Hammon, Angelina; Zinn, Sabine

Validating an Index of Selection Bias for Proportions in Non-Probability Samples

International Statistical Review

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Hammon, Angelina; Zinn, Sabine (2024) : Validating an Index of Selection Bias for Proportions in Non-Probability Samples, International Statistical Review, ISSN 1751-5823, Wiley, Hoboken, NJ, Vol. 93, Iss. 3, pp. 499-516, https://doi.org/10.1111/insr.12590

International Statistical Review

International Statistical Review (2025), **93**, 3, 499–516 doi: 10.1111/insr.12590

Validating an Index of Selection Bias for Proportions in Non-Probability Samples

Angelina Hammon[1,2] and **Sabine Zinn**[1,3]

[1]*German Socio-Economic Panel Study Department, Berlin, Germany*
[2]*Chair of Statistics and Econometrics, University of Bamberg, Bamberg, Germany*
[3]*Department of Social Sciences, Humboldt University, Berlin, Germany*
Correspondence *Angelina Hammon, German Socio-Economic Panel Study Department, Berlin, Germany. Email: ahammon@diw.de*

Summary

Fast online surveys without sampling frames are becoming increasingly important in survey research. Their recruitment methods result in non-probability samples. As the mechanism of data generation is always unknown in such samples, the problem of non-ignorability arises making vgeneralisation of calculated statistics to the population of interest highly questionable. Sensitivity analyses provide a valuable tool to deal with non-ignorability. They capture the impact of different sample selection mechanisms on target statistics. In 2019, Andridge and colleagues proposed an index to quantify potential (non-ignorable) selection bias in proportions that combines the effects of different selection mechanisms. In this paper, we validate this index with an artificial non-probability sample generated from a large empirical data set and additionally applied it to proportions estimated from data on current political attitudes arising from a real non-probability sample selected via River sampling. We find a number of conditions that must be met for the index to perform meaningfully. When these requirements are fulfilled, the index shows an overall good performance in both of our applications in detecting and correcting present selection bias in estimated proportions. Thus, it provides a powerful measure for evaluating the robustness of results obtained from non-probability samples.

Key words: non-ignorable sample selection; pattern-mixture modelling; selection not at random; sensitivity analysis.

1 Introduction

For almost a decade now, there has been a boom in companies offering non-probability sampling as an alternative to probability sampling as a cheaper and easier to implement alternative. Examples include sampling via social media such as Facebook, large global companies such as Amazon (Amazon Mechanical Turk), and companies spun off from universities such as Respondi. All samples from these companies have in common that their entities—mostly individuals or companies—self-select into these samples without relying on a frame that clearly defines the population to be studied. This means that the inclusion probabilities of the entities and the peculiarities that constitute the sample generation process cannot be determined. Also, the characteristics of the non-participating entities of a population are unknown. All these features are necessary for computing unbiased population estimates (Kish, 1965). The classical

probability sample is a special case of an ignorable sample selection mechanism (Rubin, 1976), also referred to as *selection at random* (SAR).[1] That is, sampling from a population was either due to simple random sampling or due to a selection mechanism that is completely under the control of the sample designer with all sampling and design features are known. Then, standard design-based methods for inference with survey data (adjusting for issues due to survey sampling, nonreponse, and undercoverage in statistical analysis) yield unbiased population vestimates. However, if the selection mechanism is non-ignorable, that is, *selection not at random* (SNAR) generated the sample at hand, classical survey procedures are not automatically valid anymore and might provide misleading inferences about the population. In the case of non-probability samples SNAR is highly likely because first, the population from which sample entity stems is unknown and second, entities recruit themselves into these samples. In other words, the data generation process of non-probability samples is unknown. Thus, there is no fully observed information that accounts for the process that generated the sample in the statistical analysis (Little *et al.*, 2019; Valliant *et al.*, 2018; Valliant, 2020). Other currently available methods for non-probability samples such as blended calibration (DiSogra *et al.*, 2011; Fahimi *et al.*, 2015) or pseudo-weighting procedures (Elliot, 2009; Elliott & Valliant, 2017) also assume SAR and can yield at most approximate estimates of population statistics. As there is no way testing SAR against SNAR, conducting sensitivity analyses is the only possibility to assess the robustness of analysis results under different assumptions about the selection mechanism.

To get an idea about the magnitude of non-ignorable selection bias in estimated means of non-probability samples, Little *et al.* (2019) and Andridge *et al.* (2019) developed an index based on the proxy pattern-mixture model (PPMM) (Andridge & Little, 2011, 2020). Their proposed index quantifies the robustness of estimates under varying assumptions about the selection mechanisms yielding a comprehensive sensitivity analysis (Andridge *et al.*, 2019). To compute the index, we need a set of auxiliary variables describing the variable of interest that has to be available for all units in the target population. Little *et al.* (2019), Andridge *et al.* (2019) and Boonstra *et al.* (2021) evaluated the index in different simulation studies using artificially generated data as well as empirical data sets and could show a general good performance of the measure. However, for the index to properly detect selection bias, there are some requirements on the non-probability data and auxiliary variables that may not be available in real-world applications, or at least may not be testable.

Thus far, there have been only two very recent applications of the index and its methodology to actual non-probability samples. West & Andridge (2023) applied the index to 2020 polling data, and a current study by Andridge (2024) utilised the index methodology with vaccine data. With our paper, we aim to further narrow this gap in the literature. We are the first to apply this index of non-ignorable selection bias outside its original research group to proportions estimated from data on current political attitudes obtained from a real non-probability sample (which recruits its participants via river sampling on websites).

Before doing so, we validate the proposed index with an artificial non-probability sample using data from the General Social Survey in Germany. From this validation study, we derive concrete guidelines for applying the index in practice.

The paper is structured as follows: First, we give a brief overview of the proxy pattern-mixture model in the context of sample selection. Then, we describe the index for quantifying non-ignorable selection bias. Afterwards, we conduct our validation study and derive the guidelines. The application to the real non-probabilistic data follows. We conclude with a summary, lessons learnt, and open questions where we also point to limitations of the index and of our work.

2 Methods

To assess the degree of non-ignorable selection bias in proportions, Andridge *et al.* (2019) designed an index based on a proxy pattern-mixture model (PPMM). Andridge & Little (2011, 2020) introduced this model class for handling missing data that are supposed to be *missing not at random* (MNAR) (Rubin, 1976). In the following section, we describe and recapitulate the methodology developed in Andridge *et al.* (2019) and Andridge & Little (2020) on which the index calculation is based. This concise summary of the underlying model ideas and assumptions is necessary for our subsequent validation of the index and its evaluation for use in our real-world application.

2.1 The Proxy Pattern-Mixture Model for Binary Data

The PPMM (Andridge & Little, 2011, 2020) is an extension of the bivariate normal pattern-mixture model proposed in Little (1994). However, compared with the latter, the PPMM uses a derived variable X, the so-called 'proxy', that can reflect a set of different explanatory variables Z considered important in explaining Y, rather than just one. This surrogate variable X is defined as the best predictor for Y and can be computed as the linear predictor from a fitted probit regression of Y on Z, including interactions and non-linear terms if appropriate (Andridge & Little, 2011, 2020; Andridge *et al.*, 2019; Little *et al.*, 2019). Thus, in case of the PPMM the joint distribution of the proxy X and the outcome variable Y follows the normal pattern-mixture model discussed in Little (1994) and Andridge & Little (2011, 2020). The PPMM was originally developed in the context of missing data, but can also be applied to non-ignorable sample selection by replacing the original indicator for missingness M by an indicator for selection into the sample S as the mechanism behind is structurally the same (Andridge *et al.*, 2019; Little *et al.*, 2019).

Y denotes a binary outcome variable that can only take on the values 1 or 0. To describe this binary variable, Andridge *et al.* (2019) and Andridge & Little (2020) apply the standard probit specification based on latent variable formulation that assumes that a normally distributed non-observed variable U created the binary outcome Y where $Y = 1$ when $U > 0$ and $Y = 0$ otherwise. Furthermore, S indicates the binary sampling indicator with $S = 1$ if a unit of the target population is selected into a sample and with $S = 0$ if not. Outcome Y is only observed for cases that have been selected into the sample; thus, for $S = 1$. A set of completely (i.e. in the whole target population) observed auxiliary variables Z, that are predictive for the outcome variable Y, is combined to a sole variable X that serves as surrogate variable for Y (Andridge & Little, 2020). Based on Andridge & Little (2020) and Andridge *et al.* (2019) the following bivariate pattern-mixture model (Andridge & Little, 2011; Little, 1994) can be assumed for the joint distribution of U and X conditioned on S

$$(U, X|S = j) \sim N\left(\begin{pmatrix}\mu_u^{(j)}\\ \mu_x^{(j)}\end{pmatrix}, \begin{pmatrix}\sigma_{uu}^{(j)} & \rho_{ux}^{(j)}\sqrt{(\sigma_{uu}^{(j)}\sigma_{xx}^{(j)})}\\ \rho_{ux}^{(j)}\sqrt{(\sigma_{uu}^{(j)}\sigma_{xx}^{(j)})} & \sigma_{xx}^{(j)},\end{pmatrix}\right) \tag{1}$$

where $N(\cdot)$ marks the bivariate normal distribution with $\mu_u^{(j)}$ and $\mu_x^{(j)}$ as the expected values of U and the proxy X for selection pattern $S = j$ with $j = 1$ denoting the selected and $j = 0$ the non-selected units. $\sigma_{uu}^{(j)}$ and $\sigma_{xx}^{(j)}$ denote the variances of U and X. $\rho_{ux}^{(j)}$ marks the correlation between latent U and X for pattern j, which Andridge & Little (2020) and Andridge *et al.* (2019) specify as the biserial correlation (Tate, 1955) between X and Y for pattern j. Thus, in model (1) different parameter values are possible for the distinct selection patterns j (Andridge & Little, 2011, 2020).

In this paper, our target statistics are proportions. Andridge & Little (2020) and Andridge *et al.* (2019) showed that the marginal mean of Y can be expressed as the weighted average over both patterns $j = (0, 1)$:

$$\begin{aligned}\mu_y = P(Y = 1) &= P(U > 0) \\ &= P(U > 0|S = 1) \cdot P(S = 1) + P(U > 0|S = 0) \cdot P(S = 0) \\ &= \pi\Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right) + (1 - \pi)\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) \\ &= \pi\Phi\left(\mu_u^{(1)}\right) + (1 - \pi)\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right).\end{aligned} \tag{2}$$

Here, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and π is the sampling fraction. For identification, we have to fix $\sigma_{uu}^{(1)}$ and set it to $\sigma_{uu}^{(1)} = 1$ as common for latent variables (Andridge & Little, 2020; Andridge *et al.*, 2019).

Model (1) is underidentified because the data at hand does not contain any information about the parameters related to U for the non-selected units of the target population. Getting estimates for these unidentified model parameters, requires restrictions based on the assumed nature of the sample selection. The parameters are just identified by the following general assumption about the selection mechanism (Andridge & Little, 2011, 2020; Andridge *et al.*, 2019; Little, 1994):

$$P(S = 1|U, X) = g((1 - \phi)X^* + \phi U) \tag{3}$$

where X^* is the rescaled proxy X to have the same variance as U. $\phi \in [0, 1]$ denotes a sensitivity parameter reflecting different scenarios how the selection is related to U and X (Andridge *et al.*, 2019; Little *et al.*, 2019). Thus, ϕ is 'a measure of the degree of non-random selection after conditioning on X' (Andridge *et al.*, 2019, p. 5). ϕ cannot be estimated from the data and has to be set using assumptions about the nature of the selection mechanism. If $\phi = 0$, selection only depends on the observed proxy variable X. Then the selection is SAR and the distribution of Y given X is the same for both selection patterns. If $\phi = 1$ the selection probability is solely related to U and thus to outcome variable Y. Then the selection is SNAR. Values in between represent an association with both X and U (Andridge *et al.*, 2019; Andridge & Little, 2020; Little, 1994; Little *et al.*, 2019). As X forms a proxy of U it is plausible to assume that both are positively correlated. Under this circumstance, both affect selection in the same direction. This is why ϕ is defined to be non-negative (Andridge & Little, 2011; Andridge *et al.*, 2019; Andridge & Little, 2020; Little *et al.*, 2019). For a proof and explanation why there is no need to specify function $g(\cdot)$ to make (1) identifiable, please refer to Little (1994) and Little (2011, 2020).

Given restriction (3) and the properties of the underlying bivariate normal distribution, Andridge & Little (2020) and Andridge *et al.* (2019) show that the mean and variance of U for the non-selected units can be defined by the following equations:

$$\begin{aligned}\mu_u^{(0)} &= \mu_u^{(1)} + \frac{\phi + (1 - \phi)\rho_{ux}^{(1)}}{\phi\rho_{ux}^{(1)} + (1 - \phi)} \frac{\mu_x^{(0)} - \mu_x^{(1)}}{\sqrt{\sigma_{xx}^{(1)}}} \\ \sigma_{uu}^{(0)} &= 1 + \left(\frac{\phi + (1 - \phi)\rho_{ux}^{(1)}}{\phi\rho_{ux}^{(1)} + (1 - \phi)}\right)^2 \frac{\sigma_{xx}^{(0)} - \sigma_{xx}^{(1)}}{\sigma_{xx}^{(1)}}.\end{aligned} \tag{4}$$

To get estimates for the non-selected units, mean and variance for the selected cases are shifted depending on the sensitivity parameter ϕ and the correlation between U and X in the selected sample. In addition, this shifting is affected by the selection bias in X measured by the deviations of proxy distributions X between selected and non-selected cases (Andridge & Little, 2020). The further away the overall mean of Y from the mean of the sampled units, the larger the selection bias for the estimated mean of Y (based on the selected non-probability sample) (Andridge & Little, 2020; Andridge *et al.*, 2019). Andridge & Little (2020) denote the correlation $\rho_{ux}^{(1)}$ as 'strength' of the proxy X where higher correlation values correspond to a higher predictive power of the auxiliary variables Z. The detailed derivations of the formulas (4) can be found in Andridge & Little (2020, 2011) and Sullivan & Andridge (2015).

2.2 The Index

Based on the model (1), Andridge *et al.* (2019) propose an index for quantifying non-ignorable selection bias in estimated proportions. They call this index measure of unadjusted selection bias for proportions (MUBP). It is computed by Andridge *et al.* (2019)

$$\begin{aligned} MUBP(\phi) &= \hat{\mu}_y^{(1)} - \hat{\mu}_y(\phi) \\ &= \hat{\mu}_y^{(1)} - \left[\hat{\pi}\Phi\left(\hat{\mu}_u^{(1)}\right) + (1 - \hat{\pi})\Phi\left(\hat{\mu}_u^{(0)}/\sqrt{\hat{\sigma}_{uu}^{(0)}}\right)\right]. \end{aligned} \tag{5}$$

where ˆmark estimated values. Plugging in formulas (4) finally yields

$$\begin{aligned} MUBP(\phi) = \hat{\mu}_y^{(1)} &- \hat{\pi}\,\Phi\left(\hat{\mu}_u^{(1)}\right) - (1 - \hat{\pi}) \\ &\times \Phi\left(\left\{\hat{\mu}_u^{(1)} + \frac{\phi + (1 - \phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1 - \phi)} \frac{\hat{\mu}_x^{(0)} - \hat{\mu}_x^{(1)}}{\sqrt{\hat{\sigma}_{xx}^{(1)}}}\right\}\Big/ \cdot \right. \\ &\times \left.\sqrt{\left[1 + \left\{\frac{\phi + (1 - \phi)\hat{\rho}_{ux}^{(1)}}{\phi\hat{\rho}_{ux}^{(1)} + (1 - \phi)}\right\}^2 \frac{\hat{\sigma}_{xx}^{(0)} - \hat{\sigma}_{xx}^{(1)}}{\hat{\sigma}_{xx}^{(1)}}\right]}\right) \end{aligned}$$

$\mu_x^{(j)}$ and $\sigma_{xx}^{(j)}$ can be estimated as sample means and variances for non-selected and selected units, $j = (0, 1)$. $\hat{\mu}_y^{(1)}$ is the sample proportion of binary variable Y based on the sampled units. The estimate for π is given by the sampling fraction, that is, the proportion of selected individuals from the population (Andridge & Little, 2020; Andridge *et al.*, 2019). Thus, we need to know the size of the target population that requires its exact specification. For estimating the biserial correlation $\rho_{ux}^{(1)}$, Andridge *et al.* (2019) and Andridge & Little (2020) suggest using the 'two-step' approach of Olsson *et al.* (1982) that yields as byproduct an estimate of the mean of U in the selected sample. To additionally prevent potential issues with overfitting, Andridge *et al.* (2019) recommend applying multifold cross-validation. For more details on the cross-validation and two-step estimation of the biserial correlation, please refer to Andridge & Little (2020) and Andridge *et al.* (2019).

The index (5) represents the difference between the proportion of Y estimated in the non-probability sample and the estimated proportion of Y in the overall population for a given choice of ϕ. Thus, the computation of a concrete index value involves the choice of a specific

sensitivity parameter ϕ that captures the varying assumptions about the selection mechanism. As $\phi = 0$ reflects SAR with regards to the variables used to calculate X and $\phi = 1$ means complete dependence of the selection on the outcome variable Y, that is, SNAR, Little *et al.* (2019) and Andridge *et al.* (2019) recommend to calculate the interval $[MUBP(0), MUBP(1)]$ to map the extent of possible selection bias. Setting ϕ to 1 gives an idea of the highest potential bias that might occur under an extreme non-ignorable selection mechanism. In addition, they suggest computing $MUBP(0.5)$ that serves as estimate of selection bias where X and Y are equally affecting the selection propensity. The amount of variation between the index values for different values of ϕ reflecting different assumptions about the selection mechanism indicate the extent of possible selection bias in the considered variable Y. The ideal case are low index values, that hardly differ across different ϕ values. If the index yields similar values for different ϕ, the mean of Y is robust to different selection mechanisms, which is a desirable property for the population statistics to be estimated. Be aware, that this proposed index is variable-dependent, so it has to be computed for each variable of interest separately and can turn out very differently across varying outcome variables (Andridge *et al.*, 2019; Little *et al.*, 2019). Andridge *et al.* (2019) point out that the MUBP index is not automatically monotonic over the $[0, 1]$ interval for ϕ, and give more details on this potential issue.

Andridge *et al.* (2019) and Andridge & Little (2020) also provide a fully Bayesian version of the index calculation to appropriately reflect parameter uncertainty during the estimation process. The Bayesian approach generates draws from the posterior distribution of the $MUBP$ index and enables the calculation of credible intervals for specific choices of sensitivity parameter ϕ that capture the uncertainty in the location of $MUBP(\phi)$. The Bayesian procedure also allows to draw values of ϕ, for example, by using a $Uniform(0,1)$ prior distribution. One iteration of the respective data augmentation strategy draws the latent variable U using a truncated normal distribution, creates posterior draws of the regression parameters of the probit model for proxy X, draws new parameter candidates of all pattern-mixture model parameters, and then generates a draw of the $MUBP(\phi)$ (Andridge *et al.*, 2019). For all model parameters, non-informative Jeffreys' priors are used (Andridge & Little, 2011). More details on the Bayesian computation of the MUBP index can be found in Andridge *et al.* (2019) and the exact steps of the applied Gibbs sampling routine are available in Andridge & Little (2020), Andridge & Little (2009), and Andridge (2009).

There are four requirements to keep in mind, so that the index $MUBP(\phi)$ can be applied to quantify selection bias in Y and yields reasonable results (Andridge *et al.*, 2019; Andridge & Little, 2011, 2020; Little *et al.*, 2019). Table 1 summarises these conditions and gives a brief description of each point.

In sum, the index strongly depends on the used variables Z and the model assumptions made by the underlying PPMM. Thus, $MUBP(\phi)$ should always be interpreted with caution and not be misunderstood as a global measure for selection bias because without further external assumptions, a measure like this cannot exist.

3 Validation Study

We conducted a short validation study to evaluate the performance of the MUBP index to detect selection bias in proportions estimated from a non-probability sample. In doing so, we followed the applications presented in Little *et al.* (2019) and Andridge *et al.* (2019). We used data from the German General Social Survey (GGSS) (GESIS - Leibniz-Institut für Sozialwissenschaften, 2021). The GGSS is a biennial, cross-sectional probability survey of individuals who are resident in private households in Germany and are at least 18 years old

Table 1. *Overview of the required conditions to apply the suggested index, collected from Andridge et al. (2019), Little et al. (2019) and Andridge & Little (2011, 2020).*

#	*Requirement*	*Description*
1	Moderate correlation between proxy X and U	• Weak correlations between X and Y result in increased uncertainty and very wide $[MUBP(0), MUBP(1)]$ intervals • Wide $[MUBP(0), MUBP(1)]$ intervals are not useful in practice as they fail to provide an effective indication of selection bias • Andridge & Little (2011, 2020) suggest working with a $\hat{\rho}_{xu}^{(1)}$ of at least 0.3 and Little *et al.* (2019) even propose a cutoff of 0.4
2	Certain deviation between proxy distributions of the selected and non-selected sample	• Assumption: selection bias in X indicates selection bias in Y • If the proxy distributions are too similar, the index is not able to detect still possible selection bias in Y
3	Validity of imposed structural relationship between selection mechanism and X and U	• Assumption of identifying restriction (3): selection depends on a linear combination of X and U • X and U affect the selection in the same direction
4	Information about auxiliary variables Z of the non-sampled units (only aggregates in form of means and covariances needed)	• Required to estimate the mean and variance of proxy X of the non-selected cases • In case of large target populations and negligible sampling fractions, possible to use information on the whole population (from administrative records, census data or large population surveys) • If only population means of Z available, Andridge *et al.* (2019) and Little *et al.*, (2019) recommend to assume equal proxy variances of selected and non-selected units • Bayesian approach requires microdata or variance and covariance estimates of Z

on January 1st of the year of the survey. It collects information on attitudes, behaviour, and social change in the Federal Republic of Germany. For our analysis, we pooled GGSS data for 2012, 2014, 2016 and 2018, resulting in a total sample size of 13,918 respondents. We selected a subset of variables to further work with (described below) and only kept units with complete observations on these variables. This set of 10,963 complete respondents was eventually treated as hypothetical population that allowed the calculation of true values and thus actual bias values in the considered proportions. From this hypothetical population, we created an artificial non-probability sample by selecting all individuals using internet for private purposes and with strong or very strong interest in politics. The selection fraction of the artificial non-probability sample is 0.308 (corresponding to 3,381 individuals).

We assessed the performance of the MUBP index based on ten distinct target variables Y (see also column one Table 2): 'currently unemployed', 'having a monthly net income >1,500€', 'having voted in the last federal election', 'assessing the personal economic situation as good or better', 'being a member of a labour union', 'having no trust to fellow humans', 'having own children', 'assessing freedom of expression as most important political goal', 'being a city resident' and 'being Catholic'. As auxiliary variables Z we chose female/male, German citizenship, education, employment status, marital status, region (East/West Germany), age, and household size. In this example, information on auxiliary variables Z is directly available for the non-selected cases. Hence, there is no need to use means of Z from the overall hypothetical population as sufficient statistics (which would require an additional adjustment to account for the non-negligible selection fraction, see Andridge *et al.*, 2019).

Table 2. *Analysis results for the ten binary variables of interest Y including true bias, $MUBP(\phi)$ for $\phi = (0, 0.5, 1)$ and 95% credible interval for MUBP based on the Bayesian approach.*

Binary variable of interest Y	$\hat{\boldsymbol{p}}_{\boldsymbol{pop}}$	$\hat{p}_{samp}$	*TB*	$cor(Y, X)$	d*	*MUBP (0.5)*	*[MUBP(0), MUBP(1)]*	*Proposed interval covers TB?*	*95% credible interval*	*Interval covers TB?*
Member of labour union	0.133	0.155	0.022	0.198	−0.151	0.030	[0.005, 0.107]	Yes	[−0.008, 0.096]	Yes
Freedom of expression	0.243	0.343	0.100	0.280	−0.402	0.085	[0.028, 0.165]	Yes	[0.029, 0.163]	Yes
No trust	0.370	0.252	−0.119	0.288	0.597	−0.154	[−0.041, 0.354]	Yes	[−0.324, −0.045]	Yes
City resident	0.272	0.335	0.064	0.295	−0.184	0.047	[0.014, 0.171]	Yes	[0.008, 0.151]	Yes
Good economic situation	0.665	0.760	0.095	0.376	−0.507	0.127	[0.044, 0.325]	Yes	[0.047, 0.291]	Yes
Voted	0.841	0.926	0.086	0.411	−0.358	0.061	[0.019, 0.193]	Yes	[0.020, 0.172]	Yes
Catholic	0.253	0.232	−0.022	0.468	−0.104	0.017	[0.009, 0.024]	No	[−0.005, 0.039]	No
Unemployed	0.361	0.295	−0.066	0.685	0.336	−0.091	[−0.061, −0.131]	Yes	[−0.134, −0.058]	Yes
Own children	0.714	0.702	−0.012	0.705	−0.003	0.006	[0.003, 0.010]	No	[−0.010, 0.021]	No
Income>1,500 Euro	0.438	0.617	0.179	0.736	−0.588	0.161	[0.118, 0.215]	Yes	[0.120, 0.211]	Yes

$\hat{\boldsymbol{p}}_{\boldsymbol{pop}}$ is the proportion of $Y = 1$ in the artificial population and $\hat{\boldsymbol{p}}_{\boldsymbol{samp}}$ is its proportion in the non-probability sample. **TB** (true bias) is the difference between proportions in the non-probability sample and population. $\boldsymbol{cor}(\boldsymbol{Y}, \boldsymbol{X})$ is the biserial correlation between proxy X and binary variable of interest Y in the non-probability sample. d* is the standardised difference of proxy distributions between selected and non-selected units $(\overline{x}^{(0)} - \overline{x}^{(1)})/\sqrt{S_{xx}^{(1)}}$ where $\overline{x}^{(0)}$ and $\overline{x}^{(1)}$ denote the mean of proxy X for the non-selected and selected cases, and $S_{xx}^{(1)}$ is the variance estimate of X in the non-probability sample.

International Statistical Review (2025), **93**, 3, 499–516

For each target variable Y, we calculated the proposed MUBP index with varying values of sensitivity parameter $\phi = (0, 0.5, 1)$.[2] For this purpose, we initially fitted a probit model regressing the respective variable Y on all auxiliary variables Z using the artificially created non-probability sample only. The resulting estimated regression parameters were then used to calculate the linear predictor as proxy X for the underlying latent variable U for selected and non-selected cases, respectively. To prevent issues with overfitting we used five-fold cross-validation to determine X, as described in Andridge *et al.* (2019). Then, we computed the MUBP indices based on the fitted proxy variable X. The biserial correlation between Y and X was estimated using the selected cases and the two-step approach introduced in Andridge *et al.* (2019) and Andridge & Little (2020). We estimated all parameters required to specify the MUBP index in a frequentist manner. Additionally, the fully Bayesian estimation approach was applied to get 95% credible intervals for $MUBP(\phi)$. At this, ϕ was drawn where a $Uniform$ (0,1) distribution served as prior.

Table 2 shows the results of our calculations for each binary target variable Y. To evaluate if the MUBP index is able to properly detect (non-ignorable) selection bias in the estimated proportions, we computed the true bias as difference between the proportion in the non-probability sample and the 'true' proportion in the hypothetical population. For each variable Y, we examined whether the interval $[MUBP(0),\ MUBP(1)]$ obtained from the frequentist estimation and the Bayesian credible interval are able to cover the actual bias. In addition, we evaluated if $MUBP(0.5)$ performs well in giving an approximate estimate of the true bias as it was suggested byLittle *et al.* (2019) and Andridge *et al.* (2019). Table 2 also contains the estimates of the biserial correlation between X and Y. This correlation is an important indicator of the suitability of the proxy and thus of the meaningfulness of the estimated index. The difference of the proxy distributions of non-selected units and those selected for the non-probability sample is another important indicator of the goodness of the index calculation. Therefore, we included the standardised difference $\mathrm{d}^* = (\bar{x}^{(0)} - \bar{x}^{(1)})/\sqrt{S_{xx}^{(1)}}$ in Table 2 as well.[3] Here, $\bar{x}^{(0)}$ and $\bar{x}^{(1)}$ denote the mean of proxy X for the non-selected and selected cases, respectively, and $S_{xx}^{(1)}$ is the sample variance of X in the non-probability sample. The $[MUBP(0),\ MUBP(1)]$ intervals including the true bias and $MUBP(0.5)$ estimates are also depicted in Figure 1.

For eight of the ten targets Y considered, $[MUBP(0),\ MUBP(1)]$ and the 95% credible intervals are able to cover the true selection bias in the estimated proportions, see Figure 1 and the last column of Table 2. This is even the case for the variables 'good economic situation', 'member of labour union', 'no trust', 'freedom of expression', and 'city resident' with low proxy strengths (i.e. with a biserial correlation between X and Y smaller than 0.4). However, we find that the uncertainty of the selection bias is high when a low proxy correlation is combined with a large deviation in the X distributions in the population and the non-probability sample. In our example, this concerns 'no trust' and 'good economic situation'. Both $[MUBP(0),\ MUBP(1)]$ and the 95% credible interval fail to cover the true bias for the variables 'own children' and 'Catholic'. And this despite the fact that here the biserial correlation between X and Y is relatively large, especially for 'own children'.

Andridge *et al.* (2019) describe a similar finding with some of the variables they considered in their simulation. To examine the underlying selection mechanism in more detail, they fitted probit regression models to the selection indicator of the hypothetical population using the respective binary outcome variable Y and its proxy X as predictors. They found that for these variables the estimated coefficients of Y and X had opposite signs, that is, Y and X had reverse effects on the probability of being selected into the sample. However, the underlying pattern-mixture of the index calculation explicitly assumes that the selection mechanism is a function of $(1 - \phi)X^* + \phi U$ where ϕ is assumed to be positive. Thus, the model does not

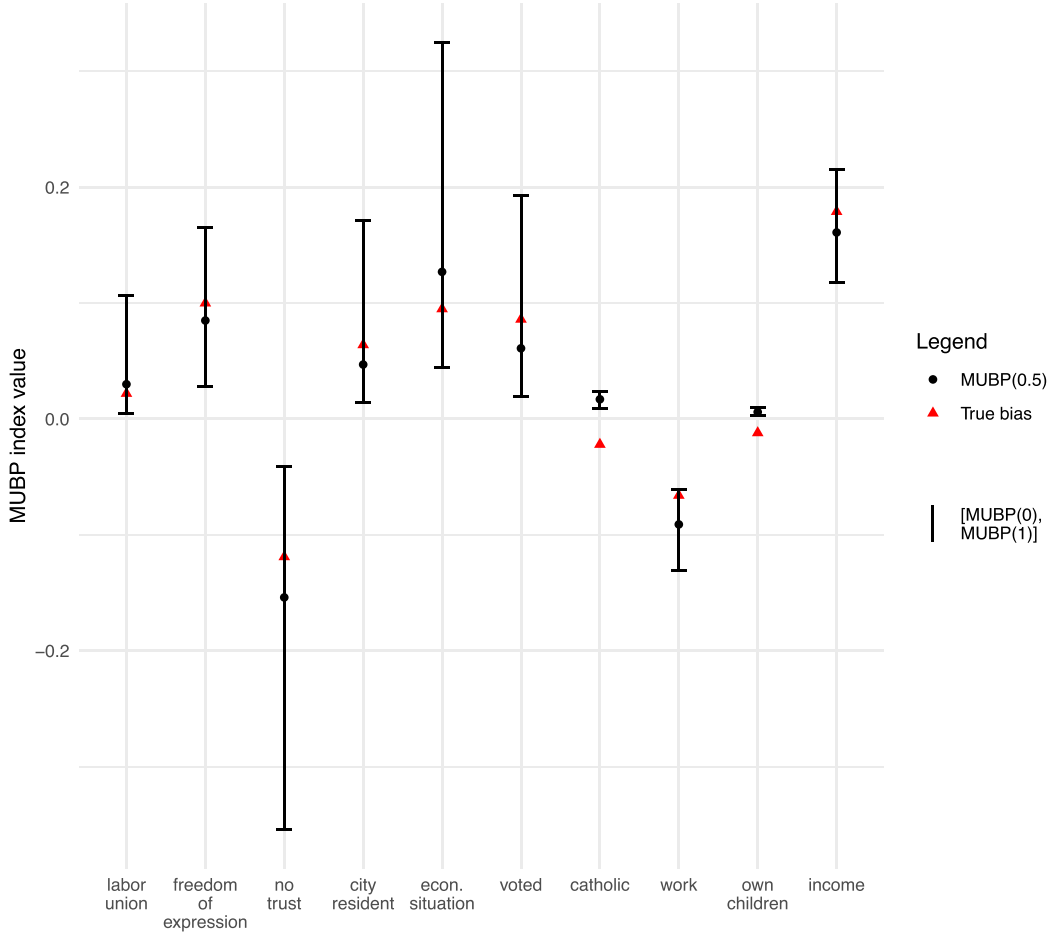

Figure 1. *Range of* $MUBP(0)$ *and* $MUBP(1)$ *index values to illustrate the interval of potential selection bias including* $MUBP(0.5)$ *as 'estimate' of the bias.*

account for cases where X and Y have opposite effects on the selection mechanism, why here the MUBP index is not able to properly reflect the selection bias.

We investigated if this issue is also present for our two variables 'own children' and 'Catholic' and found coefficients of opposite directions for both of them. This provides a possible explanation why both intervals fail to capture the actual selection bias properly. Andridge *et al.* (2019) point out that this scenario is only an issue for proxy variables that are strongly correlated with outcome Y because with weak proxies the interval $[MUBP(0),\ MUBP(1)]$ will be very wide and probably cover the true bias anyway. If the probability of selection has a positive relationship with X and a negative relationship with Y, they suggest computing $MUBP(-\infty)$ and using the interval of $[MUBP(-\infty),\ MUBP(1)]$ for specifying the range of selection bias. In their example, they were able to cover the bias that way. Unfortunately, this approach is less relevant for practice as it leads to very wide intervals that are of course less effective. We applied Andridge's et al. (2019) workaround for the two problem cases 'Catholic' and 'own children' in our validation study: For 'Catholic', the interval of $[MUBP(-\infty),\ MUBP(1)]$ indeed covers the true bias, but for 'own children' the selection bias is still not reflected properly.

Thus, there seems to be another problem here, at least with 'own children'. Having a look at the standardised difference of the proxy distributions of the non-selected and sampled cases, we find that for both variables it is quite small. Actually, it is much lower than for the remaining considered outcome variables. The underlying pattern-mixture model uses the selection bias in proxy X (i.e. the difference of distributions of X between selected and non-selected units) as measure for existing selection bias in outcome Y. Thus, if there is a lack of selection bias in X it is by definition also not present in Y. When the correlation between X and Y is additionally very high, this evidence of lack of bias in Y gets even stronger (Andridge & Little, 2011, 2020). As a consequence, the $[MUBP(0),\ MUBP(1)]$ interval gets narrower in such a case. Little *et al.* (2019) mention this potential issue and point out that it is of course still possible that Y suffers from selection bias even when the proxy distributions of X between sampled and non-sampled are very similar. Therefore, we recommend to not only presenting the biserial correlation between X and Y along with the calculated indices but also the difference of proxy distributions. When the difference in proxy distributions is very small and the correlation between Y and X is high, small index values should be interpreted with caution.

In our application, the $[MUBP(0),\ MUBP(1)]$ intervals and the 95% credible intervals do not show any difference in terms of coverage of the selection bias. However, the Bayesian approach properly reflects uncertainty in parameter estimation and the Bayesian credible intervals provide a more intuitive interpretation, why we propose to prefer the Bayesian procedure in cases where the information required for computation is accessible.

In summary, our validation study shows that the proposed index performs very well in detecting selection bias in estimated proportions if the assumptions of the underlying PPMM are fulfilled. This also means that a moderate difference of proxy distributions between non-selected and sampled cases seems is crucial to properly indicate the true bias. In our analysis, this requirement seem to be even more relevant than a very high correlation between X and Y. The latter is, however, an important condition to prevent ineffective and very wide intervals of potential selection bias. The validation study also came across the issue of opposite effects of proxy X and outcome variable Y on the selection probability—a situation precluded by the assumptions of the PPMM. Clearly, the assumption that the selection mechanism depends on X and Y in same directions is theoretically very reasonable. Nevertheless, opposite effects seem to occur in practice and can lead to misinterpreting the extent of selection bias.

4 Application to a Real Non-Probability Sample

To assess the applicability and usefulness of the suggested index in practice, we applied it to data from a large market and opinion research company that arise from a real non-probability sample. This company collects its data, or rather recruit its users by applying a non-probability online sampling procedure called river sampling (American Association for Public Opinion Research, 2013). That is, they place questions via a network of websites that cooperate with them. The questions are embedded in the online content of the respective websites, mostly in articles to different topics such as politics, so that the website user takes notice of them during reading. After a number of answered questions the respondents are invited to register for their panel. As registered user you have access to a web interface and are in principle able to give answers to all questions that are currently active. The vast majority of their data come from these registered users.

The company's target population are individuals with German citizenship aged 18 years or older. For computing population-based estimates, they adjust their data by weights. They obtain these weights via raking using selected variables and their known marginal distributions in the population. Due to the company's self-recruiting sampling strategy the application of standard

design-based methods might not be sufficient to obtain estimates extrapolating to the population level as the non-random recruting procedure possibly produces a non-ignorable selection mechanism. Under such a selection scheme, standard survey weighting procedures like raking are not assured to produce unbiased population estimates (Elliott & Valliant, 2017; Valliant, 2020; Valliant *et al.*, 2018). Against this background, this provided non-probability sample is ideal for applying (and trying out) the MUBP index. Calculating the measure requires proper population data with variables suited as predictors for calculating the proxy variables. An obvious choice is the German Socio-Economic Panel (SOEP) (Socio-Economic Panel (SOEP), 2022) that is one of the largest and longest-running household surveys worldwide and collects annually comprehensive data on persons living in private households in Germany. It is a rich data set on information about demographic, socio-economic, behavioural and attitudinal measures of Germany's resident population (Giesselmann *et al.*, 2019). The SOEP uses random sampling strategies to produce a large probability sample of approximately 20,000 households (Goebel *et al.*, 2019). For our real-world application, we use data from the non-probability sample and SOEP collected between February and August 2020. We restrict the SOEP data set to adults with German citizenship so that it corresponds to the company's target population. From the pool of questions available for the non-probability sample, we selected the following four political questions as outcome variables:

- 'Do you think that the German health system needs to be extensively reformed?' (Question 1)
- 'Do you think that the coexistence of people from different cultures enriches our society?' (Question 2)
- 'Do you think that the membership in the European Union brings more prosperity to its member states?' (Question 3)
- 'Do you think that populism threatens democracy worldwide?' (Question 4)

Originally, the four questions are ordinal-scaled where respondents could rate their answers on 5-point Likert scales. To be able to apply the MUPB index, we dichotomised each question. For this purpose, we merged the categories 'Yes, definitely' and 'Rather yes' to value 1, and the remaining categories to value 0.

For the index to perform well, proper auxiliary variables are necessary that can be used as predictors (see Section 3). Overall such task is not trivial as it means that suitable questions asked in a comparable way have to be available for the non-probability sample and the probability (or population) data set (here, the SOEP). Socio-demographic variables (such as age and sex) seems to be natural candidates. However, we note that in our case, the usage of standard socio-demographics resulted in only very weak correlations between the four binary outcomes studied and their proxies (biserial correlations among 0.15). Because we know that calculating the MUPB under this condition yields only limited meaningful results, the task was to find more suitable variables in the rich SOEP data. We finally used sex, age, employment status, marital status, region (East/West Germany), formal educational level, children in the household, concerns about personal economic situation and concerns about migration as predictor variables Z to form the proxy X for the four chosen outcome variables.

The number of respondents with observations on all auxiliary variables Z and the respective binary variable of interest Y define the final sample size for each question entering the analysis. A special challenge of working with the provided non-probabilistic data was that the company prioritises the single questions differently. As a consequence, some units do not get the opportunity to answer single polls because they are not displayed to them. Therefore, not all questions are commonly observed for all panel members and the sample sizes of the single questions are different. Question 1 and 2 have a sample size of 15, 876 and 28, 613, respectively, for Question 3 we have information of 18, 756 individuals, and for Question 4 data are available for a total of

21, 792 units. Table 3 shows some descriptive statistics for the chosen auxiliary variables to villustrate their distributions in both data sets. For the data of the non-probability sample, we examplarily only show the (unweighted) distributions of the sample of Question 1 as they did not differ essentially between the four questions. For the SOEP, we provide the weighted vpopulation estimates.

For each variable of interest Y, we used the respective units of the non-probability sample to compute a probit regression where Y is regressed on the chosen predictor variables described above. The resulting regression estimates were used to compute the linear predictor that forms the proxy variable X for these selected units. To form the proxy for the non-selected units, population estimates for the selected auxiliary variables are necessary. For this purpose, we use the SOEP data (with a sample size of 21, 590 individuals) and apply the provided individual-level weights to obtain weighted means and (co)variances of the predictor variables Z. Using the estimated coefficients, the weighted means and variances of Z, we were able to calculate the mean and variance of proxy X for the non-selected units. The sample fraction for each of the considered questions was determined by their respective final sample size used for analysis divided by the size of the company's target population (individuals with German citizenship that are at least 18 years old) in 2020. According to the Federal Statistical Office of Germany, this is a total of 60, 561, 857 people (https://www-genesis.destatis.de/genesis/online, reference date: 31.12.2020). Based on these measures we computed the $MUBP(\phi)$ posterior

Table 3. *Descriptive statistics for the selected auxiliary variables Z in both data sets (proportions for the categorical variables and means for the continuous variable).*

Variable	*Non-prob sample (Question 1)*	*SOEP (weighted)*
Sex		
Male	0.7162	0.4872
Female	0.2838	0.5128
Age	65.85	52.22
Employment status		
Full-time	0.2270	0.3990
Not full-time	0.7730	0.6010
Marital status		
Divorced	0.1059	0.1081
Married	0.6849	0.5182
Single	0.1324	0.2985
Widowed	0.0768	0.0752
Region		
East	0.2139	0.2146
West	0.7861	0.7854
Educational level		
Primary	0.0034	0.0202
Lower secondary	0.0105	0.0957
Upper secondary	0.4663	0.5743
Tertiary	0.5198	0.3098
Children in household		
Yes	0.0989	0.2071
No	0.9011	0.7929
Worried about finances		
Not concerned at all	0.4682	0.3259
Somewhat concerned	0.4215	0.4044
Very concerned	0.1103	0.2797
Worried about immigration		
Not concerned at all	0.2307	0.4722
Somewhat concerned	0.2458	0.4268
Very concerned	0.5235	0.1010

The data of the non-probability sample are not weighted.

distribution applying the fully Bayesian approach where the sensitivity parameter ϕ was drawn using a *Uniform* (0,1) prior. Following the polling application presented in West & Andridge (2023), we calculated various population estimates in our application. In this way, we are able to vperform a sensitivity analysis that maps our unawareness about the actual nature of the assumed non-ignorable selection mechanism.

Figure 2 visualises three different types of estimates of the population mean for each of the four outcome variables studied. Besides point estimates (marked as dots), we also provide measures of uncertainty in the form of confidence and credible intervals. The first estimates are the unweighted sample means with their respective 95% confidence intervals (*Un*).[4] These estimates do not consider any corrections and just use the data of the available non-probability sample.

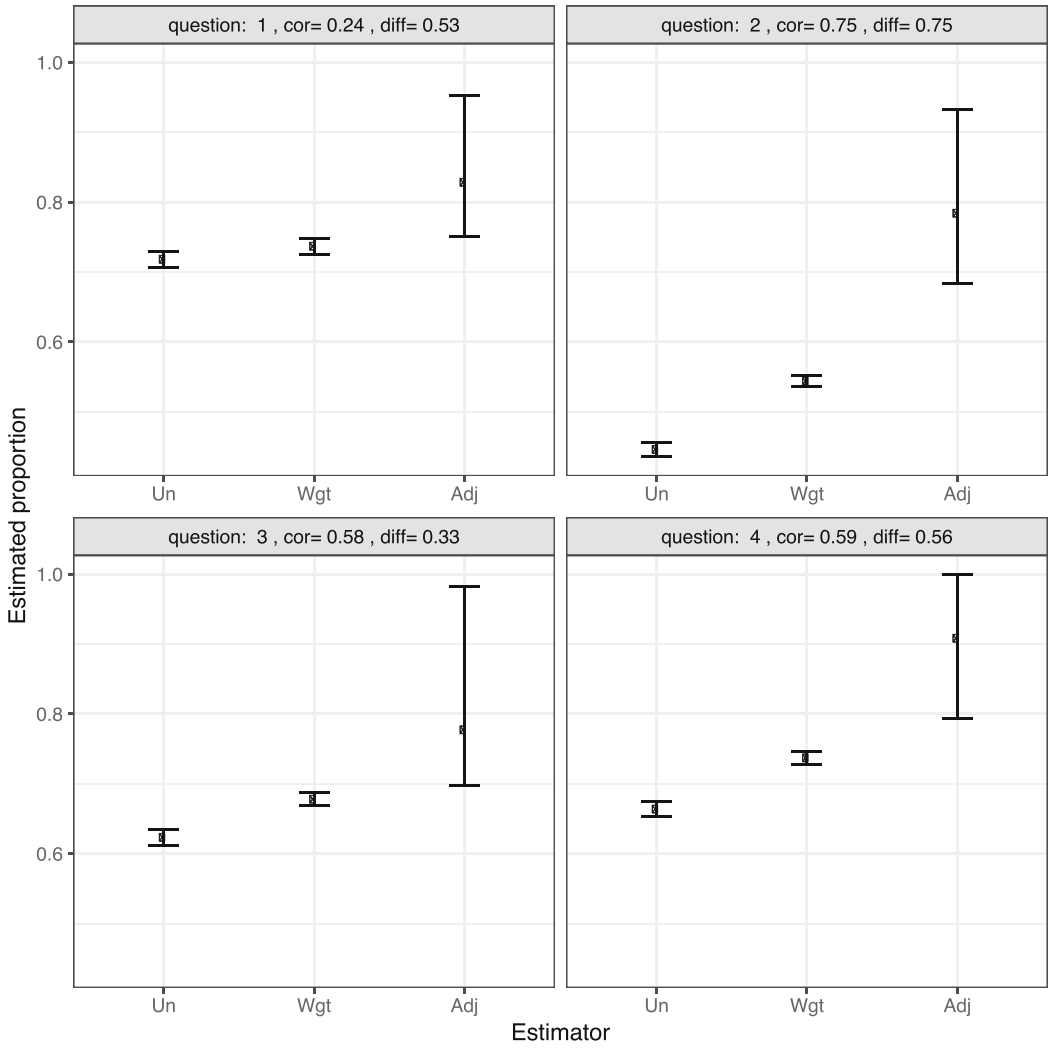

Figure 2. *Different types of estimates for the population proportions of the considered binary variables including measures of uncertainty. Un denotes the unweighted sample mean, Wgt the weighted mean using the company's current weighting strategy, and Adj is the sample mean adjusted by the median of the posterior distribution of $MUBP(\phi)$ with ϕ drawn using a Uniform(0,1) prior. Cor indicates the biserial correlation between proxy X and binary outcome Y among the selected units, and diff is the standardised difference of proxy means in the SOEP and the non-probability sample.*

The second estimates denoted by *Wgt* are the provided weighted sample means generated by the company's current weighting strategy. We constructed 95% confidence intervals for these weighted estimates using their delivered effective sample sizes for each question.[5] The third measures are the adjusted estimates along with their adjusted credible intervals. The adjusted estimates are calculated as the difference between the (unweighted) sample mean and the median of the MUBP posterior distribution, which directly results from rearranging the MUBP index formula given in Equation 5. The adjusted credible intervals are determined by the difference of the sample mean and the 2.5% and 97.5% quantiles of the MUBP posterior distribution.[6]

Contrary to the polling application of West & Andridge (2023), we do not know the true values of the our variables studied. Hence, we cannot assess the coverage of the three credible intervals. However, we can also assess the robustness of the estimated proportions to different assumptions about the selection mechanism by comparing the three different point estimates with their respective intervals. Compared with the results presented by West & Andridge (2023), our application yields very wide adjusted credible intervals for each outcome variable that reflects the high uncertainty in the estimated proportions. The difference between the MUBP-adjusted estimates and the weighted estimates is obvious indicating a downward bias in the latter. However, we also see that this downward bias is attenuated compared with the unweighted proportion estimates.

Figure 2 also gives the biserial correlations between proxy X and the binary outcome variables Y as well as the standardised differences of proxy distributions between the SOEP (used for the non-selected units) and the non-probability sample (selected units). As stated above, this correlation is a crucial measure for assessing the validity of the MUBP index. We used variables with a range of varying correlation values, however, in this case, it did not seem to have much impact on the width of the interval. This may be due to the fact that the proxy distributions of the SOEP and the non-probabilistic data are very different for all considered variables of interest. Because the MUBP index uses this deviation as indication for selection bias present in Y, we receive very wide intervals of potential selection bias for all considered variables regardless of their correlation values. Please note that, for questions 1 and 3, the relation of their interval widths seems counterintuitive if we are only looking at the proxy strength and the difference in proxy means. Having a closer look at the drivers of the index value apart from these two key quantities (using the frequentist variant of the index), we found that question 3 is associated with much larger differences in proxy variances than question 1. This difference is also negative as the proxy variance of the selected cases is higher than for the non-selected cases. In addition, the difference's amount lies between 0 and 1. This leads to a very small estimated variance of the latent variable U for the non-selected cases. Dividing $\hat{\mu}_u^{(0)}$ by the squared root of this value (in the index formula (5)) leads to a larger value entering the normal cumulative distribution function—which then results in larger estimates of the mean of Y for the non-selected cases (for $\phi = 1$). The single numerical results of the ML estimates are available in Table S.1 in the supplementary material accompanying this article. For double-checking, we computed the $[MUBP(0),\ MUBP(1)]$ intervals assuming equal variances between non-selected and selected cases that is recommended by the authors of the index in cases where no information about the variances of the non-selected cases is available. Then these 'odd' results should not occur because the differences in proxy variances do not play any role for the index computation. Figure S1 shows that these frequentist intervals behave as suspected that confirms that the counter-intuitively-looking intervals in our original figure indeed arise from the proxy variances and the numerical particularities explained above.

Altogether, we see a great variability of results depending on the assumed nature of the vselection mechanism. According to our conducted analyses, the estimated proportions of the

considered outcome variables are relatively sensitive towards the ignorability assumption of the selection mechanism. This is why they should be interpreted with caution.

Beware that this is only one specific example of sensitivity analysis. Our results strongly depend on the underlying model assumptions (as stated in Section 2), the suitability of the population data source utilised to get information about the non-selected units, as well as on the choice of auxiliary variables Z that are used to form the proxy X. Considerable effort must be made to find an appropriate data source for population data so that one is able to assess and correct for potential selection bias in non-probability samples. It is clear that the SOEP cannot capture the population figures 100% correctly. There are dropout processes and potentially selective answers as well, especially in questions on attitudes and opinions. However, survey data such as the SOEP are the only source of information we have for subjective measures at the population level. In Germany, official statistics cannot provide more reliable information either.

5 Conclusion

In the context of non-probability samples, it is not appropriate to simply assume that the underlying selection mechanism is ignorable. Biased estimates and erroneous interpretations might be the result. Thus, there is an indispensable need for a handy measure that makes non-ignorable sample selection quantifiable, and ideally also correctable. With the measure of unadjusted selection bias for proportions (MUBP), Andridge *et al.* (2019) proposed such a measure. The MUPB helps detecting (non-ignorable) selection bias in proportions estimated based on non-probability samples. However, so far, applications of the index are sparse. To the best of our knowledge, the few that exist have been performed by the developers of the index themselves. This paper now presents another application of the index. Convinced of the index's potential, we applied it to data from a non-probability sample on the political situation in Germany that resulted from river sampling. However, in our view, reexamining the fit of the MUPB to a specific application problem is absolutely necessary as it is based on a number of weighty assumptions. Furthermore, the calculation of the MUPB hinges on the availability of population data matching the scopes of the non-probability sample examined. To test the index for its usefulness for our purposes, we designed a corresponding validation study. All in all, the performance of the index was very good, but we also became aware of some conditions that must be met for it to provide valid and useful results. First, the availability of appropriate population data with good auxiliary variables (that can be used as covariates to form the proxy variable) is crucial. Finding accordant data for a specific data situation is not trivial as the non-probability sample and population data need to share a set of identically measured variables that have to be predictive for the studied variable(s). Related to this is that different variables can be used as predictors to form the proxy, which in turn can lead to (very) different index values. Therefore, this step (i.e. identification of appropriate population data and variables to derive the proxy) is the most important in the entire index building process. Second, as the index computation requires the sampling fraction of selected cases, we need an exact specification of the sample's target population. This should be self-evident when working with data, but is not always automatically fulfilled in the context of non-probability samples. Third, just like reported in Andridge *et al.* (2019), we also ran into the problem of opposite effects of proxy X and outcome variable Y on the selection mechanism. This may theoretically be implausible, but apparently occurs in the context of empirical data nevertheless. In this case, it is not assured that the MUBP index covers the bias appropriately. Forth, the index uses the differences in proxy distributions of selected and non-selected cases, that is, the selection bias in X, serves as measure for selection bias in the focal variable(s). If the proxy distributions are very similar or nearly identical, the measure will not detect a selection bias in Y, even if there is one. That

is, the index should be interpreted with caution in cases where the deviations are very small. Fifth, the correct application of the method requires a good proxy variable that has at least a moderate correlation with the variable of interest, which means that the auxiliary variables used must be predictive of Y. If this is not the case, the index calculation will not be effective leading to very wide intervals. Such intervals are not useful for real-world applications because they cannot meaningfully represent the true extent of possible selection bias.

Taking all of these potential hindrances into account, the MUBP is very powerful in mapping sample selection bias. We were able to demonstrate this with our real-world application on political attitudes collected via river sampling. Here, the index was able to meaningfully quantify the potential sample selection bias. It additionally provides a simple and sensible adjustment method with weaker assumptions about the selection mechanism than standardly applied design-based methods. Clearly, the analyses presented here apply to the specific settings of our validation study and real-world data example. General and more broader statements about the performance of the MUBP necessitate more applications to additional non-probability samples. We would therefore like to encourage other researchers to use it and try it out. For this, it is extremely helpful that its developers Little *et al.* (2019) and Andridge *et al.* (2019) have made the R code for its computation freely available and documented it very well (see https://github.com/bradytwest/IndicesOfNISB). Finally, it remains to be said that nowadays—in times of an ever increasing demand for quickly available data—data quality must not be lost sight of. Problems like non-ignorable selection bias are a serious issue especially for data that arise from non-probabilistic sampling methods and it must be considered for valid statements about an underlying target population.

ACKNOWLEDGEMENTS

This paper uses data from the German General Social Survey (GGSS): DOI: 10.4232/1.13749, and the German Socio-Economic Panel (SOEP): DOI: 10.5684/soep.core.v37eu. We want to thank the market and opinion research company for letting us use their data as example for a true non-probability sample. Open Access funding enabled and organized by Projekt DEAL.

Notes

[1]Strictly speaking, ignorability and SAR are not completely equivalent terms. While SAR is a necessary condition for the former, ignorability additionally requires that the parameters of the analysis and missing data model are distinct. As the latter condition is usually less critical, we will use the terms interchangeably in this paper. Same applies to non-ignorability and SNAR.

[2]For our calculations in Section 3 and 4, we used the R code that is publicly available on Brady West's GitHub site https://github.com/bradytwest/IndicesOfNISB.

[3]Note that this definition of the standardised difference is slightly different to earlier papers, for example, Andridge & Little (2020), where the overall mean of X is used instead of the mean of the non-selected cases. Because our outcome variables are binary we used the definition provided in Andridge *et al.* (2019).

[4]The intervals for the unweighted means were calculated by Wilson's score method (Wilson, 1927).

[5]For the weighted means, we used the traditional Wald-type interval in combination with the effective sample size. Please note that the computation of effective sample sizes is at most a crude approximation in the context of non-probability samples. The provided effective sample sizes were 5,655, 15,147, 10,412, and 9,024 for Questions 1 to 4, respectively.

[6]Examining frequentist uncertainty intervals alongside Bayesian credible intervals is generally straightforward, particularly when employing non-informative priors, as highlighted in studies such as Gray *et al.* (2015).

References

American Association for Public Opinion Research 2013. Report of the AAPOR Task Force on Non-Probability Sampling. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Andridge, R.R. (2009). Statistical methods for missing data in complex sample surveys. Ph.D. Thesis, University of Michigan.

Andridge, R.R. (2024). Using proxy pattern-mixture models to explain bias in estimates of COVID-19 vaccine uptake from two large surveys. *J. R. Stat. Soc. Ser. A: Stat. Soc.*, qnae005.

Andridge, R.R. & Little, RJA (2009). Extensions of Proxy Pattern-Mixture Analysis for Survey Nonresponse. In *Joint Statistical Meetings (JSM) Proceedings, Section on Survey Research Methods*, pp. 2468–2482.

Andridge, R.R. & Little, RJA (2011). Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Stat.*, **27**(2), 153–180.

Andridge, R.R. & Little, RJA (2020). Proxy pattern-mixture analysis for a binary variable subject to nonresponse. *J. Off. Stat.*, **36**(3), 703–728.

Andridge, R.R., West, B.T., Little, RJA, Boonstra, P.S. & Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)*, **68**(5), 1465–1483.

Boonstra, P.S., Little, RJA, West, B.T., Andridge, R.R. & Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *J. Off. Stat.*, **37**(3), 751–769.

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

Elliot, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Surv. Pract.*, **2**(6), 2982.

Elliott, M.R. & Valliant, R. (2017). Inference for nonprobability samples. *Stat. Sci.*, **32**(2), 249–264.

Fahimi, M., Barlas, F.M., Thomas, R.K. & Buttermore, N. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. *Surv. Pract.*, **8**(5), 1–11.

GESIS - Leibniz-Institut für Sozialwissenschaften 2021. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUScompact - Kumulation 1980-2018. GESIS Datenarchiv, Köln. ZA5275 Datenfile Version 1.1.0, https://doi.org/10.4232/1.13749

Giesselmann, M., Bohmann, S., Goebel, J., Krause, P., Liebau, E., Richter, D., Schacht, D., Schröder, C., Schupp, J. & Liebig, S. (2019). The individual in context(s): Research potentials of the socio-economic panel study (SOEP) in sociology. *Eur. Sociol. Rev.*, **35**(5), 738–755.

Goebel, J., Grabka, M.M., Liebig, S., Kroh, M., Richter, D., Schröder, C. & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, **239**(2), 345–360.

Gray, K., Hampton, B., Silveti-Falls, T., McConnell, A. & Bausell, C. (2015). Comparison of Bayesian credible intervals to frequentist confidence intervals. *J. Modern Appl. Stat. Methods*, **14**(1), 43–52.

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

Little, RJA (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**(3), 471–483.

Little, RJA, West, B.T., Boonstra, P.S. & Hu, J. (2019). Measures of the degree of departure from ignorable sample selection. *J. Surv. Stat. Methodol.*, **8**, 932–964.

Olsson, U., Drasgow, F. & Dorans, N.J. (1982). The polyserial correlation coefficient. *Psychometrika*, **47**(3), 337–347.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Socio-Economic Panel (SOEP) 2022. Data for years 1984–2020, SOEP-Core v37, EU Edition.

Sullivan, D. & Andridge, R. (2015). A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. *Comput. Stat. Data Anal.*, **82**, 173–185.

Tate, R.F. (1955). The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, **42**(1-2), 205–216.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *J. Survey Stat. Methodol.*, **8**(2), 231–263.

Valliant, R., Dever, J.A. & Kreuter, F. 2018. Nonprobability sampling. In *Practical Tools for Designing and Weighting Survey Samples*, Springer, pp. 565–603.

West, B.T. & Andridge, R.R. (2023). Evaluating pre-election polling estimates using a new measure of non-ignorable selection bias. *Public Opin. Quart.*, **87**(S1), 575–601.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.*, **22**(158), 209–212.

[*Received February 2023; accepted July 2024*]