**Indices of Non-Ignorable Selection Bias for Proportions**

**Estimated from Non-Probability Samples**

Rebecca R. Andridge

Brady T. West

Roderick J.A. Little

Philip S. Boonstra

Fernanda Alvarado-Leiton

**ABSTRACT**

The proportion of individuals in a finite target population that has some characteristic of interest is arguably the most commonly estimated descriptive parameter in survey research. Unfortunately, rising costs of survey data collection and declining response rates have caused researchers to turn to non-probability samples to make descriptive statements about populations. However, unlike probability samples, non-probability samples may produce severely biased descriptive estimates due to selection bias. This paper develops and evaluates a simple model-based index of the potential selection bias in estimates of population proportions due to non-ignorable selection mechanisms. The index depends on an inestimable parameter ranging from 0 to 1 that captures the amount of deviation from selection at random and is thus well-suited to a sensitivity analysis. We describe maximum likelihood and Bayesian estimation approaches and provide new and easy-to-use R functions for their implementation. We use simulation studies to evaluate the ability of the proposed index to reflect selection bias in non-probability samples and show how the index outperforms a previously proposed index that relies on an underlying normality assumption (Little et al., 2018). We demonstrate the use of the index in practice with real data from the National Survey of Family Growth.

**INTRODUCTION**

Probability sampling and corresponding design-based approaches to finite population inference provide survey researchers with a mathematical basis for making unbiased inferential statements about specific features of target populations. Arguably the most common descriptive quantity used by survey researchers to describe finite populations is a proportion, which quantifies the fraction of units in a target population that has some characteristic of interest (e.g., what proportion of female adults in the United States has ever had a major depressive episode?). Given information about the probabilities of selection for individuals in a probability sample, a binary (yes/no) survey measure of interest, and any additional information necessary to make representative population inferences (e.g., nonresponse adjustments, complex sample design features such as sampling stratum codes, replicate weights, etc.), a survey researcher can compute an unbiased estimate of a proportion and quantify the sampling variance associated with the estimate of that proportion. The random selection of elements from a population of interest into a probability sample, where all population elements have a known non-zero probability of selection, ensures (in theory) that elements included in the sample mirror the population in expectation: that is, for all variables of interest, the mechanism of selection of a subset of elements into the sample is *ignorable*, following the theoretical framework for missing data mechanisms originally introduced by Rubin (1976).

Unfortunately, the effectiveness of probability sampling for studies with these descriptive objectives has been declining in the modern survey research environment. Contacting sampled units has become quite difficult, survey response rates continue to decline in all modes of administration (face-to-face, telephone, etc.) (Brick and Williams, 2013), and the costs of collecting and maintaining probability samples are steadily rising (Presser and McCulloch, 2011). In an environment characterized by steadily declining response rates, the possibility arises that individuals who actually respond to a survey request constitute a biased selection of all individuals initially sampled, eliminating the notion that the mechanism ultimately producing the selection of survey respondents is entirely ignorable. In this case, there may be non-ignorable selection bias in survey estimates computed based on probability samples, due to non-ignorable nonresponse mechanisms.

Because of these issues with probability sampling, survey researchers are starting to turn to the “big data” generated by inexpensive non-probability samples of population elements (Wang et al., 2015; Shlomo and Goldstein, 2015; Miller et al., 2010; Bowen et al., 2007; Brooks-Pollock et al., 2011; Braithwaite et al., 2003; Eysenbach and Wyatt, 2002). These “infodemiology” data might be scraped from social media platforms such as Twitter (e.g., Myslín et al,. 2013; Nascimento et al., 2014; Reavley and Pilkington, 2014; McCormick et al., 2015; Nwosu et al., 2015), or collected from other sources such as commercial databases, online searches (Shlomo and Goldstein, 2015; DiGrazia, 2015), and online surveys (e.g., Evans et al., 2007; Brooks-Pollock et al., 2011; Heiervang and Goodman, 2011). Several researchers have started to use these data sources to estimate the prevalence of health problems in larger populations (e.g., Zhang et al., 2013; Myslín et al., 2013; Evans et al., 2007; Koh and Ross, 2006). However, these are ultimately non-probability samples, and researchers cannot employ inferential methods that assume probability sampling to make unbiased population inferences under the assumption of ignorable sample selection (Pasek and Krosnick, 2011; Yeager et al., 2011). The field would thus stand to benefit from the development of sound measures of the extent to which estimates of proportions from a given non-probability sample are affected by non-ignorable selection bias.

This paper builds on recent work by Little et al. (2018) and Andridge and Little (2011; 2018) to develop measures of non-ignorable selection bias for estimates of population proportions computed from non-probability samples. Little et al. (2018) proposed and evaluated model-based indices of non-ignorable selection bias for descriptive quantities, but these indices rely on an underlying normal pattern-mixture model for the variables of interest. While Little et al. (2018) showed that these indices performed reasonably well when used to assess the selection bias in estimates of proportions based on binary variables, the indices had much better performance for means based on continuous variables, as would be expected given the underlying model. Andridge and Little (2018) have developed estimators of proportions based on proxy pattern-mixture models with improved inferential properties in the presence of non-ignorable survey nonresponse, and this work has important implications for the estimation of proportions in the non-probability sampling context. We aim to leverage these recent developments and overcome the limitations of the indices proposed by Little et al. (2018), developing improved indices of potential non-ignorable selection bias for survey estimates of population proportions computed from non-probability samples.

**BACKGROUND: NON-IGNORABLE SAMPLE SELECTION**

Rubin (1976) was the first to define joint models for the data and the missingness mechanism, and defined sufficient conditions under which the missingness mechanism can be ignored, for likelihood and frequentist inference. This framework has subsequently been applied to sample selection, with the indicator for response being replaced by the indicator for selection into the sample (Rubin, 1978; Little, 2003).

Following Little et al. (2018), let be survey data for each unit in the population, where could be a vector. Let *Z* be a set of fully observed auxiliary or design variables, and let the sample inclusion indicators, , take the values = 1 if the unit *i* is included in the sample and 0 otherwise. Finally, based on these selection indicators we partition *Y* into , where for units in the sample (i.e., with = 1) and for units not in the sample. Without loss of generality, we assume that the survey data *Y* are binary (1, 0) indicators of having a particular characteristic.

Under a model-based (Bayesian) framework, we assume a model for the joint distribution of *Y* and *S* conditional on *Z* (Little, 2003). This joint distribution is factored into the marginal distribution of *Y* and the conditional distribution of *S* given *Y* (the selection mechanism) as

|  |  |
| --- | --- |
| . | (1) |

The density for *Y* given *Z* is indexed by unknown parameters and the density for *S* given *Y* and *Z* is indexed by unknown parameters . The full likelihood based on the observed data (*Z* and *S* for all units and *Y* for units selected into the sample only) is then given by

|  |  |
| --- | --- |
| . | (2) |

Letting be a prior distribution for the parameters, the corresponding posterior distributions for , and are:

|  |  |
| --- | --- |
|  | (3) |

Instead of specifying the selection mechanism for a non-probability sample, which is generally infeasible, it is common instead to make a so-called *ignorability assumption*, the appropriateness of which depends upon two conditions being satisfied: *Selection at Random (SAR)* and *Bayesian Distinctness*. Selection at random means that *S* and *Yexc* are independent after conditioning on *Yinc* and *Z*, i.e., for all . Bayesian distinctness means that and have independent prior distributions, i.e., . These conditions together imply that:

|  |  |
| --- | --- |
| . | (4) |

Thus, when the ignorability assumption is correct, the model for the selection mechanism (the model for *S*) does not affect inferences about the parameters .

Probability sampling is a special form of SAR, where the selection mechanism is known and may depend on auxiliary variables *Z* but not on the survey outcomes *Y*. Thus, reduces to . Probability sampling is stronger than SAR in three important respects. First, under complete response it is automatically valid, and not an assumption, if probability sampling is used to select the sample. Second, it implies that, conditional on *Z*, inclusion in the sample is independent of *Y*, and also any other unobserved variables that might be included in a model, such as latent variables in a factor analysis. Third, it implies that *S* is independent of *Yinc*, whereas SAR only requires the weaker assumption that *S* and *Yexc* are independent after conditioning on *Yinc* and *Z*. These properties make probability sampling highly desirable, but in the current survey research environment it is not always attainable.

When researchers make inferences based on a non-probability sample, they often are (perhaps implicitly) making an assumption of selection at random. However, although the SAR assumption is weaker than probability sampling, it may not always be valid for selection into a non-probability sample. The model-based indices of non-ignorable selection bias of Little et al. (2018) were designed to quantify the potential for selection bias in estimates derived from continuous survey variables. These indices use SAR as a starting point and quantify changes in estimates of the mean of *Y* if the SAR assumption does not hold (to varying degrees). In this paper we modify these indices to be specifically applicable to proportions, producing an index of potential selection bias in estimates of population proportions due to non-ignorable selection mechanisms.

**INDICES OF NON-IGNORABLE SELECTION BIAS FOR PROPORTIONS**

Let *Y* be a binary variable of interest, taking a value of 1 if a survey respondent has a characteristic of interest, and 0 otherwise. An observed value of *Y* for a survey respondent arises from an underlying latent variable *U*, which follows some continuous distribution. We assume that *Y* = 1 when *U* > 0, *U* follows a normal distribution, and that *Y* is only available for cases selected in the non-probability sample. Let *X* be a proxy variable available for all units in the target population that has a reasonably strong correlation with the latent variable *U*. Alternatively, *X* may itself be a function of a vector of auxiliary variables *Z*, as in Andridge and Little (2018). In this case, Zmust be available for all units in the non-probability sample, and either sufficient statistics (means, variances, and covariances) or microdata for the variables in *Z* are available for the non-selected units. Then, *X* is assumed to be the linear predictor from a probit regression of *Y* on *Z* fit to the non-probability sample, and the means and variances of *X* for the non-selected cases, which as we will show are sufficient for calculating our bias indices, are calculated from this estimated probit regression model. As previously defined, let *S* be an indicator of being selected for the non-probability sample. Finally, let *V* be a set of other covariates that are independent of *Y* and *X* for selected units but that may be related to selection (i.e. associated with *S*).

We assume a proxy pattern-mixture (PPM) model (Andridge and Little, 2011; 2018) for *U* and *X*, conditional on *V*, defined as follows:

|  |  |
| --- | --- |
|  | (5) |

where is the intercept, the coefficient of *V*, and the residual variance in the regression of *U* on *V* for pattern *j*. In order to identify this pattern-mixture model, we assume that selection into the sample is a function of the additional covariates *V* and a linear combination of *X* and *U*:

|  |  |
| --- | --- |
|  | (6) |

Here is the proxy *X* rescaled to have the same variance as *U* in the population of selected cases, and is a sensitivity parameter, which we assume to be between 0 and 1 (inclusive). If we also assume that *V* is uncorrelated with *X* for non-selected cases (*S* = 0) and that *X* is the best predictor of *U* for non-selected cases, then this pattern mixture model reduces to:

|  |  |
| --- | --- |
| . | (7) |

For the proof, see the **Appendix**. Without loss of generality, we set . We note that , the correlation between latent *U* and *X* for selected (*j* = 1) and non-selected (*j* = 0) samples, is the *biserial correlation* of *X* and *Y* for pattern *j* (Tate, 1955). Of primary interest is the marginal mean of *Y*, which can be expressed as a function of the pattern-mixture model:

|  |  |
| --- | --- |
|  | (8) |

where  denotes the CDF of the standard normal distribution, evaluated at *z*.

The parameters in the probit PPM model in (7) for the non-selected units (*S* = 0), , , and , are just identifiable given the assumption about the selection mechanism given in (6). Following Little et al. (2018), the parameter in the selection mechanism provides a measure of the degree of non-random selection after conditioning on *X*. If = 0, the probability of being selected in the non-probability sample depends only on *X* and *V*, and thus selection is at random (SAR) since both are fully observed. On the other hand, if = 1, the probability of being selected in the non-probability sample depends on the value of the latent variable *U* (and thus the binary variable of interest, *Y*) and on *V*, and thus selection is not at random. As described in Andridge and Little (2011; 2018), the function *g* does not have to be specified in order for estimates based on this model to be valid.

Given these restrictions, Andridge and Little (2018) show that the unidentified parameters and for a specific choice of  are given by

|  |  |
| --- | --- |
|  | (9) |

It follows that a Measure of the Selection Bias for a Proportion, MUBP(), associated with the estimate of the proportion computed using the non-probability sample () and for a given choice of , would be defined as

|  |  |
| --- | --- |
| . | (10) |

Calculation of the index in (10) for a given choice of therefore requires knowledge of the overall fraction of the population included in the non-probability sample, , which will generally be close to zero for larger populations; the estimated biserial correlation of *X* and *Y* based on the selected non-probability sample, ; and sufficient statistics for the proxy variable *X* for both the selected and non-selected portions of the target population. We note that this last piece is a stronger requirement than the indices for continuous *Y* in Little et al. (2018), where only the mean of *X* was required and not its variance. Maximum likelihood estimates of these sufficient statistics for the selected cases can easily be computed using the selected cases in the non-probability sample. For purposes of this study, we estimate using the “two-step” approach, originally proposed by Olsson et al. (1982) and evaluated empirically by Andridge and Little (2018) as being better than competing estimation approaches when *X* is not normally distributed. A desirable consequence of the two-step approach is that the mean of the latent variable *U* in the selected sample is given by , i.e., the mean of *Y* in the selected sample; alternative estimation approaches for biserial correlation coefficients do not have this property.

If *X* is computed as the linear predictor from a probit model, steps should be taken to prevent overly optimistic estimation of based on potential over-fitting of the probit model to the data from the non-probability sample. In this case, we generally recommend a conservative approach to computing based on multi-fold cross-validation. To do this, the probit model would be fit to randomly selected subsamples of the non-probability sample, and the value of *X* for each observation (whether or not it is in the subsample) calculated from each fitted model. Averaging the set of *X* values for each case produces a single *X* value for each observation; this cross-validated *X* should then be used to compute . The R functions provided in the supplementary materials and available at <https://github.com/bradytwest/IndicesOfNISB> include a function (cv.glm) implementing this cross-validation step, the output of which can then be passed to another function used for two-step estimation of the biserial correlation.

Estimates of the sufficient statistics for *X* for the non-selected sample may be less readily available, but assuming a negligible sampling fraction, reasonable estimates based on the large number of non-selected cases in the target population could be computed from large prior data collections, such as population censuses or large surveys with very small sampling error that also collected measures of *X*. If *X* is the linear predictor from a probit regression of *Y* on *Z* in the non-probability sample, the mean of *X* could be computed by applying the same probit model coefficients estimated from the non-probability sample to overall population means on the auxiliary variables in the vector *Z*. In the presence of a non-negligible sampling fraction, and given an overall marginal population mean for *X* (denoted ), the mean of *X* for non-selected cases could be approximated as . The variance of *X* for non-selected cases could be assumed to be the same as the population variance (in the absence of any additional information on changes in the element variance depending on selection).

We note that there are special cases of MUBP(). When = 0, for instance, selection into the non-probability sample is occurring at random (SAR). In this case the selection mechanism is ignorable and the bias is given by the following simplified expression:

|  |  |
| --- | --- |
|  | (11) |

When = 1, the non-ignorable selection mechanism depends entirely on *U* and *V*, but not on the proxy *X*. In this extreme case, we have

|  |  |
| --- | --- |
|  | (12) |

Following Little et al. (2018), we recommend computing the interval defined by [MUBP(0), MUBP(1)] to assess the range of potential selection bias values, depending on the choice of . As a compromise between the two extreme cases defining this interval, we recommend the computation of MUBP(0.5) as a maximum likelihood “estimate” of the bias, as this choice represents equal dependence of selection on the proxy *X* and the underlying latent value of the variable of interest *U*.

We also note that the MUBP index is not always monotonic in over the [0,1] range. This property of the MUBP index depends on the estimated values of and (i.e., the mean of *Y* and the strength of the proxy in the selected sample) and how far apart the means and variances of the proxy variable *X* are for the selected and non-selected cases. Letting the standardized differences in the selected and non-selected means and variances of *X* be denoted and , then MUBP will be non-monotonic over the [0,1] interval if and only if

This condition will be satisfied when there are extreme differences between *X* in the selected and non-selected populations, there are large differences in the variance of *X* for selected and non-selected cases, and/or weak correlation between *U* and *X*. If we assume that the proxy variances are equal for the selected and non-selected cases, as was suggested in the absence of information about the variance of *X* for the non-selected cases, then = 0, and MUBP is automatically monotone over the [0,1] interval.

Regardless of the choice of the index, at least a moderate biserial correlation between *Y* and *X* is important for the indices to effectively indicate selection bias. In the normal case, Little et al. (2018) heuristically suggested that a correlation of 0.4 or above would increase the ability of these indices to detect non-ignorable selection bias. If there is little or no information about the variable of interest in the proxy variable *X*, and selection is in fact non-ignorable, no known indices of selection bias (including those proposed here) will provide any useful information. In that scenario, the [MUBP(0), MUBP(1)] will be very wide, even sometimes reaching the bounds created by assuming all non-selected cases have *Y* = 0 (or all have *Y* = 1).

*Bayesian Inference*

We also consider a Bayesian approach to making inference about the MUBP index, accounting for uncertainty in the estimation of the coefficients of *Z* in the probit regression of *Y* on *Z* when forming the proxy variable *X*. We follow the Gibbs sampler approach outlined in Section 4.2 of Andridge and Little (2018), which like the maximum likelihood estimates described earlier requires the availability of sufficient statistics for *Z* for the selected and non-selected cases. One could follow two possible approaches here:

1. Proceed with the Gibbs sampler (see below for details) for all other parameters used to compute the MUBP index for a *given* value of , assuming non-informative prior distributions for the identified parameters. This might be done, for example, to account for uncertainty in the estimation of the parameters used to compute the MUBP index for a given. So, when computing the proposed interval [MUBP(0), MUBP(1)] one could form 95% highest posterior density (HPD) credible intervals for both MUBP(0) and MUBP(1), enabling a description of the uncertainty in each “limit” of the interval.
2. Given that there is no information about in the data, instead of fixing one could first randomly draw values of  from a UNIFORM(0,1) distribution, and then proceed with the Gibbs sampler as described in option 1) above.

To initiate the Gibbs sampler, we first fit the probit regression model to the data on *Y* and *Z* from the cases selected for the non-probability sample, which yields starting values for the regression coefficients in this model. We then create the proxy variable *X* as a function of the coefficients. An iteration of the sampler (conditional on either a random draw of or a fixed choice of) then starts with draws of the latent variable *U* from a truncated normal distribution conditional on *X* (and thus also conditional on the probit model coefficients). We then select posterior draws of the regression coefficients in the probit model given the previous augmented values for *U*, and recreate the proxy variable *X* given the current draws of the regression coefficients. This data augmentation approach in each iteration then enables posterior draws of the pattern-mixture model parameters defined in (7) and (9), following the explicit steps and constraints outlined in Andridge and Little (2011). We then generate the corresponding posterior draw of MUBP() in (10) based on the parameter draws. The Gibbs sampler then proceeds to the next iteration. Given a large number of draws of MUBP() we can then generate 95% HPD credible intervals for MUBP().

**SIMULATION STUDIES**

We now describe a simulation study designed to illustrate the ability of MUBP() to detect selection bias in estimated proportions based on simulated data and to show what can go wrong when applying the normal model of Little et al. (2018). All simulations and data analysis were performed using the R statistical computing environment (R Core Team 2018), and the code used is available upon request.

*Data Generation*

We used a latent variable framework to generate populations of size 10,000 containing a binary outcome variable *Y* and a single continuous auxiliary variable *Z* as follows. A single auxiliary variable *zi ~ N(0,1)* was generated for all units. Then for each of , a latent variable *ui* was generated as with . Then an observed binary variable *yi* was created as *yi* = 1 if *ui* > 0 and *yi* = 0 otherwise . Note that, for this simulation, is the biserial correlation between *Y* and the proxy for the entire population, not for the selected sample. In this simulation *Z* was univariate, and thus *Corr*(*U*,*X*) = *Corr*(*U*,*Z*), but more generally *Z* could be a set of auxiliary variables and *X* the linear predictor from a probit regression of *Y* on *Z* for selected cases as described earlier. We set so that . In order to assess how the indices performed for proportions of different magnitude, we simulated data using .

The sample selection indicator *Si* was generated according to a logistic model,

and values of *yi* were deleted for non-selected units, i.e., when *si* = 0. We simulated a wide range of selection bias mechanisms by varying the values of , as shown in **Table 1**, with the value of chosen to result in a 5% sampling fraction. These mechanisms ranged from depending entirely on *Z* to depending entirely on *U*, with intermediate mechanisms that depended either more heavily on *Z* than *U*, or vice-versa, or equally on *Z* and *U*. Within each mechanism, varying the values of and served the purpose of changing the magnitude of the selection bias, with larger values of or leading to larger bias. We note that the resulting bias in the selected mean varied not only by selection mechanism, but was also a function of and . Once *ui* was used for data generation and sample selection, it was discarded.

The process of generating {*zi*, *ui*, *yi*, *si*} was repeated 1,000 times for each combination of , , and . For each simulated dataset, we calculated the indices MUBP(0), MUBP(0.5), and MUBP(1) as defined in (10) through (12), using a probit model of *Y* on *Z* (for selected cases) to estimate the proxy *X* (for all cases). We used the two-step estimator to obtain an estimate of the biserial correlation among the selected cases without cross-validation, since in this controlled simulation setting there was only one auxiliary variable *Z*. We also computed credible intervals by implementing the fully Bayesian approach for the MUBP, with draws of from a Uniform(0,1) distribution, 20 burn-in draws of the Gibbs sampler, and 1,000 subsequent iterations. For comparison, we also calculated indices proposed by Little et al. (2018). Since the outcome is binary, we elected to calculate their measure of unadjusted bias, MUB(), instead of their standardized measure of unadjusted bias, SMUB(), so that it would be more directly comparable to the MUBP(). We also calculated credible intervals for the SMUB() using a uniform prior for . For both MUBP and MUB/SMUB indices, we used sufficient statistics for the auxiliary variable *Z* for the non-selected cases when calculating the indices, though with a 5% sampling fraction, results would likely not differ much if sufficient statistics for the entire population were used.

To assess performance of the indices, we calculated the correlation of each index with the true estimated bias for each simulated dataset, defined as the population mean of *Y* minus the mean of *Y* for the selected cases. We also assessed the ability of the intervals [MUBP(0), MUBP(1)] and [MUB(0), MUB(1)] to cover the true estimated bias, as well as the coverage of the Bayesian intervals for MUBP and SMUB; coverage for SMUB was based on the standardized true estimated bias, i.e., the true estimated bias divided by the population standard deviation.

**Results**

To help visualize the differences in the two models, **Figure 1** shows intervals for the selection bias indices for the two models for a single simulated data set from each combination of proxy strength and selection model, for the scenarios with E[*Y*]=0.3 and the middle set of for the selection model. In all cases, the [MUB(0), MUB(1)] intervals based on the normal model are substantially wider than both the [MUBP(0), MUBP(1)] and Bayesian intervals based on the probit model. This exaggeration is most pronounced for selection mechanisms that depend more on *Z* than on *U* and for weaker proxies (lower correlations). All sets of intervals cover the true bias in most cases, with the Bayesian intervals doing slightly better.

The median estimated index values across replicates for MUBP() and MUB() for ={0,0.5,1} are shown in **Figure 2**, for the scenarios with E[*Y*]=0.3. For all selection mechanisms and correlations between the proxy and the outcome, both sets of indices “track” with the estimated bias; as the estimated bias goes up, so does the index. When selection is a function of *Z* only, both MUBP(0) and MUB(0) produce unbiased estimates of bias for all proxy strengths (lines overlap on the plot). When selection is only a function of *U*, MUBP(1) is approximately unbiased and there is a substantial upward bias in MUB(1). More interesting, however, are the intermediate mechanisms, where selection is a function of both *Z* and *U*. In these cases, the intervals [MUBP(0), MUBP(1)] and [MUB(0), MUB(1)] cover the truth, with =0.5 coming closest to the truth most of the time. However, the interval widths are much wider for the normal model (MUB) than for the probit model (MUBP), even when the proxy variable is highly correlated with the outcome. Interestingly, the intervals based on the normal model are more exaggerated when selection depends more heavily on *Z*, the fully observed variable. Importantly, for weaker proxies (lower correlations), the normal model intervals have an implausible bound for =1, i.e. produce estimates of E[*Y*] that are outside the (0,1) interval, whereas the probit model intervals bound at the upper limit (i.e. E[*Y*]=1). In **Figure 2**, the hitting of the upper bound can be seen by the curving of the solid MUBP(1) line for selection based on *Z* and a weak proxy. In practice, one does not know the true selection mechanism, but using the probit model will give tighter intervals and produce index values that more closely reflect the bias, with both strong and weak proxies. Similar patterns are seen with E[*Y*]=0.1 and E[*Y*]=0.5 (**Supplemental Figures 1 and 2**).

Not surprisingly, all indices have higher correlation with the true estimated bias for stronger proxies than for weaker proxies, as shown in **Figure 3**. Generally, the patterns of correlations seen are similar across selection mechanisms, though there is more separation between the models (probit versus normal) for selection mechanisms that have larger dependence on *Z*. For rare outcomes (E[*Y*]= 0.1), the MUBP() index has a higher correlation with the estimated bias than the MUB() index does across all selection mechanisms and proxy strengths. Strikingly, when E[*Y*]=0.1 and the proxy is weak, MUB(1) has essentially zero correlation with the truth, whereas MUBP() has a noticeably higher correlation. This dramatic difference between the two models appears to be reduced when the mean of *Y* nears 0.5; some differences are still seen for E[*Y*]= 0.3, but there are very few differences when E[*Y*]=0.5.

Since the [MUB(0), MUB(1)] intervals are considerably wider than the [MUBP(0), MUBP(1)] intervals across all scenarios, coverage of these normal-based intervals is generally higher than for the probit-based intervals (**Supplemental Figure 3**). At the two extremes of the selection models (based on *Z*, based on *U*), coverage is only around 50% for the probit model regardless of proxy strength, and is often larger for the normal model. This makes sense, since in these cases MUBP(0) and MUBP(1) are actually unbiased estimates, so naturally the interval would only cover the truth about 50% of the time. Additionally, coverage of the probit intervals does not depend on E[*Y*], but coverage for the normal model does. For stronger proxies, coverage is worse for the normal model as E[*Y*] moves away from 0.5, more so for mechanisms that depend more heavily on *Z*. Conversely, for weaker proxies and a nonignorable selection mechanism, coverage is better for lower proportions; this reflects the fact that in these cases the intervals are enormous. The magnitude of the true estimated bias also has an impact on coverage for both models, with coverage improving as the estimated bias increases. The exception is for the normal model with the strongest proxy, smaller E[*Y*], and selection mechanisms that depend more heavily on Z, where coverage actually decreases as true estimated bias increases.

The coverage of the Bayesian intervals is much better for both models (**Supplemental Figure 4**). As with the intervals based on ∈[0,1], coverage does not depend on E[*Y*] for the probit model but does for some scenarios for the normal model, particularly when selection is more heavily dependent on *Z*. The probit model has high coverage for all the intermediate mechanisms (3*Z*+*U*, *Z*+*U*, *Z*+3*U*), but coverage decreases with increasing true bias for the cases with selection based only on *Z* and only on *U*. As with the MLE intervals, this is not surprising; in these cases the endpoints MUBP(0) and MUBP(1) are unbiased, respectively, and thus an interval with a uniform prior on will not have high coverage. Of note, the decrease in coverage is not as severe for the Bayesian intervals as for the MLE intervals (in most cases, it does not go as low as 50%). The normal model has high coverage for all selection mechanisms except selection based only on *Z*. In these cases, coverage decreases drastically with increasing bias. There is also a downward trend in coverage for the strongest proxy with E[*Y*]=0.1 for the intermediate selection mechanisms that depend more heavily on *Z* (3*Z*+*U*, *Z*+*U*).

Overall, the MUBP indices perform well overall across a variety of selection mechanisms. These indices based on the probit model provide a more precise estimate of bias compared to the SMUB/MUB indices based on the normal model and do not return implausible estimates. As was suggested in Little et al. (2018) for the normal-based indices, at least a moderately strong predictor of *Y* is necessary for MUBP to be useful. In the simulation, scenarios with biserial correlations of 0.5 or 0.8 had stronger correlations between the estimated bias and the true bias than scenarios with a biserial correlation of 0.2. Note, however, that the biserial correlation is always greater than the Pearson correlation between *X* and binary *Y*, and how much larger it is depends on the mean of *Y*. In this simulation, the Pearson correlation ranged from 0.12 to 0.64, and a correlation between *Y* and *X* of 0.3 or greater appears to provide reasonable estimates of the selection bias.

**APPLICATION**

We now revisit an analysis of real survey data from the National Survey of Family Growth (NSFG) presented in Little et al. (2018). In this analysis, the authors used the publicly available NSFG sample as a hypothetical population, and took the sub-sample of smartphone users as a hypothetical non-probability sample. They calculated their normal model-based selection bias indices, SMUB(), to evaluate potential selection bias in sample means for a variety of different variables. Importantly, though the pattern-mixture model underlying the SMUB indices is bivariate normal, the SMUB() index was applied to means estimated for a mixture of different types of survey variables, including binary variables. Of the 16 proportions analysed, the [SMUB(0), SMUB(1)] interval only “covered” the actual bias in the smartphone proportions 8 times (given knowledge of the true proportions based on the full NSFG sample). These results suggested that there was room for improvement in the performance of these indices for these binary variables. In the present application, we seek to evaluate the improvement in coverage of actual bias based on the MUBP measures proposed in the present study.

For each of the 16 binary variables in the NSFG data, we initially fitted probit regression models to the data from the smartphone sample, regressing the binary variable *Y* on the same covariates *Z* that were considered by Little et al. (2018). Values of the linear predictor *X* for the underlying variable *U* were then computed for both the selected cases and the non-selected cases, and the five-fold cross-validation approach described earlier was used for two-step estimation of the biserial correlation for each variable. This process was facilitated using the aforementioned R code provided in the supplementary materials. Given these data, we then proceeded to compute the MUBP indices defined in (10) through (12). Because we had access to data on *Y* for the non-selected cases in this hypothetical illustration, we could compute the difference between the proportion in the smartphone sample and that for the full “population” to get an estimate of the “true” bias in the smartphone proportion. We then compared the values of the computed indices to this “true” bias measure.

We also implemented the fully Bayesian inference approach for the MUBP index described earlier, with draws of from a UNIFORM(0,1) distribution, 20 burn-in draws of the Gibbs sampler, and 2,000 subsequent iterations. The resulting simulated samples of the MUBP index were used to form 95% HPD credible intervals for MUBP in the case of each proportion. We then examined whether the credible intervals covered the “true” bias, expecting that coverage may improve from exploitation of the uncertainty in the estimated parameters enabled by the presence of sufficient statistics for *Z* on the non-selected NSFG cases.

**Table 2** compares the previously reported performance of the SMUB indices for these 16 proportions to that of the MUBP indices. Notably, the selection fractions for this hypothetical application were quite different from zero: for variables measured on males, the selection fraction was 0.788 (6,942 smartphone users out of 8,809 males), and for variables measured on females, the selection fraction was 0.817 (8,981 smartphone users out of 10,991 females). **Table 2** also includes the cross-validated “two-step” estimates of the biserial correlations of the proxy variable *X* with the outcome *Y* among the selected cases.

The performance of the proposed intervals based on the MUBP index is generally improved relative to that of the proposed SMUB intervals derived from the underlying normal pattern-mixture model. First and foremost, the MUBP intervals are significantly narrower than the intervals for the same proportions based on the SMUB index, reflecting the sensitivity of the MUBP index (derived from the probit model) to the limited range and discrete nature of the binary survey variables. The proposed intervals therefore provide a more precise sense of the potential selection bias associated with these estimates of the proportions than the estimates based on the SMUB index, and this result holds regardless of the biserial correlation. These results are consistent with our findings in the simulation studies presented earlier.

Second, 10 of the 16 estimated bias values are either directly covered or very nearly covered by the proposed [MUBP(0), MUBP(1)] interval, representing a slight improvement over the SMUB approach that was based on the normal pattern-mixture model. For example, considering the binary indicator of children being present in the household for males, we see that accounting for the uncertainty in the input estimates via the Bayesian approach for the fixed choices of 0 and 1 for would result in coverage of the estimated bias. The results are also quite similar when applying the fully Bayesian approach with Uniform draws for . Furthermore, as was noted by Little et al. (2018), a moderate biserial correlation (say, greater than 0.3) ensures that the proposed interval does a good job of covering the estimated bias; this was true for 9 out of 12 proportions where the biserial correlation was 0.3 or larger in this illustration.

There are several cases where no approach to constructing an interval for MUBP covers the estimated bias, despite the fact that the biserial correlation between *X* and *U* is relatively large (e.g., the indicator of being between the ages of 30 and 44, with biserial correlation = 0.65). Since we had *Y* available for the entire NSFG “population” in this example, we were able to fit a probit regression model to the selection indicator, regressing the indicator of owning a smartphone (“selection”) on both *X* and *Y* to further investigate the “true” selection mechanism. Surprisingly, we found that the estimated coefficient for *X* was positive while the estimated coefficient for *Y* was negative, and thus the probability of being selected into the NSFG smartphone “sample” was a positive function of *X* and a *negative* function of *Y*. In our model, we assume in (6) that the selection mechanism is a function of with restricted to be non-negative, and thus a selection mechanism that depends on *X* and *Y* in opposite directions will not be covered by the [MUBP(0), MUBP(1)] interval.

Little (1994), who defined the probability of non-selection underlying the PMM in (7) as with , suggested that was a plausible value for this mechanism; in this case, selection would depend on the *difference* between *X* and *U*. Following our approach, would imply that . We subsequently computed MUBP() for the age 30-44 indicator for males as an illustration and found that the resulting value was -0.024. Taken together with the MUBP() values in **Table 2**, we find that the interval of [MUBP(), MUBP(1)] for this proportion is [-0.024, 0.039] which does in fact cover the small estimated bias (-0.002). So while this resulting interval is relatively wide, it does allow for the unusual but not implausible possibility that the probability of selection has a positive relationship with the proxy variable *X* and a negative relationship with *U*. Analysts can easily perform this computation [calculating MUBP()] using the R functions available in the supplementary materials to assess the implications of this plausible scenario for potential selection bias. We also note that this scenario is only a problem with strong proxy variables *X*, having a moderate-to-large biserial correlation with *U*. With weak proxies, the proposed interval will basically cover the two extremes -- the selection bias if all non-selected cases were 1s, and the selection bias if all non-selected cases were 0s.

**DISCUSSION**

We have proposed simple model-based indices measuring the potential selection bias in proportions estimated based on non-probability samples, where the selection mechanism underlying the realized non-probability sample may be non-ignorable. These “measure of selection bias” (MUBP) indices are easy to compute using the R functions provided in the supplementary materials and also freely available at <https://github.com/bradytwest/IndicesOfNISB>. Via empirical simulation studies and an application to smartphone users in a real survey setting, we have demonstrated the ability of the MUBP indices to effectively indicate potential selection bias for estimated proportions. Notably, the indices enable sensitivity analyses, allowing users to vary their assumptions about the amount of non-ignorability in the underlying selection mechanism.

The proposed indices also have a dual benefit in that the underlying methodology can be used to make inferences about the estimated proportions based on a non-probability sample. Making inference when following this approach requires means, variances, and covariances for the auxiliary variables *Z* in the non-selected sample that are used to form the auxiliary proxy that is key to the effectiveness of this methodology. While these sufficient statistics (and specifically the variances and covariances) may be difficult to obtain for non-selected cases in practice, one could at least assume that the variances and covariances are similar to those observed for the non-probability sample. In the absence of this information, and given that the auxiliary proxy *X* has a moderately strong (cross-validated) biserial correlation with the binary variable of interest *Y*, one could still use our methodology to identify those estimates at the highest risk of selection bias.

The MUBP indices could also be used during an ongoing data collection to identify estimates that are becoming more and more prone to selection bias as the data collection proceeds. In this sense, the indices could be used to inform adaptive survey designs that prioritize subgroups of cases which are predicted to have unique values on the binary variable of interest that may be under-represented in the responding sample. We feel that future research could focus on this potential utility of the proposed indices to reduce selection bias in a real-time fashion.

There are three key avenues for extending this work in the future. First, the pattern-mixture model here can be extended to estimated proportions for *ordinal* categorical variables (e.g., self-rated health) in a straightforward manner, as outlined in Andridge and Little (2018). In this case there would not be a single MUBP() but a value of MUBP() for each level of the outcome; future work could develop measures that combine these values into one (for each value of ). Another important area of research is whether the MUBP() index can be extended for *multinomial* categorical variables (e.g., political party preference). Finally, the development of measures of selection bias for other estimands besides the population proportion, e.g. for estimated regression coefficients in logistic regression models, is also necessary.

**APPENDIX**

To show that (5) reduces to (7) under the stated assumptions, note first that since *V* is assumed to be uncorrelated with *X* for both selected and non-selected cases, . Secondly, since *X* is assumed to be the best predictor of *Y* for both selected and non-selected cases,

|  |  |
| --- | --- |
|  | (A1) |

for *j* = 0, 1. Thus the model in (5) reduces to

|  |  |
| --- | --- |
|  | (A2) |

which is the pattern-mixture model in (7) with , , , , and .

**ACKNOWLEDGMENTS**

This work was supported by an R21 grant from NIH (PI: West; NIH Grant No. 1R21HD090366-01A1). The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC’s) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan’s Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS nor the other funding agencies.

**Table 1:** Values of (log-odds ratios) that determine the selection mechanism for the simulation study.

|  |  |
| --- | --- |
| **Selection**  **Mechanism** | **Values of** |
| *Z* | {.1, 0}, {.2, 0}, {.3, 0}, {.4, 0}, {.5, 0} |
| *3Z + U* | {.075, .025}, {.15, .05}, {.225, .075}, {.3, .1}, {.375, .125} |
| *Z + U* | {.05, .05}, {.1, .1}, {.15, .15}, {.2, .2}, {.25, .25} |
| *Z + 3U* | {.025, .075}, {.05, .15}, {.075, .225}, {.1, .3}, {.125, .375} |
| *U* | {0, .1}, {0, .2}, {0, .3}, {0, .4}, {0, .5} |

**Table 2:** Computed values of the [SMUB(0), SMUB(1)] intervals based on the normal pattern-mixture model for the 16 NSFG proportions (from Little et al. 2018), in addition to measures of standardized true estimated bias (STEB) for each estimated proportion1, along with the true estimated bias, the [MUBP(0), MUBP(1)] intervals, and 95% highest posterior density (HPD) credible intervals for MUBP based on the fully Bayesian approach for the same 16 proportions based on the probit pattern-mixture model1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Binary NSFG Variable (Males / Females) | Interval Based on Normal Model:  [SMUB(0), SMUB(1)] | STEB | Normal Interval Cover STEB? | Population Proportion (95% CI) | Smartphone Proportion (95% CI) | True Estimated Bias | Cross-Validated  Biserial Correlation (*Y*, *X*) | Interval Based on Probit Model:  [MUBP(0), MUBP(1)], Bayesian CIs for Selected Limits | Probit Interval Cover Bias?2 | 95% Credible Interval for MUBP: Bayesian Approach | Credible Interval Cover Bias?2 |
| Never been married (M) | [-17, -42] | -24 | Y | 0.566 | 0.555 | -11 | 0.817 | [-10, -14] | Y | [-16, -7] | Y |
| Never been married (F) | [-3, -8] | -3 | Y | 0.468 | 0.466 | -2 | 0.726 | [-2, -5] | Y | [-7, 1] | Y |
| Age = 30-44 (M) | [32, 94] | -4 | N | 0.435 | 0.433 | -2 | 0.654 | [16, 39],  [(13,19), (31,46)] | N | [14, 38] | N |
| Age = 30-44 (F) | [17, 59] | -15 | N | 0.467 | 0.460 | -8 | 0.612 | [8, 24],  [(6,11), (17,31)] | N | [7, 24] | N |
| Currently employed (M) | [37, 130] | 86 | Y | 0.689 | 0.729 | 40 | 0.603 | [16, 46] | Y | [16, 45] | Y |
| Children present in HU (M) | [-5, -20] | -12 | Y | 0.371 | 0.366 | -5 | 0.573 | [-2, -4],  [(-5,0), (-15,7)] | C | [-10, 3] | Y |
| Currently employed (F) | [26, 156] | 63 | Y | 0.626 | 0.657 | 31 | 0.482 | [12, 50] | Y | [11, 47] | Y |
| Children present in HU (F) | [-21, -152] | -19 | N | 0.548 | 0.538 | -10 | 0.454 | [-10, -47] | Y | [-10, -45] | Y |
| “Other” race (F) | [23, 172] | 17 | N | 0.553 | 0.562 | 9 | 0.451 | [11, 54],  [(9,13), (42,65)] | C | [9, 51] | Y |
| “Other” race (M) | [30, 276] | 12 | N | 0.590 | 0.596 | 6 | 0.410 | [14, 102],  [(12,17), (78,129)] | N | [11, 85] | N |
| Education: “Some coll.” (M) | [10, 143] | 50 | Y | 0.299 | 0.322 | 23 | 0.368 | [5, 17],  [(3,7), (5,29)] | C | [3, 21] | C |
| Education: “Some coll.” (F) | [3, 47] | 29 | Y | 0.328 | 0.342 | 14 | 0.340 | [2, 16] | Y | [0, 16] | Y |
| Region = "south" (F) | [-7, -130] | 14 | N | 0.438 | 0.445 | 7 | 0.274 | [-3, -36],  [(-5,-2), (-52,-20)] | N | [-34, -2] | N |
| Region = "south" (M) | [-3, -62] | 26 | N | 0.418 | 0.431 | 13 | 0.253 | [-1, -20],  [(-3,1), (-46,10)] | N | [-26, 4] | N |
| Income: $20K-$59,999 (M) | [-6, -146] | 10 | N | 0.417 | 0.422 | 5 | 0.249 | [-3, -123],  [(-5,-1), (-130,-38)] | N | [-120, -1] | N |
| Income: $20K-$59,999 (F) | [1, 70] | 11 | Y | 0.388 | 0.393 | 5 | 0.156 | [1, 72] | Y | [-2, 72] | Y |

1 Values multiplied by 1,000.

2 Y = Yes; C = Close, allowing for uncertainty in the input estimates (see Bayesian CIs for selected limits); N = No

**Figure 1:** Intervals for the selection bias for a single simulated dataset for each combination of proxy strength (rows) and selection model (columns), with E[*Y*]=0.3 and the middle set of for each selection mechanism. Intervals shown are: Normal: [MUB(0), MUB(1)], Probit: [MUBP(0), MUBP(1)], and Bayesian Probit with uniform prior on . Symbol in the middle of each interval is the estimate with = 0.5 for the MLE intervals and the median for the Bayesian intervals. Grey horizontal line is the true estimated bias in each dataset.

****

**Figure 2:** MUBP() from the probit model (solid lines/solid symbols) and MUB() from the normal model (dotted lines/open symbols) versus the true estimated bias, shown for combinations of the biserial correlation *Corr*(*U*, *X*) = (rows) and the selection mechanism (columns), for E[*Y*] = 0.3. Grey dashed line is equality (index = estimated bias). Results are medians across 1000 simulated data sets for each scenario.



**Figure 3:** Correlation between MUBP() and true estimated bias, and between MUB() and true estimated bias, versus the biserial correlation *Corr*(*U*,*X*) =, for combinations of selection mechanism (columns), (rows), and (shape). Results from all estimated biases (all values of and ) are all plotted together. Correlations are estimated from 1000 simulated data sets for each scenario.



**REFERENCES**

Andridge, R.R. and Little, R.J.A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Stats.* 27; 153-180.

Andridge, R.R. and Little, R.J.A. (2018). Proxy pattern-mixture analysis for a binary survey variable subject to nonresponse. Submitted to *J. Off. Stats*.

Bowen, D.J., Bradford, J., and Powers, D. (2007). Comparing Sexual Minority Status across Sampling Methods and Populations. *Women and Health*, 44(2), 121-134.

Braithwaite, D., Emery, J., de Lusignan, S., and Sutton, S. (2003). Using the Internet to Conduct Surveys of Health Professionals: A Valid Alternative? *Family Practice*, 20(5), 545-551.

Brick, J.M. and Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*, 645, 36-59.

Brooks-Pollock, E., Tilston, N., Edmunds, W.J., and Eames, K.T.D. (2011). Using an Online Survey of Healthcare-seeking Behaviour to Estimate the Magnitude and Severity of the 2009 H1N1v Influenza Epidemic in England. *BMC Infectious Diseases*, 11, 68.

Brooks-Pollock, E., Tilston, N., Edmunds, W.J., and Eames, K.T.D. (2011). Using an Online Survey of Healthcare-seeking Behaviour to Estimate the Magnitude and Severity of the 2009 H1N1v Influenza Epidemic in England. *BMC Infectious Diseases*, 11, 68.

DiGrazia, J. (2015). Using Internet Search Data to Produce State-Level Measures: The Case of Tea Party Mobilization. *Sociological Methods and Research*. DOI: 10.1177/0049124115610348.

Evans, A.R., Wiggins, R.D., Mercer, C.H., Bolding, G.J., and Elford, J. (2007). Men Who Have Sex with Men in Great Britain: Comparison of a Self-Selected Internet Sample with a National Probability Sample. *Sexually Transmitted Infections*, 83, 200-205.

Eysenbach, G. and Wyatt, J. (2002). Using the Internet for Surveys and Health Research. *Journal of Medical Internet Research*, 4(2).

Heiervang, E. and Goodman, R. (2011). Advantages and Limitations of Web-Based Surveys: Evidence from a Child Mental Health Survey. *Social and Psychiatric Epidemiology*, 46, 69-76.

Koh, A.S. and Ross, L.K. (2006). Mental Health Issues: A Comparison of Lesbian, Bisexual, and Heterosexual Women. *Journal of Homosexuality*, 51(1), 33-57.

Little, R.J.A. (1994). A Class of Pattern-Mixture Models for Normal Incomplete Data. Biometrika, 81, 471–483.

Little, R.J.A. (2003), “The Bayesian Approach to Sample Survey Inference,” in *Analysis of Survey Data*, eds. R. L. Chambers, and C. J. Skinner, pp. 49-57, Wiley: New York.

Little, R.J.A., West, B.T., Boonstra, P., Hu, J. (2018) Measures of the Degree of Departure from Ignorable Sample Selection. Submitted to *J. Surv. Stat. Meth.*

McCormick, T.H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E.S. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods and Research*. DOI: 10.1177/0049124115605339.

Miller, P.G., Johnston, J., Dunn, M., Fry, C.L., and Degenhardt, L. (2010). Comparing Probability and Non-Probability Sampling Methods in Ecstasy Research: Implications for the Internet as a Research Tool. *Substance Use and Misuse*, 45, 437-450.

Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research*, *15*(8), e174. doi:10.2196/jmir.2534.

Nascimento, T. D., DosSantos, M. F., Danciu, T., DeBoer, M., van Holsbeeck, H., Lucas, S. R., et al. (2014). Real-Time Sharing and Expression of Migraine Headache Suffering on Twitter: A Cross-Sectional Infodemiology Study. *Journal of Medical Internet Research*, *16*(4), e96. doi:10.2196/jmir.3265.

Nwosu, A.C., Debattista, M., Rooney, C., and Mason, S. (2015). Social media and palliative medicine: a retrospective 2-year analysis of global Twitter data to evaluate the use of technology to communicate about issues at the end of life. *BMJ Support Palliat Care*, 5(2), 207-212. doi: 10.1136/bmjspcare-2014-000701.

Olsson, U., Drasgow, F., and Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika*, 47, 337-347.

Pasek, J. and Krosnick, J.A. (2011). Measuring Intent to Participate and Participation in the 2010 Census and Their Correlates and Trends: Comparisons of RDD Telephone and Non-Probability Sample Internet Survey Data. *Statistical Research Division of the U.S. Census Bureau*, 15.

Presser, S. and McCulloch, S. (2011). The Growth of Survey Research in the United States: Government-sponsored Surveys, 1984-2004. *Social Science Research*, 40(4), 1019-1024.

Reavley, N. J., & Pilkington, P. D. (2014). Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, *2*, e647. doi:10.7717/peerj.647.

Rubin, D.B. (1976). Inference and Missing Data (with Discussion). *Biometrika*, 63, 581-592.

Shlomo, N. and Goldstein, H. (2015). Editorial: Big Data in Social Research. *Journal of the Royal Statistical Society, Series A*, 178(4), 787-790.

Tate, R.F. (1955). The Theory of Correlation Between Two Continuous Variables When One is Dichotomized. *Biometrika*, 42, 205–216.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting*, 31(3), 980-991.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., and Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, 75(4), 709-747.

Zhang, N., Campo, S., Janz, K. F., Eckler, P., Yang, J., Snetselaar, L. G., & Signorini, A. (2013). Electronic Word of Mouth on Twitter About Physical Activity in the United States: Exploratory Infodemiology Study. *Journal of Medical Internet Research*, *15*(11), e261. doi:10.2196/jmir.2870.