Universiteit
Leiden
The Netherlands

Estimating the bias and variance of imputed population totals

Mürk, Anete

Citation
Mürk, A. (2025). *Estimating the bias and variance of imputed population totals*.

Version:	Not Applicable (or Unknown)
License:	License to inclusion and publication of a Bachelor or Master Thesis, 2023
Downloaded from:	https://hdl.handle.net/1887/4258422

Note: To cite this publication please use the final published version (if applicable).

Estimating the bias and variance of imputed population totals

Anete Mürk

Thesis advisor: Dr. Sander Scholtus, Statistics Netherlands
Thesis advisor: Dr. Arnout van Delden, Statistics Netherlands
Thesis advisor: Dr. Julian Karch, Leiden University

Defended on 7 July, 2025

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

Official statistics are increasingly being produced by integrating different data sources, which speeds up production while reducing costs and alleviating the response burden on participants. A commonly used approach is to integrate survey and administrative data by mass imputing a missing target variable in administrative data based on a model trained on survey data.

Accuracy estimation of statistical output based on imputed data, however, remains an open problem, and several approaches have been proposed. The design-based variance estimator by Scholtus and Daalmans (2021) is limited to estimating imputation-specific error in categorical target variables and cannot estimate bias. The global mean squared error ($GMSE$) estimator (Alleva et al., 2021; Deliu et al., 2025) can address these issues but may potentially undermine the extent of sampling error and be vulnerable to model misspecification.

Both approaches require the derivation of formulae specific to parametric imputation models, which can get computationally expensive. In this thesis, a new method is proposed upon adapting the misclassification error framework by van Delden et al. (2016) to evaluate the bias and variance of statistical output based on an imputed categorical variable. This method will be termed the prediction error modelling approach (PEM) and can provide fast estimates without access to a parametric model form.

This thesis aims to evaluate the robustness of the three approaches in estimating the bias and/or variance of statistical output based on a mass-imputed categorical target variable.

A simulation study demonstrated that, under varying amounts of bias and variance, all estimators remained robust in estimating the variance of population totals. Furthermore, the estimators remained relatively robust even when their assumptions were violated. Meanwhile, PEM was the only approach capable of estimating bias, albeit to a limited extent. The main results of the simulation study were replicated in a case study based on the Dutch Educational Attainment file. However, a tendency was observed for the design-based variance estimator to provide more conservative estimates. In addition, PEM provided unreliable estimates in extremely small domains.

The study demonstrated that $GMSE$ is a robust alternative to the design-based variance estimator. On the other hand, it was shown that the accuracy of imputed estimates can be quantified using the relatively simple PEM approach. Future research is encouraged to explore robustness under more complex sampling designs and the estimation of non-sampling errors beyond imputation-specific errors. Additionally, estimating bias remains an important challenge.

Contents

1 Introduction **5**

2 Background **8**

2.1 Qualitative assessment of error in the mass-imputed estimator 8

2.2 Quantitative assessment of error in the mass-imputed estimator 10

2.2.1 Setup and notation 10

2.2.2 MSE$\left(\hat{Y}_U\right)$ 12

2.3 Existing approaches to approximate MSE($\hat{Y}_U$) 15

2.3.1 V_{design} 15

2.3.2 $GMSE$ 16

3 Prediction error modelling approach **18**

3.1 Estimating the bias and variance of statistical output under misclassification error 18

3.2 Estimating the bias and variance of the mass-imputed estimator under imputation error . 19

4 Simulation study **23**

4.1 Methods 23

4.1.1 Simulation procedure 23

4.1.2 Benchmark estimators 24

4.1.3 Relative estimates 24

4.1.4 The data 25

4.1.5 The experimental conditions 26

4.2 Results 31

4.2.1 Rationale for performance evaluation 31

4.2.2 Change in benchmark estimators 31

4.2.3 Performance in the baseline condition 32

4.2.4 Performance in the multinomial condition 36

4.2.5 Performance in the variance conditions 38

4.2.6 Performance in the bias and interaction conditions 40

4.2.7 Conclusions from the simulation study 42

5 Case study **44**

5.1 Methods 44

5.1.1 The target variable 44

5.1.2 Setup 45

5.1.3 Imputation models 46

5.2 Results . . . 46
5.2.1 The whole population . . . 48
5.2.2 Internal domain - income class . . . 49
5.2.3 External domains . . . 50
5.2.4 Computational complexity . . . 52

6 Discussion **53**
6.1 Key findings . . . 53
6.2 Strengths and Limitations . . . 54
6.3 Conclusion . . . 55

A Code **61**

B Details on estimating V_{design} for binomial and multinomial model **62**
B.1 Logistic regression . . . 62
B.2 Multinomial logistic regression . . . 62

C Details on estimating GMSE for binomial and multinomial target variables **64**
C.1 Multinomial logistic regression . . . 64
C.2 Logistic regression . . . 65

D Derivation of PEM **66**
D.1 Binary target variable . . . 66
D.2 Multinomial target variable . . . 69

E True and estimated domain sizes across superpopulation 1 and 2 **71**

F CSRMs for benchmarks and accuracy estimators across the experimental conditions with the binomial target variable **74**

G CSRMs for benchmark and accuracy estimators in the multinomial condition **79**

H Benchmark estimators across the sampling distribution **81**

I Simulation results for the female subdomain **82**

Chapter 1

Introduction

National Statistical Institutes (NSIs) strive to produce high-quality official statistics that provide reliable insights into the development and well-being of populations. The European Statistical Code of Practice (CoP) defines quality in terms of five principles: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity (Eurostat, 2018). This thesis focuses on evaluating the accuracy of statistical output.

Quantifying the accuracy of statistical output has become more complicated due to the increasing use of multiple data sources in the production of official statistics (Ascari et al., 2020). Traditionally, population means and totals, alongside their variances, have been estimated based on purposefully designed probability surveys. Over time, official statistics has moved beyond using survey data to integrating available data when estimating population parameters (Tillé et al., 2022). For example, several countries have abandoned dedicated census questionnaires, combining pre-existing surveys and administrative records to describe the population instead (Daalmans, 2017; S. Falorsi, 2017; Lundy, 2022). Combining available data sources enables a quicker production of statistics while reducing production costs and the response burden. However, estimating population parameters based on multiple data sources introduces additional errors in the estimates that traditional survey methods do not account for.

Mass imputation is a commonly used technique to estimate population totals based on integrated sample and administrative data (De Waal et al., 2011). It involves imputing the target variable in the population based on a model trained on sample data, resulting in a complete microdata file where the totals can be estimated by simply counting the units. This technique is used by Statistics Netherlands for the estimation of specific business statistics and, more recently, in the estimation of educational attainment as part of the Dutch Virtual Population and Housing census (e.g., Daalmans, 2017; De Waal and Daalmans, 2018). Mass imputation is also a key technique used in the Italian Permanent Census, which can now be produced annually based on a system of integrated survey and administrative data, replacing the previously decennial census survey (S. Falorsi, 2017).

Estimation of accuracy, or the bias and variance of imputed totals, hereby termed the mass-imputed estimator, however, is a complex task due to the need to account for various sources of error that occur during the process. Traditional methods in sampling theory only account for the error due to random sampling of the units. In contrast, the mass-imputed estimator contains additional errors resulting from the estimation of model parameters and the imputation process.

While standard methods, such as multiple imputation and bootstrapping, can be adapted for the estimation of accuracy in this case, alternatives are sought to address the shortcomings of these methods, including the creation of multiply imputed microdata files and the computational burden associated with resampling (Scholtus & Daalmans, 2021). A faster but more methodologically complicated approach

is to derive a formula for the mean squared error (MSE) of the mass-imputed estimator. Two kinds of formulae have been discussed in the literature that estimate either the variance or the MSE of the mass-imputed estimator (Alleva et al., 2021; Deliu et al., 2025; Scholtus & Daalmans, 2021). Due to differences in the assumptions made in the deriving the formulae, each approach has its unique advantages and disadvantages. Furthermore, the different approaches have not been compared, which makes it unclear which methodological choices are best to adopt in practice and for further research regarding the evaluation of the accuracy of the mass-imputed estimator.

Scholtus and Daalmans (2021) derived a formula to estimate the variance of the mass-imputed estimator based on a categorical target variable. This approach, hereby termed V_{design}, is effective if detailed information on the sampling design is available and all domain variables, or the subgroups where the totals are estimated, and their interactions, are included in the imputation model. However, these conditions can be challenging to fulfill in practice when sampling designs become complicated, imputation models become large, and estimation proceeds across highly granular domain variables, such as all cities in the Netherlands. When these conditions are not fulfilled, both the mass-imputed estimator and the estimated variance can become biased. The second formula, first discussed in Alleva et al. (2021) and further developed in Deliu et al. (2025), can address these issues. Their approach termed the global or generalised mean squared error ($GMSE$), approximates the total effect of the bias and variance of the mass-imputed estimator and can adapt to both numerical and categorical target variables. Furthermore, $GMSE$ represents a methodological approach that can be extended to account for errors beyond the immediate imputation process, such as coverage and non-response errors affecting administrative records and surveys. This is a significant advancement given that the methods have so far been developed under the assumption that these errors are negligible.

However, simplifications are made during the derivation of $GMSE$ that might not always hold well in practice. Namely, it is assumed that the sampling design can be ignored when modelling the target variable and that the imputation model is correctly specified with respect to the target variable. These assumptions can be met with many commonly used sampling designs upon careful modelling. On the other hand, $GMSE$ might not adequately account for other commonly used sampling designs that oversample certain groups of individuals. Furthermore, misspecification of the imputation model can occur easily since, in practice, one never has access to all possible variables that influence the distribution of the target variable. The variance formula by Scholtus and Daalmans (2021), on the other hand, accounts for all kinds of sampling designs without assuming a correctly specified imputation model, thereby presenting a good standard upon which to compare $GMSE$ in the case where the bias is negligible.

A third approach to estimating the bias and variance of the mass-imputed estimator is proposed in this thesis. Namely, the methodology developed by van Delden et al. (2016) to evaluate the effect of misclassification error on the bias and variance of population totals can be adapted to evaluate the effect of imputation error. This method has never been evaluated in this context before but promises to be a fast and simple alternative to the other formulae that require complex unit-level matrix computations. This approach, hereby termed the prediction error modelling approach (PEM), consists of estimating the probabilities of correctly imputing the target variable in the population based on sample data. This corresponds to a simple estimation of the proportion of true positives and negatives, which, unlike the other two formulae, does not require access to the form of a parametric imputation model. The estimated probabilities are dependent on the other sources of error and can be estimated while taking the sampling design into account. In turn, this approach takes into account the same errors as $GMSE$ and the design-based approach.

This thesis aims to compare the robustness of the three approaches in evaluating the variance and, if appropriate, the bias of the mass-imputed estimator based on a categorical target variable. Robustness

hereby refers to the ability to correctly estimate the bias and/or variance of the mass-imputed estimator under varying levels of either. The results of this study will contribute to the growing literature on evaluating the accuracy of statistical output derived from integrated data.

Robustness will be evaluated in a simulation study using synthetic data. Specifically, the performance of the different approaches is compared across various experimental conditions that induce changes in bias and/or variance in the mass-imputed estimator due to different reasons that are likely to occur in practice, but may violate the assumptions of an accuracy estimator to some extent. Furthermore, it will be investigated if the results of the simulation study generalise to practice. For this, a case study will be conducted using sample data from the Labour Force Survey (LFS) and administrative data from the Dutch Educational Attainment file (EAF).

The remainder of the thesis is structured as follows. Chapter 2 provides the background regarding the errors affecting the mass-imputed estimator and the existing approaches to quantify them. Chapter 3 outlines the novel PEM method. Sections 4 and 5 describe the methods and results of the simulation and the case study respectively. Section 6 concludes with a discussion.

Chapter 2

Background

The accuracy of estimates in official statistics is expressed as a function of sampling and non-sampling error they contain (Eurostat, 2020). Non-sampling error refers to any error that arises from sources other than the random variation introduced by selecting a sample according to a defined sampling design. Estimates based on multiple data sources involve a complex production chain, which introduces error from the individual data sources and integration strategies in the final statistic (Rocci et al., 2022; Zhang, 2012). This section begins with an explanation of the errors that need to be quantified during the evaluation of the accuracy of the mass-imputed estimator. It then proposes a suitable measure of accuracy and discusses existing methods for approximating it.

2.1 Qualitative assessment of error in the mass-imputed estimator

Error frameworks are commonly used tools that facilitate the quantitative evaluation of error by identifying all possible sources of error. A suitable framework for the mass-imputed estimator is the Total Process Error Framework (TPE) described in Rocci et al. (2022), which builds on the seminal two-phase framework by Zhang (2012). Table 1 outlines the adaptation of the TPE for the case of the mass-imputed estimator.

Table 1: Qualitative assessment of error in the mass-imputed estimator according to the Total Process Error approach (Rocci et al., 2022)

Step	**Phase 1**	**Phase 2**	
		Phase 2a	**Phase 2b**
1	General quality assessment of the sample and the administrative data source. Detection and/or evaluation and/or correction of source-specific error.		
2		Quality assessment of the sample and the administrative data source with respect to target statistical product. Comparison of suitable data integration techniques. Determining integration-specific error.	
3			Quality assessment of the integrated data file. Evaluation of cumulative error in the imputed data file.

The TPE evaluates error in statistical output based on integrated sources in a step-wise manner across the different phases of integration. The purpose of the first phase in TPE is to evaluate the quality of the data as it is affected by source-specific error. This phase enables to answer the question: "How well does each source measure what they are supposed to measure?". In the case of the mass-imputed estimator, errors that need to be considered in this phase are specific to survey and administrative data. Common non-sampling errors associated with these sources are coverage error, measurement error and processing error (Zhang, 2012). Non-response error is also a growing issue affecting surveys (Beaumont, 2020).

The methods evaluated in this thesis simplify the error structure of the mass-imputed estimator by ignoring the source-specific non-sampling errors. Instead, they focus on the errors arising from the imputation procedure in phase 2. However, it is essential to note that, in practice, these errors cannot be easily neglected, as demonstrated by the fact that even small coverage errors can have a significant impact on population estimates (Meng, 2018). $GMSE$ can incorporate these errors, but has not been fully formalised in this regard.

The purpose of the second phase is to evaluate the error arising from the data integration process. It consists of two phases. Phase 2a identifies the error that needs to be evaluated based on a chosen integration strategy, while Phase 2b determines the cumulative error resulting from all the previous steps.

Phase 2a deserves further attention since it is essential to acknowledge that alternative strategies for integrating sample and administrative data without relying on mass imputation exist. These methods enable avoiding imputation-specific errors, and methods for assessing their accuracy have been developed. However, these alternative strategies have limitations in practice (De Waal et al., 2011, pp. 244–263; De Waal, 2016). They involve reweighting survey units based on administrative data, which can lead to unstable estimates in small domains, difficulties in combining survey weights with those in administrative

data (for an example using weighted administrative data, see Linder et al., 2011), and ensuring numerical consistency of cross-tabulations across several domains. Numerical consistency means that cross-classified totals remain the same across different tabulations—for example, totals by gender and employment status (e.g., employed/unemployed) are consistent whether estimated in a 2×2 or a more detailed, 5x2x2 table (e.g., region x gender x employment status). Mass imputation yields numerically consistent estimates by default, provided that the imputation model includes the relevant domain variables. On the other hand, if alternative methods are viable, they can be more efficient. More importantly, they are better for avoiding biased estimates in domains that were not included in the imputation model (Kooiman, 1998).

The non-sampling errors to be evaluated upon choosing mass imputation as an integration strategy stem from the estimation of the imputation model and imputing units in the administrative data (Scholtus & Daalmans, 2021). Estimation of the imputation model introduces model parameter error, which describes the variability in the model parameters estimated from sample data. For generalized linear models such as the multinomial regression model, this variability can be approximated by the negative inverse of the information matrix (e.g., Agresti, 2013). Model parameter error introduces variability in the predicted probabilities, which are a non-linear function of the model parameters (see Section 2.2 for details).

The predicted probabilities can immediately be used as imputed values, resulting in non-integer totals (e.g., Kim and Rao, 2012). This approach has the benefit of not introducing further errors in the mass-imputed estimator. In contrast, imputation can proceed in a stochastic manner, whereby the imputed value is drawn randomly based on the predicted probabilities. This approach was adopted during the Dutch census mentioned earlier and will be chosen for imputation in the current study. This error will be termed an imputation error. The benefit of stochastic imputation is that the imputed values will follow the modelled distribution of the target variable instead of falling directly on the decision boundary. Alternatively, non-parametric approaches such as random hot-deck imputation can be used, but these can become computationally expensive for frequent use at NSIs (Scholtus & Pannekoek, 2015).

2.2 Quantitative assessment of error in the mass-imputed estimator

Before formulating the quantitative assessment of accuracy of the mass-imputed estimator as a function of sampling and imputation-specific non-sampling errors, the notation specific to the setup of the mass-imputed estimator in the current thesis will be introduced.

2.2.1 Setup and notation

We consider a finite population $U = \{1, \ldots, N\}$ represented by the administrative dataset R with no coverage error, so that each unit $k \in U$ is recorded in R. A probability sample $s \subseteq U$ has been obtained by means of a known sampling design P. s has not been affected by selective non-response and can be described by the vector of first-order inclusion probabilities $\boldsymbol{\pi}_k = \{\pi_1, \ldots, \pi_N\}$ and second-order inclusion probabilities $\boldsymbol{\pi}_{k,l} = \{k, l = 1, \ldots, N\}$, altogether defining the set of all possible samples $S = \{s_1, \ldots, s_{|S|}\}$. The inclusion probabilities determine the vector $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_N\}$ characterising sample membership, such that $\lambda_k = 1$ if $k \in s$, and $\lambda_k = 0$ otherwise.

For each $k \in s$, the values of a categorical target variable $\boldsymbol{y} = \{y_{c,1}, \ldots, y_{c,k}\}$ are observed, where $y_{c,k} \in \{1, \ldots, C\}$ with each of the C classes being mutually exclusive. In addition, we observed for each $k \in s$ a set of covariates $\boldsymbol{x}_k = \{x_{k1}, \ldots, x_{kJ}\}$. It is assumed that $\boldsymbol{x_k}$ is also available for each $k \in R$, and

that $\boldsymbol{x}$ contains no measurement error. Similarly to sample membership, each unit belongs to a domain d with $d \in \{1, \ldots, D\}$, so that the domain membership is represented by $\gamma_{d,k} = 1$ if unit k belongs to domain d, and $\gamma_{d,k} = 0$ otherwise.

Two cases are considered regarding the target variable—namely, the multinomial and the binomial case. The binomial case is a special case of the multinomial distribution, with $C = 2$ and $y_{c,k} \in \{0, 1\}$. We are interested in estimating the total number of units of each category of $\boldsymbol{y}$ in domain d within the population U. Note that d may or may not be included among the covariates in $\boldsymbol{x_k}$.

The finite population parameter of interest is defined as

$$Y_{Ud,c} = \sum_{k \in U} \gamma_{d,k}\, y_{c,k}. \tag{2.1}$$

We want to estimate $Y_{U_{d,c}}$ from R. Since $\boldsymbol{y}$ is not observed in R, we want to impute it. Therefore, we have to estimate an imputation model for $\boldsymbol{y}$. Since $\boldsymbol{y}$ is a categorical variable, the model can be defined as:

$$\begin{aligned} y_{c,k} &= \tilde{y}_{c,k} + e_{c,k}, \\ \tilde{y}_{c,k} &= f(\boldsymbol{x}_k; \vartheta), \\ f(\boldsymbol{x}_k; \vartheta_m) &= \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{C-1} \exp(\boldsymbol{x}_k^T \boldsymbol{\beta}_j)} & \text{for } c = C, \\ \dfrac{\exp(\boldsymbol{x}_k^T \boldsymbol{\beta}_c)}{1 + \sum_{j=1}^{C-1} \exp(\boldsymbol{x}_k^T \boldsymbol{\beta}_j)} & \text{for } c = 1, \ldots, C-1, \end{cases} \\ \text{reducing to} \quad f(\boldsymbol{x}_k; \vartheta_b) &= \frac{1}{1 + \exp(-\boldsymbol{x}_k^T \boldsymbol{\beta})} \quad \text{if } C = 2. \end{aligned} \tag{2.2}$$

In both cases, the error term $e_{c,k}$ is defined as:

$$e_{c,k} = \begin{cases} 1 - \tilde{y}_{c,k} & \text{with probability } \tilde{y}_{c,k} \\ -\tilde{y}_{c,k} & \text{with probability } 1 - \tilde{y}_{c,k} \end{cases}$$

so that $y_{c,k} \in \{0, 1\}$ with $P(y_{c,k} = 1 \mid \tilde{y}_{c,k}) = \tilde{y}_{c,k}$.

The function $f(\boldsymbol{x}_k; \vartheta_m)$ describes a multinomial logistic regression with parameters $\vartheta_m = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{C-1}^T)^T$. In the binomial case, $f(\boldsymbol{x}_k; \vartheta_b)$ describes a logistic regression with parameters $\vartheta_b = (\beta_1, \ldots, \beta_J)^T$. Note that in both cases, $\sigma^2_{y,c,k} = V(e_{c,k}) = \tilde{y}_{c,k}(1 - \tilde{y}_{c,k})$, and $\mathrm{Cov}(e_{c,k}, e_{c,r}) = 0$ for $k \neq r$.

Sample data is used to estimate the model parameters ϑ. Let the vectors $\hat{\vartheta}_m = (\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_{C-1}^T)^T$ and $\hat{\vartheta}_b = (\hat{\beta}_1, \ldots, \hat{\beta}_J)^T$ denote the estimated model parameters in the multinomial and binomial case, respectively. For each person $k \in R$, the imputation model is used to draw a vector $(\hat{y}_{1k}, \ldots, \hat{y}_{Ck})$ with predicted probabilities $(\hat{p}_{1k}, \ldots, \hat{p}_{Ck})$ such that exactly one of the values $\hat{y}_{c,k}$ is 1 and the others are 0. The imputed values can therefore be defined as:

$$
\begin{aligned}
\hat{y}_{c,k} &= \hat{p}_{c,k} + \hat{e}_{c,k}, \\
\hat{p}_{c,k} &= f(\boldsymbol{x}_k; \hat{\vartheta}),
\end{aligned}
$$

$$
f(\boldsymbol{x}_k; \hat{\vartheta}_m) = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{C-1} \exp(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_j)} & \text{for } c = C, \\ \dfrac{\exp(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_c)}{1 + \sum_{j=1}^{C-1} \exp(\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_j)} & \text{for } c = 1, \ldots, C-1, \end{cases} \tag{2.3}
$$

$$
\text{reducing to} \quad f(\boldsymbol{x}_k; \hat{\vartheta}_b) = \frac{1}{1 + \exp(-\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}})} \quad \text{if } C = 2,
$$

$$
\hat{e}_{c,k} = \begin{cases} 1 - \hat{p}_{c,k} & \text{with probability } \hat{p}_{c,k}, \\ -\hat{p}_{c,k} & \text{with probability } 1 - \hat{p}_{c,k}. \end{cases}
$$

The resulting estimator for $Y_{U_{d,c}}$ is:

$$
\hat{Y}_{U_{d,c}} = \sum_{k \in R} \gamma_{d,k} \hat{y}_{c,k} \tag{2.4}
$$

Note that in the following subsections, the domain and class subscripts d and c are sometimes omitted to simplify notation.

2.2.2 MSE$\left(\hat{Y}_U\right)$

Estimators in survey methodology are often evaluated in terms of their variance since inclusion probabilities enable the construction of unbiased estimates. However, estimators that involve modelling are typically biased (Lohr, 2021, Chapter 4). Hence, an ideal measure of accuracy for the mass-imputed estimator incorporates both bias and variance. The MSE of the mass-imputed estimator is formulated as

$$
\begin{aligned}
\text{MSE}\left(\hat{Y}_U\right) &= \mathbb{E}\left[\left(\hat{Y}_U - Y_U\right)^2\right] \\
&= \left\{\mathbb{E}\left[\hat{Y}_U - Y_U\right]\right\}^2 + \text{Var}\left(\hat{Y}_U - Y_U\right) \\
&= \left\{\text{Bias}\left(\hat{Y}_U\right)\right\}^2 + \text{Var}\left(\hat{Y}_U - Y_U\right).
\end{aligned} \tag{2.5}
$$

The bias and variance of an estimator describe the systematic deviation of the estimate from its true value and variability around its expected value, respectively. The bias and variance are evaluated in expectation based on the assumed distribution that generated the target variable. The assumptions regarding the distribution of the target variable define the frame of inference, which affects how the different sources of error in the mass-imputed estimator are quantified. In survey methodology, two types of inference are commonly used—design-based and model-based inference (see Lohr, 2021 and Beaumont and Haziza, 2022 for a recent comprehensive review). The design-based inference treats y as fixed, whereas model-based inference treats it as random. The three approaches under comparison in this thesis differ in their frames of inference. V_{design} reflects a design-based approach, while $GMSE$ reflects a combination

of design- and model-based approach. The new PEM approach will be formulated under design-based assumptions.

Each frame of inference has its advantages and disadvantages when it comes to describing the accuracy of statistical output. Specifically, the frame of inference can have a significant impact on the reliability and accuracy of the output. The following subsections will provide further explanation and comparison of the concepts. While design-based inference is the standard in the majority of official statistics, the discussion begins with model-based inference since this aligns with the traditional frequentist framework. Subsections 2.2.2.1 and 2.2.2.2 are primarily based on various chapters in Lohr (2021) and Beaumont and Haziza (2022). Seminal work on the model-based approach can be found in Valliant et al. (2000).

2.2.2.1 Model-based inference

To estimate a population value, one needs to infer the unobserved values of $\boldsymbol{y}$ in the population from s. Under the model-based approach, it is assumed that $\boldsymbol{y}$ in the finite population is randomly generated by superpopulation M, so that $\boldsymbol{y} \sim M$. In the current setup, $M = f(\boldsymbol{x}_k; \vartheta)$ and $V_M(\boldsymbol{y}) = \sigma^2$. The sampled $\boldsymbol{y}$ are used to estimate ϑ to predict the unobserved $\boldsymbol{y}$ in the population.

The MSE of the mass-imputed estimator under model-based inference is defined as:

$$\mathbb{E}_M \left[(\hat{Y}_U - Y_U)^2 \mid S = s \right] = \left\{ \mathbb{E}_M \left[\hat{Y}_U - Y_U \mid S = s \right] \right\}^2 + \mathrm{V}_M \left(\hat{Y}_U - Y_U \mid S = s \right) \tag{2.6}$$

Note that subscript M implicitly refers to conditioning on the obtained sample, which will be left out from the notation of the following formulae.

In model-based inference, S is treated as fixed, while $\boldsymbol{y}$ varies according to $V_M(\boldsymbol{y}) = \sigma^2$. If one assumes that M can be perfectly described by a multinomial logistic regression model with parameter estimates $\hat{\boldsymbol{\vartheta}}_m = \left(\hat{\boldsymbol{\beta}}_1^\top, \ldots, \hat{\boldsymbol{\beta}}_{C-1}^\top \right)^\top$, then the bias is zero, and the $\mathrm{V}_M \left(\hat{Y}_{U_d} - Y_{U_d} \right)$ corresponds to the combination of model parameter error, imputation error, and model error. Model error refers to the variability of the finite population total $V_M(Y_{U_d})$, model parameter error refers to $V_M(\hat{\vartheta})$ and imputation error reflects the random drawing of integer values according to $V_M(e)$.

Overall, the model-based approach aligns well with the mass-imputation framework. However, it relies on a strong assumption that the superpopulation is correctly estimated from the sample data. While this assumption can be reasonable with careful modelling, it can become problematic during periods of societal or structural change (Hansen et al., 1983). In such contexts, infrequently conducted surveys may fail to capture current population characteristics, leading to misspecified models and, in turn, unreliable estimates. Nevertheless, model-based approaches are key to inference in some areas of official statistics—such as small area estimation (e.g., Rao & Molina, 2015) and when working with non-probability data (e.g., Rao, 2021)—where traditional design-based methods become unreliable or infeasible.

2.2.2.2 Design-based inference

Design-based inference, formally defined as randomization theory, treats $\boldsymbol{y}$ as fixed and S as random. The unobserved values can be inferred based on $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_{k,l}$, which gives information on the values that could have been observed had we obtained a different sample. Instead of $\boldsymbol{y} \sim M$, we have $\boldsymbol{y} \sim P$.

The MSE of the mass-imputed estimator under design-based inference is defined as

$$E_P \big[\hat{Y}_U - Y_U \big]^2 = \big\{ E_P \big[\hat{Y}_U - Y_U \mid y = y_k \big] \big\}^2 + V_P \big(\hat{Y}_U - Y_U \mid y = y_k \big). \tag{2.7}$$

where subscript P refers to conditioning on the fixed $\boldsymbol{y}$ in the finite population.

According to the design-based approach, the imputation model is a working model leveraging the relationships between $\boldsymbol{y}$ and $\boldsymbol{x_k}$ without assuming that these are true on a superpopulation level. Put differently, the imputation model is only a tool for predicting the missing values in the finite population, not for simultaneous inference regarding $\boldsymbol{y}$ in the superpopulation. $\hat{\vartheta}$ describes the relationships between $\boldsymbol{y}$ and $\boldsymbol{x_k}$ according to P and can be estimated based on the pseudo-maximum likelihood approach (Chambers & Skinner, 2003, pp. 22–26).

Bias according to the design-based approach occurs when the inclusion probabilities are not taken into account when estimating the model parameters and estimation proceeds in domains not included in the imputation model, hereby termed as external domains. Kim and Rao (2012) define the relative design-bias of the mass-imputed estimator as

$$\mathrm{RB}(\hat{Y}_{U_d}) = \frac{\mathrm{cov}(\gamma_{d,k}, r_k)}{\hat{Y}_{U_d}}, \tag{2.8}$$

where $r_k = y_k - f(x_k; \hat{\vartheta})$ is the residual. Design-bias occurs for external domains due to residual correlation with the domain variable. Typically, small bias can be ignored in practice, which presents when the domain variable and the target variable are unrelated (Kim & Rao, 2012). Note that if the latter is not the case, the mass-imputed estimator can be significantly biased from both the model-based and the design-based perspective.

Since variance under the design-based approach is evaluated with respect to P, a misspecified imputation model will not affect validity of the variance estimate. This is because it is not assumed that the distribution of $\boldsymbol{y}$ given the imputation model corresponds to the actual distribution of $\boldsymbol{y}$. Finally, under negligible design-bias, $V_P\big(\hat{Y}_U - Y_U\big)$ consists of sampling error $V_P(Y_U)$, model parameter error $V_P(\hat{\vartheta})$ and imputation error $V_P(e)$.

The design-based approach has the advantage that, even under model misspecification, confidence intervals remain valid—ensuring that the true finite population value falls within the interval at the nominal confidence level (Lohr, 2021, p. 438). In turn, assuming that there is no selective non-response affecting the sample, inference is reliable even if the model is wrong. However, if the model is correct, design-based variance estimates can be more conservative in comparison to model-based estimates (Beaumont & Haziza, 2022).

2.2.2.3 Design- and model-based inference

Based on the preceding discussion, it can be hypothesised that perhaps the most robust estimates of population values may be obtained by accounting for both the sampling and the model distribution of the target variable. This idea underlies the concept of anticipated variance, which refers to the variance of an estimator with respect to the joint distribution of M and P (Isaki & Fuller, 1982). Anticipated variance was introduced to facilitate the development of sampling designs that acknowledge the relationships between the target variable and covariates. Subsequent work has described algorithms capable of finding the optimal inclusion probabilities to reduce the anticipated variance of the desired estimate (e.g., P. D. Falorsi & Righi, 2015).

$GMSE$ builds on the concept of anticipated variance to evaluate the MSE of the mass-imputed estimator (Alleva et al., 2021). $GMSE$ is defined as

$$E_P E_M\big(\hat{Y}_U - Y_U\big)^2. \tag{2.9}$$

Equation 2.9 evaluates sampling and model error, in addition to model parameter and imputation error with respect to the joint distribution of P and M. The advantage of this approach is that it enables to comprehensively evaluate errors in the mass-imputed estimator without ignoring any source of variability. Furthermore, Alleva et al. (2021) propose that $GMSE$ is suitable for evaluating source-specific errors such as coverage and measurement error upon extending the concatenation of expectations with respect to the additional error, e.g.

$$E_{NR}E_PE_M\big(\hat{Y}_U - Y_U\big)^2 \tag{2.10}$$

where E_{NR} represents the expectation with respect to a distribution explaining non-response, which could be summarized with a model.

However, simplifying assumptions are made during derivation of the exact equation for approximating Equation 2.9, which might not guarantee robust estimation in several practical scenarios. These will be discussed in the proceeding sections detailing how $GMSE$ and the other two approaches propose to approximate the MSE functions described above.

2.3 Existing approaches to approximate MSE($\hat{Y}_U$)

Common approaches to evaluating the accuracy of estimators based on complex functions in survey methodology include Taylor linearisation and resampling or bootstrapping (Lohr, 2021, Chapter 7). Linearisation-based methods estimate the function with small-order derivatives of its parameters. While computationally efficient, deriving the linearised equation can become difficult and requires access to the exact form of the imputation model. Furthermore, approximation by linearisation may not be effective in small samples. Resampling-based approaches, on the other hand, generate an empirical distribution of the estimator based on replicating the steps that affect its variability a large number of times. The bias and variance of the estimator can then be assessed based on the empirical distribution. Resampling is relatively easy and reliable, but it can incur an unacceptably large computational burden for statistical offices that are interested in performing these computations on a daily basis.

Linearised equations for evaluating the variance component in Equation 2.7 and the MSE as defined in Equation 2.9 have been derived by Scholtus and Daalmans (2021) for models satisfying Equations 2.2 defined earlier in Subsection 2.2.1, and by Alleva et al. (2021) (see also Deliu et al., 2025) for a wide range of models. Design-based variance estimation of imputed population totals has also been discussed in the context of combining several samples through mass imputation (e.g., Chipperfield et al., 2012; Golini & Righi, 2024; Kim & Rao, 2012; Kim et al., 2021). However, because integrating sample and administrative data differs from integrating two samples, the work is not directly applicable to the mass-imputed estimator as defined in the current project.

2.3.1 V_{design}

In the case where design-bias in the mass-imputed estimator is negligible, Equation 2.7 reduces to the variance component, which can be approximated according to the following equation:

$$V_{\text{design}}(\hat{Y}_{U_{d,c}}) = E\left\{\sum_{k\in R}\gamma_{d,k}\hat{p}_{c,k}(1-\hat{p}_{c,k})\right\} + \sum_{k\in R}\sum_{l\in R}\gamma_{d,k}\gamma_{d,l}\,\text{cov}\{\hat{p}_{c,k},\hat{p}_{c,l}\}. \tag{2.11}$$

derived by Scholtus and Daalmans (2021). They derived the formula using the law of total variance,

expressing $V_P(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}})$ as

$$E_P\left[V_P\left(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}} \mid S\right)\right] + V_P\left[E_P\left(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}} \mid S\right)\right]. \tag{2.12}$$

which simplifies to Equation 2.11 assuming that each unit is independently imputed, the target variable follows the multinomial distribution, and that $\boldsymbol{y}$ is fixed. The first term of the equation represents the average error of the mass-imputed estimator within different realisations of the sampling design, i.e., the variance not explained by the sampling design. This corresponds to imputation error under the assumption that $E_P(\hat{y}_{c,k} \mid S) = \hat{p}_{c,k}$ and $V_P(\hat{y}_{c,k} \mid S) = \hat{p}_{c,k}(1 - \hat{p}_{c,k})$. The second term reflects the variability in error between the different realisations of the sampling design, i.e., the variance explained by the sampling design. This corresponds to sampling error and model parameter error.

Equation 2.11 is consistent if the imputation model contains only categorical variables, the domain variables, any variables used to define the sampling design and all higher-order interactions of these variables. According to Scholtus and Daalmans (2021), the equation will hold reasonably well for large samples if the imputation model contains the interactions of the domain variable and any variables used to define the sampling design. Evaluating Equation 2.11 based on binomial and multinomial models is described in Appendix B.

Scholtus and Daalmans (2021) compared the analytical approach described in Equation 2.11 to a bootstrap approach using a simulation study. They found that both approaches were good in approximating the simulated true variance of the mass imputed estimator, while the analytical approach was slightly more precise and imposed far less computational burden. Interestingly, it was reported that the analytical approach, in comparison to the bootstrap approach, underestimated variance in larger domains. According to the authors, this was due to the omission of higher-order interactions, which created bias in the cross-tabulated estimates, becoming more notable as the number of affected units grew. On the other hand, modelling the interactions caused the analytical approach to severely overestimate variance in smaller domains due to increased model parameter error.

It should be noted that V_{design} can be estimated based on multiple imputation (Rubin, 1987). However, besides the burden of having to store the multiply imputed micro-data files, Scholtus and Daalmans (2021) further explained that multiple imputation is undesirable to develop in the case of mass-imputation as it is not broadly applicable to more complex population estimators since one can only evaluate the variance of the multiple imputation estimator.

2.3.2 *GMSE*

Equation 2.9 for the evaluation of the MSE of the mass-imputed estimator was approximated in Alleva et al. (2021) by

$$\text{GMSE}(\hat{Y}_{U_{d,c}}) \approx E_P\left[V_M\left(\hat{Y}_{U_{d,c}} \mid \lambda\right)\right] + V_M\left(Y_{U_{d,c}}\right) - 2\ \text{Cov}_M\left[E_P\left(\hat{Y}_{U_{d,c}} \mid y\right), Y_{U_{d,c}}\right]. \tag{2.13}$$

This expression results from simplifying the squared expression upon adding and subtracting the mean:

$$E_P E_M\left(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}}\right)^2 = E_P E_M\left(\hat{Y}_{U_{d,c}} - E_P E_M\left(\hat{Y}_{U_{d,c}}\right) + E_P E_M\left(\hat{Y}_{U_{d,c}}\right) - Y_{U_{d,c}}\right)^2. \tag{2.14}$$

The first and dominant component reflects the variability of the mass-imputed estimator as both y and S are random. The second component refers to model error and the final term refers to the covariance of the mass-imputed estimator and the true finite population total under the model distribution. It is important to note that these components do not independently express the bias or variance. Bias and variance are approximated upon jointly considering all the components.

Alleva et al. (2021) specify that Equation 2.13 reduces to

$$\text{GMSE}(\hat{Y}_{U_{d,c}}) \approx E_P\left[V_M\left(\hat{Y}_{U_{d,c}} \mid \lambda\right)\right] - V_M\left(Y_{U_{d,c}}\right) \tag{2.15}$$

when the mass-imputed estimator is design-unbiased, i.e., $E_P\left(\hat{Y}_{U_{d,c}} \mid y\right) = Y_{U_{d,c}}$. This follows from $\text{Cov}_M\left(Y_{U_{d,c}}, Y_{U_{d,c}}\right) = V_M\left(Y_{U_{d,c}}\right)$.

Equation 2.13 does not yet account for imputation error due to stochastic imputation and hence needs to be extended as

$$\text{GMSE}(\hat{Y}_{U_{d,c}}) \approx \text{GMSE}(\hat{Y}_{U_{d,c}})_{\text{Eq.2.15}} + \sum_{k=1}^{N} \gamma_{d,k}\tilde{y}_{c,k}\left(1 - \tilde{y}_{c,k}\right) \tag{2.16}$$

(Deliu et al., 2025).

$GMSE$ is derived assuming that the estimated superpopulation parameters are unbiased. Recall from Sections 2.2.2.2 and 2.2.2.1 that in this case, bias under the model-based approach is zero and negligible under the design-based approach. This suggests that $GMSE$ might, in practice, give estimates that are very close to V_{design}. Larger differences may occur if the model is misspecified, leading to an inaccurate accuracy estimate. However, since Equation 2.13 involves expectation over the sample distribution, $GMSE$ could remain robust. On the other hand, $GMSE$ does not require that the model is fitted according to the pseudo-maximum likelihood approach. This is done under the assumption that the sampling design is ignorable, meaning that the distribution of the target variable is not influenced by the sampling design (Sugden & Smith, 1984). This occurs when inclusion probabilities are close to 0 or 1. In fact, in that case, model parameter error under regular maximum likelihood and pseudo-maximum likelihood-based approach converge (Chambers & Skinner, 2003, pp. 26). According to Alleva et al. (2021), this is common enough in practice, thereby justifying the assumption.

The authors suggest using the dominant component as an upper bound for estimating $GMSE$ in practice, which will be adopted in the current thesis. They propose a four-step linearisation strategy to evaluate this component, followed by a simplified two-step linearisation procedure for categorical outcomes in a more recent publication (Deliu et al., 2025). See Appendix C for linearised forms of 2.13.

$GMSE$ has been evaluated across two simulation studies by the authors (Alleva et al., 2021; Deliu et al., 2025). Alleva et al. (2021) found that the linearised estimator for $GMSE$ provided good approximations to the true simulated $GMSE$. It was also reported that the first component results in a downward approximation of the simulated $GMSE$ when the imputation model is design-biased. On the other hand, it was found that approximation improves regardless of design bias if the domain size increases. Similar findings were reported by (Deliu et al., 2025). Deliu et al. (2025) compared the upper bound and a bootstrap approach whilst varying the population size. It was found that the linearised approach was less precise in smaller domains and registries, regardless of design bias. Similarly to Scholtus and Daalmans (2021), the bootstrap estimator was less precise and more computationally intensive.

Chapter 3

Prediction error modelling approach

van Delden et al. (2016) (see also Burger et al., 2015 for earlier work on the topic) developed an approach to evaluate the effect of misclassification errors in administrative data on population totals. Misclassification error occurs when some categorical domain variable in the register contains measurement error. For example, in their case study, the focus was on evaluating the effect of misclassifications in the type of businesses on the total revenue estimated per business type.

The following sections outline the misclassification error approach and its adaptation to imputation error under design-based inference.

3.1 Estimating the bias and variance of statistical output under misclassification error

The misclassification error approach utilises the properties of the parametric bootstrap method (e.g., Tibshirani and Efron, 1993) to evaluate the bias and variance of a population total θ. A model is assumed to describe the distribution of misclassification error. Repeated sampling for $b = 1, \ldots, B$ times from that model produces the sampling distribution of the domain total as affected by misclassification error. Bias and variance of the domain total can be assessed based on this sampling distribution by

$$\begin{aligned} \widehat{Bias}_B(\hat{\theta}) &= m_B(\hat{\theta}) - \hat{\theta}, \\ \widehat{V}_B(\hat{\theta}) &= \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b - m_B(\hat{\theta}) \right)^2, \\ \text{where} \quad m_B(\hat{\theta}) &= \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b. \end{aligned} \tag{3.1}$$

In this case, the model assumed is

$$p_{gh} = P(\hat{s}_k = h \mid s_k = g) \tag{3.2}$$

where $s_k = g$ represents the true, but unknown, fixed domain category and $\hat{s}_k = h$ represents the observed, but random, domain category. The model thus describes the probabilities of observing domain category h given true domain category g. The model can be summarised by a transition matrix P with probabilities p_{gg} on the main diagonal and p_{gh} on the off-diagonals. P can be estimated from audit data, that is, a sample where both g and h are observed.

Furthermore, van Delden et al. (2016) showed that the bootstrap bias and variance of domain total $\hat{\theta}$ can equivalently be estimated by the following set of equations as $B \to \infty$:

$$\begin{aligned} \widehat{Bias}(\hat{\theta}) &= \sum_{k=1}^{N} \left\{ (\hat{p}_{hh} - 1)\hat{a}_{h,k} t_k + \sum_{\substack{g=1 \\ g \neq h}}^{H} \hat{p}_{gh} \hat{a}_{g,k} t_k \right\}, \\ \widehat{V}(\hat{\theta}) &= \sum_{k=1}^{N} \sum_{g=1}^{H} \hat{p}_{gh}(1 - \hat{p}_{gh}) \hat{a}_{g,k} t_k^2 \end{aligned} \tag{3.3}$$

where $\hat{a}_k = (\hat{a}_{1k}, \ldots, \hat{a}_{Hk})^T$ is a vector of indicator variables such that $\hat{a}_{h,k} = 1$ if $\hat{s}_k = h$ and 0 otherwise. Here, t_k is a numerical target variable.

3.2 Estimating the bias and variance of the mass-imputed estimator under imputation error

To adapt the misclassification error approach to the case of imputation error, misclassification probabilities can be modelled in the sense of imputation error. To this end, sample data is used to estimate the model

$$p_{gh} = P\big(\hat{y}_k = h \mid y_k = g\big), \quad (g \in \{1, \ldots, C\},\ h \in \{1, \ldots, C\}) \tag{3.4}$$

where p_{gh} is the prediction error probability. Below are shown the key steps for estimating the bias and variance based on prediction error probabilities for the multinomial target variable. Full steps of derivation of the formulae under the design-based inference are shown in Appendix D.

Let $\hat{p}_{c,k} = P(\hat{y}_{c,k} = 1 \mid x_k, s)$ be the predicted probability based on the imputation model. The bias of the mass-imputed estimator can be expressed as follows:

$$E_P\big(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}}\big) = E_P\left\{ \sum_{k \in U} \gamma_{d,k} \big(\hat{y}_{c,k} - y_{c,k}\big) \right\} = \sum_{k \in U} \gamma_{d,k} \big\{ E_P(\hat{p}_{c,k}) - y_{c,k} \big\} \tag{3.5}$$

under the assumption that $E_P\big[E(\hat{y}_{c,k} \mid s)\big] = E_P(\hat{p}_{c,k})$, resembling the deterministic imputation approach. For the variance, we can write analogously to Scholtus and Daalmans (2021):

$$\begin{aligned} V_P\big(\hat{Y}_{U_{d,c}} - Y_{U_{d,c}}\big) &= V_P\left\{ \sum_{k \in U} \gamma_{d,k} \big(\hat{y}_{c,k} - y_{c,k}\big) \right\} \\ &= E_P\left\{ \sum_{k \in U} \gamma_{d,k}\, V\big(\hat{y}_{c,k} - y_{c,k} \mid s\big) \right\} + V_P\left\{ \sum_{k \in U} \gamma_{d,k} \big[E(\hat{y}_{c,k} \mid s) - y_{c,k}\big] \right\} \\ &= \sum_{k \in U} \gamma_{d,k}\, E_P(\hat{p}_{c,k}) \big\{ 1 - E_P(\hat{p}_{c,k}) \big\} + V_P\left(\sum_{k \in U} \gamma_{d,k} \hat{p}_{c,k} \right) - \sum_{k \in U} \gamma_{d,k}\, V_P(\hat{p}_{c,k}) \end{aligned} \tag{3.6}$$

where the final term is negligible in practice.

In equations 3.5 and 3.6, assuming that unit-level probabilities can be well approximated by $p_{gh,d}$ or the average probability of correctly imputing a unit in domain d from the sample, substituting $p_{gh,d}$ in the equations results in the following approximation for design-based MSE

$$\mathrm{PEM}_{\mathrm{MSE}}\big(\hat{Y}_{U_{d,c}}\big) = \big(\widehat{\mathrm{Bias}}\big(\hat{Y}_{U_{d,c}}\big)\big)^2 + \widehat{V}\big(\hat{Y}_{U_{d,c}}\big)$$

where

$$\begin{aligned}\widehat{\mathrm{Bias}}\big(\hat{Y}_{U_{d,c}}\big) &\approx N_{dc}\big(p_{cc,dU} - 1\big) + \sum_{\substack{g=1\\g\neq c}}^{C} N_{dg} p_{gc,dU},\\ \widehat{V}\big(\hat{Y}_{U_{d,c}}\big) &\approx \frac{N}{n}\sum_{g=1}^{C} N_{dg} p_{gc,dU}(1 - p_{gc,dU})\end{aligned} \tag{3.7}$$

with

$$N_{dg} = \sum_{k\in U} \gamma_{d,k} y_{g,k}$$

In the binomial case, the above formulae reduce to

$$\begin{aligned}\widehat{\mathrm{Bias}}\big(\hat{Y}_{U_{d,c}}\big) &\approx N_{d1}(p_{11,dU} - 1) + N_{d0}(1 - p_{00,dU}),\\ \widehat{V}\big(\hat{Y}_{U_{d,c}}\big) &\approx \frac{N}{n}\left[N_{d1} p_{11,dU}(1 - p_{11,dU}) + N_{d0} p_{00,dU}(1 - p_{00,dU})\right]\end{aligned} \tag{3.8}$$

In practice, we can estimate $p_{gh,d}$ by

$$\hat{p}_{gh,dS} = \frac{\sum_{k\in S} \gamma_{d,k} w_k y_{g,k} \hat{p}_{h,k}}{\sum_{k\in S} \gamma_{d,k} w_k y_{g,k}} \tag{3.9}$$

where w_k is the sampling weight. Note that these weights can be ignored under simple random sampling.

The N_d terms can be estimated by

$$\hat{N}_{dgU} = \sum_{k\in U} \gamma_{d,k} \hat{p}_{g,k}$$

In essence, PEM represents a simplification of V_{design}, whereby the complex covariance of unit-level probabilities is replaced by estimating the variability of p_{gh}. Covariance of the class-specific prediction error probabilities as per Equation 3.4 is approximately zero since these are estimated from disjoint subsets of the sample.

It is unclear whether the above assumptions prove reasonable in practice, resulting in a robust estimate of MSE. The key difference between PEM and the other two estimators is that PEM simplifies the complex error structure associated with model parameters and unit-level probabilities. This is instead attempted to be captured though prediction error probabilities, which are dependent on this error structure, but it remains to be seen whether simplifying it in this manner is adequate. The key advantage arising from this simplification is that the estimation of the probabilities does not need knowledge of the imputation model form. Furthermore, the computation of the formulae is a simple sum, which is very fast in comparison to the unit-level matrix multiplications needed for the computation of the other approaches. However, alike the other approaches, it could be expected that a large sample size is important for the performance of PEM as this enables to estimate the prediction error probabilities more accurately.

Before proceeding with reporting from the simulation and case study, Table 2 summarising the key characteristics of the three approaches under evaluation is provided to facilitate a quick comparison of the different methods and their advantages and disadvantages.

Table 2: Summary of the accuracy estimators

Estimator	$V_{\mathbf{design}}$	**GMSE**	**PEM**
Measure of accuracy	Variance	MSE	Bias and variance
Frame of inference	Design-based	Design- and model-based	Design-based
Errors	Sampling error, model parameter error, imputation error	Sampling error, model parameter error, imputation error, model error	Imputation error
Method of estimation	Linearization	Linearization	Aggregation
Advantages	• Directly evaluates multiple errors • Robust to misspecification of the imputation model	• Approximates design-bias • Directly evaluates multiple errors • Methodology extendable to further errors • If model-based conditions hold, provides a simpler and less conservative estimate than V_{design}	• Does not require a parametric model form • Fast
Disadvantages	• Ignores design-bias • Requires knowledge of second-order inclusion probabilities • Possibly more conservative estimate	• Possibly oversimplifies complex sampling designs • Vulnerable to model misspecification	• Directly evaluates only imputation error • Possibly oversimplifies the error structure

Chapter 4

Simulation study

4.1 Methods

To evaluate the behaviour of the estimators under varying levels of bias and variance, a Monte Carlo (MC) simulation study was conducted. The simulation design, inspired mainly by Alleva et al. (2021) and Deliu et al. (2025), enables the evaluation of the performance of the accuracy estimators under both the joint and the sampling distribution. The study was implemented in R version 4.4.0 (R Core Team, 2021) using a modified version of the synthetic dataset *Samplonia* (Bethlehem, 2009, pp. 11–13), previously used by Scholtus and Daalmans (2021). Details of the simulation flow, modifications to the dataset, and experimental conditions are provided in subsections 4.1.1–4.1.5.

The link to all the code scripts for replication of the simulation study is provided in Appendix A. Note that the code for the computation of $GMSE$ was adapted from the materials provided in Deliu et al. (2025).

4.1.1 Simulation procedure

Generation of the joint distribution can be summarised as follows:

Table 3: Simulation steps for the generation of the joint distibution

for each $m \in M$
Generate y according to ϑ (model error)
Record Y_{U_d}
for each $p \in P$
Generate p using a probability sampling design (sampling error)
Estimate $\hat{\beta}$ based on p (model parameter error)
Generate $\hat{y}_{c,k}$ (imputation error)
Estimate $\hat{Y}_{U_d}$
Estimate the difference: $\hat{Y}_{U_d} - Y_{U_d}$
Compute V_{design}, GMSE, PEM

P and M were set at 100, consistent with the number of iterations used in previous studies (Alleva et

al., 2021; Deliu et al., 2025; Scholtus & Daalmans, 2021). Note that the current simulation design implies reversing the order of the expectations in comparison to Equation 2.9. However, given non-informative sampling, this was not expected to affect on the results (Deliu et al., 2025). Furthermore, the current simulation logic was also adopted in Alleva et al. (2021).

4.1.2 Benchmark estimators

Benchmark estimators for MSE, bias, and variance were computed from the two distributions and used as the basis for comparing the three estimators. The benchmarks were computed as:

$$\widehat{\text{MSE}}_{\text{design}} = \frac{1}{P}\sum_{p=1}^{P}\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k}^{(p)} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}\right)^2 \tag{4.1}$$

$$\widehat{\text{MSE}}_{\text{joint}} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{P}\sum_{p=1}^{P}\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k}^{(p,m)} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}^{(m)}\right)^2\right) \tag{4.2}$$

For design-based benchmarks, the sampling distribution was based on a realisation of one finite population, specifically $m = 100$. The benchmark estimators for bias and variance were defined similarly upon evaluation of the average difference and the variance of the difference under the two distributions:

$$\widehat{\text{Bias}}_{\text{design}} = \frac{1}{P}\sum_{p=1}^{P}\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k}^{(p)} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}\right) \tag{4.3}$$

$$\widehat{\text{Bias}}_{\text{joint}} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{P}\sum_{p=1}^{P}\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k}^{(p,m)} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}^{(m)}\right)\right) \tag{4.4}$$

$$\widehat{\text{Var}}_{\text{design}} = \text{Var}_P\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}\right) \tag{4.5}$$

$$\widehat{\text{Var}}_{\text{joint}} = \text{Var}_{MP}\left(\sum_{k=1}^{N}\gamma_{d,k}\hat{y}_{c,k}^{(p,m)} - \sum_{k=1}^{N}\gamma_{d,k}y_{c,k}^{(m)}\right) \tag{4.6}$$

4.1.3 Relative estimates

The coefficient of variation (CV), relative root mean squared error (RRMSE), and relative bias (RB) are commonly used dimensionless measures to evaluate the size of the variance relative to the estimate, defined as

$$\text{CV}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{V(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.7}$$

$$\text{RRMSE}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{\text{MSE}(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.8}$$

$$\mathrm{RB}(\hat{Y}_{R_{c,d}}) = \frac{\mathrm{Bias}(\hat{Y}_{R_{c,d}})}{E(\hat{Y}_{R_{c,d}})} \tag{4.9}$$

Note that bias has a direction, indicating whether the mass-imputed estimator is larger (positive sign) or smaller (negative sign) than the true total in expectation. Altogether, these measures were adopted as benchmarks for the variance, MSE, and bias respectively. Recall that only PEM could provide approximation to all three relative measures. Subsequently, the benchmark estimators were approximated based on the the accuracy estimators as follows:

$$\widehat{\mathrm{CV}}_{V_{\mathrm{design}}}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{V_{\mathrm{design}}(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.10}$$

$$\widehat{\mathrm{CV}}_{V_{\mathrm{PEM}}}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{\mathrm{PEM}_{\mathrm{Var}}(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.11}$$

$$\widehat{\mathrm{RRMSE}}_{\mathrm{GMSE}}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{\mathrm{GMSE}(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.12}$$

$$\widehat{\mathrm{RRMSE}}_{\mathrm{PEM}}(\hat{Y}_{R_{c,d}}) = \frac{\sqrt{\mathrm{PEM}_{\mathrm{MSE}}(\hat{Y}_{R_{c,d}})}}{E(\hat{Y}_{R_{c,d}})} \tag{4.13}$$

$$\mathrm{RB}_{\mathrm{PEM}}(\hat{Y}_{R_{c,d}}) = \frac{\mathrm{PEM}_{\mathrm{bias}}(\hat{Y}_{R_{c,d}})}{E(\hat{Y}_{R_{c,d}})} \tag{4.14}$$

Estimates 4.10 to 4.14 will hereby be referred to as component specific relative measures (CSRMs). As common in practice, these were expressed as percentages.

4.1.4 The data

4.1.4.1 The target variable

The synthetic dataset *Samplonia* contains 6 variables, of which the level of education for the population aged 15 years and older was used as the target variable - similarly to Scholtus and Daalmans (2021). The level of education consists of 8 categories, $C = \{\text{none}, \text{basic education}, \text{VMBO}, \text{havo/VWO}, \text{MBO}, \text{HBO}, \text{WO-Bachelor}, \text{WO-Master}\}$.

For computational reasons, the simulation study was carried out by recoding the multinomial education variable as a binomial variable, with levels 0 = no higher education = {none, basic education, VMBO, havo/VWO, MBO} and 1 = has higher education = {HBO, WO-Bachelor, WO-Master}.

The multinomial variable was nevertheless evaluated in a separate experimental condition upon re-

coding the original variable into three categories:

$$\begin{cases} 1 = \text{low educational level} = \{\text{none}, \text{basic education}\} \\ 2 = \text{middle educational level} = \{\text{VMBO}, \text{havo/VWO}, \text{MBO}\} \\ 3 = \text{high educational level} = \{\text{HBO}, \text{WO-Bachelor}, \text{WO-Master}\} \end{cases}.$$

However, due to computational reasons, only the sampling distribution of the recoded multinomial variable was evaluated. To strengthen the evidence from this condition, the number of iterations was increased to $P = 1000$.

4.1.4.2 The superpopulations

Two superpopulations were created that formed the basis of further modifications throughout the experimental conditions. The first superpopulation was defined by fitting a logistic regression model with gender, age, and income as covariates to estimate its parameters. Age and income were included in the model after preprocessing. Age was log-transformed to reduce skewness and subsequently centred and scaled to have a mean of 0 and a standard deviation of 1. Income, originally a continuous variable, was recoded into a three-level categorical variable with the categories *low*, *middle*, and *high*. The first superpopulation was characterized by high model error since the covariates, while significantly related to the target variable, did not explain much variance, as indicated by the AIC value (see Tables 4 and 5). The second superpopulation was hence defined based on covariates with a stronger association with the target variable. Since the original dataset did not contain such variables, two independent synthetic variables were generated. First, a new three-level income variable was generated by sampling the categories with replacement according to the vector of probabilities $p = \{0.1, 0.2, 0.7\}$ if a person has higher education, and $p = \{0.7, 0.2, 0.1\}$ if the person does not have higher education. Similarly, a new four-level region variable with levels {North, West, South, East} was generated with $p = \{0.15, 0.7, 0.1, 0.05\}$ if $y = 1$, and $p = \{0.05, 0.1, 0.7, 0.15\}$ if $y = 0$. In addition, a quadratic effect of the transformed age variable was included in the model. As indicated by a lower AIC value in superpopulation 2, these modifications resulted in lower model error.

Logistic regression models were fitted on an enlarged version of the original *Samplonia* dataset, created by multiplying the subset of individuals over 15 years old (745 rows) by 125 to approximate the size of real-life administrative records better. In addition, the people with higher education were oversampled to balance the two classes in the superpopulation. This step was motivated by the idea of establishing a baseline condition whereby the estimation of the imputation model is easier. Therefore, the proportion of people with higher education was increased from 0.34 to 0.5 upon sampling with replacement from the rows of people with higher education, resulting in a final population size $U = 122,250$. Estimating the superpopulation parameters did not require extensive model fitting, as the goal was not prediction but rather defining the distribution of the target variable. Summaries of the resulting superpopulations are presented in Tables 4 and 5.

4.1.5 The experimental conditions

The experimental conditions were defined based on the strengths and limitations of the three estimators outlined in Table 2, while taking practical relevance into account. Each condition was designed to

Table 4: Model coefficients and summary statistics for superpopulation 1

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	-0.6286	0.0134	-46.874	< 0.0001
genderF	0.5631	0.0150	37.558	< 0.0001
age	0.4279	0.0065	65.646	< 0.0001
incomeMiddle	1.0550	0.0175	60.189	< 0.0001
incomeHigh	0.6848	0.0195	35.129	< 0.0001
Model Fit Statistics				
Deviance	161789			
Null Deviance	169474.5			
AIC	161799			

Table 5: Model coefficients and summary statistics for superpopulation 2

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	1.1851	0.0329	35.998	< 0.0001
genderF	0.1343	0.0202	6.661	< 0.0001
age	0.2620	0.0107	24.572	< 0.0001
income_synLow	-3.8787	0.0266	-146.016	< 0.0001
income_synMiddle	-1.9049	0.0257	-74.086	< 0.0001
region_synNorth	2.1565	0.0391	55.113	< 0.0001
region_synSouth	-0.8914	0.0318	-28.049	< 0.0001
region_synWest	3.0354	0.0329	92.150	< 0.0001
I(age^2)	-0.4248	0.0110	-38.711	< 0.0001
Model Fit Statistics				
Deviance	65330.9			
Null Deviance	169474.5			
AIC	65348.9			

influence primarily the bias or the variance of the mass-imputed estimator. All experimental conditions were evaluated across the two superpopulations, which enabled the evaluation of potential interactions of increased bias or variance in different conditions with model error. In addition to the exploration of independent effects of increased bias and variance, one condition was set up to explore the effects of increased bias and variance.

The experiment focused on exploring the independent effects of the different conditions while changing one key parameter relative to the baseline at a time. Sometimes, a few parameters had to be changed simultaneously to keep another parameter fixed (for example, the condition "Small population", see Table 7). While a full-factorial design would have enabled a more thorough examination of robustness, this was not pursued due to the computational burden of $M \times P = 10,000$ simulation steps across a relatively large population. Keeping the sample size large was hence prioritised over the exploration of interaction effects, in the hope of producing fewer but more practical and robust findings. Furthermore, the interactions of some conditions may have become challenging to define due to issues with model convergence, such as estimating a large number of model parameters from a small sample.

4.1.5.1 The baseline condition

Changes in bias and variance were evaluated with respect to a baseline condition. Key parameters that were modified across the experimental conditions are summarised in Table 6. The baseline condition was designed to provide an ideal scenario where all estimators were most likely to approximate the benchmarks well. First, a relatively large sample based on the standards in official statistics was drawn according to simple random sampling without replacement (SRSWOR), which corresponds to an ignorable design.

Second, the imputation model was defined using the same covariates as the superpopulation model, resulting in negligible bias in the mass-imputed estimator. In addition, the mass-imputed estimator was estimated in the domain of gender, which was roughly equally spread across the target variable. Since gender was included in the set of covariates, it represented an internal domain.

Table 6: Parameters of the baseline condition

Parameter	**Notation**	**Description**
Sampling fraction	n/N	0.05
Population size	N	122,250
Sampling design (inclusion probabilities)	π	0.05 (SRSWOR)
Imputation model parameters vs. Superpopulation parameters	$\hat{\vartheta} = \vartheta$	Equal, no bias
Target variable	y	$\{0, 1\}$
Domain size (has higher education)	N_{d1}	$\sim 30,000$

4.1.5.2 Affecting the variance component

Changes in the variance component reflect variations in model error, sampling error, model parameter error, and imputation error, all of which are interlinked. The cascade of effects begins with model error, which affects the distribution of the target variable to which the sampling design is applied. The sample size and design, in turn, affect model parameter error, which directly impacts imputation error. Therefore, instead of attempting to increase any of these interconnected sources of error in isolation, the variance conditions focused on increasing the total variance of the mass-imputed estimator by triggering the cascade through several practical scenarios expected to challenge one or more of the accuracy estimators. Five such conditions were defined.

4.1.5.2.1 Small sample size First, the sample size was reduced by decreasing the sampling fraction from 5% to 1%. A smaller sample size increases the variability of the target variable across the samples and the variability of the estimated model parameters. This was expected to have a negative impact on all the estimators.

4.1.5.2.2 Small population Second, the population size was reduced while keeping the sampling fraction fixed. This increased model error by increasing the variability of the target variable generated from the superpopulation, which in turn impacted other errors. This condition was mainly expected to negatively impact $GMSE$ based on the results discussed in section 2.3.2.

4.1.5.2.3 Non-ignorable sampling Third, a negative impact on $GMSE$ was expected upon violating the non-ignorable sampling assumption. In superpopulation 1, a stratified sample was drawn using simple random sampling without replacement across strata defined by income, resulting in oversampling of the middle and higher income groups. This design reduces sampling error of survey-based estimates by lowering variance within strata, typically chosen based on covariates strongly linked to the target variable (e.g., Lohr, 2021, Chapter 3). The total sample size was kept fixed by redistributing the sample across the strata according to fractions $F = \{0.2, 0.3, 0.5\}$ with respect to the original sample size. This results in inclusion probabilities $\pi = \{0.016, 0.068, 0.152\}$. In superpopulation 2, stratification was done based on the synthetic region variable instead according to $F = \{0.4, 0.4, 0.1, 0.1\}$, resulting in $\pi = \{0.197, 0.201, 0.013, 0.013\}$. In both superpopulations, the benchmark estimators were computed according to a weighted imputation model, with V_{design} and PEM adjusted for the sampling design. The

weights of the imputation model were defined as the inverse of the inclusion probabilities. $GMSE$, on the other hand, was applied according to instructions with a non-weighted imputation model while taking the correct inclusion probabilities into account.

4.1.5.2.4 Overparameterisation Fourth, based on the results reported in Section 2.3.1, a negative impact on all estimators was expected when including noise variables in the imputation model alongside the true population parameters. Additional variables in the imputation model that lack predictive power have a small impact on the estimation of the true predictive variables due to affecting the likelihood surface from which the model is estimated. In practice, one might want to include such variables or interactions of the variables in the imputation model to completely eliminate design-bias. Variable province with 6 levels from the original dataset and a synthetic variable "random" with 4 levels generated according to $p = \{0.25, 0.25, 0.25, 0.25\}$ if y=1 or 0 were added to the imputation, but not superpopulation model.

4.1.5.2.5 Multinomial outcome Lastly, the superpopulation parameters were estimated using a multinomial logistic regression model with a three-level target variable, as described in Section 4.1.4. The same set of covariates was used otherwise, except for the superpopulation 2 where a synthetic gender variable was introduced to reduce model error further. The synthetic gender variable was generated according to $p = \{0.2, 0.8\}$ if y = Low, $p = \{0.5, 0.5\}$ if y = Middle and $p = \{0.8, 0.2\}$ if y = High. A multinomial logistic regression model requires estimating more parameters, which increases model parameter error. Additionally, model error remained relatively high because the classes were not balanced, resulting in some domain totals being much smaller than others. Altogether, all estimators were expected to perform slightly worse than in the baseline condition.

4.1.5.3 Affecting the bias component

Three conditions were designed to examine the effects of model and design bias.

4.1.5.3.1 Model bias Model bias was examined across two levels of severity. In the first model bias condition, the domain variable gender was omitted from the imputation model. This introduced a slight bias in the mass-imputed estimator since, according to the superpopulation model, there were slightly more males than females with higher education (see Appendix E). It was expected that the performance of all estimators would worsen. To put PEM on equal standing with the other estimators, the prediction error probabilities were not estimated domain-specifically, mimicking a scenario with a lack of access to important covariates in the sample data. In the second model bias condition, the model bias was increased further. This was achieved by redefining the superpopulation upon replacement of the original gender variable with an independently generated synthetic gender variable that was defined according to $p = \{0.6, 0.4\}$ if y=1 and $p = \{0.4, 0.6\}$ if y=0.

4.1.5.3.2 Design bias Design bias was examined by creating a synthetic variable, "random domain," that was generated independently of the covariates and the target variable based on a binomial draw with $p = \{0.5, 0.5\}$. This was expected to have a mildly negative effect on V_{design}.

4.1.5.3.3 Affecting the bias and variance component The effect of increased bias and variance was investigated by crossing the conditions of small sample size and large model bias.

Table 7: Expected effects under different conditions

Main effect	**Condition**	**Parameters**	**Expected effects**
Variance	Small sample size	$n/N = 0.01$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} = \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of all estimators
Variance	Small population size	$n/N = 0.05$ $N = 48,900$ $\pi = 0.05$ $\hat{\beta} = \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 12,000$	Deterioration of all estimators, specifically GMSE
Variance	Non-ignorable sampling design	$n/N = 0.05$ $N = 122,250$ $\pi = \{0.016, 0.068, 0.152\}$ (sup 1) $\pi = \{0.197, 0.201, 0.013, 0.013\}$ (sup 2) $\hat{\beta} = \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of GMSE
Variance	Overparameterization	$n/N = 0.05$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} \neq \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of all estimators
Variance	Multinomial outcome	$n/N = 0.05$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} = \vartheta$ $y = \{1, 2, 3\}$ $N_{d1} = [7304, 28693]$	Deterioration of all estimators
Bias	Model bias (small)	$n/N = 0.05$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} \neq \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of all estimators
Bias	Model bias (large)	$n/N = 0.05$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} \neq \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of all estimators
Bias	Design bias	$n/N = 0.05$ $N = 122,250$ $\pi = 0.05$ $\hat{\beta} \neq \vartheta$ $y = \{0, 1\}$ $N_{d1} \approx 30,000$	Deterioration of V_{design}

4.2 Results

The following subsections present the results of the simulation study. To begin, a brief overview is provided to outline the rationale behind the interpretation of the findings. The section then proceeds to examine the behaviour of the benchmark estimators across the experimental conditions, followed by an evaluation of the accuracy estimators. The section concludes with an interpretation of the overall findings.

4.2.1 Rationale for performance evaluation

The following analysis aims to evaluate the performance of the accuracy estimators in estimating the benchmark values across the experimental conditions. It is expected that in the baseline condition, GMSE provides consistently good estimation of at least RRMSE_{joint} and V_{design} provides consistently good approximation of CV_{design}. For PEM to be considered an acceptable measure of accuracy, it should provide a good approximation of RRMSE_{design}, RB_{design}, and CV_{design}.

Ideal performance is reflected in the centring of the MC distribution of the accuracy estimators around the benchmark value with minimal spread. Performance is acceptable if the mean of the distribution is close to the benchmark. A large spreading of the distribution and divergence of the mean and the interior quartiles from the benchmark value, however, suggests systematic under- or overestimation. While this rationale corresponds to significance testing with respect to the benchmark value, the proceeding analysis will utilize a series of boxplots instead. Formal testing in the current study is limited due to the need to set a stringent correction on the p value, given the large number of tests required. Indeed, significance testing in simulation studies of this kind is uncommon and has not been applied in the studies on which the current simulation is based.

Since the aim of the thesis is to inform practice, it is also valuable to examine the performance when translating percentages into real units. For example, the RRMSE_{design} ranges from 1.3% to 24% in the first superpopulation and 0.7% to 9% in the second superpopulation (see Tables 20 and 21 in F). This represents a percentage of the domain total in the category "has higher education", suggesting that the mass-imputed estimator is estimated with an accuracy of 210 people at best and 7,200 at worst, given domain category totals of around 30,000 (see Appendix E). Consequently, differences of less than 1% or approximately 300 people might be argued to be negligible in practice, as this is virtually close to the lowest possible variance across the two superpopulations. In the condition with the multinomial target variable, a similar interpretation applies. However, slightly larger differences are also acceptable due to larger inherent variance from a more complicated model.

4.2.2 Change in benchmark estimators

Figure 1 presents changes in the benchmark estimators for the joint distribution across the different experimental conditions. The condition with the multinomial target variable is presented separately in Figure 4 and will be discussed in Subsection 4.2.4. The exact values for the benchmark and accuracy estimators across all conditions are outlined in Tables 20 and 21 in Appendix F.

The estimators are virtually equivalent under both the joint and the sampling distributions across all experimental conditions (see Appendix H for the plots based on the sampling distribution). This result follows from the use of an ignorable sampling design throughout the simulation. In the condition with a non-ignorable sampling design, the benchmarks were computed based on a weighted imputation model,

which likely reduced any differences between the two distributions.

Figure 1 shows that across the variance conditions, the accuracy of the mass-imputed estimator was influenced most by a reduction in sample size, followed by a reduction in population size and a non-ignorable sampling design, with the latter having a more pronounced effect in the female subdomain of superpopulation 1. This is likely due to a small interaction between gender and income, which was neither modelled nor taken into account when specifying the strata. No further domain effects can be observed across the variance conditions, which is to be expected given that the domain sizes were roughly equal. Over-parameterisation of the imputation model did not affect the variance of the mass-imputed estimator in superpopulation 1; however, it had a small effect in superpopulation 2. This is likely due to the large inherent variability in the target variable in superpopulation 1, resulting in negligible noise in the imputation model from the additional variables. In fact, $RRMSE_{design}$ even improved slightly in comparison to the baseline in superpopulation 1 (see Table 20 in Appendix F). The pattern of increased variance across the other variance conditions appears similar across the two superpopulations, indicating no interaction between the different sources of increased variance and the variability of the target variable.

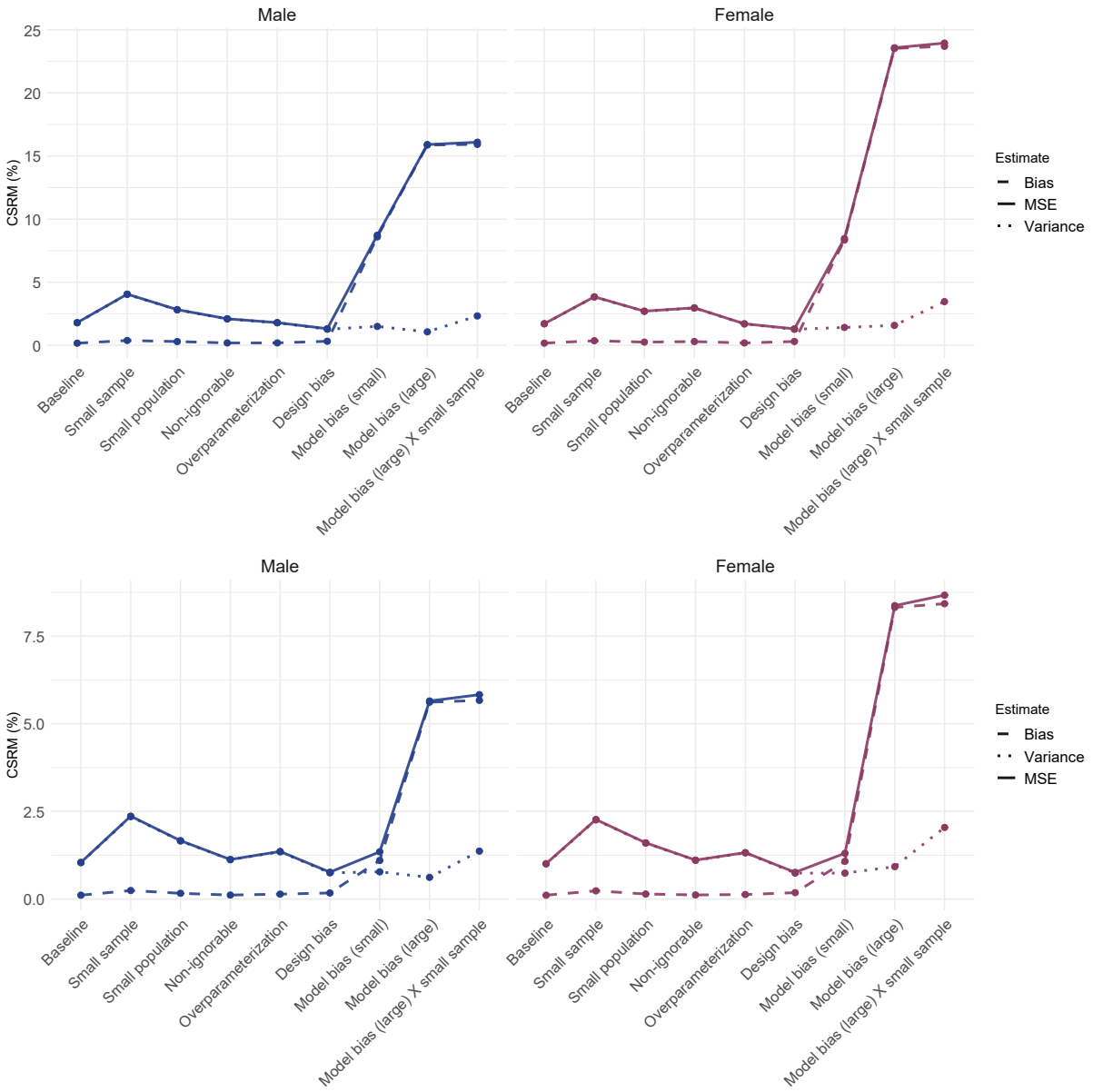
Across the bias conditions, $RRMSE$ was most affected by model bias in both superpopulations. The effect is pronounced in the female subdomain under large model bias condition, since the synthetic gender variable created a larger male subdomain. The largest $RRMSE$ can be observed in the interaction condition across both superpopulations, suggesting that model bias and increased variance due to a small sample size contribute additively to the MSE. However, as can be expected from the bias-variance tradeoff, the increase in variance is slightly limited by a simultaneous increase in bias. This is indicated by a slightly higher variance in the small sample condition.

4.2.3 Performance in the baseline condition

Figure 2 shows the performance of the accuracy estimators against $RRMSE_{design}$ and $RRMSE_{joint}$ in the baseline condition. The results are presented for the male subdomain, as the pattern is similar across the female subdomain (see Figure 9 in Appendix I). According to expectations, V_{design} and $GMSE$ provide good approximations of the benchmark estimators under ideal conditions. Recall that $GMSE$ was computed according to its upper bound, which is reflected in a slight overestimation of $RRMSE_{joint}$. This pattern is also more consistent across the joint distribution, which aligns with the definition of $GMSE$, and in the slightly larger female subdomain (see Figure 9 Appendix I). On the other hand, PEM is consistently overestimating the benchmark value. This behaviour is more pronounced across the joint distribution and in superpopulation 2. The same pattern occurs for $PEM(bias)$ in estimating RB_{design} (see Figure 3).

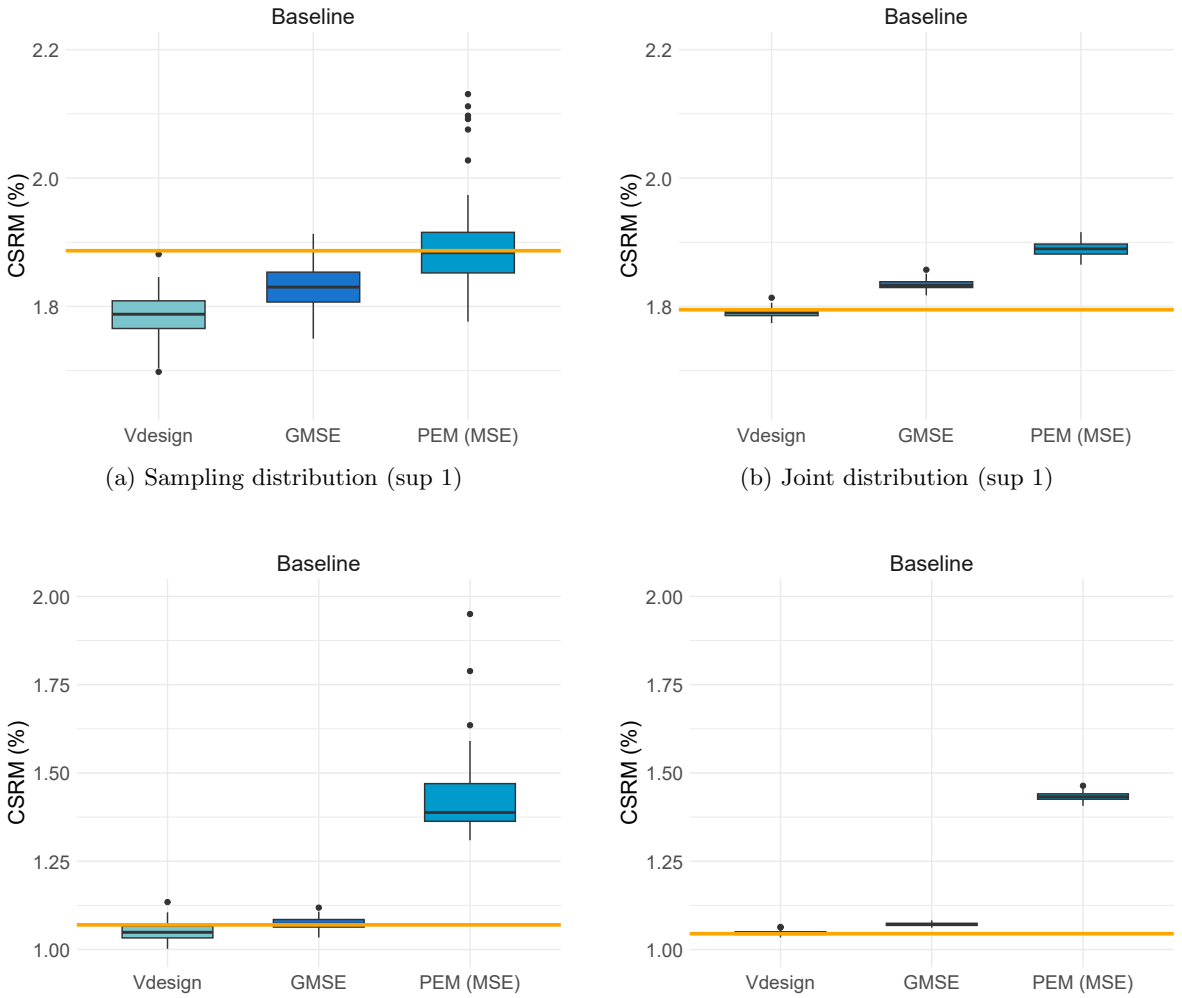
V_{design} and $GMSE$ are both very precise estimators, which is reflected in little spread in their distributions. PEM, on the other hand, is consistently the most variable estimator. Additionally, all accuracy estimators exhibit greater variability across superpopulation 1, reflecting the larger inherent variability of the target variable. The estimators are also more variable across the sampling distribution. This result likely reflects stochastic noise that is reduced upon averaging over 100 populations.

Figure 1: Change in CSRMs across the experimental conditions in superpopulation 1 (top) and 2 (bottom) for the binomial target variable



Note: This figure indicates the change in benchmark estimators for the joint distribution across the experimental conditions. The benchmark estimators can be distinguished by linetype. The conditions are ordered starting with the baseline, followed by the variance, bias, and interaction conditions.

Figure 2: Performance in the baseline condition for male subdomain across superpopulations 1 and 2



(a) Sampling distribution (sup 1)

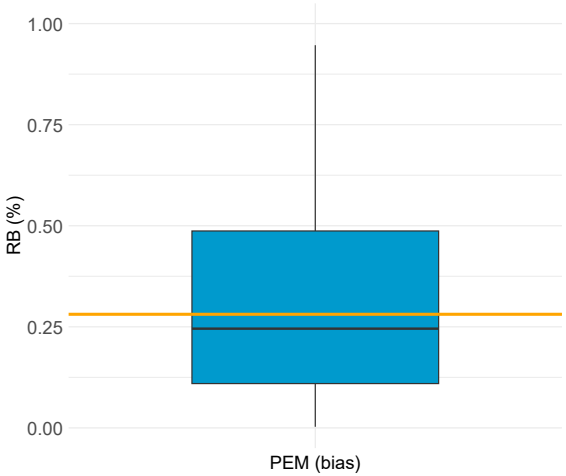
(b) Joint distribution (sup 1)

(c) Sampling distribution (sup 2)

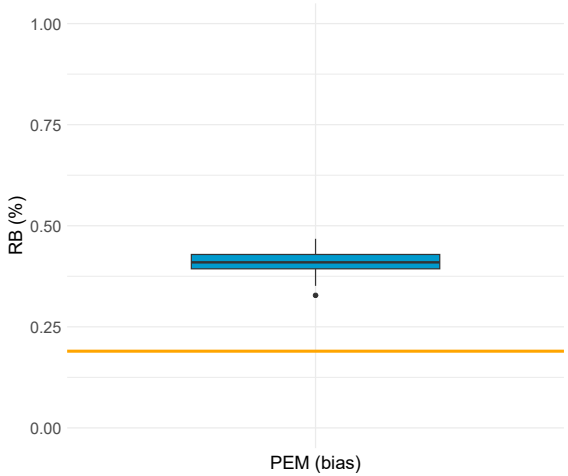
(d) Joint distribution (sup 2)

Note. The panels show the distribution of the estimators in the baseline condition across the two superpopulations in the male subdomain. The solid orange lines refer to $\text{RRMSE}_{\text{design}}$ in the sampling distribution and $\text{RRMSE}_{\text{joint}}$ in the joint distribution.

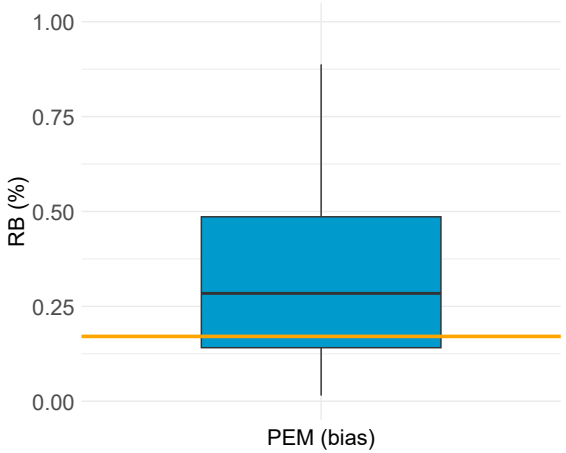
Figure 3: Performance of $PEM(bias)$ in the baseline condition for male subdomain across superpopulations 1 and 2.



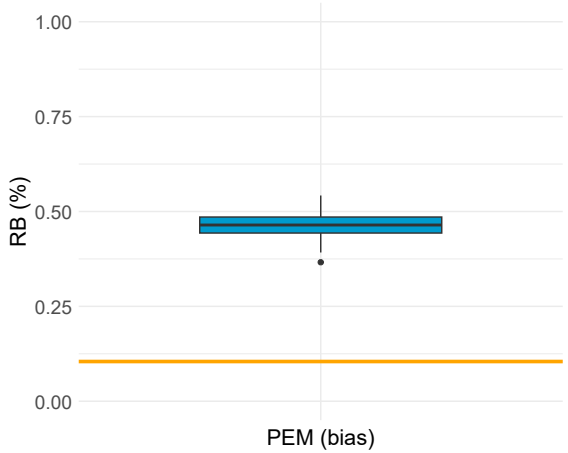
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



(c) Sampling distribution (sup 2)



(d) Joint distribution (sup 2)

Note. The panels show the distribution of PEM (bias) in the baseline condition in the male subdomain. The solid lines refer to $\text{RB}_{\text{design}}$ in the sampling distribution and RB_{joint} in the joint distribution. For clarity, the percentages are presented as absolute values.

4.2.4 Performance in the multinomial condition

Table 8: Distribution of education levels for superpopulations 1 and 2 in the multinomial condition

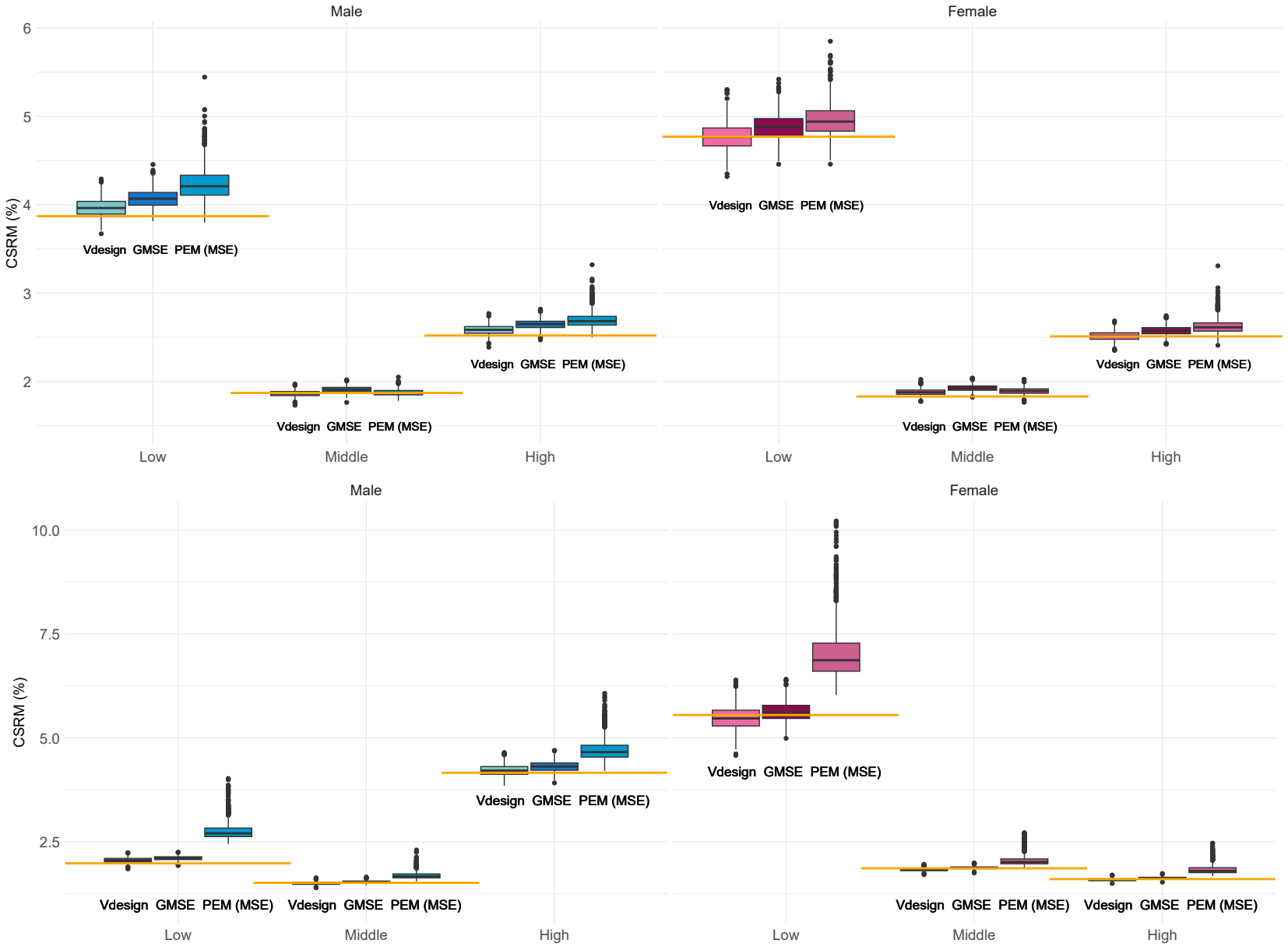
Gender	**Superpopulation 1**			**Superpopulation 2**		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
Male	9722	28693	19035	13664	27973	7769
Female	7304	27615	19381	3419	28047	30878

Table 8 presents the true population totals across the two superpopulations in the multinomial condition. In superpopulation 1, the sizes of categories per domain increase in increments of 10,000, with the category Middle being the largest, followed by High and Low. The target variable is spread similarly across the two subdomains in superpopulation 1. In contrast, the introduction of the synthetic gender variable in superpopulation 2 resulted in larger Male-Low and Female-High domains.

Figure 4 presents the performance of the accuracy estimators in the multinomial condition. Similarly to the binomial condition, $GMSE$ and PEM consistently overestimate the benchmark. Overestimation of PEM is again pronounced in superpopulation 2. Similarly to the binomial baseline condition, PEM is the most variable estimator. All estimators are prone to overestimation in smaller domains to a different extent, whereas in the larger domains, differences between the estimators are greatly reduced.

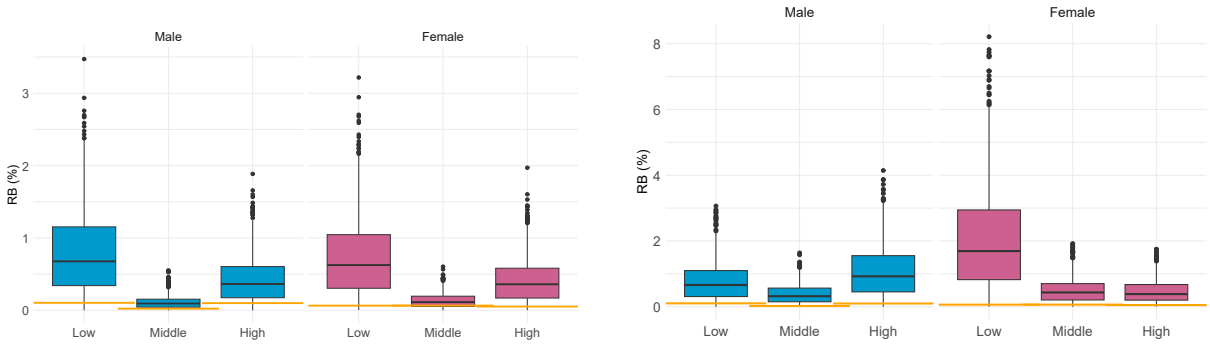
Similarly to the binomial baseline condition, PEM is overestimating RB_{design} (see Figure 5). Similarly to its MSE component, this effect is more pronounced in smaller domains.

Figure 4: Performance of the accuracy estimators in the multinomial condition in superpopulation 1 (top) and 2 (bottom)



Note: The figure displays performance of the estimators across the superpopulations for each category of the multinomial target variable. The solid orange line refers to $\text{RRMSE}_{\text{design}}$.

Figure 5: Performance of $PEM(bias)$ in the multinomial condition in superpopulation 1 (left) and 2 (right)



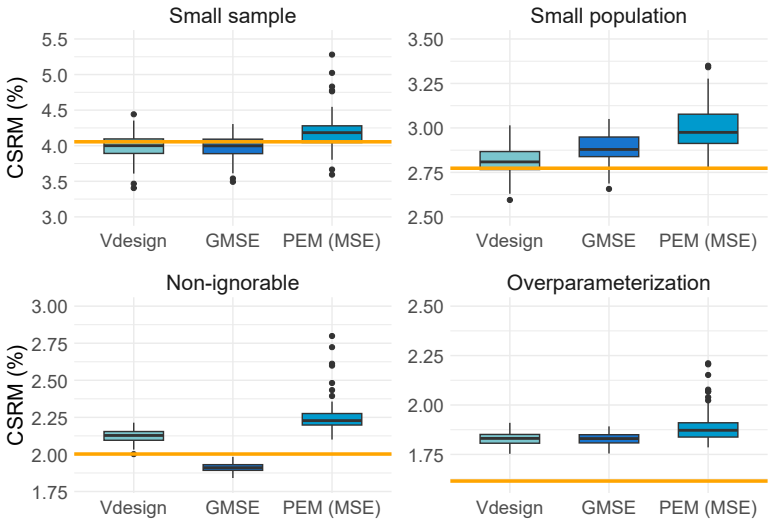
Note: The figure displays performance of PEM (bias) across the superpopulations for each category of the multinomial target variable. The solid line refers to $\text{RB}_{\text{design}}$. For clarity, the percentages are presented as absolute values.

4.2.5 Performance in the variance conditions

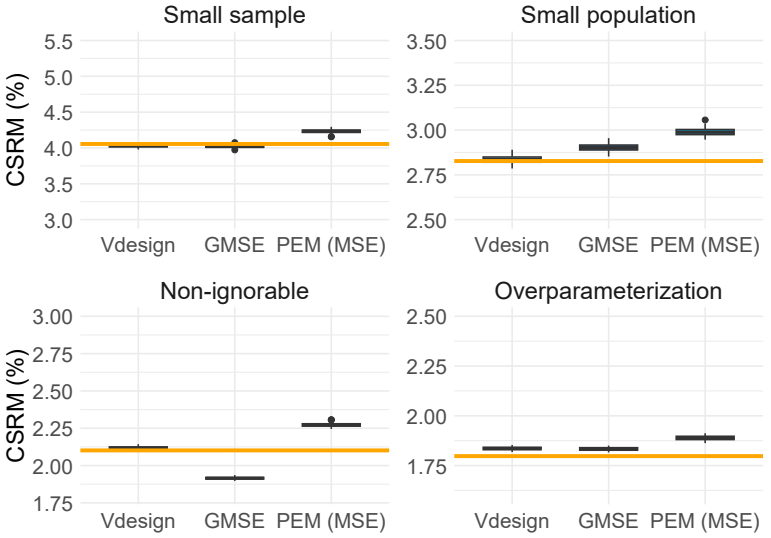
Figure 6 displays the performance of the accuracy estimators across the variance conditions in the male subdomain. Across both superpopulations, the largest negative effect on all estimators is observed in the condition of overparameterisation. Another expected effect is the deterioration of $GMSE$ in the non-ignorable sampling condition, while V_{design} appears to be the most robust estimator in this condition. $GMSE$ consistently underestimates the benchmark under a non-ignorable sampling design, with the effect being diminished in superpopulation 2. Similarly to the baseline condition, PEM is prone to overestimation.

Reductions in sample and population size have affected the variability of all accuracy estimators. In the small sample size condition, $GMSE$ appears to have been most negatively affected, as its distribution now lies consistently at or below the benchmark line. On the other hand, V_{design} appears to be the most robust estimator in this condition. Contrary to expectations, this pattern of results is reversed in the small population condition, with V_{design} prone to overestimating the benchmark variance, while $GMSE$ appears to be unaffected. PEM also appears to be more negatively affected in the small population condition in comparison to the small sample condition, suggesting a pattern of design-based estimators being more vulnerable in this condition. The pattern, however, is diminished in superpopulation 2 and in the female subdomain (see Figure 11 in Appendix I), suggesting an interaction with model error and domain size.

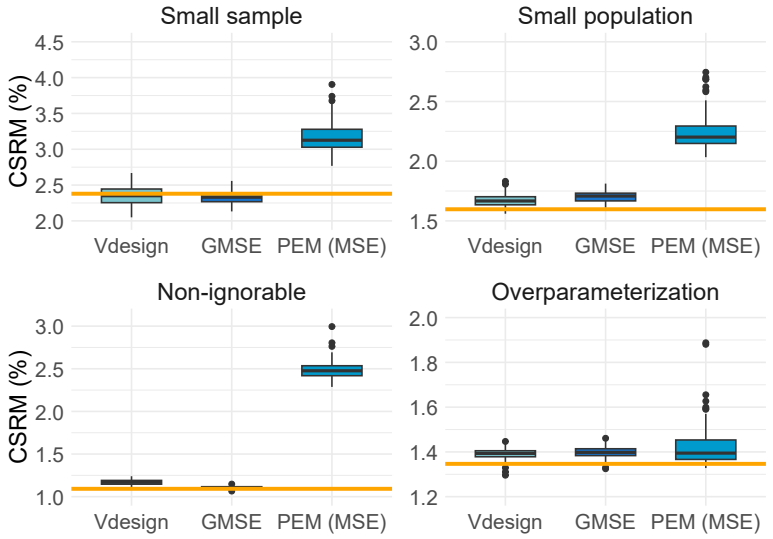
Figure 6: Performance of the estimators across the variance conditions with the binomial target variable in the male subdomain



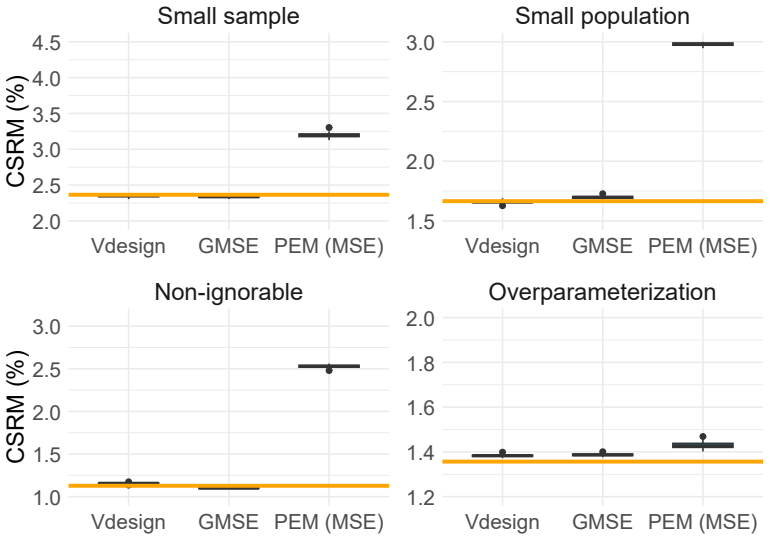
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



(c) Sampling distribution (sup 2)



(d) Joint distribution (sup 2)

Note: The figures above show the performance of the estimators relative to benchmark $\text{RRMSE}_{\text{design}}$ and $\text{RRMSE}_{\text{joint}}$ shown in solid orange line.

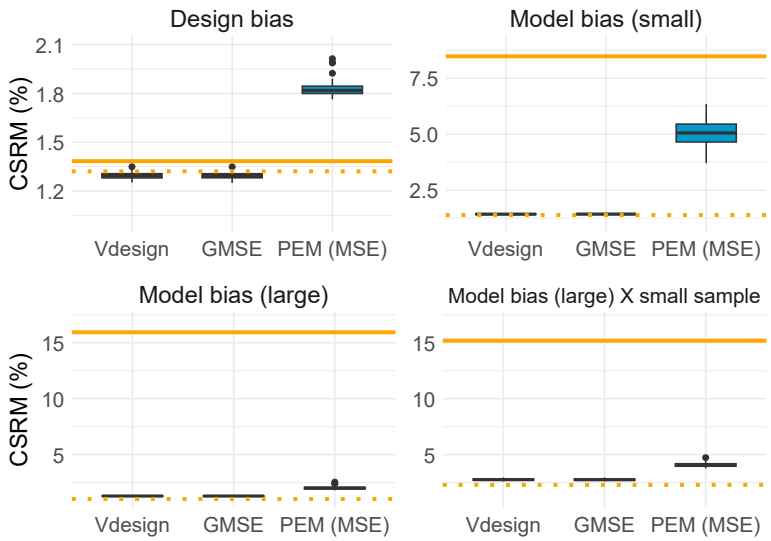
4.2.6 Performance in the bias and interaction conditions

Figure 7 shows the performance of the estimators in the bias and interaction conditions across the two superpopulations in the male subdomain. Contrary to expectations, $GMSE$ behaves virtually identically to V_{design} in the design bias condition. In superpopulation 1, $GMSE$ is underestimating the RRMSE benchmarks, but provides a good approximation of CV benchmarks alongside V_{design}. In superpopulation 2, the pattern reverses, with V_{design} overestimating the CV benchmarks alongside GMSE. PEM is overestimating both components. Recall that domains were not taken into account during the computation of PEM across the bias and interaction conditions, which appears to pronounce the systemic overestimation.

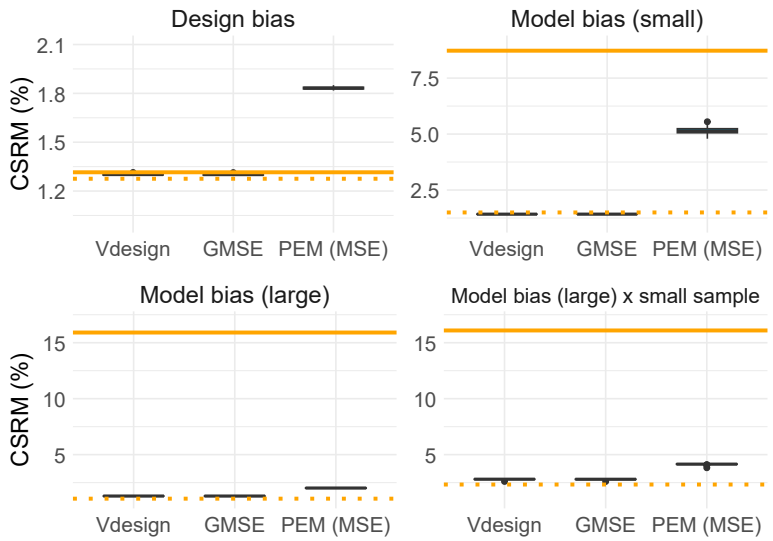
V_{design} and $GMSE$ also perform very similarly in the conditions with increased model bias. PEM continues to overestimate the variance component; however, it now provides a downwards approximation of RB (see Table 20 and 21 Appendix F). The downwards estimation is relatively good in some conditions, as the distribution of PEM lies close to the benchmark RRMSE values in superpopulation 2 and in superpopulation 1 if the bias is small.

Under increased bias and variance, no estimator can estimate $RRMSE$ well. Furthermore, all three approaches consistently overestimate the variance component in the male subdomain, while $GMSE$ and V_{design} underestimate the variance component in the female subdomain (see Figure 12 in Appendix I). This pattern of results indicates a complex interaction between domain size and bias. It appears that as bias increases and domain size decreases, $GMSE$ and V_{design} are more prone to underestimating the variance component and vice versa. Recall that in the model bias (large) condition, the synthetic gender variable introduced a smaller female subdomain.

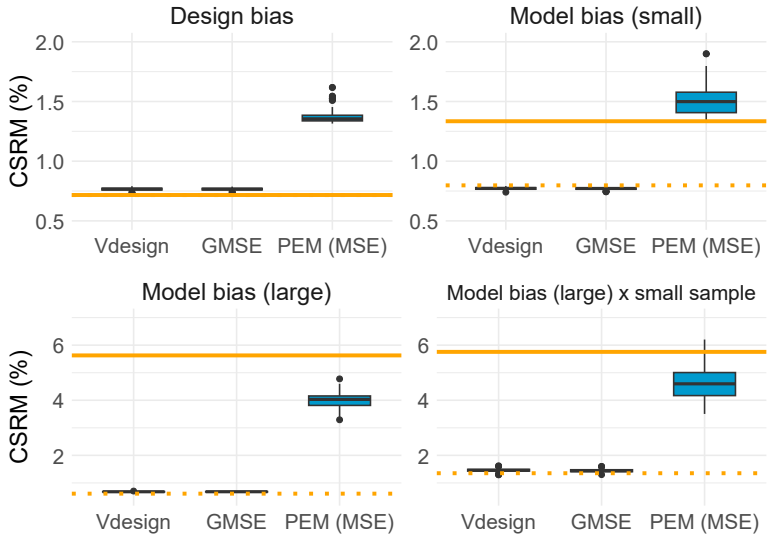
Figure 7: Performance of the estimators across bias and interaction conditions in the male subdomain



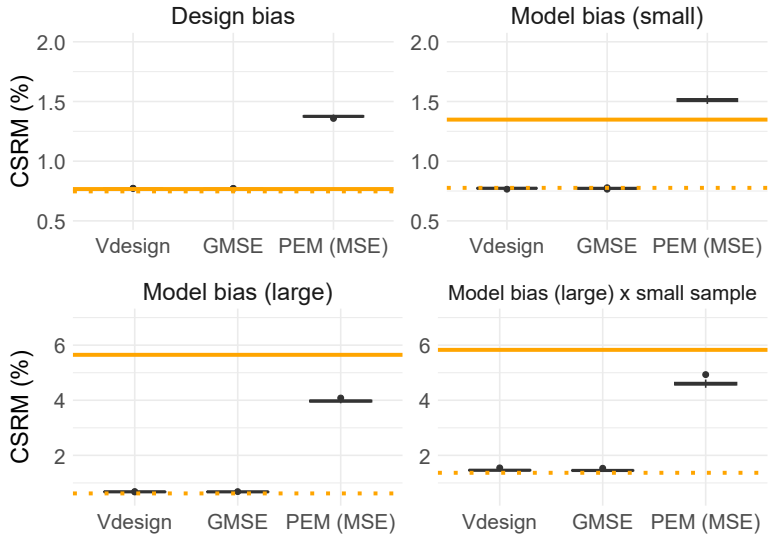
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



(c) Sampling distribution (female)



(d) Joint distribution (female)

Note: The figures above show the performance of the estimators relative to $\text{RRMSE}_{\text{design}}$ and $\text{RRMSE}_{\text{joint}}$ (shown in solid orange line) and $\text{CV}_{\text{design}}$ and CV_{joint} (shown in dotted orange line).

4.2.7 Conclusions from the simulation study

Several expected patterns in performance were observed in the simulation study. First, it was shown that under ideal circumstances, all estimators provided good or acceptable estimates of the benchmark values. PEM was consistently the most variable estimator, which could be associated with its links to bootstrap estimators. This corroborates the findings from previous studies, which show that bootstrap estimators are more variable compared to linearised estimators (Deliu et al., 2025; Scholtus & Daalmans, 2021). Interestingly, a pattern of overestimation was observed for all components of PEM, which was pronounced in superpopulation 2. This is most likely due to the behaviour of the function $p(1-p)$, which forms the key component of PEM (see Equation 3.7). This function reaches its maximum if $p = 0.5$. Since averaging unit-level probabilities tends to move p closer to 0.5, PEM can be expected to provide an upwardly biased accuracy estimate most of the time. This effect becomes more pronounced as the prediction error probabilities approach 0 and 1 since the function changes more rapidly at extreme values. For instance, replacing the unit-level probability $p = 0.9$ by $p = 0.8$ causes a larger change in the function (0.09 vs 0.16) than replacing $p = 0.6$ by $p = 0.5$ (0.24 vs 0.25). This first case occurred in superpopulation 2, where the prediction error probabilities are generally above 0.8 or below 0.2, thereby explaining the stronger overestimation observed in that setting.

The multinomial condition revealed a domain effect consistent with previous studies. Upon closer examination of the results of Deliu et al. (2025), it can be observed that $GMSE$ was prone to overestimation in smaller domains (see Table 4, Deliu et al., 2025). This was observed for all estimators in the current study. Interestingly, in the small sample condition, $GMSE$ was underestimating the benchmark. This suggests an interaction between sample size and domain size, whereby $GMSE$ is more affected by the reduction in sample size. Hence it is likely that the previous results by Deliu et al. (2025), whereby the performance of $GMSE$ improved upon increasing the population size, might be better explained by increased sample size since the sampling fraction was kept constant in their study.

In contrast, the design-based estimators appear more affected by the reduction in domain size, given that they overestimated more consistently in small domains and in the small population condition, but not in the small sample condition. Altogether, it appears that the conservative nature of design-based estimators becomes more pronounced in small domains, while the more liberal nature of $GMSE$ leads to vulnerability upon reduced sample size, possibly due to increased difficulty in learning the target-covariate relationships from the data.

While the reduction in sample and population size primarily increased the variability of the estimators, all estimators exhibited a systematic pattern of overestimation upon overparameterisation of the imputation model. This finding is consistent with the results of Scholtus and Daalmans (2021) and suggests that other accuracy estimators may be similarly affected. PEM appears to be the least affected since its general tendency to overestimate did not seem pronounced in this condition. This is likely due to PEM not explicitly evaluating model parameter error via an estimated variance-covariance matrix like V_{design} and $GMSE$, suggesting that PEM is the most robust estimator under this condition.

Per expectations, $GMSE$ demonstrated deteriorated performance under a non-ignorable sampling design. However, this result could be partially attributed to the reversed order of expectations in the simulation study in comparison to the definition of $GMSE$. Furthermore, the effect was greatly reduced in superpopulation 2. On the other hand, design-based estimators were prone to slight overestimation in this condition, suggesting that the robustness of all estimators was affected in this condition.

Contrary to expected effects, V_{design} and $GMSE$ performed virtually identically across all conditions with increased bias. While some similarity in performance was anticipated (as discussed in Section 2.3.2),

the current findings suggest that the principal practical value of $GMSE$ lies in variance estimation. Notably, both estimators approximated the variance well under conditions of increased bias even though the assumptions underlying both methods were violated. This suggests that the assumption of model-unbiasedness for $GMSE$ may not significantly impact its performance, likely because it retains the ability to rely on its design-based inference component. Likewise, V_{design} appears robust to imputation model misspecification from a design-based perspective. However, given that the interaction condition showed accuracy estimators to be prone to overestimation under large model bias, it follows that, in many practical applications, the linearised estimators are more likely to provide conservative variance estimates, except if the domain size is small.

Surprisingly, PEM was the most robust estimator across all conditions with increased bias. This suggests that PEM might present a simpler alternative to $GMSE$ that circumvents the assumption of no model bias. However, domain-specific estimation is important to prevent severe overestimation of the variance.

Altogether, although systematic patterns emerged that demonstrate unique vulnerabilities of the estimators, from a practical point of view, all estimators can be considered robust estimators of variance.

Chapter 5

Case study

To strengthen the evidence for the robustness of the three estimators, it should be demonstrated that they also perform comparably on a non-synthetic dataset. For this purpose, the estimators will be compared in a case study using the EAF. The following subsections describe the dataset, methods and results of the study.

5.1 Methods

5.1.1 The target variable

The target variable of this study is the mother's educational level of primary school children in the EAF. This choice is motivated by a project in collaboration between the Dutch Ministry of Education, Culture, and Science and Statistics Netherlands, which seeks to reduce educational disadvantages by increasing financial support to primary schools with more disadvantaged children (e.g., Posthumus et al., 2019). To determine which schools should receive the support, Statistics Netherlands has developed an indicator of educational disadvantage, which essentially reflects a linear regression model with several variables reflecting the socio-economic background of children as input. Crucial variables are those describing parental educational attainment, which even formed the sole basis of the weighing scheme for budgetary allocation before Statistics Netherlands development. However, this information is missing for children in specific subpopulations, for one or both parents, in the EAF. This is because the EAF is primarily based on different sources of administrative data, which provide good coverage for young adults but have under coverage for older people (Linder et al., 2011). Furthermore, administrative data is selective and lacks information, for example, on people who obtained their education abroad or immigrated to the Netherlands later in life. Where possible, these gaps are filled by linking administrative records with the Dutch Labour Force Survey (LFS) (Linder et al., 2011). The annual Dutch Labour Force survey dates back to 1987. It has been the basis for information on educational attainment in the Netherlands until the creation of EAF, which enables the provision of more accurate estimates due to the steep decline in response rates since the beginning of the millennium and generally large sampling errors in small subpopulations.

The missing parental education is imputed based on a series of multinomial logistic regression models according to the latest developments in the project (Statistics Netherlands, 2024). These models are trained on the LFS data and applied sequentially so that first, the mother's education is imputed based on the subpopulation from LFS for whom both parental education is known. This can then be used to impute education in other subpopulations. The remaining input variables are obtained from the

population registry and are fully observed. Therefore, the key source of uncertainty in the score for educational disadvantage comes from the imputed educational attainment variables. In turn, it is possible to transform the estimators under study so that it becomes possible to evaluate the accuracy of the educational disadvantage scores. Although this is outside the scope of the current project, the estimators will be applied to the relevant target variable to motivate such developments in the future.

5.1.2 Setup

The case study will be set up based on two imputation models used to impute the mother's education in the EAF. These imputation models correspond to those described in the latest report on the project (Statistics Netherlands, 2024), with a few modifications outlined below. The primary distinction between the two models is that the second model does not include the father's education as a covariate, which is strongly correlated with the mother's education. This is because the second imputation model is used to impute the mother's education in a subpopulation where the father's education is unknown; therefore, using it as a covariate would involve assuming that father's education in this group of children is similar to the group where father's education is known. For the case study, this presents a good opportunity to evaluate the performance of the estimators with a worse imputation model that is potentially more biased.

The focus is limited to the subgroup of children who are also registered in the population register. The missingness pattern of parental education in this group of children can be described by distinguishing between 4 subpopulations based on which parent's educational attainment is known (see Table 9).

Table 9: Missingness pattern of parental education of Dutch children

Subpopulation	**Missingness pattern**	**0-27 years**	**0-12 years**
A	No information on parental education	15.5%	7.9%
B	Mother's education known	17.2%	13.2%
C	Father's education known	11.7%	7.8%
D	Both parents' education known	55.6%	71.1%

Note. Percentages refer to the extent of missing data in the subpopulation. Column headers 0-27 years and 0-12 years refer to the age of the child. The table has been adapted from Statistics Netherlands (2024).

This case study will focus on imputing mother's education in subpopulations A and C. The imputation model for subpopulation A does not include the father's education. To align with the setup described in Section 2.2.1, two "populations" will be defined, which will be treated as population registers. The first population is defined based on sample data (LFS data) from subpopulation D, where the mother's education is observed and the entire subpopulation C. Subsequently, sample data where both the mother's and father's education is observed will be selected and treated as an SRSWOR from this population. This is a simplification of the more complex sampling design used in LFS data, but it is reasonable given that the same assumption is applied when comparing all estimators (Scholtus & Daalmans, 2021). The assumption for the first population is, therefore, that the target variable distribution can be well described by the mother's and father's education in subpopulation D. The second population is defined similarly but uses subpopulation A instead of C as the basis. The sample data consists of all children whose mothers' education is known. Hence, it is assumed that the target variable can be well described by the mother's education in subpopulation D.

The inclusion probabilities refer to the fraction of sample data in either of the populations. These were

$\pi_C = \frac{88{,}107}{762{,}859} \approx 0.017$ based on 88,107 units from a population of $U = 762{,}859$, and $\pi_A = \frac{187{,}662}{959{,}678} \approx 0.037$ based on 187,662 units from a population of $U = 959{,}678$, respectively.

5.1.3 Imputation models

Educational attainment in the EAF is a multinomial target variable with 8 categories. The training of the imputation models is described extensively in the corresponding project report by Statistics Netherlands (Statistics Netherlands, 2024; see also Posthumus et al., 2016). In short, a multinomial logistic regression model was trained using stepwise selection based on a list of covariates that describe the demographic and socio-economic status of the parents. Models for subpopulations A and C have the following covariates in common:

- Income of the mother (21 levels)
- Country of origin (8 levels)
- Age of the mother
- Urbanisation level of the area of the household (6 levels)
- Civil status of the mother (4 levels)
- Type of economic activity of the mother (13 levels)

Initially, the imputation model for subgroup C was fitted separately for each level of the father's education. For the sake of simplicity, the current study included the father's educational level as a covariate in the imputation model for subpopulation C.

Further simplifications were made due to computational reasons. First, both mothers' and fathers' educational levels were recoded as a 3-level categorical variable with levels low, middle, and high. Second, instead of using the 13-level categorical variable that describes the mother's economic activity, a two-level variable indicating whether the mother works or not was used. Finally, the numerical age variable was standardized.

The imputation models were fitted on the sample data as described in the previous section without sampling weights due to assumption of SRSWOR. Note that the sampling weights were taken into account when estimating the accuracy estimators, an approach also applied throughout the simulation study. The first imputation model fits the data better as expected due to the inclusion of father's education, as indicated by a higher $R^2_{McFadden}$ value ($R^2_{McFadden} = 0.163$ vs $R^2_{McFadden} = 0.131$). Pseudo-R-squared has been computed according to McFadden's formula (McFadden, 1972).

5.2 Results

Accuracy estimation was fourfold in both populations, focusing on comparing the performance in a mix of large and small, internal and external domains. First, the estimates were obtained for the whole population. Across both populations, the middle education level is the largest domain, followed by low and high education (see Table 10). Performance in an internal domain was evaluated using the 21-category mother's income variable, which comprises a mix of small and large domains based on increasing levels of income (see Table 13). Income is related to the target variable across both populations, as it can be observed that the domains in the largest income classes are larger in the high education category. Performance was also evaluated in external domains with different levels of granularity. The larger

external domain (see Table 11) corresponds to the type of school the child is attending at the time of measurement. This domain variable has nine levels, but the estimation focused on two levels - children attending primary school and children attending special needs education. Across both populations, fewer students attend schools providing special needs education. Overall, it appears that the target variable is similarly distributed across the types of schools in both populations, suggesting that there is no strong association between the type of school attended and the mother's educational attainment, given the covariates in the imputation model. The final, smaller external domain refers to the specific school that the child is attending. Since there are well over 7,000 schools represented in the population, a selection was made based on schools with the largest number of sampled units. Five schools were chosen, and the corresponding imputed totals are presented in Table 12. The spread of the target variable is again similar across both populations. No school appears to stand out based on a different spread of the target variable, suggesting no strong relationships between the specific school and the mother's educational level.

Table 10: Distribution of mother's education levels in populations A and C

Population	**Low**	**Middle**	**High**
A	241797	535498	182383
C	182669	430266	149924

Table 11: Distribution of education levels by school type for Populations A and C

School Type	**Population A**			**Population C**		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
Primary school	34585	69308	20997	30820	74272	20477
Special needs	1087	1935	460	1185	2042	367

Table 12: Distribution of education levels by school for Populations A and C

School	**Population A**			**Population C**		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
S1	70	155	18	24	41	2
S2	16	41	5	12	40	4
S3	43	73	9	20	60	7
S4	39	113	11	36	46	7
S5	63	177	13	26	57	6

Table 13: Distribution of education levels by mother's income level for Populations A and C

Income class	Population A			Population C		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
1	8980	12856	4832	7235	9732	3551
2	14280	18475	5112	11078	13373	3640
3	25542	26915	8163	15471	17753	3705
4	9179	12955	2383	6266	9414	1607
5	15016	24131	3647	11351	18647	2830
6	10429	14788	2626	8345	13106	1808
7	13212	26283	3294	10162	21400	2881
8	13011	31647	4129	10324	25398	3634
9	12037	35550	4804	9663	29608	4033
10	10180	38287	5826	7594	31214	5460
11	8400	39647	7197	6777	34242	6631
12	6724	36022	8722	5581	30431	8185
13	5786	30889	9419	4385	26124	9021
14	4360	26689	10363	3772	22316	9920
15	3276	22545	10111	3094	19261	9260
16	2635	18322	11400	2193	16541	10566
17	2514	15546	11909	2121	13319	11258
18	1873	12386	12150	1550	10967	10794
19	1162	7755	12713	1029	7317	11253
20	839	4045	13934	793	4378	12887
unknown	72362	79765	29649	53885	55725	17000

5.2.1 The whole population

Based on the results of the simulation study, it is expected that both $GMSE$ and V_{design} will provide an estimate of the CV. Table 14 presents the CSRMs expressed as percentages based on all the estimators computed for the whole population across the two populations. All accuracy estimators indicate that the mass-imputed estimator is very precise, varying by less than 1% of the total. The effect of domain size can also be observed, with the estimated variance being smallest in the largest middle category, followed by the smaller low and high attainment categories.

While the estimators all indicate low variability, notable differences emerge between V_{design} and the other two estimators. Interestingly, although V_{design} performed very similarly to $GMSE$ across the simulation study, its estimate is now 30-40% larger than $GMSE$. The differences are slightly smaller in population C, suggesting that the assumed variability of the target variable could be driving some of the differences. On the other hand, while PEM was more prone to overestimating the variance in the simulation study, it is now performing extremely close to $GMSE$, especially in population C. These differences, although unexpected, can partly be explained by the fact that both $GMSE$ and PEM simplify the design-based variance. The results also confirm that PEM performs better as the sample and population size get larger.

Another reason for V_{design} providing a larger variance estimate is due to the vast number of parameters in the imputation model (2x38 in subpopulation A and 2x40 in subpopulation C). Though $GMSE$ was also similarly affected in the simulation study, it could be that overparameterisation has a larger impact on V_{design} due to having to take inclusion probabilities into account when estimating the model parameters and their variance.

PEM indicates a relatively large bias component in the low education category in population A, and

high education category in population C. This could be because there are more highly educated mothers in the sample from population A, whereas the pattern is reversed in population C (22% vs 24% in the low category and 19% vs 16% in the high category). Since this assumes a reversal of the distribution in the target variable, it is likely the reason for the pattern observed in bias. Given that PEM was prone to overestimating the bias in the simulation study, these values can be interpreted with a modest level of confidence.

Table 14: CSRM (%) for mother's education levels by based on the different accuracy estimators in Populations A and C

Estimator	**Population A**			**Population C**		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
Vdesign	0.607	0.295	0.666	0.741	0.382	0.828
GMSE	0.423	0.221	0.490	0.570	0.292	0.620
PEM (MSE)	12.164	5.141	1.146	0.823	4.219	12.788
PEM (bias)	-12.157	5.137	1.036	0.570	4.209	-12.773
PEM (variance)	0.395	0.212	0.490	0.593	0.292	0.623

5.2.2 Internal domain - income class

Table 15 presents the results from the different estimators in the income domain for every fifth class. The pattern of results is more similar to that observed in the simulation study, with V_{design} and $GMSE$ giving very similar estimates and PEM providing a slightly larger variance estimate. In many domains, $GMSE$ now provides a slightly larger estimate. Similarly to the simulation study, larger variance is observed in smaller domains. It could be that the results in this domain are comparable to the simulation study due to more similar domain sizes.

The bias estimates are of similar magnitude across the income domain as in the entire population, with smaller domains exhibiting a more pronounced bias. The difference noted in the previous subsection between the distributions of the target variable across the populations also has a slight effect, which is to be expected given that income and educational attainment are related.

Table 15: CSRMs (%) based on the different accuracy estimators for mother's education level cross-classified by income class (every 5th class) in subpopulations A and C.

Income Class	Estimator	Population A			Population C		
		Low	Middle	High	Low	Middle	High
1	Vdesign	2.258	1.629	3.374	2.708	2.217	4.971
	GMSE	2.158	1.579	3.192	2.730	2.191	4.322
	PEM (MSE)	4.454	5.741	8.840	2.819	7.377	20.768
	PEM (bias)	-3.452	5.346	-7.810	0.514	7.053	-20.377
	PEM (variance)	2.815	2.093	4.141	2.772	2.162	4.013
5	Vdesign	1.650	1.059	3.975	2.090	1.370	5.549
	GMSE	1.662	1.069	3.961	2.240	1.447	5.350
	PEM (MSE)	3.582	2.766	6.508	5.985	1.538	18.910
	PEM (bias)	-2.860	2.387	-4.017	5.545	-0.602	-18.278
	PEM (variance)	2.157	1.398	5.120	2.252	1.416	4.847
10	Vdesign	2.193	0.675	3.021	2.948	0.916	3.943
	GMSE	2.265	0.700	3.117	3.157	0.931	3.710
	PEM (MSE)	4.211	1.512	4.840	11.652	0.947	17.531
	PEM (bias)	-3.045	1.208	-2.620	11.199	0.284	-17.198
	PEM (variance)	2.908	0.909	4.069	3.217	0.903	3.400
15	Vdesign	4.272	0.971	1.982	4.926	1.373	2.638
	GMSE	4.312	1.015	2.081	5.160	1.331	2.485
	PEM (MSE)	6.936	1.351	2.967	13.924	4.387	13.244
	PEM (bias)	-4.253	0.162	1.018	12.881	4.191	-13.021
	PEM (variance)	5.479	1.342	2.787	5.288	1.298	2.424
20	Vdesign	8.426	3.275	1.017	10.229	4.163	1.459
	GMSE	8.593	3.479	1.065	10.636	3.824	1.329
	PEM (MSE)	12.574	4.635	1.460	33.431	15.132	7.030
	PEM (bias)	-6.435	0.238	0.318	31.319	14.594	-6.885
	PEM (variance)	10.802	4.629	1.425	11.693	3.997	1.421
unknown	Vdesign	1.008	0.916	1.903	1.222	1.216	2.593
	GMSE	0.721	0.664	1.308	0.942	0.948	2.038
	PEM (MSE)	1.187	3.956	12.223	1.519	1.305	7.025
	PEM (bias)	0.710	3.857	-12.110	1.181	0.912	-6.734
	PEM (variance)	0.952	0.878	1.660	0.956	0.933	2.002

5.2.3 External domains

5.2.3.1 School type

Table 16 presents the results from the different estimators across the two populations in the external domain "school type". Similarly to results from other domains, V_{design} provides a larger variance estimate relative to $GMSE$. Interestingly, the estimates align better in the smaller special needs education subdomain than in the larger primary education subdomain. Given that the bias in this condition is relatively large, the smaller subdomain might reflect the results from the simulation study whereby larger bias in small domains led to underestimation of variance by both $GMSE$ and V_{design}.

PEM variance estimates are closely aligned with V_{design} in the larger subdomain in superpopulation C. Note that the prediction error probabilities were estimated domain-specifically in this condition, which highlights the importance of this action, as it appears to reduce the overestimation observed for external

domains in the simulation study. Nevertheless, in the smaller subdomain, PEM is severely overestimating the variance in comparison to the other estimators, suggesting that PEM becomes unreliable as the domain size gets very small.

Table 16: CSRM (%) based on the different accuracy estimators for mother's education level cross-classified by school type in subpopulations A and C.

		Population A			**Population C**		
Domain	**Estimator**	**Low**	**Middle**	**High**	**Low**	**Middle**	**High**
Primary education (BO)	Vdesign	1.046	0.543	1.371	1.285	0.628	1.626
	GMSE	0.708	0.388	0.938	0.864	0.423	1.077
	PEM (MSE)	19.002	13.653	13.905	4.090	7.642	21.902
	PEM (bias)	-18.951	13.629	-13.772	-3.834	7.613	-21.842
	PEM (variance)	1.394	0.806	1.910	1.422	0.668	1.615
Special needs (SBO)	Vdesign	2.557	1.503	4.386	2.404	1.485	4.691
	GMSE	2.432	1.448	4.194	2.244	1.400	4.502
	PEM (MSE)	21.261	15.460	20.094	9.910	8.933	24.075
	PEM (bias)	-19.788	14.687	-15.023	-7.193	7.893	-20.693
	PEM (variance)	7.775	4.827	13.346	6.816	4.183	12.305

5.2.3.2 Specific schools

Table 17 presents the estimates in the external domain differentiating between specific schools. Results are in line with those observed in the small external subdomain special needs education, whereby V_{design} and $GMSE$ are well aligned, whereas PEM is severely overestimating the variance. It is important to note that this time, PEM was not estimated domain-specifically, since there was not enough sample data across all categories of the target variable in all schools. Therefore, a significant amount of overestimation is likely due to non domain-specific estimation. However, it can be concluded that PEM provides unreliable estimates in extremely small domains.

Table 17: CSRM (%) based on the different accuracy estimators for mother's education level cross-classified by five anonymous school indicators in subpopulations A and C.

School Type	Estimator	Population A			Population C		
		Low	Middle	High	Low	Middle	High
03HU	Vdesign	9.853	4.702	19.329	15.234	9.358	83.412
	GMSE	9.848	4.700	19.324	15.246	9.364	83.391
	PEM (MSE)	34.477	16.370	67.495	47.670	28.382	223.215
	PEM (bias)	7.303	-3.785	4.191	6.568	-2.157	-34.595
	PEM (variance)	33.695	15.926	67.365	47.216	28.299	220.518
03IK	Vdesign	21.265	8.866	38.204	26.501	8.463	40.423
	GMSE	21.263	8.866	38.202	26.499	8.462	40.405
	PEM (MSE)	75.662	30.731	122.734	92.565	27.680	104.185
	PEM (bias)	13.670	-4.505	-6.897	34.185	-8.129	-21.266
	PEM (variance)	74.412	30.399	122.540	86.021	26.460	101.992
04LD	Vdesign	12.054	7.374	29.509	20.556	7.280	31.167
	GMSE	12.051	7.372	29.506	20.539	7.274	31.138
	PEM (MSE)	39.762	24.531	96.598	67.610	22.252	78.937
	PEM (bias)	-5.611	3.634	-2.671	20.484	-3.241	-30.746
	PEM (variance)	39.364	24.261	96.561	64.432	22.015	72.703
04QV	Vdesign	14.172	5.259	28.277	11.594	9.561	28.509
	GMSE	14.168	5.258	28.275	11.593	9.560	28.493
	PEM (MSE)	51.460	18.266	91.074	37.550	29.481	75.095
	PEM (bias)	14.085	-3.717	-11.754	9.856	-5.022	-17.684
	PEM (variance)	49.495	17.883	90.313	36.234	29.050	72.983
06JP	Vdesign	11.223	4.245	30.202	16.116	7.736	34.380
	GMSE	11.219	4.243	30.196	16.108	7.732	34.362
	PEM (MSE)	38.480	14.233	96.768	50.855	23.480	88.980
	PEM (bias)	4.733	-0.397	-17.525	7.995	-0.903	-26.070
	PEM (variance)	38.188	14.227	95.168	50.222	23.463	85.076

5.2.4 Computational complexity

As explained in the earlier sections, the key motivation for evaluating the performance of PEM is its computational speed. Across the relatively large populations distinguished based on EAF data, PEM provided results extremely fast in comparison to the other two approaches. Computation of $GMSE$ in the entire population took 2 hours for population A, while V_{design} took 20 minutes, and PEM took approximately a second. This needs to be evaluated while considering two key points. First, the code for $GMSE$ is not the most optimal. It contains several nested loops and could be parallelized in parts. Second, both V_{design} and $GMSE$ contain computational bottlenecks, the main component of which comes from the estimation of the variance-covariance matrix of the imputation model. This, however, needs to be completed only once, and all domains of interest can be computed based on this much faster. Across both internal and external domains, although V_{design} was faster than $GMSE$, all computations remained under a minute. PEM remained around a second.

Chapter 6

Discussion

The thesis aimed to evaluate the robustness of three different approaches in estimating the bias and/or variance of the mass-imputed estimator. Two approaches from published literature and a novel, simpler approach were compared in a simulation and a case study.

6.1 Key findings

The simulation study demonstrated that all estimators can be considered robust variance estimators. Across several experimental conditions where the variance of the mass-imputed estimator increased while bias remained negligible, the accuracy estimators approximated the true simulated variance well. This was the case even when the conditions put the estimators in an unfavourable position due to the violation of assumptions. Minor systematic deviations from the benchmark values were observed, which, though practically negligible, are consistent with findings from previous research. Consistent with previous studies examining the behaviour of V_{design} by Scholtus and Daalmans (2021) and $GMSE$ by Alleva et al. (2021) and Deliu et al. (2025), estimation was less precise in small domains. In small domains, all estimators were more variable and prone to overestimation. A novel finding from the current study is that the reverse is true if the bias increases, whereby estimators become prone to underestimation in small domains. Another result consistent with Scholtus and Daalmans (2021) is that overparameterising the imputation model by estimating parameters with little predictive power leads to overestimation of variance, which was found to affect both V_{design} and $GMSE$, while PEM was the least affected given that it does not directly estimate model parameter error.

For the first time, it was demonstrated that the novel PEM approach could be considered a robust estimator of variance and bias, particularly in large domains and if the bias was not too large. It was found that generally, PEM tends to overestimate both bias and variance due to relying on averaged unit-level probabilities. This was pronounced if the estimated prediction error probabilities became closer to 0 and 1, as the variance function of multinomial variables is more sensitive to extreme probabilities. This was also pronounced if the domains were not taken into account when estimating the prediction error probabilities. This suggests that for the best performance of PEM, prediction error probabilities need to be estimated domain-specifically. Furthermore, PEM can likely be improved in general by estimation at even more granular levels, e.g., based on a combination of several domain variables.

Surprisingly, V_{design} and $GMSE$ performed similarly even if the assumptions for V_{design} were not met and the mass-imputed estimator was biased. This suggests that V_{design} is a more robust variance estimator than thought. On the other hand, due to the assumptions made during derivation, $GMSE$ appears to be limited in estimating bias, leading to the estimation of practically negligible bias at best. Furthermore, a

slight underestimation was observed under a non-ignorable sampling design, suggesting that, in practice, V_{design} is a more appropriate choice in this case. However, it may yield a more conservative estimate.

Encouragingly, the results from the simulation study were mainly replicated in the case study. Minor differences were observed in the pattern of the estimates provided by the three estimators compared to the simulation study. In the simulation study, V_{design} provided consistently smaller estimates than $GMSE$ and PEM, while the reverse was true in the case study. This could be due to the latter two estimators simplifying the effect of the sampling design on the variance estimate. Even though the sampling design was ignorable in theory, V_{design} incorporates it in the estimation of model parameter error, leading to larger standard errors of the parameters, which possibly resulted in a relatively larger variance estimate in the case study since the imputation model was much more extensive in comparison to the simulation study. On the other hand, the differences could partly be due to the simplified actual sampling design in the case study.

A finding from the case study that was not investigated during the simulation study revealed that in extremely small domains of less than 1000 units, PEM becomes unreliable due to a severe overestimation of variance compared to the other two approaches. This suggests that PEM is not appropriate for accuracy estimation in such domains. On the other hand, it presents a good alternative to the other approaches in large domains, especially given that it is easier and much faster to compute. Furthermore, it is likely that estimation in extremely small domains improves significantly if prediction error probabilities can be estimated at a more granular level.

6.2 Strengths and Limitations

The current study is the first to compare different accuracy estimators for the mass-imputed estimator. Furthermore, it is the first study to compare the estimators under suboptimal conditions, which can yet be expected to occur in practice. Given that the comparison proceeded across a relatively large set of experimental conditions defined based on realistic sample and population sizes, the results are a good guide for choosing between accuracy estimators in practice.

Unfortunately, the study was not conducted under a full factorial experimental design primarily due to computational reasons. For the same reason, the conditions were defined mostly in terms of one level, which did not enable the analysis of the gradient of the effects when, for example, the sample size gets incrementally smaller. As a result, it was tricky to disentangle some effects when analysing the behaviour of the estimators. For example, it remains somewhat unclear why V_{design} overestimated the variance in the case study. While the explanation provided above is plausible, several effects co-occur, such as the large number of parameters and omission of key interactions from the imputation model. In addition, several interesting extreme situations, such as non-ignorable sampling and small sample size, remained unexplored. Future research is encouraged to explore the effects at a more incremental level using a full factorial experimental design. Besides optimising the code for the estimators, the computational efficiency of such an experiment can be increased by ignoring the joint distribution, given that all estimators are essentially interested in estimation at the finite population level. Furthermore, there was limited evidence that estimators differ beyond stochastic noise at the level of the joint distribution.

Another limitation is that the analysis of the simulation study was based on comparing the estimators against the benchmark value. While informative, analysing the coverage rates of estimated confidence intervals might provide more conclusive results in terms of the robustness of estimation with respect to sampling or joint distribution. This type of analysis requires more simulation iterations and, therefore, was not conducted in the current study. Future research is encouraged to explore this at the level of the sampling distribution.

The final note concerns the evaluation of complex sampling designs. While the estimators were investigated under stratified random sampling, the purpose was to test the robustness of the estimators under non-ignorable sampling. Given that $GMSE$ appears relatively robust under this condition, future research could explore how $GMSE$ behaves under more complex sampling designs, whereby V_{design} might become cumbersome to compute.

6.3 Conclusion

The current study yielded encouraging results for the production of official statistics based on mass imputation. The variance of the mass-imputed estimator can be reliably estimated using several approaches, from which an optimal method can be chosen based on the specifics of the sample and the target variable. Furthermore, several interesting directions for future research can be suggested based on the results. It was established that $GMSE$ is a robust estimator of imputation-specific errors; therefore, research can now explore adapting it to other types of non-sampling errors. This suggestion also extends to PEM, which warrants further exploration due to its demonstrated speed and robustness as a fast alternative to other approaches. In addition, PEM was the only approach that showed promise in estimating the bias of the mass-imputed estimator, which remains an important challenge.

Acknowledgements

I want to thank my supervisors for their support throughout this thesis. Sander and Arnout, I will look back fondly to our weekly meetings, as they had an immense impact on shaping my reasoning, making me a more thorough, independent and patient thinker. Julian, I am very grateful for your help with presenting complicated concepts clearer.

Finally, I am grateful for Statistics Netherlands for creating opportunities for students to learn from impactful projects in a kind and productive atmosphere.

References

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.

Alleva, G., Falorsi, P. D., Petrarca, F., & Righi, P. (2021). Measuring the accuracy of aggregates computed from a statistical register. *Journal of Official Statistics*, *37*(2), 481–503. https://doi.org/10.2478/jos-2021-0021

Ascari, G., Blix, K., Brancato, G., Burg, T., McCourt, A., van Delden, A., Krapavickaitė, D., Ploug, N., Scholtus, S., Stoltze, P., et al. (2020). Quality of multisource statistics–the komuso project. *The Survey Statistician*, *81*, 36–51. https://etalpykla.vilniustech.lt/handle/123456789/148714

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, *46*(1), 1–29.

Beaumont, J.-F., & Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and bayesian approaches. *Canadian Journal of Statistics*, *50*(4), 1186–1212. https://doi.org/10.1002/cjs.11717

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. John Wiley & Sons.

Burger, J., van Delden, A., & Scholtus, S. (2015). Sensitivity of mixed-source statistics to classification errors. *Journal of Official Statistics*, *31*(3), 489–506. https://doi.org/10.1515/jos-2015-0029

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. John Wiley & Sons.

Chipperfield, J., Chessman, J., & Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, *54*(2), 223–238. https://doi.org/10.1111/j.1467-842X.2012.00666.x

Daalmans, J. (2017). *Mass imputation for census estimation* (Discussion Paper). Statistics Netherlands. https://www.cbs.nl/en-gb/background/2017/11/mass-imputation-for-census-estimation

De Waal, T. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the IAOS*, *32*(2), 231–243. https://doi.org/10.3233/SJI-150950

De Waal, T., & Daalmans, J. (2018). *Mass imputation for census estimation: Methodology* (tech. rep.). Statistics Netherlands.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. John Wiley & Sons.

Deliu, N., Falorsi, P. D., Falorsi, S., Chianella, D., & Alleva, G. (2025). Assessing the accuracy of multi-source register-based official statistics for multinomial outcomes. *arXiv preprint arXiv:2502.10182*. https://doi.org/10.48550/arXiv.2502.10182

Eurostat, E. C. (2018). European statistics code of practice: For the national statistical authorities and eurostat (eu statistical authority). https://doi.org/10.2785/798269

Eurostat, E. C. (2020). Quality assurance framework of the european statistical system: Version 2.0. https://doi.org/10.2785/847733

Falorsi, P. D., & Righi, P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey methodology*, *41*(1), 215–236. https://www.istat.it/wp-content/uploads/2023/12/B_14149-eng.pdf

Falorsi, S. (2017). Census and social surveys integrated system. *19th Meeting of the Group of Experts on Population and Housing Censuses, Geneva, Switzerland*, 4–6. https://www.istat.it/wp-content/uploads/2018/11/FalorsiS_original-paper.pdf

Golini, N., & Righi, P. (2024). Integrating probability and big non-probability samples data to produce official statistics. *Statistical Methods & Applications*, *33*(2), 555–580. https://doi.org/10.1007/s10260-023-00740-y

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*(384), 776–793. https://doi.org/10.1080/01621459.1983.10477018

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*(377), 89–96. https://doi.org/10.1080/01621459.1982.10477770

Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(3), 941–963. https://doi.org/10.1111/rssa.12696

Kim, J. K., & Rao, J. N. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, *99*(1), 85–100. https://doi.org/10.1093/biomet/asr063

Kooiman, P. (1998). Massa-imputatie: Waarom niet!? [Internal note, Statistics Netherlands].

Linder, F., Van Roon, D., & Bakker, B. F. (2011). Combining data from administrative sources and sample surveys; the single variable case. In *Essnet data integration. wp4 case studies* (pp. 39–97). Eurostat.

Lohr, S. L. (2021). *Sampling: Design and analysis*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429298899

Lundy, E. R. (2022). Predicting the quality and evaluating the use of administrative data for the 2021 canadian census of population. *Statistical Journal of the IAOS*, *38*(4), 1177–1183. https://doi.org/10.3233/SJI-220082

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, *12*(2), 685–726.

Posthumus, H., Bakker, B., van der Laan, J., de Mooij, M., Scholtus, S., Tepic, M., van den Tillaart, J., & de Vette, N. (2016). *Herziening gewichtenregeling primair onderwijs – fase i* (Technical Report) (In Dutch). Statistics Netherlands. https://shorturl.at/CWT9P

Posthumus, H., Scholtus, S., & Walhout, J. (2019). *Nieuwe onderwijsachterstandenindicator primair onderwijs: Samenvattend rapport* (Technical Report). Statistics Netherlands. https://shorturl.at/Z6qpj

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rao, J. N. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*(1), 242–272. https://doi.org/10.1007/s13571-020-00227-w

Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Rocci, F., Varriale, R., & Luzi, O. (2022). Total process error: An approach for assessing and monitoring the quality of multisource processes. *Journal of Official Statistics*, *38*(2), 533–556. https://doi.org/10.2478/jos-2022-0025

Rubin, D. B. (1987). Multiple imputation for survey nonresponse.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Scholtus, S. (2018). *Variances of census tables after mass imputation of educational attainment* (Discussion Paper). Statistics Netherlands. https://www.cbs.nl/en-gb/background/2018/49/variances-of-census-tables-after-mass-imputation

Scholtus, S., & Daalmans, J. (2021). Variance estimation after mass imputation based on combined administrative and survey data. *Journal of Official Statistics*, *37*(2), 433–459. https://doi.org/10.2478/jos-2021-0019

Scholtus, S., & Pannekoek, J. (2015). *Mass-imputation of educational levels (in dutch)* (Internal report). Statistics Netherlands.

Statistics Netherlands. (2024). *Onderzoek herijking risico-indicator onderwijsachterstanden – fase 1* (Technical Report). Statistics Netherlands. https://shorturl.at/fPVQh

Sugden, R., & Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, *71*(3), 495–506. https://doi.org/10.1093/biomet/71.3.495

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, *57*(1), 1–436.

Tillé, Y., Debusschere, M., Luomaranta, H., Axelson, M., Elvers, E., Holmberg, A., & Valliant, R. (2022). Some thoughts on official statistics and its future (with discussion). *Journal of Official Statistics*, *38*(2), 557–598. https://doi.org/10.2478/jos-2022-0026

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: A prediction approach*. Wiley New York.

van Delden, A., Scholtus, S., & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of official statistics*, *32*(3), 619–642. https://doi.org/10.1515/jos-2016-0032

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

Appendix A

Code

All the data and code to replicate the simulation study can be found from this GitHub link. Since the case study has been carried out using confidential data, the relevant scripts can be requested privately.

Appendix B

Details on estimating V_{design} for binomial and multinomial model

This appendix is based on notes provided by Dr. Sander Scholtus, used with permission.

B.1 Logistic regression

Given independent imputations, the first term in Equation 2.11 can be computed as is. The second term referring to design-based covariance of predicted probabilities can be estimated upon first-order Taylor series approximation of the logistic function $\frac{1}{1+\exp(-\boldsymbol{x}_k^T\boldsymbol{\beta})}$ (see Scholtus, 2018 and Scholtus and Daalmans, 2021 for details), resulting in

$$\hat{cov}_P(\hat{p}_k, \hat{p}_l) = \hat{p}_k(1-\hat{p}_k)\,\boldsymbol{x_k'}(\boldsymbol{X'\hat{\Delta}_w X})^{-1}\hat{\boldsymbol{\Gamma}}(\boldsymbol{X'\hat{\Delta}_w X})^{-1}\boldsymbol{x_l}\,\hat{p}_l(1-\hat{p}_l) \tag{B.1}$$

where $X'\hat{\Delta}_w X$ is the information matrix of a weighted logistic regression model,

$$\mathcal{I} = -\frac{\partial^2 \ell}{\partial\beta\,\partial\beta'} = \sum_{k\in s}\hat{p}_k(1-\hat{p}_k)w_k\,\mathbf{x}_k\mathbf{x}_k' = \mathbf{X}'\Delta_w\mathbf{X} \tag{B.2}$$

with $w_k = 1/\pi_k$ if $k \in s$ and $w_k = 0$ otherwise, and

$$\hat{\Gamma} = \sum_{k\in S}\sum_{l\in S}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\,\boldsymbol{x_k}(y_k-\hat{p}_k)(y_l-\hat{p}_l)\boldsymbol{x_l'}$$

which is the standard Horvitz-Thompson estimator for $var(\sum_{k\in s} w_k(y_k-p_k)\mathbf{x}_k)$ (Särndal et al., 1992, p.48).

B.2 Multinomial logistic regression

For a multinomial target variable, Equation B.1 takes a similar form, except that the information matrix needs to now account for $C-1$ sets of parameters (Agresti, 2013, Chapter 8). In addition, we need to linearise the probabilities with respect to the softmax function.

The information matrix for a multinomial logistic regression model takes a block-form,

$$I = \begin{pmatrix} I_{1,1} & \cdots & I_{1,C-1} \\ \vdots & \ddots & \vdots \\ I_{C-1,1} & \cdots & I_{C-1,C-1} \end{pmatrix}$$

where $I_{c,d} = -\frac{\partial^2\ell}{\partial\boldsymbol{\beta}_c\,\partial\boldsymbol{\beta}_d^T} = \mathbf{X}^T\Delta_{c,d}\mathbf{X}$, with

$$(\Delta_{c,d})_{jj} = \begin{cases} \hat{p}_{ck}(1-\hat{p}_{ck})w_k & \text{if } c = d \\ -\hat{p}_{ck}\hat{p}_{dk}w_k & \text{if } c \neq d \end{cases}$$

First-order Taylor linearisation of the softmax function results in following expressions for the predicted probabilities:

$$\hat{p}_{ck} \approx \tilde{y}_{ck} + \tilde{y}_{ck}(1-\tilde{y}_{ck})\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c) - \sum_{\substack{d=1\\d\neq c}}^{C-1} \tilde{y}_{ck}\tilde{y}_{dk}\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d), \quad c = 1,\ldots,C-1,$$

$$\hat{p}_{Ck} \approx \tilde{y}_{Ck} - \sum_{c=1}^{C-1} \tilde{y}_{Ck}\tilde{y}_{ck}\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}_c)$$

the variances and covariances of which can be approximated by

$$\begin{aligned}
\operatorname{var}(\hat{p}_{ck}) &\approx \operatorname{var}\left(\sum_{d=1}^{C-1} (\Delta_{c,d})_{kk}\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d)\right)\\
&= \sum_{d=1}^{C-1}\sum_{d'=1}^{C-1} (\Delta_{c,d})_{kk}\,\mathbf{x}_k^T \operatorname{cov}(\hat{\boldsymbol{\beta}}_d, \hat{\boldsymbol{\beta}}_{d'})\,\mathbf{x}_k\,(\Delta_{c,d'})_{kk}\\
&= \sum_{d=1}^{C-1}\sum_{d'=1}^{C-1} (\Delta_{c,d})_{kk}\,\mathbf{x}_k^T V_{d,d'}\,\mathbf{x}_k\,(\Delta_{c,d'})_{kk}
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}(\hat{p}_{Ck}) &\approx \operatorname{var}\left(\sum_{d=1}^{C-1} \tilde{y}_{Ck}\tilde{y}_{dk}\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d)\right)\\
&= \sum_{d=1}^{C-1}\sum_{d'=1}^{C-1} \tilde{y}_{Ck}\tilde{y}_{dk}\,\mathbf{x}_k^T V_{d,d'}\,\mathbf{x}_k\,\tilde{y}_{Ck}\tilde{y}_{d'k}
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(\hat{p}_{ck}, \hat{p}_{c'l}) &\approx \operatorname{cov}\left(\sum_{d=1}^{C-1} (\Delta_{c,d})_{kk}\,\mathbf{x}_k^T(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d), \sum_{d=1}^{C-1} (\Delta_{c',d})_{ll}\,\mathbf{x}_l^T(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d)\right)\\
&= \sum_{d=1}^{C-1}\sum_{d'=1}^{C-1} (\Delta_{c,d})_{kk}\,\mathbf{x}_k^T V_{d,d'}\,\mathbf{x}_l\,(\Delta_{c',d'})_{ll}
\end{aligned}$$

where $V_{d,d'}$ is the estimated covariance matrix of the pair of estimated parameter vectors $\hat{\boldsymbol{\beta}}_d$ and $\hat{\boldsymbol{\beta}}_{d'}$, which is derived using the information matrix I, similar to the binomial case. Note that if c and/or $c' = C$, then $(\Delta_{c,d})_{kk}$ and $(\Delta_{c',d'})_{kk}$ should be replaced by $-\hat{p}_{Ck}\hat{p}_{dk}w_k$ and $-\hat{p}_{Ck}\hat{p}_{d'k}w_k$ respectively.

Appendix C

Details on estimating GMSE for binomial and multinomial target variables

C.1 Multinomial logistic regression

The two-step linearisation procedure is described in detail in Deliu et al. (2025). The notation below is adapted for consistency with the current thesis.

The two-step linearisation proceeds by first linearising the mass-imputed estimator with respect to the estimated model parameters under the model distribution and then linearising the estimated model parameters with respect to their expected value under the sampling distribution, resulting in the following expression for $GMSE$ based in stochastic imputation:

$$\text{GMSE}(\hat{Y}_{U_{d,c}}) = \boldsymbol{\gamma}_d^T \mathbf{F}_c \left(\sum_{i=1}^{N} \pi_k \bar{U}_k \Delta_{c,d} \bar{U}_k \right) \mathbf{F}_c^T \boldsymbol{\gamma}_d + \sum_{k=1}^{N} \gamma_{d,k} \hat{p}_{c,k} \left(1 - \hat{p}_{c,k}\right), \quad \text{for } c = 1, \ldots, C$$

where $\mathbf{F}_c$ is an $N \times H$ matrix where $H = C \times J$ composed of the first-order derivatives of the softmax function evaluated at the expected value of the model parameters, so that each element in $\mathbf{F}_c$ is

$$\mathbf{F}_{c,k} = \begin{cases} \boldsymbol{x_{kj}}\, \hat{p}_{ck}(1 - \hat{p}_{ck}), & \text{if } c = c' \\ -\boldsymbol{x_{kj}}\, \hat{p}_{ck}\hat{p}_{ck}, & \text{if } c \neq c \end{cases}$$

$\bar{U}_k$ is an $H \times C$ matrix equivalent to $I^{-1}X_k$ where X_k is an $H \times C$ matrix

$$\mathbf{X}_k = \begin{bmatrix} x_{k1} & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ x_{kJ} & \cdots & \cdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \vdots & \cdots & \cdots & x_{k1} \\ \vdots & \cdots & \cdots & \vdots \\ 0 & \vdots & 0 & x_{kJ} \end{bmatrix}$$

$\Delta_{c,d}$ is a $C \times C$ variance-covariance matrix of Y with elements $\hat{p}_{ck}(1 - \hat{p}_{ck})$ on the main diagonal and $-\hat{p}_{ck}\hat{p}_{dk}$ on the off diagonals.

C.2 Logistic regression

For logistic regression, $\mathbf{F}_{c,k}$ reduces to $\mathbf{F}_k = \boldsymbol{x_k}\,\hat{p}_k(1-\hat{p}_k)$, I reduces to Equation B.2 and the variance-covariance matrix reduces to Δ with elements $\hat{p}_k(1-\hat{p}_k)$ on the main diagonal.

Appendix D

Derivation of PEM

This appendix is based on notes provided by Dr. Sander Scholtus, used with permission.

D.1 Binary target variable

Recall from Section 2.2.1 that for each unit k we obtain a predicted probability $\hat{p}_k = P(\hat{y}_k = 1 \mid \boldsymbol{x}_k, s)$. Let p_k present the true underlying probability. In what follows, conditioning on the auxiliary variables $\boldsymbol{x}_k$ is excluded to simplify notation, unless it is explicitly needed, but is supposed to be there throughout.

As a simplification, it is assumed that s is a SRSWOR. Additionally, let $\hat{Y}_{U_d} \approx \sum_{k \in s} \gamma_{d,k} y_k + \sum_{k \in U \setminus s} \gamma_{d,k} \hat{y}_k$, that is to say that we ignore the difference between imputing the entire population or only the non-sampled part of the population, which is reasonable if the sampling fraction is small since the finite population correction factor becomes negligible.

From a design-based point of view, the bias and variance of $\hat{Y}_{U_d}$ as an estimator of Y_{U_d} are:

$$\mathbb{E}_P(\hat{Y}_{U_d} - Y_{U_d}) = \mathbb{E}_P \left\{ \sum_{k \in U} \gamma_{d,k} (\hat{y}_k - y_k) \right\};$$

$$\mathbb{V}_P(\hat{Y}_{U_d} - Y_{U_d}) = \mathbb{V}_P \left\{ \sum_{k \in U} \gamma_{d,k} (\hat{y}_k - y_k) \right\}.$$

For the bias, we can write:

$$\mathbb{E}_P(\hat{Y}_{U_d} - Y_{U_d}) = \sum_{k \in U} \gamma_{d,k} \left\{ \mathbb{E}_P(\hat{y}_k) - y_k \right\} = \sum_{k \in U} \gamma_{d,k} \left\{ \mathbb{E}_P \left[\mathbb{E}(\hat{y}_k \mid s) \right] - y_k \right\},$$

which yields

$$\mathbb{E}_P(\hat{Y}_{U_d} - Y_{U_d}) = \sum_{k \in U} \gamma_{d,k} \left\{ \mathbb{E}_P(\hat{p}_k) - y_k \right\}. \tag{D.1}$$

For the variance, it follows analogously to Scholtus and Daalmans (2021) that

$$\mathbb{V}_P(\hat{Y}_{U_d} - Y_{U_d}) = \mathbb{E}_P \left\{ \sum_{k \in U} \gamma_{d,k} \, \mathbb{V}(\hat{y}_k - y_k \mid s) \right\} + \mathbb{V}_P \left\{ \sum_{k \in U} \gamma_{d,k} \left[\mathbb{E}(\hat{y}_k \mid s) - y_k \right] \right\}$$

$$= \sum_{k \in U} \gamma_{d,k} \, \mathbb{E}_P \left\{ \hat{p}_k (1 - \hat{p}_k) \right\} + \mathbb{V}_P \left(\sum_{k \in U} \gamma_{d,k} \hat{p}_k \right)$$

$$= \sum_{k \in U} \gamma_{d,k} \left\{ \mathbb{E}_P(\hat{p}_k) - \left[\mathbb{E}_P(\hat{p}_k) \right]^2 - \mathbb{V}_P(\hat{p}_k) \right\} + \mathbb{V}_P \left(\sum_{k \in U} \gamma_{d,k} \hat{p}_k \right).$$

We write this as

$$\mathbb{V}_P(\hat{Y}_{U_d} - Y_{U_d}) = \sum_{k \in U} \gamma_{d,k}\, \mathbb{E}_P(\hat{p}_k)\,[1 - \mathbb{E}_P(\hat{p}_k)] + \mathbb{V}_P\left(\sum_{k \in U} \gamma_{d,k} \hat{p}_k\right) - \sum_{k \in U} \gamma_{d,k}\, \mathbb{V}_P(\hat{p}_k). \tag{D.2}$$

Below it will be shown that the final term is negligible in practice.
In Equations D.1 and D.2, we can replace all instances of $\hat{p}_k$ by $y_k \hat{p}_k + (1 - y_k)\hat{p}_k$. In addition, assume that it is reasonable to use the following approximation for all units in domain d:

$$\hat{p}_k = P(\hat{y}_k = 1 \mid x_k, s) \approx P(\hat{y}_k = 1 \mid y_k, s) \approx \begin{cases} p_{11,ds} & \text{if } y_k = 1 \\ 1 - p_{00,ds} & \text{if } y_k = 0 \end{cases} \tag{D.3}$$

where

$$p_{11,ds} = \frac{\sum_{k \in s} \gamma_{d,k}\, y_k\, p_k}{\sum_{k \in s} \gamma_{d,k}\, y_k}; \quad p_{00,ds} = \frac{\sum_{k \in s} \gamma_{d,k}\, (1 - y_k)(1 - p_k)}{\sum_{k \in s} \gamma_{d,k}\, (1 - y_k)}. \tag{D.4}$$

resulting in

$$\hat{p}_k = y_k \hat{p}_k + (1 - y_k)\hat{p}_k \approx y_k\, p_{11,ds} + (1 - y_k)(1 - p_{00,ds}). \tag{D.5}$$

According to standard results from sampling theory (see, e.g., Särndal et al., 1992, Section 5.8 and Exercise 5.34), it holds that

$$\begin{aligned}
\mathbb{E}_P(p_{11,ds}) &\approx \frac{\sum_{k \in U} \gamma_{d,k}\, y_k\, p_k}{\sum_{k \in U} \gamma_{d,k}\, y_k} \equiv p_{11,dU}, \\
\mathbb{E}_P(p_{00,ds}) &\approx \frac{\sum_{k \in U} \gamma_{d,k}\, (1 - y_k)(1 - p_k)}{\sum_{k \in U} \gamma_{d,k}\, (1 - y_k)} \equiv p_{00,dU}, \\
\mathbb{V}_P(p_{11,ds}) &\approx \frac{N}{N_{d1}} \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) p_{11,dU}(1 - p_{11,dU}), \\
\mathbb{V}_P(p_{00,ds}) &\approx \frac{N}{N_{d0}} \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) p_{00,dU}(1 - p_{00,dU}), \\
\mathbb{C}_P(p_{11,ds}, p_{00,ds}) &\approx 0.
\end{aligned} \tag{D.6}$$

where N and n denote the size of U and s. The covariance in the final line is approximately zero because the two estimators are computed on disjoint subsets of sample s.

Using approximations D.5 and D.6, we obtain

$$\mathbb{E}_P(\hat{p}_k) \approx y_k\, \mathbb{E}_P(p_{11,ds}) + (1 - y_k)\, \mathbb{E}_P(1 - p_{00,ds}) \approx y_k\, p_{11,dU} + (1 - y_k)(1 - p_{00,dU})$$

and

$$\begin{aligned}\mathbb{V}_P\left(\sum_{k\in U}\gamma_{d,k}\hat{p}_k\right) &\approx \mathbb{V}_P\left\{\sum_{k\in U}\gamma_{d,k}\left[y_k\, p_{11,ds} + (1-y_k)(1-p_{00,ds})\right]\right\}\\ &= \mathbb{V}_P\left\{N_{d1}\, p_{11,ds} + N_{d0}\,(1-p_{00,ds})\right\}\\ &= N_{d1}^2\,\mathbb{V}_P(p_{11,ds}) + N_{d0}^2\,\mathbb{V}_P(p_{00,ds})\\ &\quad + 2N_{d1}N_{d0}\,\mathbb{C}_P(p_{11,ds}, p_{00,ds})\\ &\approx \frac{N}{n}\left(1-\frac{n}{N}\right)\left\{N_{d1}p_{11,dU}(1-p_{11,dU}) + N_{d0}p_{00,dU}(1-p_{00,dU})\right\}.\end{aligned}$$

Moreover,

$$\mathbb{V}_P(\hat{p}_k) \approx \frac{N}{n}\left(1-\frac{n}{N}\right)\left\{\frac{1}{N_{d1}}y_k\, p_{11,dU}(1-p_{11,dU}) + \frac{1}{N_{d0}}(1-y_k)\, p_{00,dU}(1-p_{00,dU})\right\}.$$

From the last two expressions it is clear that $\sum_{k\in U}\gamma_{d,k}\,\mathbb{V}_P(\hat{p}_k) \ll \mathbb{V}_P\left(\sum_{k\in U}\gamma_{d,k}\hat{p}_k\right)$, which shows that the final term in Equation D.2 may be ignored in practice.

Substituting the approximate results for the expected value and variance of $\hat{p}_k$ into Equations D.1 and D.2, we have for the following result for the bias

$$\mathbb{E}_P\big(\hat{Y}_{U_d} - Y_{U_d}\big) \approx \sum_{k\in U}\gamma_{d,k}\left\{y_k\big(p_{11,dU}-1\big) + (1-y_k)(1-p_{00,dU})\right\} \approx N_{d1}\,(p_{11,dU}-1) + N_{d0}\,(1-p_{00,dU}) \tag{D.7}$$

For Equation D.2 for the variance, it is seen that the first term $\sum_{k\in U}\gamma_{d,k}\mathbb{E}_P(\hat{p}_k)\,\{1-\mathbb{E}_P(\hat{p}_k)\}$ may be approximated by

$$\begin{aligned}&\sum_{k\in U}\gamma_{d,k}\left[y_k p_{11,dU} + (1-y_k)(1-p_{00,dU})\right]\left\{1-\left[y_k p_{11,dU} + (1-y_k)(1-p_{00,dU})\right]\right\}\\ &= N_{d1}p_{11,dU}(1-p_{11,dU}) + N_{d0}p_{00,dU}(1-p_{00,dU})\end{aligned}$$

where we used that $y_k^2 = y_k$, $(1-y_k)^2 = (1-y_k)$, and $y_k(1-y_k) = 0$.

It follows that

$$\begin{aligned}\mathbb{V}_P\left(\hat{Y}_{U_d} - Y_{U_d}\right) &\approx N_{d1}p_{11,dU}(1-p_{11,dU}) + N_{d0}p_{00,dU}(1-p_{00,dU})\\ &\quad + \frac{N}{n}\left(1-\frac{n}{N}\right)\left\{N_{d1}p_{11,dU}(1-p_{11,dU}) + N_{d0}p_{00,dU}(1-p_{00,dU})\right\}.\end{aligned}$$

Finally, using that $1 + \frac{N}{n}\left(1-\frac{n}{N}\right) = \frac{N}{n}$, we obtain:

$$\mathbb{V}_P\left(\hat{Y}_{U_d} - Y_{U_d}\right) \approx \frac{N}{n}\left\{N_{d1}p_{11,dU}(1-p_{11,dU}) + N_{d0}p_{00,dU}(1-p_{00,dU})\right\}. \tag{D.8}$$

The above results could be generalized relatively easily to the case where the sampling design of s is more complicated than SRSWOR. In the general case, D.4 should incorporate the inclusion weights $w_k = 1/\pi_k$

of sample s:

$$p_{11,ds} = \frac{\sum_{k\in s} \gamma_{d,k} w_k \, y_k \, p_k}{\sum_{k\in s} \gamma_{d,k} w_k \, y_k}; \quad p_{00,ds} = \frac{\sum_{k\in s} \gamma_{d,k} w_k \, (1-y_k)(1-p_k)}{\sum_{k\in s} \gamma_{d,k} \, w_k (1-y_k)}.$$

In Equation D.6, the variance formula based on SRSWOR should be replaced by the general variance formula of a ratio of Horvitz-Thompson estimators, and the subsequent expressions need to be adjusted accordingly.

To estimate the above bias and variance approximations in practice, the following approach seems attractive. First, to estimate $p_{11,ds}$ and $p_{00,ds}$, one could use

$$\hat{p}_{11,ds} = \frac{\sum_{k\in s} \gamma_{d,k} w_k \, y_k \, \hat{p}_k}{\sum_{k\in s} \gamma_{d,k} w_k \, y_k}; \quad \hat{p}_{00,ds} = \frac{\sum_{k\in s} \gamma_{d,k} w_k \, (1-y_k)(1-\hat{p}_k)}{\sum_{k\in s} \gamma_{d,k} \, w_k (1-y_k)}.$$

In the case of simple random sampling, the weights $w_k = N/n$ may be left out. Next, formulas D.7 and D.8 could be estimated by

$$\hat{\mathbb{E}}_P \left(\hat{Y}_{U_d} - Y_{U_d} \right) = \hat{N}_{d1U} \left(\hat{p}_{11,ds} - 1 \right) + \hat{N}_{d0U} \left(1 - \hat{p}_{00,ds} \right),$$

$$\hat{\mathbb{V}}_P \left(\hat{Y}_{U_d} - Y_{U_d} \right) = \frac{N}{n} \left\{ \hat{N}_{d1U} \hat{p}_{11,ds} (1 - \hat{p}_{11,ds}) + \hat{N}_{d0U} \hat{p}_{00,ds} (1 - \hat{p}_{00,ds}) \right\},$$

with $\hat{N}_{d1U} = \sum_{k\in U} \gamma_{d,k} \hat{p}_k$, and $\hat{N}_{d0U} = \sum_{k\in U} \gamma_{d,k} (1 - \hat{p}_k)$.

In all of these expressions, $\hat{p}_k$ could also be replaced by $\hat{y}_k$, but this would add additional noise due to stochastic imputation of $\hat{y}_k$, potentially leading to less precise estimates in the case of $\hat{p}_{11,ds}$ and $\hat{p}_{00,ds}$, and even biased estimates in the case of $\hat{N}_{d1U}$ and $\hat{N}_{d0U}$ (van Delden et al., 2016). Alternatively, we could estimate N_{d1} and N_{d0} in EquationsD.7 and D.8 directly from the sample by $\hat{N}_{d1s} = \sum_{k\in s} \gamma_{d,k} w_k y_k$ and $\hat{N}_{d0s} = \sum_{k\in s} \gamma_{d,k} w_k y_k$. This has the advantage that it does not rely on the imputation model being correct, but also the disadvantage of being potentially inaccurate if the sample size is small.

D.2 Multinomial target variable

The approach can be extended to $C \geq 2$. Introducing $\hat{p}_{c,k} = P(\hat{y}_{c,k} = 1 \mid \boldsymbol{x}_k, s)$, we now obtain instead of D.1 and D.2

$$\mathbb{E}_P \left(\hat{Y}_{U_d,c} - Y_{U_d,c} \right) = \sum_{k\in U} \gamma_{d,k} \left\{ \mathbb{E}_P \left(\hat{p}_{c,k} \right) - y_{c,k} \right\} \tag{D.9}$$

and

$$\mathbb{V}_P \left(\hat{Y}_{U_d,c} - Y_{U_d,c} \right) = \sum_{k\in U} \gamma_{d,k} \, \mathbb{E}_P \left(\hat{p}_{c,k} \right) \left\{ 1 - \mathbb{E}_P \left(\hat{p}_{c,k} \right) \right\} + \mathbb{V}_P \left(\sum_{k\in U} \gamma_{d,k} \hat{p}_{c,k} \right) - \sum_{k\in U} \gamma_{d,k} \, \mathbb{V}_P \left(\hat{p}_{c,k} \right) \tag{D.10}$$

Approximation D.5 is now replaced by

$$\hat{p}_{c,k} = \sum_{g=1}^{C} y_{g,k} \hat{p}_{c,k} \approx \sum_{g=1}^{C} y_{g,k} \, p_{gc,ds} \tag{D.11}$$

where for all combinations of $g \in \{1, \ldots, C\}$ and $h \in \{1, \ldots, C\}$

$$p_{gh,ds} = \frac{\sum_{k \in S} \gamma_{d,k}\, y_{g,k}\, p_{h,k}}{\sum_{k \in S} \gamma_{d,k}\, y_{g,k}}. \tag{D.12}$$

Instead of D.6, we obtain

$$\begin{aligned}
\mathbb{E}_P\left(p_{gh,ds}\right) &\approx \frac{\sum_{k \in U} \gamma_{d,k}\, y_{g,k}\, p_{h,k}}{\sum_{k \in U} \gamma_{d,k}\, y_{g,k}} \equiv p_{gh,dU}, \\
\mathbb{V}_P\left(p_{gh,ds}\right) &\approx \frac{N}{N_{dg}} \cdot \frac{1}{n}\left(1 - \frac{n}{N}\right) p_{gh,dU}(1 - p_{gh,dU}), \\
\mathbb{C}_P\left(p_{gh,dS}, p_{g'h',ds}\right) &\approx 0 \quad \text{if } g \neq g', \\
\text{with } N_{dg} &= \sum_{k \in U} \gamma_{d,k}\, y_{g,k}.
\end{aligned} \tag{D.13}$$

Instead of D.1 and D.2 we obtain

$$\mathbb{E}_P\big(\hat{Y}_{U_d,c} - Y_{U_d,c}\big) \approx N_{dc}\left(p_{cc,dU} - 1\right) + \sum_{\substack{g=1 \\ g \neq c}}^{C} N_{dg}\, p_{gc,dU} \tag{D.14}$$

$$\mathbb{V}_P\left(\hat{Y}_{U_d,c} - Y_{U_d,c}\right) \approx \frac{N}{n} \sum_{g=1}^{C} N_{dg}\, p_{gc,dU}\left(1 - p_{gc,dU}\right) \tag{D.15}$$

which can be estimated in practice by

$$\begin{aligned}
\hat{\mathbb{E}}_P\left(\hat{Y}_{U_d,c} - Y_{U_d,c}\right) &= \hat{N}_{dcU}\left(\hat{p}_{cc,ds} - 1\right) + \sum_{\substack{g=1 \\ g \neq c}}^{C} \hat{N}_{dgU}\, \hat{p}_{gc,ds}, \\
\hat{\mathbb{V}}_P\left(\hat{Y}_{U_d,c} - Y_{U_d,c}\right) &= \frac{N}{n} \sum_{g=1}^{C} \hat{N}_{dgU}\, \hat{p}_{gc,ds}\left(1 - \hat{p}_{gc,ds}\right),
\end{aligned} \tag{D.16}$$

with

$$\hat{p}_{gh,ds} = \frac{\sum_{k \in s} \gamma_{d,k} w_k y_{g,k} \hat{p}_{h,k}}{\sum_{k \in s} \gamma_{d,k} w_k y_{g,k}}, \quad \hat{N}_{dgU} = \sum_{k \in U} \gamma_{d,k} \hat{p}_{g,k}.$$

Appendix E

True and estimated domain sizes across superpopulation 1 and 2

Table 18: Domain sizes for binomial target variable for category "has higher education" in superpopulation 1

Condition	Domain	Distribution	True total	Mass-imputed estimator
Baseline	Male	Joint	30 154.42	30 156.44
		Sampling	30 100.00	30 100.35
	Female	Joint	30 961.23	30 965.78
		Sampling	30 754.00	30 815.81
Small sample	Male	Joint	30 168.02	30 179.52
		Sampling	30 354	30 491.87
	Female	Joint	30 956.37	30 955.99
		Sampling	31 076	31 165.25
Non-ignorable	Male	Joint	30 163.53	30 160.17
		Sampling	30 028	30 130.49
	Female	Joint	30 942.56	30 949.39
		Sampling	30 684	30 517.57
Small population	Male	Joint	12 069.28	12 068.32
		Sampling	11 934	11 971.61
	Female	Joint	12 381.79	12 383.95
		Sampling	12 480	12 418.34
Overparameterization	Male	Joint	30 180.92	30 183.98
		Sampling	30 202	30 258.56
	Female	Joint	30 960.72	30 953.10
		Sampling	30 992	31 016.30
Design bias	A	Joint	30 388.58	30 395.37
		Sampling	30 372	30 364.34
	B	Joint	30 747.76	30 742.18
		Sampling	30 654	30 705.62
Model bias (small)	Male	Joint	30 165.11	32 757.35

Condition	Domain	Distribution	True total	Mass-imputed estimator
		Sampling	30 206	32 730.56
	Female	Joint	30 944.87	28 363.50
		Sampling	31 013	28 464.86
Model bias (large)	Male	Joint	36 530.24	30 731.35
		Sampling	36 475	30 676.31
	Female	Joint	24 614.01	30 404.74
		Sampling	24 680	30 353.78
Model bias (large) x small n	Male	Joint	36 569.79	30 745.29
		Sampling	36 618	31 120.68
	Female	Joint	24 591.00	30 418.50
		Sampling	24 847	30 817.01

Table 19: Domain sizes for binomial target variable for category "has higher education" in superpopulation 2

Condition	Domain	Distribution	True total	Mass-imputed estimator
Baseline	Male	Joint	30 154.42	30 156.44
		Sampling	30 100	30 100.35
	Female	Joint	30 961.23	30 965.78
		Sampling	30 754	30 815.81
Small sample	Male	Joint	30 163.07	30 155.82
		Sampling	30 354	30 187.12
	Female	Joint	30 953.59	30 952.74
		Sampling	31 076	30 828.56
Non-ignorable	Male	Joint	30 176.77	30 180.64
		Sampling	30 102	30 009.65
	Female	Joint	30 948.79	30 948.71
		Sampling	30 857	30 863.13
Small population	Male	Joint	12 078.28	12 078.76
		Sampling	12 089	12 087.54
	Female	Joint	12 374.18	12 373.56
		Sampling	12 364	12 362.93
Overparameterization	Male	Joint	30 174.23	30 170.0

Condition	Domain	Distribution	True total	Mass-imputed estimator
		Sampling	30 026	30 022.46
	Female	Joint	30 962.10	30 959.4
		Sampling	31 099	31 116.52
Design bias	Male	Joint	30 388.58	30 395.37
		Sampling	30 372	30 364.34
	Female	Joint	30 747.76	30 742.18
		Sampling	30 654	30 705.62
Model bias (small)	Male	Joint	30 162.08	30 492.80
		Sampling	30 188	30 511.03
	Female	Joint	30 948.30	30 616.72
		Sampling	31 033	30 651.17
Model bias (large)	Male	Joint	36 539.05	34 487.37
		Sampling	36 536	34 491.95
	Female	Joint	24 594.75	26 640.22
		Sampling	24 579	26 669.45
Model bias (large) x small n	Male	Joint	36 547.23	34 474.56
		Sampling	36 502	34 459.62
	Female	Joint	24 568.52	26 632.75
		Sampling	24 590	26 570.60

Appendix F

CSRMs for benchmarks and accuracy estimators across the experimental conditions with the binomial target variable

Table 20: CSRM (%) for Superpopulation 1

Condition	Domain	Distribution	MSE	Bias	Variance	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
Baseline	Male	Joint	**1.795**	**0.281**	**1.787**	1.790	1.834	1.889	0.410	1.844
		Sampling	**1.887**	**-0.281**	**1.862**	1.786	1.830	1.891	−0.317	1.847
	Female	Joint	**1.714**	**0.175**	**1.705**	1.717	1.759	1.811	0.397	1.767
		Sampling	**1.738**	**0.214**	**1.734**	1.735	1.778	1.804	−0.336	1.760
Small sample	Male	Joint	**4.056**	**0.382**	**4.038**	4.031	4.022	4.233	−0.948	4.125
		Sampling	**4.055**	**0.454**	**4.030**	4.0	3.989	4.184	0.687	4.084
	Female	Joint	**3.846**	**0.363**	**3.829**	3.861	3.856	4.055	0.901	3.953
		Sampling	**3.900**	**0.287**	**3.890**	3.842	3.835	4.018	−0.689	3.926
Small population	Male	Joint	**2.827**	**-0.278**	**2.811**	2.835	2.902	2.990	0.655	2.917
		Sampling	**2.681**	**0.315**	**2.773**	2.838	2.904	3.001	0.490	2.940
	Female	Joint	**2.707**	**0.286**	**2.695**	2.722	2.788	2.862	−0.620	2.793
		Sampling	**2.592**	**-0.494**	**2.791**	2.682	2.748	2.855	−0.498	2.786

Condition	Domain	Distribution	RRMSE	RB	CV	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
Non-ignorable	Male	Joint	**2.102**	**-0.185**	**2.093**	2.118	1.915	2.272	0.585	2.195
		Sampling	**2.003**	**-0.204**	**1.973**	2.125	1.914	2.254	−0.462	2.184
	Female	Joint	**2.971**	**0.286**	**2.956**	2.936	2.798	3.037	0.605	2.976
		Sampling	**3.193**	**0.111**	**3.147**	2.977	2.835	3.037	−0.508	2.976
Overparameterization	Male	Joint	**1.797**	**0.192**	**1.787**	1.836	1.833	1.890	−0.413	1.843
	Male	Sampling	**1.615**	**0.187**	**1.605**	1.828	1.827	1.895	0.360	1.838
	Female	Joint	**1.707**	**-0.189**	**1.697**	1.759	1.757	1.811	0.400	1.767
		Sampling	**1.665**	**0.078**	**1.663**	1.756	1.753	1.806	−0.318	1.763
Design bias	A	Joint	**1.315**	**-0.321**	**1.276**	1.301	1.301	1.832	0.291	1.809
		Sampling	**1.384**	**-0.291**	**1.360**	1.298	1.298	1.835	0.266	1.805
	B	Joint	**1.304**	**0.302**	**1.269**	1.299	1.299	1.825	−0.291	1.802
		Sampling	**1.496**	**0.684**	**1.249**	1.295	1.295	1.826	−0.267	1.797
Model bias (small)	Male	Joint	**8.596**	**7.913**	**1.498**	1.421	1.422	5.147	−4.858	1.700
		Sampling	**8.471**	**8.358**	**1.384**	1.423	1.424	5.040	−4.739	1.702
	Female	Joint	**8.461**	**-8.342**	**1.415**	1.600	1.600	5.942	5.620	1.930
		Sampling	**8.325**	**-8.216**	**1.343**	1.591	1.588	5.763	5.427	1.924
Model bias (large)	Male	Joint	**15.910**	**-15.874**	**1.066**	1.300	1.230	2.015	−0.919	1.793
		Sampling	**15.931**	**-15.897**	**1.034**	1.304	1.303	2.026	−0.901	1.797
	Female	Joint	**23.581**	**23.528**	**1.581**	1.317	1.317	2.041	0.925	1.820
		Sampling	**23.035**	**22.989**	**1.457**	1.320	1.320	2.033	0.862	1.823
Model bias (large) x small n	Male	Joint	**16.094**	**-15.925**	**2.329**	2.800	2.796	4.149	−1.067	4.009

Condition	Domain	Distribution	RRMSE	RB	CV	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
		Sampling	**15.187**	**-15.013**	**2.296**	2.768	2.763	4.089	−0.897	3.961
	Female	Joint	**23.951**	**23.699**	**3.464**	2.836	2.832	4.210	1.08	4.069
		Sampling	**24.256**	**24.027**	**3.322**	2.802	2.798	4.148	0.910	4.016

Table 21: CSRM (%) for Superpopulation 2

Condition	Domain	Distribution	RMSE	RB	CV	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
Baseline	Male	Joint	**1.045**	**0.105**	**1.039**	1.047	1.071	1.433	0.465	1.355
		Sampling	**1.090**	**-0.1701**	**1.070**	1.050	1.067	1.422	0.361	1.354
	Female	Joint	**1.010**	**-0.095**	**1.004**	1.010	1.034	1.382	−0.448	1.307
		Sampling	**1.049**	**-0.044**	**1.028**	1.009	1.030	1.376	0.357	1.301
Small sample	Male	Joint	**2.364**	**-0.244**	**2.351**	2.351	2.337	3.196	−1.053	3.016
		Sampling	**2.379**	**0.090**	**2.377**	2.353	2.325	3.130	−0.880	3.000
	Female	Joint	**2.269**	**-0.235**	**2.257**	2.271	2.257	3.078	−1.012	2.906
		Sampling	**2.232**	**-0.659**	**2.133**	2.269	2.257	3.064	−0.805	2.902
Small population	Male	Joint	**1.665**	**0.177**	**1.657**	1.661	1.697	3.196	0.732	3.017
		Sampling	**1.597**	**-0.012**	**1.597**	1.670	1.703	2.242	0.540	2.137
	Female	Joint	**1.604**	**-0.163**	**1.598**	1.595	1.631	3.078	−0.710	2.907
		Sampling	**1.618**	**-0.009**	**1.616**	1.599	1.639	2.156	0.536	2.052
Non-ignorable	Male	Joint	**1.811**	**-0.179**	**1.803**	1.819	1.461	2.528	−0.674	2.436
		Sampling	**2.016**	**-0.298**	**2.012**	1.817	1.460	2.487	0.459	2.423

Condition	Domain	Distribution	RMSE	RB	CV	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
	Female	Joint	**1.775**	**-0.138**	**1.766**	1.753	1.401	2.437	0.655	2.347
		Sampling	**1.753**	**0.187**	**1.746**	1.749	1.398	2.414	−0.477	2.336
Overparameterization	Male	Joint	**1.357**	**-0.101**	**1.349**	1.383	1.387	1.429	−0.460	1.352
		Sampling	**1.347**	**-0.006**	**1.347**	1.391	1.397	1.427	0.340	1.361
	Female	Joint	**1.325**	**-0.095**	**1.319**	1.340	1.333	1.379	0.448	1.304
		Sampling	**1.287**	**0.097**	**1.286**	1.339	1.329	1.367	−0.341	1.302
Design bias	A	Joint	**0.767**	**0.173**	**0.747**	0.765	0.764	1.375	0.330	1.335
		Sampling	**0.717**	**-0.025**	**0.716**	0.766	0.765	1.371	0.246	1.336
	B	Joint	**0.762**	**-0.183**	**0.740**	0.760	0.760	1.365	0.329	1.325
		Sampling	**0.725**	**0.168**	**0.715**	0.761	0.761	1.359	−0.239	1.326
Model bias (small)	Male	Joint	**1.348**	**1.102**	**0.776**	0.772	0.772	1.512	0.688	1.346
		Sampling	**1.334**	**1.070**	**0.780**	0.772	0.771	1.512	0.648	1.342
	Female	Joint	**1.306**	**-1.077**	**0.740**	0.757	0.757	1.486	−0.686	1.318
		Sampling	**1.443**	**-1.230**	**0.754**	0.756	0.755	1.465	−0.594	1.313
Model bias (large)	Male	Joint	**5.649**	**-5.615**	**0.619**	0.678	0.677	3.972	−3.794	1.177
		Sampling	**5.628**	**-5.595**	**0.610**	0.679	0.679	4.000	−3.821	1.181
	Female	Joint	**8.369**	**8.318**	**0.924**	0.883	0.882	5.132	4.899	1.530
		Sampling	**8.556**	**8.505**	**0.932**	0.884	0.884	5.155	4.920	1.534
Model bias (large) x small n	Male	Joint	**5.829**	**-5.666**	**1.367**	1.461	1.453	4.603	−3.782	2.624
		Sampling	**5.756**	**-5.595**	**1.352**	1.456	1.442	4.587	−3.758	2.604
	Female	Joint	**8.668**	**8.424**	**2.043**	1.903	1.892	5.953	4.880	3.410

Condition	Domain	Distribution	RMSE	RB	CV	Vdesign	GMSE	PEM (MSE)	PEM (bias)	PEM (variance)
		Sampling	**8.299**	**8.054**	**2.001**	1.895	1.878	5.832	4.720	3.389

Appendix G

CSRMs for benchmark and accuracy estimators in the multinomial condition

Table 22: CSRM (%) in multinomial condition for superpopulation 1

Estimator	Domain	Low	Middle	High
RRMSE	Male	3.865	1.865	2.523
	Female	4.765	1.825	2.508
RB	Male	−0.0492	0.013	0.005
	Female	−0.166	0.103	−0.084
CV	Male	3.865	1.865	2.523
	Female	4.765	1.825	2.508
Vdesign	Male	3.968	1.863	2.584
	Female	4.769	1.879	2.514
GMSE	Male	4.069	1.909	2.647
	Female	4.881	1.925	2.574
PEM (MSE)	Male	4.240	1.873	2.697
	Female	4.957	1.892	2.625
PEM (bias)	Male	−0.998	−0.142	0.527
	Female	0.923	−0.168	−0.515
PEM (variance)	Male	4.121	1.868	2.645
	Female	4.871	1.885	2.574

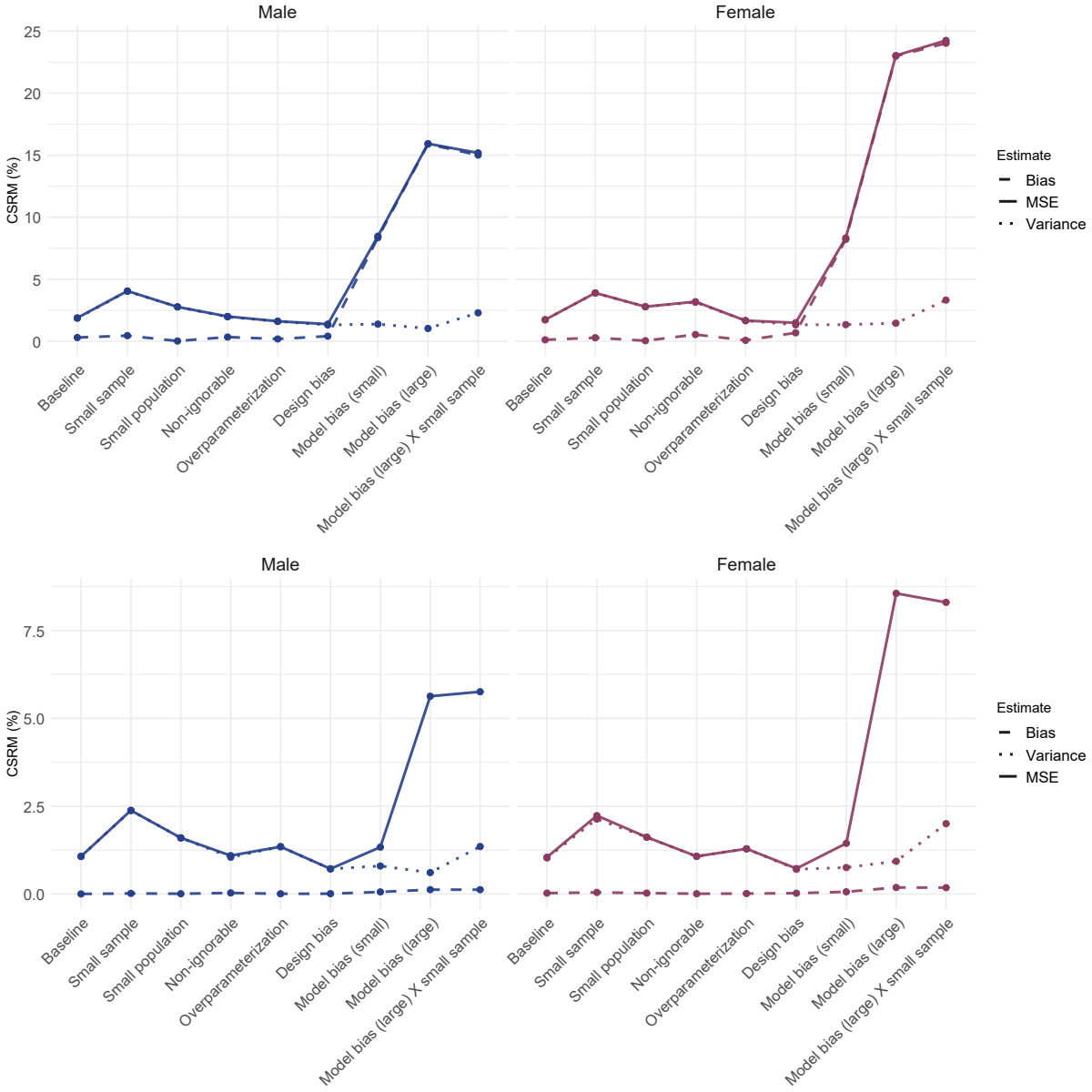
Table 23: CSRM (%) in multinomial condition for superpopulation 2

Estimator	Domain	Low	Middle	High
RRMSE	Male	1.979	1.512	4.164
	Female	5.547	1.861	1.598
RB	Male	0.104	−0.023	−0.099
	Female	0.064	−0.065	0.052
CV	Male	1.977	1.511	4.162
	Female	5.546	1.861	1.598
Vdesign	Male	2.056	1.499	4.216
	Female	5.488	1.828	1.584
GMSE	Male	2.104	1.533	4.310
	Female	5.627	1.872	1.622
PEM (MSE)	Male	2.767	1.694	4.719
	Female	7.044	2.051	1.832
PEM (bias)	Male	−0.941	0.481	1.316
	Female	−2.533	0.616	−0.562
PEM (variance)	Male	2.602	1.624	4.532
	Female	6.573	1.957	1.743

Appendix H

Benchmark estimators across the sampling distribution

Figure 8: Change in CSRMs across the experimental conditions in superpopulation 1 (top) and 2 (bottom)

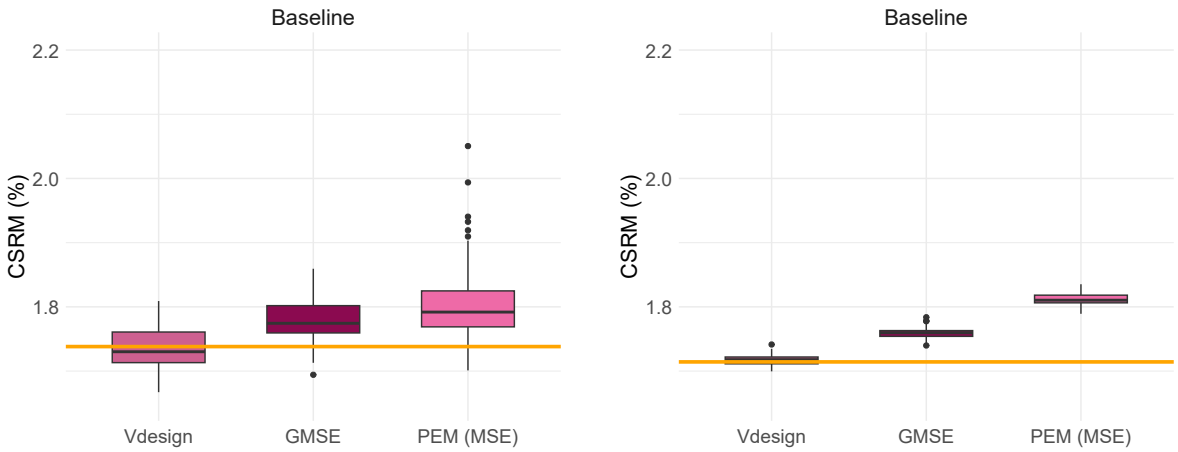


Note: This figure indicates the change in benchmark estimators for the sampling distribution across the experimental conditions. The benchmark estimators can be distinguished by linetype. The conditions are ordered starting with the baseline, followed by the variance, bias, and interaction conditions.

Appendix I

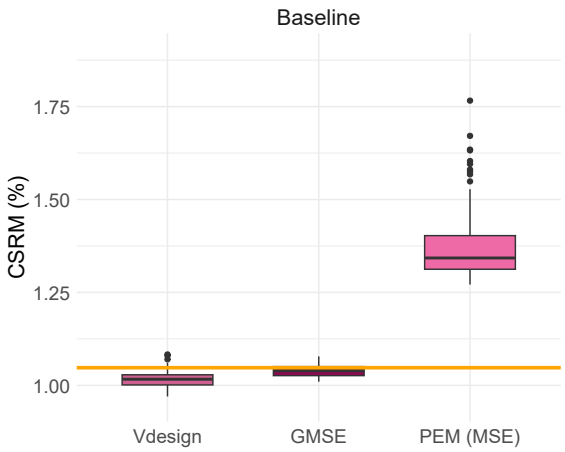
Simulation results for the female subdomain

Figure 9: Performance in the baseline condition for female subdomain across superpopulations 1 and 2

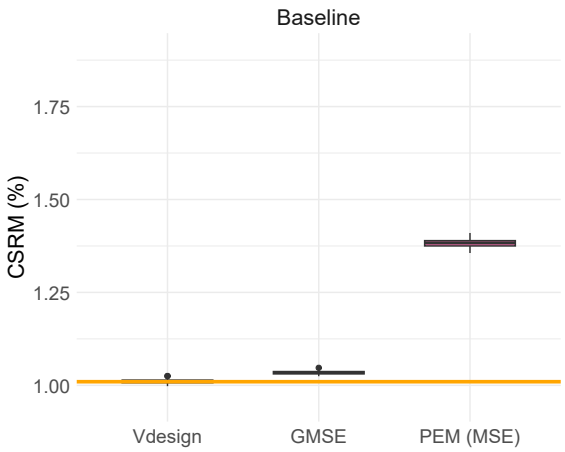


(a) Sampling distribution (sup 1)

(b) Joint distribution (sup 1)



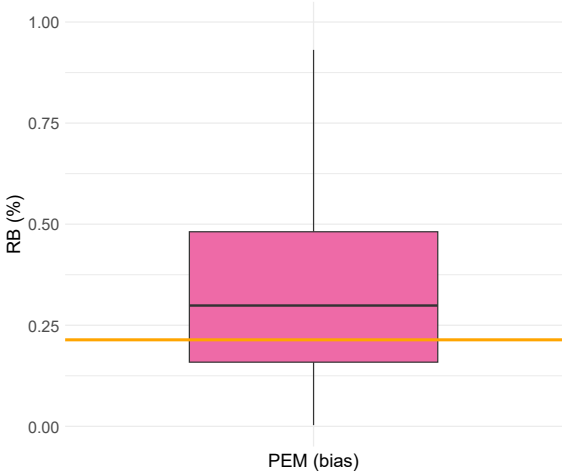
(c) Sampling distribution (sup 2)



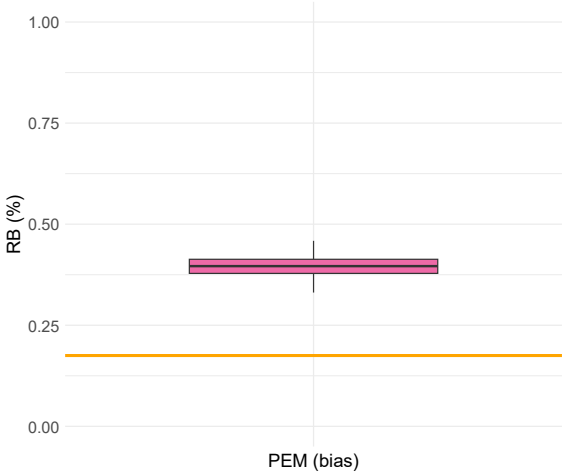
(d) Joint distribution (sup 2)

Note. The panels show the distribution of the estimators in the baseline condition across the two superpopulations in the female subdomain. The solid orange lines refer to $\text{RRMSE}_{\text{design}}$ in the sampling distribution and $\text{RRMSE}_{\text{joint}}$ in the joint distribution.

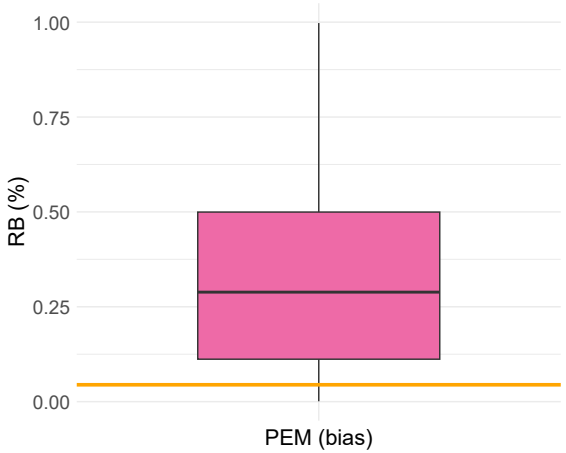
Figure 10: Performance of $PEM(bias)$ in the baseline condition for female subdomain across superpopulations 1 and 2.



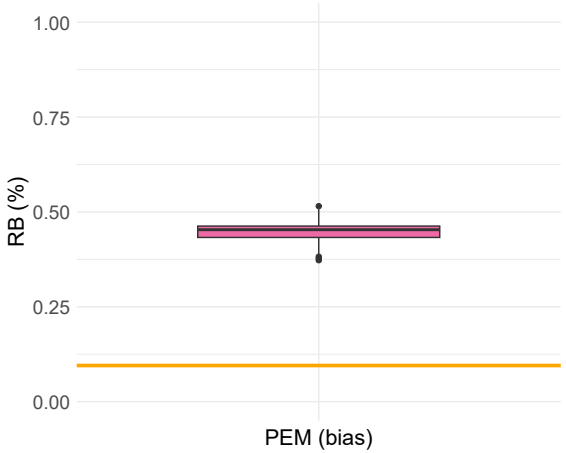
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



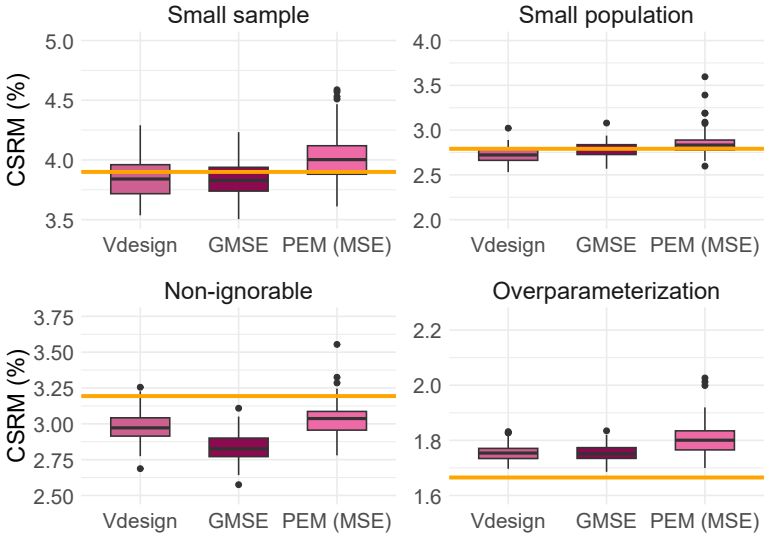
(c) Sampling distribution (sup 2)



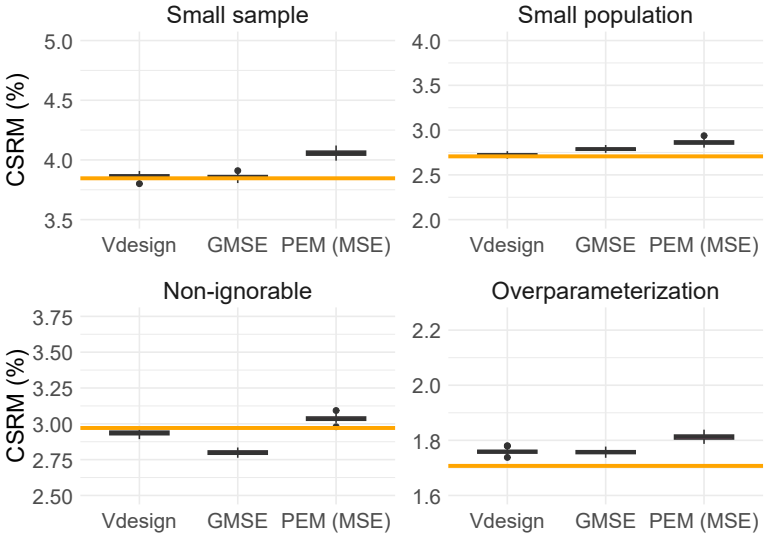
(d) Joint distribution (sup 2)

Note. The panels show the distribution of PEM (bias) in the baseline condition in the female subdomain. The solid lines refer to $\text{RB}_{\text{design}}$ and RB_{joint}

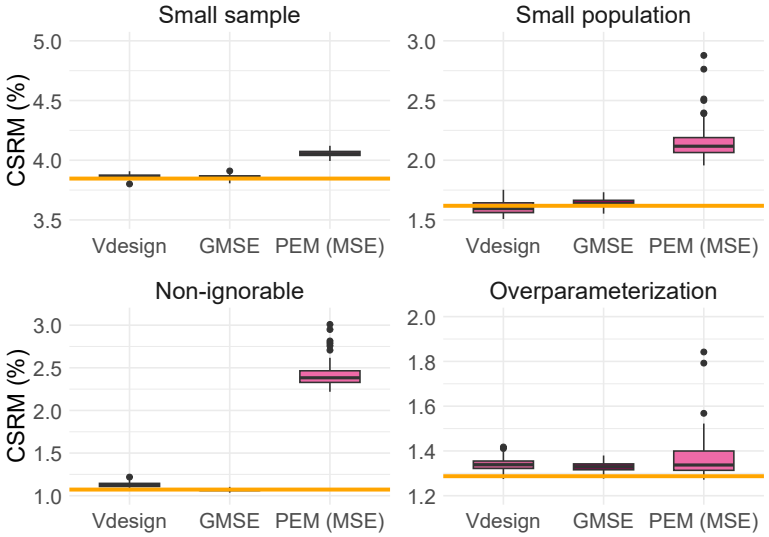
Figure 11: Performance of the accuracy estimators across the variance conditions in the female subdomain in superpopulations 1 and 2 with binomial target variable.



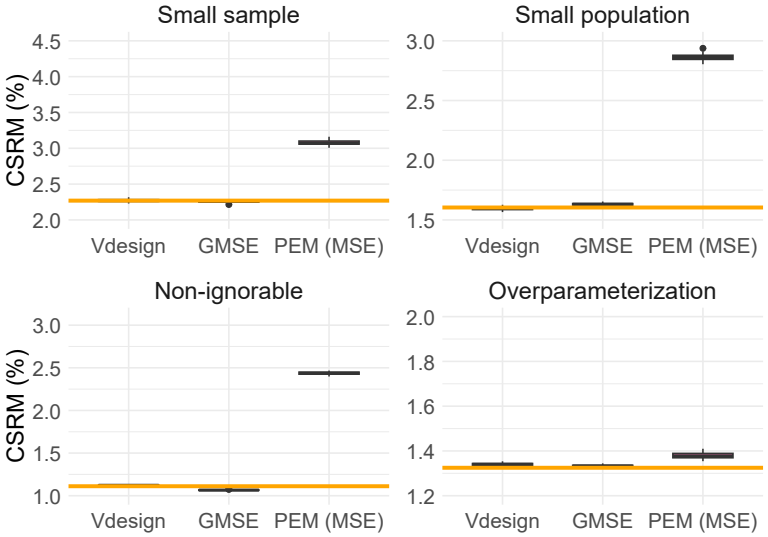
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



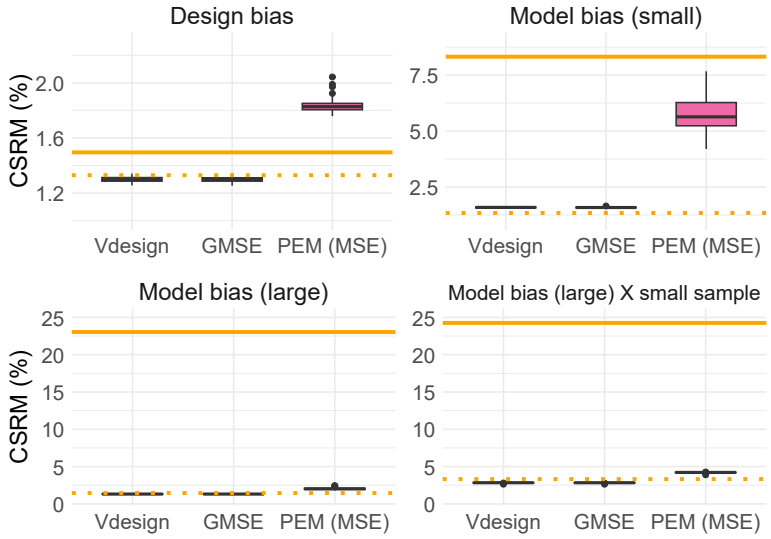
(c) Sampling distribution (sup 2)



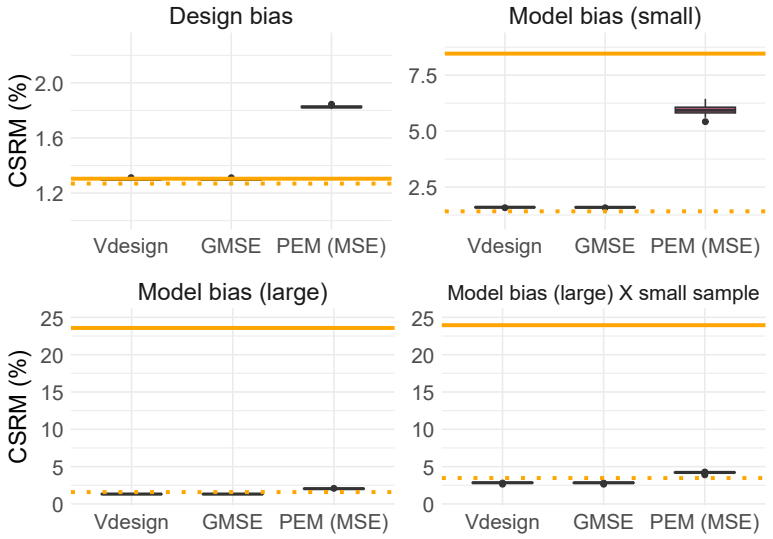
(d) Joint distribution (sup 2)

Note: The figures above show the performance of the estimators relative to benchmark $\text{RRMSE}_{\text{design}}$ and $\text{RRMSE}_{\text{joint}}$ shown in orange line.

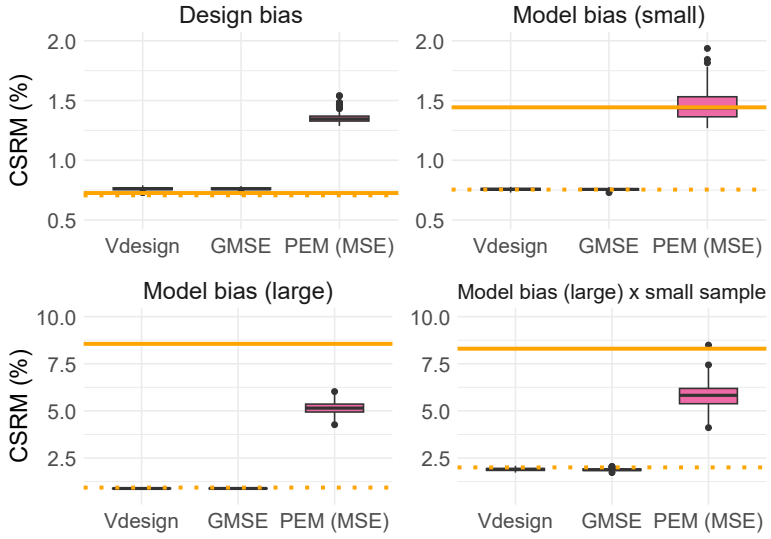
Figure 12: Performance of the estimators across bias and interaction conditions in the female subdomain in superpopulations 1 and 2.



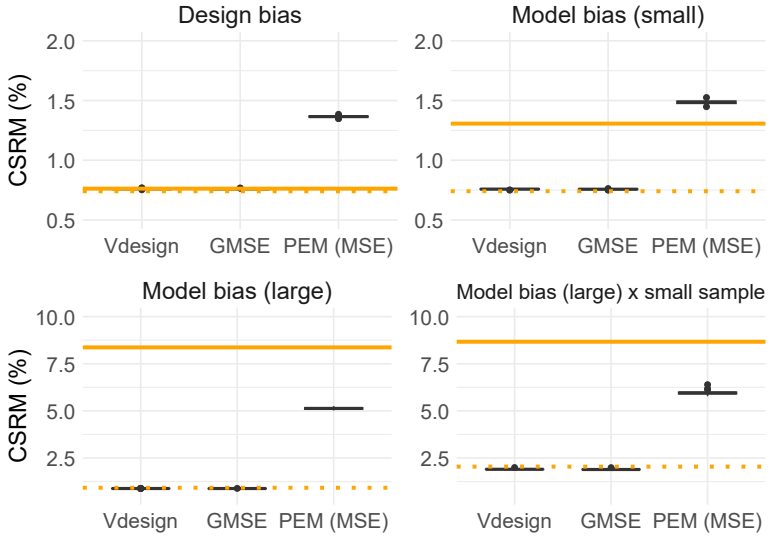
(a) Sampling distribution (sup 1)



(b) Joint distribution (sup 1)



(c) Sampling distribution (female)



(d) Joint distribution (female)

Note: The figures above show the performance of the estimators relative to $\text{RRMSE}_{\text{design}}$ and $\text{RRMSE}_{\text{joint}}$ (shown in solid orange line) and $\text{CV}_{\text{design}}$ and CV_{joint} (shown in dotted orange line).