distance between the two ML estimates, the tighter the prior distribution and, thus, larger potential gains in reducing posterior variance. A potential limitation of this approach is that if the distance is zero, that is, the corresponding probability and nonprobability estimates are equal, then the hyperparameter will be set to zero and shrink the location parameter β_j to a fixed value being $\hat{\beta}_j^{NP}$. However, in practice, such an event has virtually zero probability.When the distance is extremely small, it may severely reduce the variance of the posterior distribution for the parameter, especially when the probability sample size is very small. The next model we consider is free from this shortcoming.

By using the probability-based estimator to construct the prior distribution, the question of using data twice arises. We address this issue by pointing out that the ML estimator from the probability sample (a measure of central tendency) is used to inform the variance, rather than the mean. Further, we use the information from the probability data only in relative comparison to the nonprobability sample. Hence, any potential shrinkage in posterior variance depends on the combination of both data sets, rather than the probability data alone.

In **Model 3** we use a bootstrap procedure instead of the squared distance to derive information about the variance hyperparameter in (1). The bootstrap method has been used in many contexts and was originally proposed by Efron (1979). The general approach is to draw random subsamples with replacement from the full sample a large number of times and estimate the statistic of interest in each subsample before combining them using a bootstrap estimator. We implement the procedure by drawing 1,000 bootstrap samples from the nonprobability survey data, estimating the regression coefficient in each of them, and then calculating the variance $\left(\hat{\sigma}_{\beta_j}^{BNP}\right)^2$ across all regression coefficients. We then set the variance hyperparameter in (1) to the estimated variance and the prior distribution for

the regression coefficient in **Model 3**:

$$\beta_j \sim N\left(\hat{\beta}_j^{NP}, \left(\hat{\sigma}_{\beta_j}^{BNP}\right)^2\right), \tag{3}$$

with mean being the ML coefficient calculated using the nonprobability sample (the mean of all bootstrapped coefficients will converge to it). This approach is an alternative to calculating the uncertainty around the nonprobability-based regression coefficient and ensures it is always positive. The method is limited in a sense that the hyperparameter relies on the bootstrapped nonprobability sample which may propagate its unrepresentativeness and homogeneity, especially when very large nonprobability sample sizes are used, again leading to a false sense of certainty about the regression coefficient. However, analogous to the distance approach, this effect is reduced the larger the size of the probability sample.

For the variance of the regression model, σ^2, we first transform it to a precision, that is, inverse variance (σ^{-2}), and set $\sigma^{-2} \sim \Gamma(r, m)$ where $\Gamma(\cdot, \cdot)$ denotes a Gamma distribution with hyperparameters r being a shape and m being a rate. In our application, we set these hyperparameters to be $r = m = 10^{-3}$. This specification for the precision parameter is approximately noninformative and gives preference to the data (Gelman et al. 2013, 128). It remains the same for Models 1 through 3, which ensures comparability of the results.

3. Simulated Data Inference

In this section, we demonstrate how the proposed methods work under various assumptions regarding bias and sample size introduced through simulated data. First, we investigate the effect of bias on the regression coefficients of the model (part A of the simulation), and second, we analyze to what extent the bias affects model-based predictions of the outcome variable (part B).

The analysis was implemented in OpenBUGS (Spiegelhalter et al. 2007) and R (R Core Team 2016) using the library *r2openbugs* (Sturtz et al. 2005). We also use *MCMCpack* to summarize the results of the simulations, *boot* package for bootstrapping, as well as *ggmcmc* and *lattice* packages for visualization. In the simulations, the posterior distributions were obtained using three MCMC chains with samples of 2,000 each and 500 burn-in samples which ensured convergence of all chains.

To generate the data, we first assume the true values of the parameters in a linear regression model with intercept $\beta_1 = 5$, two parameters $\beta_2 = 0.5$ and $\beta_3 = 1$, and standard deviation of the outcome being $\sigma_y = 5$. Predictors x_1 and x_2 have means 0 and 5, respectively, standard deviations 4 and 0.5, and are correlated with correlation $\rho = 0.1$. These assumptions yield the mean response being $\bar{y} = 10$.

To introduce bias, we multiply the true parameter $\beta_3 = 1$ by 0.5, 1 (i.e., unbiased sample), 1.5, 2, 2.5, and 3 when generating the nonprobability samples (part A of simulation). For testing the effect of bias in nonprobability samples on the predicted outcomes (part B), we generate a predictive posterior distribution for a fixed probability test sample of size 500 using coefficients generated in part A. Bias introduced in this way is quite significant. For instance, when coefficient β_3 is doubled, the expected outcome

increases to 15. These scenarios are relatively extreme to real-life applications, but aim to demonstrate the limits of the proposed methods.

In the simulation we assume three nonprobability sample sizes $NPS \in \{1000, 10000, 50000\}$ and probability sample sizes $PS \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000\}$. In each simulation, for each PS we generate 100 sets of data with each combination of bias level and NPS. In total, it yields 27,000 data sets.

3.1. Model Evaluation

First, we evaluate the performance of the three modelling approaches by calculating the bias, variance, and mean-squared error (MSE; the sum of variance and squared bias) of the posterior means of the coefficients estimated using Models 1, 2 and 3. This permits an assessment of the effect of bias in nonprobability-based informative priors on all of the model coefficients.

Second, to evaluate model-based predictions, we split the probability survey data randomly into two parts: a training set (denoted by $\mathbf{y}$) and a test set ($\tilde{\mathbf{y}}$). We then use the training set to fit the models specified in Subsection 2.1. Next, we predict the outcome variable in the test set $\tilde{\mathbf{y}}$. We do so by applying posterior distributions of model parameters estimated using $\mathbf{y}$ to the covariates in the test set. The resulting distributions are called posterior predictive distributions, that is, posteriors for each data point.

Next, to evaluate the error properties of the predictions for the three models, we calculate the bias, variance, and MSE of the means, denoted by $\bar{\tilde{\mathbf{y}}}$, of the posterior predictive distributions for $\tilde{\mathbf{y}}$. In the simulation, we define the MSE as:

$$MSE(\bar{\tilde{\mathbf{y}}}) = \mathbb{E}\left[(\bar{\tilde{\mathbf{y}}} - \tilde{\mathbf{y}})^2\right],$$

which can be decomposed into variance and bias $MSE(\bar{\tilde{\mathbf{y}}}) = Bias^2(\bar{\tilde{\mathbf{y}}}) + Var(\bar{\tilde{\mathbf{y}}})$.

We compute the bias as the difference between the mean of the posterior means, $\bar{\tilde{\mathbf{y}}}$, and the mean of the test sample outcome $\tilde{\mathbf{y}}$, i.e., $Bias(\bar{\tilde{\mathbf{y}}}) = \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}} - \frac{1}{n}\Sigma\tilde{\mathbf{y}}$ whereas $Var(\bar{\tilde{\mathbf{y}}})$ is the unbiased estimator of the variance of $\bar{\tilde{\mathbf{y}}}$.

We calculate the bias, variance, and MSE of the posterior predictive means for the three models described in Subsection 2.1 under different probability sample size scenarios. To accomplish this, we run the models on training sets ranging in size from 50 to 600 cases with intervals of 50, and from 600 to 1,000 with intervals of 100. The samples are constructed cumulatively so that the same cases used in the smaller samples are also included in the larger samples.

3.2. Results

Having generated the artificial probability and nonprobability samples for each size and level of bias as described in the previous section, we applied the three modelling approaches (Model 1, 2, and 3) as described in Subsection 2.1 to produce posterior distributions of model parameters and predictive distributions for the test sample in simulation part B. We then compare the effect of bias introduced in the nonprobability sample on bias, variance, and MSE of the coefficients and means of the posterior

predictive distributions as defined in Subsection 3.1. The bias, variance, and MSE are averaged over 100 simulated data sets.

3.2.1. Part A: Regression Coefficients

Figure 1 presents the bias, variance, and MSE for the three coefficients, where β_3 has been generated with bias in the nonprobability (*NP*) sample. First, we observe that Model 2 does not lead to bias in the coefficients and performs similarly to Model 1, which relies on weakly-informative priors without information from the *NP* samples. It also leads to improvements in variance (middle panel of Figure 1) and MSE (lower panel). For Model 3, we observe larger improvements in variance compared to Models 1 and 2. However, in the presence of bias, the MSE tends to be dominated by it. This results from the fact that the

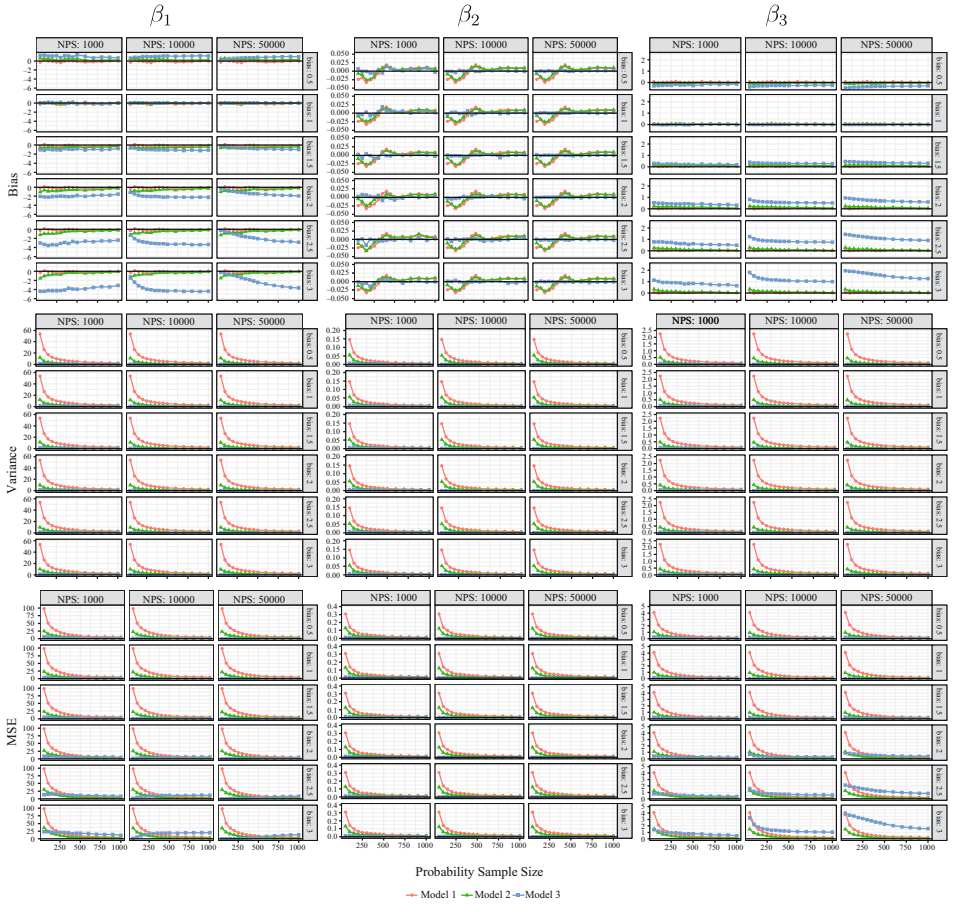

Fig. 1. Effect of bias in nonprobability samples on regression coefficient. Note: Regression parameters are in columns; measures in rows, where top row is bias (difference between the posterior mean from the model and the true coefficient), middle row is variance, and bottom row is mean-squared error (MSE), averaged over 100 simulations. Each panel shows the combination of three nonprobability sample sizes (NPS) and six levels of bias introduced to β_3 in nonprobability sample (bias:1 denotes unbiased coefficient, i.e., $\beta_3 \times 1$), with probability sample size on the x-axis.

prior in Model 3 relies on the size and variability of the *NP* sample and does not protect against bias present in it.

More precisely, in Model 3 the positive bias in the posterior mean of β_3 (top right panel) is increasing with the introduced bias (difference between posterior mean of the coefficient and the true coefficient) and it is more persistent with larger nonprobability sample sizes (*NPS*). This is offset by the negative bias in the intercept β_1 as the regression equation needs to be consistent with the expectation of the outcome in the probability sample, $E[y] = 10$. However, for large *NPS* (10,000 or 50,000), the prior for β_1 is relatively tight and it dominates the posterior of β_1 for small probability sample sizes (*PS*), which subsequently leads to bias in the predictions of the outcome (see Figure 2 and the following Subsection 3.2.2). With an increase in *PS*, the posterior becomes more and more dominated by the unbiased probability sample, which first increases the bias in the posterior of β_1 and decreases in β_3 (e.g., $NPS = 10{,}000$ and $bias = 3$ in top left and right panel of Figure 1) to gradually decrease bias in both coefficients (e.g., $NPS = 1{,}000$ and $bias = 2.5$) and output predictions (left panel in Figure 2). Coefficient β_2 remains unaffected by bias.

3.2.2. Part B: Model-Based Predictions

Figure 2 shows the effect of bias introduced in the nonprobability samples on the predictive ability of the models when priors are based on those samples. We average over means of posterior predictive distributions (referred to as predictions for brevity) for 500 generated outcome data points. In all comparisons, we utilize the true generated outcome.

In Figure 2 we observe that Model 2, compared with the weakly informative Model 1 without input from nonprobability samples, yields mostly unbiased predictions. For Model 3, as indicated in the previous section, the bias in predictions changes with the size of bias in β_3. A large bias in the coefficient yields larger prediction bias, larger variance, and larger MSE. Also, for larger nonprobability sample sizes (*NPS*), the bias persists for larger probability sample sizes (*PS*). However, for a moderate bias (β_3 multiplied by 0.5 to 1.5), Model 2 and Model 3 show a reduction in the prediction variance and MSE (presented on log scale) compared with Model 1 and for nonprobability sample sizes of 1,000 and 10,000.

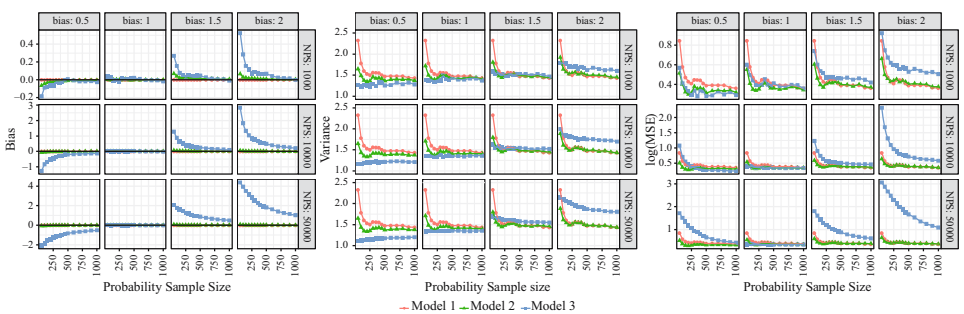

Fig. 2. Effect of bias in nonprobabiliy samples on predicted outcome. Note: Left panel shows bias (difference between the average predicted outcome and the average true generated outcome), middle panel shows variance, and right panel shows the logarithm of mean-squared error (MSE), averaged over 100 simulations. Each panel shows a combination of four levels of bias in β_3 (Beta:1 denotes no bias, i.e., $\beta_3 \times 1$) and three nonprobability sample sizes (NPS), with probability sample size on the x-axis.

For $NPS = 50{,}000$ and larger amounts of bias, reductions in variance and MSE are observed only for Model 2 and they are relatively small compared with Model 1 predictions.

4. Real-Data Application

To demonstrate the proposed methods on actual survey data, we make use of two probability-based surveys: the German Internet Panel (GIP) and the GESIS Panel, and eight nonprobability surveys. Each survey implemented the same questionnaire to respondents during overlapping field periods. Relevant details of the surveys are provided below.

We demonstrate the proposed Bayesian method on two continuous outcome measures: an additive index of a subset of Big Five (BIG-5; Digman 1990; Goldberg 1993) personality items and an additive index of a subset of Need for Cognition (NFC; Cacioppo and Petty 1982) scale items. The Big Five index included four items related to "trust of people", "artistic interests", "finding fault in others", and having an "active imagination" with each item using a 5-point response scale from "strongly disagree" to "strongly agree." The distribution of additive values approximately followed a normal distribution. The NFC index included four items about "knowing answers without understanding their rationale", "being confronted with tricky tasks to solve", "preferring to solve complex to simple problems", and "thinking only because one has to." Each item used a 7-point response scale from "strongly disagree" to "strongly agree." A square-root transformation was applied to the index to achieve approximate normality.

4.1. German Internet Panel

The GIP is an ongoing individual-level longitudinal online survey, which is designed to be representative of the population aged 16–75 in Germany. It is the central data collection project of the Collaborative Research Center 884 "Political Economy of Reforms" funded by the German Research Foundation (DFG). In 2012 and 2014, the GIP recruited sample members by means of a 3-stage stratified probability area sample and face-to-face recruitment interviews. At the first sampling stage, a random sample of areas was drawn from a database of 52,947 areas in Germany, each containing approximately equal numbers of households. Within each PSU, listers recorded every household along a predefined random route. Subsequently, a random sample of households to be interviewed drawn. All age-eligible members of sampled households were invited to become online panelists (Blom et al. 2015). The GIP covers individuals without computer and/or internet access by equipping them with the necessary devices (Blom et al. 2016a; Herzing and Blom 2019). The first recruitment process, which took place in 2012, yielded a recruitment rate of 18.5% (also based on Response Rate 2; AAPOR 2016) and in the second recruitment process in 2014 a recruitment rate of 20.5% (also based on AAPOR Response Rate 2) was achieved. Every two months, all panel members are invited to take part in an online survey of about 20–25 minutes on various social, economic, and political topics. The questionnaire module used in the present study was implemented 1–31 March 2015. Out of 4,989 original panel members, 3,426 completed the survey for a completion rate of 68.7%. Despite the low recruitment rate, the representativeness of the GIP compares well to other probability-based surveys in Germany (Blom et al. 2017).

4.2. GESIS Panel

Like the GIP, the GESIS Panel is an ongoing individual-level probability-based longitudinal survey. It is designed to be representative of the German-speaking population aged between 18 and 70 years, permanently residing in Germany. The sample was drawn from municipal population registers using a stratified multistage sampling procedure. All sample members who were interviewed with face-to-face recruitment interviews were asked to participate in the panel. The recruitment process, which took place in 2013/14, yielded a panel registration rate of 28.4% (based on Response Rate 1; AAPOR, 2016). Subsequent interviews are conducted on a bi-monthly basis using a mix of mail and web data collection. Mail questionnaires are sent to participants who are unable or unwilling to complete the interview online. Interviews are divided into two parts: a 15-minute interview on modules submitted by external researchers and a five-minute interview devoted to longitudinal core study topics developed by GESIS. The core study covers a range of topics, including values, political behavior, well-being, and usage of information technology. The questionnaire module we use was approved by the GESIS Panel team and fielded 18 February–14 April 2015. Out of 6,210 original panel members, 3,822 completed the interview (61.5%). More details of the GESIS Panel methodology can be found in Bosnjak et al. (2017), where they show the representativeness of the panel to be similar to other probability-based surveys in Germany (see also Blom et al. 2016b).

4.3. Nonprobability Surveys

The eight nonprobability web surveys were conducted by different commercial vendors. The vendors were recruited through a call for tender published in November 2014. The tender call sought to implement a ten-minute questionnaire on a sample of approximately 1,000 respondents in three waves of data collection. Initial data collection was to take place in March 2015 with two follow-up surveys in September 2015 and March 2016. The primary stipulation was that the sample should be representative of the general population aged 18–70 years living in Germany. Exactly how representativeness was to be achieved (e.g., quota sampling) was left to the discretion of each vendor. Out of 17 bids, seven commercial vendors were commissioned based on technical requirements and budgetary considerations. An eighth commercial vendor, upon learning about the study goals of the project, voluntarily offered to participate without compensation. Further details of each nonprobability survey, including cost information, is provided in Table 1. To maintain confidentiality, we do not identify the commercial vendors by name and simply refer to the nonprobability surveys by number, that is, Survey 1, Survey 2, and so on. The actual cost of the commercial surveys (excluding the gratis survey) ranged from EUR 5,392.97 to EUR 10,676.44. The average cost per respondent therefore ranged from EUR 5.40 to EUR 10.29. We do not have cost information for the GIP and GESIS Panel surveys.

4.4. Comparison of Outcome Variables Between Surveys

Here, we examine the extent to which the outcome variables differ within and between the probability and nonprobability surveys. Figure 3 displays estimated means and 95%

Table 1. List of probability and nonprobability surveys.

Survey	No. respondents	Quota variables	Fieldwork period	Total cost (in Euros)	Average cost per respondent (in Euros)
GIP	3,426	N/A	1st–31st March 2015	Unavailable	Unavailable
GESIS	3,822	N/A	18th February–14th April 2015	Unavailable	Unavailable
1	1,012	Age, gender, region, education	1st–31st March 2015	0 (pro bono)	N/A
2	1,000	Age, gender, region	5th–18th March 2015	5,392.97	5.40
3	999	Age, gender, region	2nd–11th March 2015	5,618.57	5.63
4	1,000	Age, region	1st–18th March 2015	7,061.11	7.07
5	994	Age, gender, region	2nd–16th March 2015	7,411.00	7.46
6	1,002	Age, gender, region, education	25th March–1st April 2015	7,636.22	7.62
7	1,000	Age, gender, region	3rd–9th March 2015	8,380.46	8.39
8	1,038	Age, gender, region	5th–11th March 2015	10,676.44	10.29

confidence intervals (CIs) for the BIG-5 (left panel) and NFC (right panel) outcome variables in the GIP and GESIS Panel surveys.

The figures show very little difference between the GIP and GESIS Panel estimates of BIG-5 and NFC. Both probability surveys yield mean estimates that overlap by their respective confidence intervals. Larger differences are apparent between the probability and nonprobability surveys. For the BIG-5 variable, all nonprobability surveys yield mean estimates that fall outside of the GIP and GESIS Panel confidence intervals. All but one of the nonprobability-based means is lower than the GIP and GESIS Panel means. Differences between the nonprobability surveys are less pronounced, as most of the estimates are relatively homogeneous and lie within a close range. For the NFC variable, the nonprobability mean estimates are larger than the corresponding GIP and GESIS Panel estimates. All but two of the nonprobability surveys yield mean estimates that lie outside of the GIP and GESIS Panel CIs. Analogous to the BIG-5 estimates, most of the nonprobability NFC estimates are similar to each other. In summary, it is apparent that differences in the means exist between the probability and nonprobability surveys, but differences are less apparent between the nonprobability surveys.

4.5. Comparison of Regression Coefficients Between Surveys

Next, we compare the ML estimates of regression coefficients of BIG-5 and NFC obtained from the probability and nonprobability surveys. Control variables include age (four categories), sex (binary), marital status (three categories), occupation (four categories), secondary education certificate (three categories), region of residence (binary), internet access (binary), and housing tenure (binary). We also include a survey weight variable,

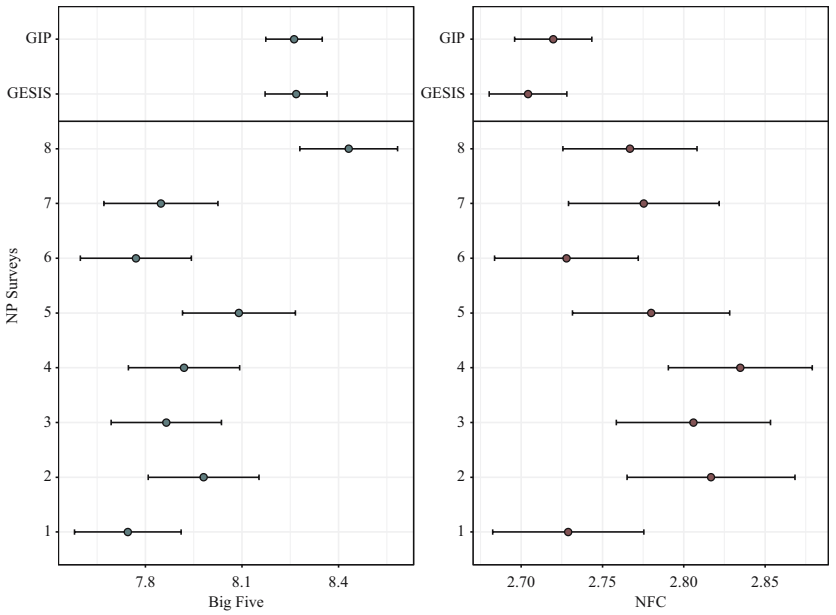

Fig. 3. Means and 95% confidence intervals for BIG-5 (left panel and Need for Cognition (NFC) (right panel) on the probability (GIP and GESIS Panel) and eight nonprobability (NP) surveys.

which was produced to reduce bias through a raking adjustment to population benchmarks (Blom et al. 2017), as a covariate in the regression. For the regression analysis of the GESIS Panel and the nonprobability surveys, we use the same independent variables, minus region and the survey weight, which were both unavailable.

Figure 4 shows the regression coefficients and 95% CIs from the BIG-5 model estimated from the GIP Panel with corresponding coefficients estimated from the nonprobability surveys. The conclusions for the GESIS Panel (not shown) are virtually the same. Overall, there is a close correspondence between the probability and nonprobability coefficients across the models. Very few of the nonprobability estimates lie outside of the CI ranges of the probability estimates. The results contrast with the results presented in Subsection 4.4, where differences in the outcome variable between the probability and nonprobability surveys were more pronounced. Our finding that regression coefficients are less affected by bias than univariate estimates in nonprobability samples is consistent with other work (Ansolabehere and Schaffner 2014; Pasek 2016).

5. Application Results

5.1. Evaluation and Efficiency

In this section, we evaluate the performance of the three modelling approaches on the GIP and GESIS Panel data by using the model-based predictions as described in Subsection 3.1. Splitting the probability survey data into training and test sets in the applicaton is done for evaluation purposes only and takes advantage of the abundant number of probability

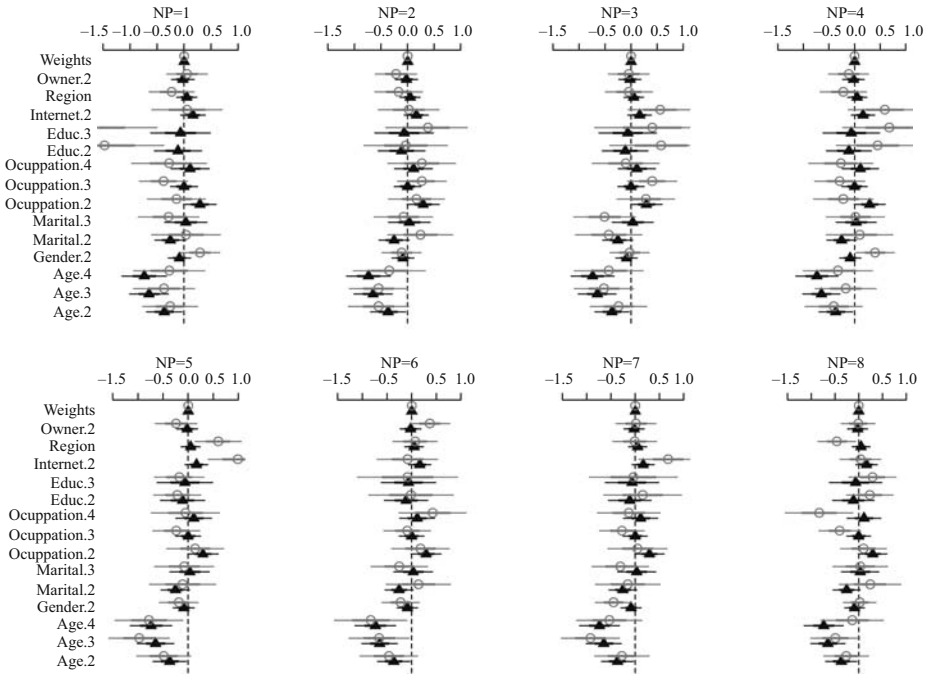

Fig. 4. Comparison of OLS regression coefficients and 95% confidence intervals for BIG-5 in the German Internet Panel (triangles) and eight nonprobability surveys (circles).

cases we have at our disposal. In practice, we envision the practitioner would only have access to a small probability sample and therefore this evaluation step would not be feasible. We then use the training set to fit the models specified in Subsection 2.1.

After excluding cases with missing data and assigning 1,000 cases from the probability survey to the training set, the remaining cases are assigned to the test set. For the BIG-5 outcome, the test set includes 1,924 and 2,150 cases for the GIP and GESIS Panel surveys, respectively. For the NFC outcome, the respective sample sizes are 1,891 and 2,088 cases. The nonprobability sample sizes are not altered.

In the application, we use $MSE(\bar{\tilde{\mathbf{y}}}) = \mathbb{E}\left[\left(\bar{\tilde{\mathbf{y}}} - \bar{\tilde{\mathbf{y}}}_{adj}^{IS}\right)^2\right]$, where $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ are the model-adjusted, in-sample (superscript *IS*) predictions in the test set of the probability survey. These predictions are adjusted by (i) applying the regression model with the same covariates as in Models 1, 2, and 3 exclusively to the test set, with noninformative Jeffrey's priors, and then (ii) computing posterior predictive means and using them as $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$. By using the adjusted predictions rather than the original observations, we account for the fact that our model may be unrealistic and explain only a small part of data variability. An important distinction between $\bar{\tilde{\mathbf{y}}}$ and $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ is that the former are out-of-sample predictions made by using one of the three specifications of models described in Subsection 2.1 on the training set, whereas the latter are in-sample predictions informed exclusively by the withheld test set. Analogously, the bias here is the difference between the mean of the posterior means, $\bar{\tilde{\mathbf{y}}}$, and the mean of the model-adjusted predictions $\bar{\tilde{\mathbf{y}}}_{adj}^{IS}$, that is, $Bias(\bar{\tilde{\mathbf{y}}}) = \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}} - \frac{1}{n}\sum \bar{\tilde{\mathbf{y}}}_{adj}^{IS}$ (cf. Subsection 3.1).

Finally, to assess the efficiency of the two models informed by the nonprobability data (Models 2 and 3) relative to the reference model (Model 1), which is based on only weakly-informative priors, we examine the ratio of the variances of the posterior predictive means:

$$\epsilon\left(Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right), Var\left(\bar{\tilde{\mathbf{y}}}_{Model2}\right)\right) = \frac{Var\left(\bar{\tilde{\mathbf{y}}}_{Model2}\right)}{Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right)},$$

$$\epsilon\left(Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right), Var\left(\bar{\tilde{\mathbf{y}}}_{Model3}\right)\right) = \frac{Var\left(\bar{\tilde{\mathbf{y}}}_{Model3}\right)}{Var\left(\bar{\tilde{\mathbf{y}}}_{Model1}\right)}.$$

Analogously, we examine the ratio of the MSEs of the posterior predictive means. If the value of any ratio is less than 1, then the informative model is more efficient than the reference model. Conversely, if the ratio is equal to or greater than 1 then the informative models do not produce efficiency gains over the reference model.

5.2. *Variance, Bias, and MSE*

This section presents the results of the three modeling approaches (Model 1, 2, and 3) implemented on the GIP and GESIS Panel surveys. The variance, bias, and MSE as defined above are computed for the posterior predictive means (hereinafter referred to simply as the mean estimates) of the two outcome variables produced under each model. The entire procedure of splitting the probability data into training and test sets was conducted 100 times to produce 100 estimates of variance, bias, and MSE for each probability sample size. The forthcoming results report the averages of these 100 repetitions. Each of the models was fitted using the independent variables described in Subsection 4.4.

The posterior characteristics were computed, as in Section 3, using three MCMC chains with samples of 1,000 and a 100 iteration burn-in sample. This ensured convergence of all chains used for creating the posterior distributions. We investigated the convergence using a larger number of iterations and found that the results were robust with respect to the number of iterations used.

Results for the BIG-5 and NFC means are shown for both GIP and GESIS Panel data in Figure 5. For brevity, we show the results using only one nonprobability survey, NP = 5, the middle-priced of the seven paid-for nonprobability surveys. Similar results (not shown) were found when the other nonprobability surveys were used.

Models 2 and 3 yield very similar variance estimates and are virtually indistinguishable in the figures. For the smallest probability sample sizes, both models yield substantially smaller variance estimates compared to the reference model (Model 1). Maximum variance efficiency is achieved with a probability sample size of 50, while efficiency gains tend to diminish as the sample size increases. All three models converge to variance equivalency at about n = 500. What is most striking is that the variance estimates produced under Models 2 and 3 for the smallest sample sizes are approximately equivalent to the variance estimates produced under the reference model for the largest probability sample size of 1,000. In other words, a probability sample size of only 50 cases with a supplement of 1,000 nonprobability cases achieves roughly the same variance as a much larger probability sample size of 1,000 does on its own.

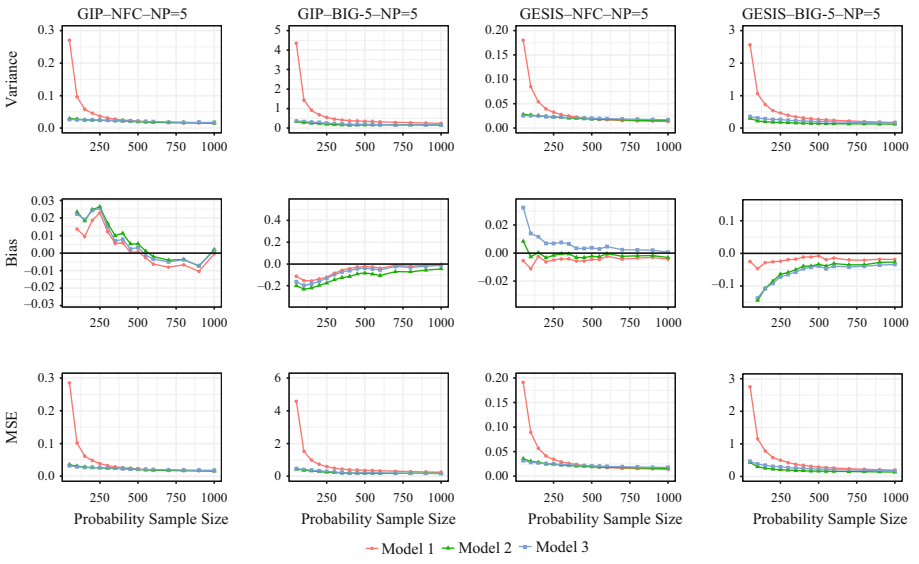

Fig. 5. Variance, bias, and mean-squared error (MSE) for estimates of BIG-5 and need for Cognition (NFC) in the GESIS Panel and GIP. Note: Results shown for one nonprobability survey (NP = 5) only. Similar results were found when other nonprobability surveys were used.

Concerning bias, as expected, the majority of plots show a slightly larger bias in Models 2 and 3 relative to the reference model for the smallest probability sample sizes, where the nonprobability-based priors have their strongest influence on the posterior estimations. In general, however, the magnitude of the bias is quite small, which is consistent with the results of the comparison of regression coefficients in Subsection 4.5.

In terms of MSE, the figures reveal that for small probability sample sizes Models 2 and 3 yield MSE values that are substantially smaller than those of the reference model. These MSE reductions persist at a decreasing rate until the probability sample size reaches about 500, at which point the values from all three models converge. The results clearly indicate that any increase in bias due to using the nonprobability-based priors is offset by the reduction in variance. Analogous to the variance results, the MSE values under Models 2 and 3 remain similarly small across the sample size spectrum. The practical implication is that the same MSE achieved through a large probability sample can be roughly achieved by supplementing a very small probability sample (e.g., 50–100 cases) with a larger nonprobability sample.

5.3. Model Efficiency and Cost Implications

In the final analysis, we summarize the MSE/variance efficiencies achieved through Models 2 and 3 and examine whether they would have likely resulted in a cost saving compared to Model 1 for a given MSE. Figure 6 presents efficiency ratios of MSE and variance for mean estimates of BIG-5 (upper panel) and NFC (lower panel) for the GIP and GESIS Panel surveys. The ratios are averaged across all eight nonprobability surveys (with equal weight given) to provide an overall summary measure of model efficiencies.

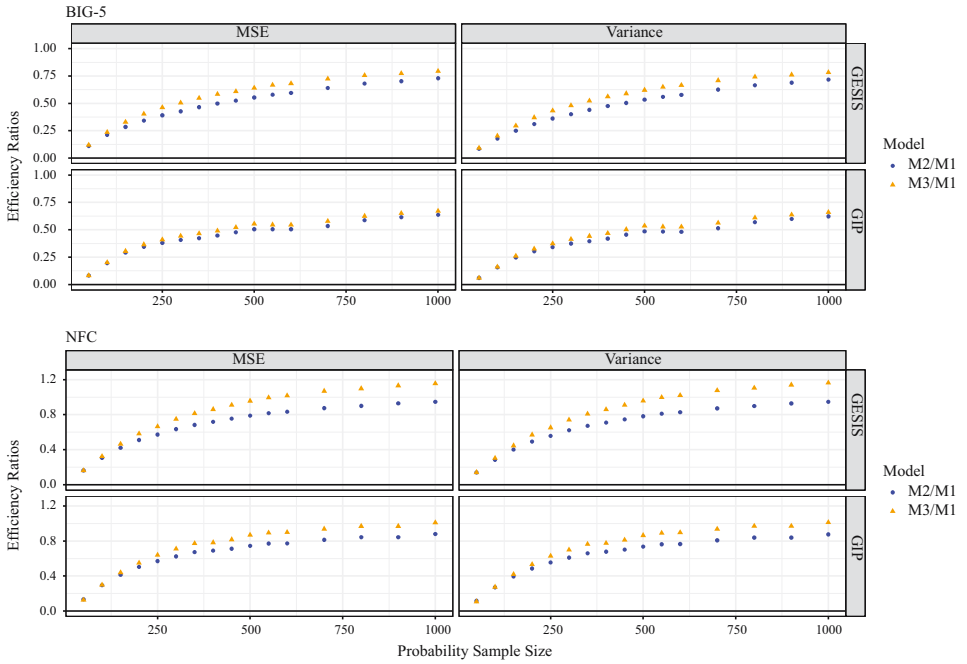

Fig. 6. Efficiency ratios of mean-squared error (MSE) and variance (columns) for mean estimates of BIG-5 (upper panel) and Need for Cognition (NFC, lower panel) in the GESIS Panel and GIP (rows). Note: Ratios are averaged across all eight nonprobability surveys.

Four observations can be made from Figure 6: (i) as observed in the previous analyses, MSE/variance efficiency gains are largest for the smallest probability sample sizes. For example, Models 2 and 3 reduce MSE and variance by more than 80%, on average, compared to Model 1 for the smallest sample size of 50. Even when the sample size is doubled to 100 cases, MSE/variance reductions of at least 70% are observed; (ii) the BIG-5 variable experiences larger efficiency gains than the NFC variable, and both variables yield slightly larger efficiency gains in the GIP than in the GESIS Panel; (iii) gains in variance efficiency are only slightly larger than gains in MSE efficiency, which indicates that the bias due to utilizing nonprobability-based priors is marginal compared to the corresponding variance reduction; and (iv) Models 2 and 3 yield very similar gains in MSE and variance efficiency with slightly larger gains achieved under Model 2.

To demonstrate the cost implications (and potential cost savings) of the different models, we utilize actual cost data for the nonprobability surveys (see Table 1) and hypothetical cost data for the probability-based GIP survey. For the GIP survey, we assume a cost per respondent of 22 euros, which is roughly 2 and 4 times larger than the most and least expensive nonprobability surveys (excluding the gratis survey), respectively. Using these data, we perform a crude estimation of the expected cost of performing a probability-only survey (under Model 1) that would achieve the same MSE that was actually achieved under Model 3 – the more conservative of the two models utilizing nonprobability-based priors. We then compare the estimated

Table 2. Estimated cost differences (and percent cost savings) between model 1 and model 3 for fixed mean-squared error (MSE) values achieved under model 3 for mean estimates of BIG-5 in the German Internet Panel.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 50							
MSE (Model 3)	0.288	0.300	0.521	0.462	0.423	0.406	0.222
Cost (Model 3)	6,492.97	6,718.57	8,161.11	8,511.00	8,736.22	9,480.46	11,776.44
Est. cost (Model 1; same MSE)	14,267.37	13,904.09	8,811.64	9,917.81	10,739.26	11,122.18	16,473.01
Cost difference (Est. M1 – M3)	7,774.40	7,185.52	650.53	1,406.81	2,003.04	1,641.72	4,696.57
Est. cost savings (%)	54.49	51.68	7.38	14.18	18.65	14.76	28.51
GIP sample size = 100							
MSE (Model 3)	0.249	0.288	0.433	0.408	0.358	0.370	0.212
Cost (Model 3)	7,592.97	7,818.57	9,261.11	9,611.00	9,836.22	10,580.46	12,876.44
Est. cost (Model 1; same MSE)	15,526.23	14,267.37	10,521.24	11,076.31	12,292.75	11,987.14	16,840.42
Cost difference (Est. M1 – M3)	7,933.26	6,448.80	1,260.13	1,465.31	2,456.53	1,406.68	3,963.98
Est. cost savings (%)	51.10	45.20	11.98	13.23	19.98	11.73	23.54
GIP sample size = 150							
MSE (Model 3)	0.227	0.262	0.397	0.373	0.346	0.345	0.203
Cost (Model 3)	8,692.97	8,918.57	10,361.11	10,711.00	10,936.22	11,680.46	13,976.44
Est. cost (Model 1; same MSE)	16,292.76	15,092.86	11,331.37	11,912.13	12,607.50	12,634.15	17,179.17
Cost difference (Est. M1 – M3)	7,599.79	6,174.29	970.26	1,201.13	1,671.28	953.69	3,202.73
Est. cost savings (%)	46.65	40.91	8.56	10.08	13.26	7.55	18.64

Table 2. Continued.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 200							
MSE (Model 3)	0.213	0.235	0.346	0.336	0.315	0.321	0.194
Cost (Model 3)	9,792.97	10,018.57	11,461.11	11,811.00	12,036.22	12,780.46	15,076.44
Est. cost (Model 1; same MSE)	16,803.26	16,009.08	1,2607.50	12,876.97	13,464.97	13,293.84	17,525.78
Cost difference (Est. M1 – M3)	7,010.29	5,990.51	1,146.39	1,065.97	1,428.75	513.38	2,449.34
Est. cost savings (%)	41.72	37.42	9.09	8.28	10.61	3.86	13.98
GIP sample size = 250							
MSE (Model 3)	0.194	0.209	0.317	0.294	0.277	0.283	0.194
Cost (Model 3)	10,892.97	11,118.57	12,561.11	12,911.00	13,136.22	13,880.46	16,176.44
Est. cost (Model 1; same MSE)	17,525.78	16,952.48	13,407.64	14,084.37	14,610.07	14,421.98	17,525.78
Cost difference (Est. M1 – M3)	6,632.81	5,833.91	846.53	1,173.37	1,473.85	541.52	1,349.34
Est. cost savings (%)	37.85	34.41	6.31	8.33	10.09	3.75	7.70

Table 3. Estimated cost differences (and percent cost savings) between model 1 and model 3 for fixed mean-squared error (MSE) values achieved under model 3 for mean estimates of need for cognition (NFC) in the German Internet Panel.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 50							
MSE (Model 3)	0.055	0.039	0.033	0.032	0.025	0.034	0.021
Cost (Model 3)	6,492.97	6,718.57	8,161.11	8,511.00	8,736.22	9,480.46	11,776.44
Est. cost (Model 1; same MSE)	4,470.38	7,104.11	8,560.89	8,837.23	11,098.38	8,294.81	12,695.68
Cost difference (Est. M1 – M3)	−2,022.59	385.54	399.78	326.23	2,362.16	−1,185.65	919.24
Est. cost savings (%)	0	5.43	4.67	3.69	21.28	0	7.24
GIP sample size = 100							
MSE (Model 3)	0.048	0.033	0.029	0.029	0.022	0.029	0.018
Cost (Model 3)	7,592.97	7,818.57	9,261.11	9,611.00	9,836.22	10,580.46	12,876.44
Est. cost (Model 1; same MSE)	5,441.20	8,560.89	9,732.26	9,732.26	12,272.42	9,732.26	14,071.49
Cost difference (Est. M1 – M3)	−2,151.77	742.32	471.15	121.26	2,436.20	−848.20	1,195.05
Est. cost savings (%)	0	8.67	4.84	1.25	19.85	0	8.49
GIP sample size = 150							
MSE (Model 3)	0.042	0.029	0.027	0.027	0.020	0.026	0.017
Cost (Model 3)	8,692.97	8,918.57	10,361.11	10,711.00	10,936.22	11,680.46	13,976.44
Est. cost (Model 1; same MSE)	6,488.51	9,732.26	10,388.86	10,388.86	13,136.09	10,736.71	14,568.12
Cost difference (Est. M1 – M3)	−2,204.46	813.69	27.75	−322.14	2,199.87	−943.75	591.68
Est. cost savings (%)	0	8.36	0.27	0	16.75	0	4.06

Table 3. Continued.

	Nonprobability surveys						
	2	3	4	5	6	7	8
GIP sample size = 200							
MSE (Model 3)	0.039	0.028	0.025	0.027	0.019	0.026	0.017
Cost (Model 3)	9,792.97	10,018.57	11,461.11	11,811.00	12,036.22	12,780.46	15,076.44
Est. cost (Model 1; same MSE)	7,104.11	10,054.22	11,098.38	10,388.86	13,594.43	10,736.71	14,568.12
Cost difference (Est. M1 – M3)	−2,688.86	35.65	−362.73	−1,422.14	1,558.21	−2,043.75	−508.32
Est. cost savings (%)	0	0.35	0	0	11.46	0	0
GIP sample size = 250							
MSE (Model 3)	0.037	0.026	0.024	0.026	0.018	0.025	0.018
Cost (Model 3)	10,892.97	11,118.57	12,561.11	12,911.00	13,136.22	13,880.46	16,176.44
Est. cost (Model 1; same MSE)	7,553.97	10,736.71	11,474.45	10,736.71	14,071.49	11,098.38	14,071.49
Cost difference (Est. M1 – M3)	−3,339.00	−381.86	−1,086.66	−2,174.29	935.27	−2,782.08	−2,104.95
Est. cost savings (%)	0	0	0	0	6.65	0	0

Model 1 costs with the actual and estimated costs of Model 3 for the fixed MSE. The analysis is conducted in two steps. First, a linear regression model of GIP costs (log-transformed) on MSE (and MSE squared) is fitted using the Model 1 MSE results. Next, we plug-in the MSE values achieved under Model 3 into the fitted model to estimate the (back-transformed) cost of collecting a probability-only sample. Lastly, we calculate differences between the estimated Model 1 costs and the actual/estimated Model 3 costs for each realized MSE and compute the expected cost savings (in percentages) under Model 3.

Tables 2 and 3 show the estimated cost differences between Model 1 and Model 3 for the BIG-5 and NFC outcomes, respectively. The cost differences are shown for the five smallest probability sample sizes (50, 100, 150, 200, and 250). Regarding the BIG-5 outcome, cost savings are evident for each sample size. In general, the largest cost savings occur for the smallest sample size of 50, followed by 100, and so on, which is consistent with the MSE reductions observed in the previous analyses. However, there is large variation in the amount of cost savings across the seven (paid-for) nonprobability surveys. For example, when the two least expensive nonprobability surveys (surveys 2 and 3) are used to construct the priors then estimated cost savings of about 55% and 52% are achieved, respectively, for the BIG-5 outcome with a probability sample size of 50. The other, more expensive, nonprobability surveys yield cost savings ranging from about 7% to 29% for the same sample size. For larger probability sample sizes of 100 and 150, the range of cost savings for the BIG-5 outcome is slightly reduced to between 12% and 51%, and 8% to 47%, respectively, across all nonprobability surveys. Beyond 150 probability cases, the two least expensive nonprobability surveys continue to achieve significant cost savings (greater than 30%), but as for the more expensive nonprobability surveys, the cost savings are more modest (less than 15%).

Cost savings for the NFC outcome are much less pronounced. Only nonprobability survey 6 yields a modest cost savings (about 21%) for a probability sample size of 50. The remaining nonprobability surveys produce cost savings of less than 8% for the same sample size, and some surveys achieve no cost savings at all. With a probability sample size greater than 150 cases, the majority of nonprobability surveys yield no cost savings. Thus, the cost-effectiveness of Model 3 appears to be sensitive to the probability sample size, differences in per respondent costs between the probability and nonprobability surveys, and the outcome variable of interest.

6. Discussion

This study demonstrated a novel method of using Bayesian inference to supplement small- and modest-sized probability samples with nonprobability samples in a way that can improve the cost and error properties of survey estimates. Specifically, we proposed two ways of constructing informative nonprobability-based priors. We then showed that using these priors to inform estimates derived from small probability samples yields substantially lower mean-squared errors (MSEs) compared to estimates derived from probability-only samples. Moreover, applying these informative priors to small probability samples (e.g., 50 or 100 cases) through a real-data application yielded estimates that were approximately as efficient as estimates based on much larger

probability-only samples (e.g., 1,000 cases). Reductions in MSE were primarily driven by large reductions in variability which completely offset any increases in bias. By using simulated data, we also demonstrated general applicability of the method and its mechanism for various sample sizes and levels of bias in the nonprobability samples.

Using actual cost data for several nonprobability surveys and a plausible assumed cost for a probability survey, we showed that the method can lead to large expected cost savings (up to 55% in our application) compared to a probability-only sample for a given MSE. However, the extent of cost savings depended on the outcome variable of interest and the nonprobability sample costs which varied across the survey vendors used. The largest cost savings tended to occur when the per-respondent costs were about four times greater in the probability survey than in the nonprobability survey.

At a time when many survey researchers are shifting away (or abandoning altogether) probability samples and embracing less-expensive nonprobability samples despite their known caveats, our results suggest that it is possible to retain the benefits of both sampling approaches in a way that is beneficial from both a cost and error perspective. The proposed method is ideally suited for tight survey budgets in which only a small probability sample (e.g., 50–100 cases) can be afforded alongside a larger nonprobability sample. The finding that the method can yield estimates that are just as efficient as estimates derived from very large probability samples is a particularly attractive feature for survey practice.

However, there are potential issues with the Bayesian method that should be considered. First, it is possible that some nonprobability samples may contain large biases that, when utilized as prior distributions, could negate reductions in variability and yield larger MSEs compared to probability-only samples. We did not face this issue in our application, as the estimated regression coefficients used in our models were not substantially different between the probability and nonprobability surveys. When using simulated data, we found that if the interest is in the size of the effect (regression coefficient), the combination of probability and nonprobability samples yields reductions in variance and MSE of that effect with minimal amount of bias. However, using nonprobability-based priors for model-based predictions or imputation of a missing outcome variable may not produce desired improvements if bias in the nonprobability sample is substantial (in our simulation study a bias of around 50% of the outcome variable). Thus, it would be prudent for the researcher to adjust the nonprobability sample data in advance of constructing priors to minimize bias at the outset, especially if prediction is the ultimate objective.

A further issue with the Bayesian approach is the vast number of modeling specifications and prior configurations that one could employ. We deliberately kept the modeling and prior specification as basic as possible. This sometimes required choosing simplicity over complexity in order to facilitate implementation and minimize computation time. Further refinements of the modeling approach could be developed to account for more complex data structures, such as categorical outcome variables. In addition, adapting the modeling approach to incorporate complex sample design features (e.g., stratum, cluster indicators) is an area for future work.

In conclusion, we find that augmenting a probability sample with a nonprobability sample under a Bayesian framework can produce survey estimates with smaller MSE and

potentially large cost savings relative to probability-only samples. The proposed method, which turns the usual approach of treating a probability sample as an unbiased prior for a nonprobability sample "on its head" as one reviewer put it, could be a useful import to survey practice where cost-saving measures and error-reduction tools are in high demand. However, despite the advantages of the method, survey organizations using nonprobability samples may still be skeptical to the idea of fielding a small probability sample survey in parallel when the nonprobability sample will likely dominate the inference. Here, we would contend that adopting a system of estimation that accounts for both sampling streams, yet incentivizes probability-based observations and allows for the direct quantification of uncertainty in survey estimates is a more defensible strategy than one that renounces probability sampling entirely along with all of its attractive theoretical properties. Moreover, the idea of enhancing a small, but carefully designed, probability sample with abundant but potentially error-prone data is not a new idea and is a widely accepted strategy in small area applications where sparse probability samples are routinely supplemented with alternative data sources to improve the cost and error properties of population estimates (Marchetti et al. 2016; Porter et al. 2014; Briggs et al. 2007; Schmertmann et al. 2013).

7. References

AAPOR. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed July 2019).

Ansolabehere, S. and D. Rivers. 2013. "Cooperative Survey Research." *Annual Review of Political Science* 16: 307–329. Doi: https://doi.org/10.1146/annurev-polisci-022811-160625.

Ansolabehere, S. and B.F. Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22(3): 285–303. Doi: https://doi.org/10.1093/pan/mpt025.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. *Report of the AAPOR Task Force on Non-Probability Sampling*. American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf (accessed July 2019).

Blom, A.G., D. Ackermann-Piek, S.C. Helmschrott, C. Cornesse, and J.W. Sakshaug. 2017. "The Representativeness of Online Panels: Coverage, Sampling and Weighting." *Paper Presented at the General Online Research Conference*.

Blom, A.G., C. Gathmann, and U. Krieger. 2015. "Setting Up an Online Panel Representative of the General Population: The German Internet Panel." *Field Methods* 27(4): 391–408. Doi: https://doi.org/10.1177/1525822X15574494.

Blom, A.G., J.M.E. Herzing, C. Cornesse, J.W. Sakshaug, U. Krieger, and D. Bossert. 2016a. "Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German

Internet Panel." *Social Science Computer Review* 35(4): 498–520. Doi: https://doi.org/10.1177/0894439316651584.

Blom, A.G., M. Bosnjak, A. Cornilleau, A.-S. Cousteaux, M. Das, S. Douhou and U. Krieger. 2016b. "A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe." *Social Science Computer Review* 35(1): 8–25. Doi: https://doi.org/10.1177/0894439315574825.

Bosnjak, M., T. Dannwolf, T. Enderle, I. Schaurer, B. Struminskaya, A. Tanner, and K.W. Weyandt. 2017. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel." *Social Science Computer Review* 36(1): 103–115. Doi: https://doi.org/10.1177/0894439317697949.

Briggs, D., D. Fecht, and K. De Hoogh. 2007. "Census Data Issues for Epidemiology and Health Risk Assessment: Experiences from the Small Area Health Statistics Unit." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170(2): 355–378. Doi: https://doi.org/10.1111/j.1467-985X.2006.00467.x.

Cacioppo, J.T. and R.E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42(1): 116. Doi: https://doi.org/10.1037/0022-3514.42.1.116.

Callegaro, M., A. Villar, J. Krosnick, and D. Yeager. 2014. "A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels." In *Online Panel Research. A Data Quality Perspective*, edited by M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Goeritz, J.A. Krosnick, and P.J. Lavrakas, 23–53. Chichester, UK: John Wiley & Sons. Doi: https://doi.org/10.1002/9781118763520.ch2.

Chang, L. and J.A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4): 641–678. Doi: https://doi.org/10.1093/poq/nfp075.

Digman, J.M. 1990. "Personality Structure: Emergence of the Five-factor Model." *Annual Review of Psychology* 41(1): 417–440. Doi: https://doi.org/10.1146/annurev.ps.41.020190.002221.

DiSogra, C., C. Cobb, E. Chan, and J. Dennis. 2012. "Using Probability-Based Online Samples to Calibrate Non-Probability Opt-in Samples." *Presentation at: 67th Annual Conference of the American Association for Public Opinion Research (AAPOR)*. Available at: http://www.websm.org/uploadi/editor/1361444163DiSogra_et_al_2012_Using_Probability_Based_Online_Samples.ppt (accessed July 2019).

Dutwin, D. and T.D. Buskirk. 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. Doi: https://doi.org/10.1093/poq/nfw061.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, 1–26. Doi: https://doi.org/10.1007/978-1-4612-4380-9_41.

Elliott, M.N. and A. Haviland. 2007. "Use of a Web-based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33(2): 211–215. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10498-eng.pdf?st=A8NHMZ2v (accessed July 2019).

Elliott, M.R. 2013. "Combining Data from Probability and Non-probability Samples Using Pseudo-weights." *Survey Practice* 2(6). Doi: https://doi.org/10.29115/SP-2009-0025.

Erens, B., S. Burkill, M.P. Couper, F. Conrad, S. Clifton, C. Tanton, A. Phelps, J. Datta, C.H. Mercer, P. Sonnenberg, et al. 2014. "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey." *Journal of Medical Internet Research* 16(12). Doi: https://doi.org/10.2196/jmir.3382.

Fahimi, M., F.M. Barlas, W. Gross, and R.K. Thomas. 2014. "Towards a New Math for Nonprobability Sampling Alternatives." *Presented at the 69th Annual Conference of the American Association for Public Opinion Research (AAPOR)*.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2013. *Bayesian Data Analysis*, Third Edition. Boca Raton, FL, USA: Chapman & Hall/CRC. ISBN: 9781439840955.

Gelman, A., S. Goel, D. Rothschild, and W. Wang. 2016. "High-frequency Polling with Non-representative Data." In Political Communication in Real Time: Theoretical and Applied Research Approaches (eds. D. Schill, R. Kirk, and A.E. Jasperson). Routledge, 117–133.

Goldberg, L.R. 1993. "The Structure of Phenotypic Personality Traits." *American Psychologist* 48(1): 26. Doi: https://doi.org/10.1037/0003-066X.48.1.26.

Herzing, J.M.E. and A.G. Blom. 2019. "The Influence of a Person's IT Literacy on Unit Nonresponse and Attrition in an Online Panel." *Social Science Computer Review* 37(3): 404–424. Doi: https://doi.org/10.1177/0894439318774758.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. Evaluating Online Nonprobability Surveys. Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics, Pew Research Center. Available at: http://www.pewresearch.org/2016/05/02/evaluatingonline-nonprobability-surveys/ (accessed July 2019).

Lee, S. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2): 329. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/propensity-score-adjustment-as-a-weighting-scheme-for-volunteer-panel-web-surveys.pdf (accessed July 2019).

Lee, S. and R. Valliant. 2009. "Estimation for Volunteer Panel Web Surveys using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods & Research* 37(3): 319–343. Doi: https://doi.org/10.1177/0049124108329643.

MacInnis, G., J.A. Krosnick, S. Ho, and M.J. Cho. 2018. "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension." *Public Opinion Quarterly*. Volume 82, Issue 4, 707–744. Doi: https://doi.org/10.1093/poq/nfy038.

Malhotra, N. and J.A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis*, 286–323. Doi: https://doi.org/10.1093/pan/mpm003.

Marchetti, S., C. Giusti, and M. Pratesi. 2016. "The Use of Twitter Data to Improve Small Area Estimates of Households' Share of Food Consumption Expenditure in Italy." *AStA Wirtschafts-und Sozialstatistisches Archiv* 10(2–3): 79–93. Doi: https://doi.org/10.1007/s11943-016-0190-4.

Mercer, A.W., F. Kreuter, S. Keeter, and E.A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." *Public Opinion Quarterly* 81(S1): 250–271. Doi: https://doi.org/10.1093/poq/nfw060.

Pasek, J. 2016. "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28(2): 269–291. Doi: https://doi.org/10.1093/ijpor/edv016.

Pennay, D.W., D. Neiger, P.J. Lavrakas, K.A. Borg, S. Mission, and N. Honey. 2018. "The Online Panels Benchmarking Study: a Total Survey Error Comparison of Findings from Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia." *Australian National University, Centre for Social Research and Methods Paper NO. 2/2018*. Available at: http://csrm.cass.anu.edu.au/sites/default/files/docs/2018/12/CSRM_MP2_2018_ONLINE_PANELS.pdf (accessed July 2019).

Porter, A.T., S.H. Holan, C.K. Wikle, and N. Cressie. 2014. "Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates." *Spatial Statistics* 10: 27–42. Doi: https://doi.org/10.1016/j.spasta.2014.07.001.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/ (accessed July 2019).

Rao, J.N. 2003. *Small-Area Estimation*. Wiley Online Library. Doi: https://doi.org/10.1002/0471722189.

Rivers, D. 2007. "Sampling for Web Surveys." Presented at the *Joint Statistical Meetings*. Available at: https://pdfs.semanticscholar.org/fffa/a7e52c5d163a0944974a68160ee6e0a6b481.pdf (accessed July 2019).

Rivers, D. and D. Bailey. 2009. "Inference from Matched Samples in the 2008 US National Elections." In *Proceedings of the Joint Statistical Meetings*, Volume 1, 627–639. Palo Alto, CA: YouGov/Polimetrix. Available at: https://pdfs.semanticscholar.org/e566/fb48f88ae34640b729387cbd4006249f8c45.pdf (accessed July 2019).

Schmertmann, C.P., S.M. Cavenaghi, R.M. Assunção, and J.E. Potter. 2013. "Bayes Plus Brass: Estimating Total Fertility for Many Small Areas from Sparse Census Data." *Population Studies* 67(3): 255–273. Doi: https://doi.org/10.1080/00324728.2013.795602.

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2007. OpenBUGS user manual, version 3.0.2. *MRC Biostatistics Unit, Cambridge*.

Sturtz, S., U. Ligges, A. Gelman, et al. 2005. "R2WinBUGS: A Package for Running WinBUGS from R." *Journal of Statistical Software* 12(3): 1–16. Doi: https://doi.org/10.18637/jss.v012.i03.

Tourangeau, R. and T. Plewes. 2013. *Nonresponse in Social Science Surveys: A Research Agenda*. National Academies Press. Doi: https://doi.org/10.17226/18293.

Valliant, R. and J.A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys." *Sociological Methods & Research* 40(1): 105–137. Doi: https://doi.org/10.1177/0049124110392533.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. "Forecasting Elections with Non-representative Polls." *International Journal of Forecasting* 31(3): 980–991. Doi: https://doi.org/10.1016/j.ijforecast.2014.06.001.

Yeager, D.S., J.A. Krosnick, L. Chang, H.S. Javitz, M.S. Levendusky, A. Simpser, and R. Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-probability Samples." *Public Opinion Quarterly* 75(1): 709–747. Doi: https://doi.org/10.1093/poq/nfr020.

Received November 2018
Revised February 2019
Accepted April 2019