Universiteit Leiden
The Netherlands

Centraal Bureau voor de Statistiek

Correcting for measurement error in the Employment Register and the Labour Force Survey using latent variable models

Elisca Mastenbroek

Primary thesis advisors: Dr. S. Scholtus and Dr. R. Stoel
Statistics Netherlands

Secondary thesis advisor: Dr. S.J.W. Willems
Methodology & Statistics, Institute of Psychology, Leiden University

Defended on August 8, 2023

MASTER THESIS
STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

The Employment Register (ER) and the Labour Force Survey (LFS) measure the labour contract of Dutch citizens. However, both sources provide different results. One possible explanation is that both sources contain measurement error (ME). Previous research has used hidden Markov models (HMMs) to estimate and correct for ME in linked data from the ER and the LFS. The HMMs did, however, have some limitations. For example, the HMMs used a suboptimal approach to include covariates that were missing for observations for whom one particular contract type was observed by the ER. In this thesis, these covariates are referred to as missing covariates.

To overcome the limitations of the HMMs, this thesis compared the performance of three different latent variable methods (LVMs), namely latent class (LC) analysis, latent class tree (LCT) analysis and tree-multiple imputation of latent classes (tree-MILC) analysis, to correct for ME in the ER and the LFS. For this purpose, two simulation studies were conducted: one without and one with missing covariates. For the second simulation study, a new approach was developed in which missing covariates were included using direct effects and parameter restrictions. In the end, LC and tree-MILC analysis was performed on real data from the ER and the LFS for respondents in the age of 25 to 55 in the first quarters of 2016, 2017 and 2018 to compare the estimates to the original HMM estimates.

In the simulation studies, little differences were found between the methods. The results showed that all model-based estimators were often considerably biased in conditions with two indicators. Although the bias and the variance decreased when one or two missing covariates were added, the largest decreases in bias and variance were observed when a third indicator was added. Furthermore, the analyses of the real data showed that the LC estimates, the tree-MILC estimates, and the original HMM estimates were different from each other. Nevertheless, the differences were smaller than the original differences between the ER and the LFS.

Future research that aims to correct for ME in the ER and the LFS could use the approach proposed in this study to include missing covariates. In addition, to enhance the accuracy of the estimates, the current findings suggest that it may be beneficial for Statistics Netherlands to find a third indicator measuring the contract types of Dutch citizens. Finally, LVMs could potentially be used for the production of official statistics. However, to implement these methods in practice, further research is needed on both a methodological and an organisational level.

Contents

List of symbols

X	A latent variable,
N	The number of units in the population,
n	The number of units in a sample,
i	A particular unit in a sample or the population,
K	The number of latent classes,
J	The number of indicators,
k, s, l	A particular latent class,
Y_j	A particular indicator,
y_j	A particular response on indicator Y_j,
$\mathbf{Y}$	A complete response vector on the indicators $Y_1, ..., Y_J$,
$\mathbf{y}$, $\mathbf{y}_i$	A complete response vector on the indicators $Y_1, ..., Y_J$ for one observation,
C	The number of covariates,
Z	A particular covariate,
$\mathbf{Z}$	A complete response pattern on the covariates $Z_1, ..., Z_C$,
$\mathbf{z}$, $\mathbf{z}_i$	A complete response vector on the covariates $Z_1, ..., Z_C$ for one observation,
W	An assigned latent class,
$I(.)$	An indicator that equals 1 if the argument is true and 0 otherwise,
X_i	The latent variable X for observation i,
Y_{ij}	The indicator Y_j for observation i,
P_k	The proportion of elements in the population that belong to the true class k,
$P_{l\|k}^{Y_j}$	The proportion of elements in the population for whom class l is observed by indicator Y_j, but who belong to the true class k,
X_{parent}	A particular parent node,
X_{child}	A particular child node,
$w_{i\|X_{parent}}$	The weight for observation i at X_{parent},
M	The number of bootstrap samples,
m	A particular bootstrap sample,
$\hat{\theta}_m$	An estimate of interest based on bootstrap sample m,
$\overline{\theta}$	A pooled estimate,
V_T	The total variance of the pooled estimates,
S_i^m	An imputed latent class for observation i based on bootstrap sample m,
Y_1, Y_2, Y_3, Y_4	A trichotomous categorical indicator,
Q	A dichotomous categorical non-missing covariate,
q	A particular response on covariate Q,
N_{Sim}	The number of simulation replications,
r	A particular simulation replication,
Z_1, Z_2	A trichotomous categorical missing covariate.

Chapter 1

Introduction

The Dutch labour force consists of people with various types of employment. Some people may for instance be employed with a permanent or a flexible contract, whereas others may be self-employed or not employed at all. How these contract types are distributed in society provides important information about the Dutch labour market (Pavlopoulos & Vermunt, 2015; Rubery et al., 2016). Accurate statistics on employment contract types are therefore important for debates on both an economic and a political level (Smits et al., 2021).

Statistics Netherlands (CBS) has two data sources that measure the contract types of Dutch citizens. These are the Labour Force Survey (LFS) and the Employment Register (ER). The LFS is a quarterly panel survey that is conducted by CBS on the labour status of Dutch citizens. The ER, on the other hand, is a register of all Dutch insured employees that is maintained on a monthly basis by the Dutch Employee Insurance Agency (UWV). However, both sources provide different results. For example, the proportion of employees with a flexible contract was estimated by the ER and LFS as respectively 35.7% and 25.2% in the fourth quarter of 2018 (Bakker et al., 2021).

Previous research has proposed two explanations to explain the differences between the ER and the LFS. One is that both sources contain measurement error. Measurement error may for instance be caused by administrative delays or by respondents who provide incorrect responses. Another potential explanation is that both sources do not measure the same concept. In contrast to the ER, the LFS may for instance not necessarily measure people's legal contract type, but also take verbal agreements between employers and employees into account. Both explanations have been studied by applying latent variable models (LVMs) to linked data from both sources (Oberski, 2017). For this purpose, the true contract type was considered as a latent variable that consisted of the categories 'permanent', 'flexible' and 'other' (Bakker et al., 2021). Here, the latter category referred to a collection of various subcategories, including those who are self-employed, unemployed, or director-major shareholders (DGAs).

Using this approach, Pavlopoulos and Vermunt (2015), Pankowska et al. (2018) and Bakker et al. (2021) estimated and corrected for measurement error in the ER and LFS across a variety of time periods. In each of these studies, hidden Markov models (HMMs) were used to estimate the real contract types while also modelling the development over time. In addition, several covariates were included that could affect the probability of having a certain contract type, such as a respondent's number of contract hours, or job duration. The results showed that for some contract types, the proportion of incorrectly measured contracts was estimated to be as high as 37.7% (Bakker et al., 2021).

Furthermore, Restrepo Estrada (2023) used latent class (LC) analysis in his master thesis to study whether the ER and LFS measured the same concept from 2016 to 2018. For this purpose, it was studied whether direct effects were present of the earlier mentioned covariates on the observed contract types. The results were, however, inconclusive. As a result, no evidence was found against the hypothesis that the ER and LFS did not measure the same concept.

Although the HMMs were used to estimate and correct for the amount of measurement error in the ER and LFS, they did have a number of limitations. For example, the models were quite complex and had assumptions about the development of the real and observed contract types over time that were difficult to verify in practice. In addition, some covariates were missing for a significant portion of observations for whom the contract type 'other' was observed by the ER. Note that throughout this thesis, these covariates are referred to as missing covariates. To prevent multicollinearity, all of these observations were assigned to an existing covariate category that seemed most suitable. Nevertheless, inaccurate assignments may have affected the accuracy of the HMM estimates.

No research has yet been conducted to estimate and correct for measurement error in the ER and LFS using other LVMs. However, the limitations of the HMMs could potentially be overcome with other methods, such as LC (Lazarsfeld, 1950), latent class trees (LCT) (Van Den Bergh, 2018) and tree-multiple imputation of latent classes (tree-MILC) analysis (Boeschoten et al., 2017; Remmerswaal, 2022). An advantage of these LVMs is that they are less complex and thus more parsimonious as compared to the HMMs. In addition, they have assumptions that are easier to verify in practice. Furthermore, in contrast to LC analysis and the HMMs, LCT and tree-MILC analysis allow for more flexibility in the inclusion of covariates due to their hierarchical structures. This suggests that LCT and tree-MILC analysis could potentially overcome the limitations of LC analysis and the HMMs with regard to the covariates.

For these reasons, the current thesis will aim to achieve the following objectives:

1. Compare the performance of LC, LCT and tree-MILC analysis in a simulation study without missing covariates (Chapter 4),
2. Develop a new approach to include missing covariates in LC, LCT and tree-MILC analysis without assigning observations to existing categories (Chapter 5),
3. Compare the performance of LC, LCT and tree-MILC analysis in a simulation study with missing covariates (Chapter 5),
4. Apply the best performing method(s) to real data from the ER and the LFS for respondents in the age of 25 to 55 in the first quarters of 2016, 2017 and 2018 to compare the estimates to the original HMM estimates (see Chapter 6).

In this thesis, Chapter 2 will firstly provide some background information on the identification of error in single-source data. Secondly, Chapter 3 will provide a detailed description of LC, LCT and tree-MILC analysis. Thirdly, Chapters 4-6 will describe how the research aims were achieved. Lastly, Chapter 7 will discuss the results.

Chapter 2

Background

2.1 Identifying error in single-source data

National statistical institutes (NSIs) have traditionally produced statistics from a single data source (De Waal et al., 2019). Most often, surveys have been used as a single data source. However, in recent years, administrative data have also become increasingly utilised. Administrative data offer several advantages over survey data, such as larger sample sizes, lower costs, and a lower respondent burden (Groen, 2012). Nevertheless, both sources are not without error (Bakker, 2011a).

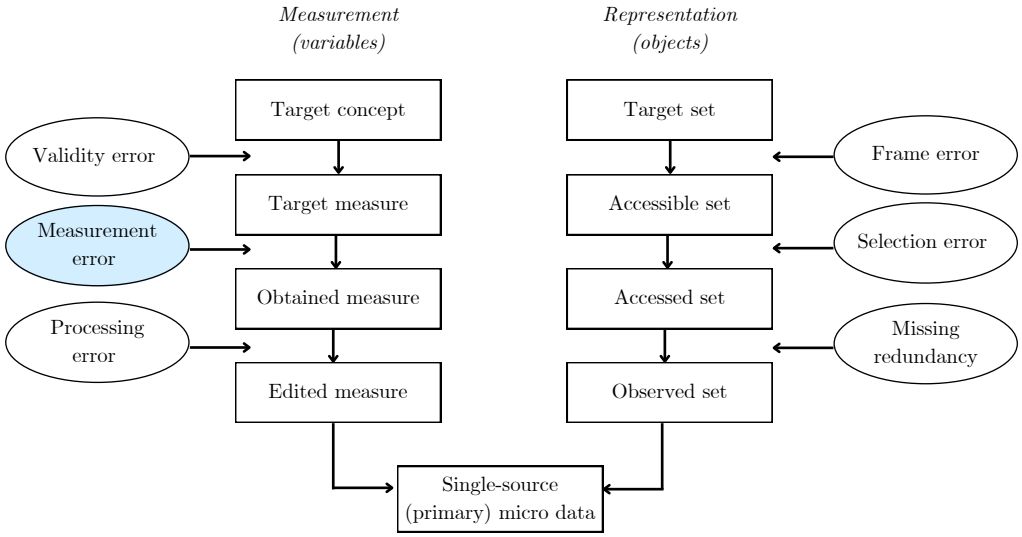
To identify potential sources for error in single-source data, the Total Survey Error (TSE) framework can be used (Groves et al., 2004) (see Figure 2.1). This framework is based on the concept of TSE, which refers to the accumulation of all errors that may occur during the design, collection, processing, and analysis of survey data (Biemer, 2010). The TSE framework is designed to identify and reduce these errors while taking constraints with regard to costs and timeliness into account. Although the TSE framework is originally developed for surveys, it is generally assumed to hold for register data as well (Bakker, 2011b; Zhang, 2012).

According to the TSE framework, error may be introduced through either measurement or representation (Groves et al., 2004). Measurement related errors occur when the attributes of a population are not measured correctly. For example, a measurement may measure a concept that is different from the concept it intends to measure (i.e. validity error), or it may differ from the real value as a result of errors that occur during the measurement process (i.e. measurement error) (see also Section 2.2). Alternatively, the variable used in estimation may differ from that provided by the respondents (i.e. processing error). Representation related errors occur when the measured attributes of a population differ from those of the target population. These differences may be caused by a gap between the target population and the sampling frame (i.e. frame error), or by a gap between the sampling frame and the sample (i.e. selection error). Lastly, the data set may contain data that is either missing or redundant (i.e. missingness or redundancies).

2.2 From single-source to multi-source data

As mentioned in Chapter 1, the current thesis will focus on the identification and reduction of measurement error (see Figure 2.1).

Figure 2.1
A schematic representation of the Total Survey Error (TSE) framework



Note. All types of error that can occur in single-source (primary) micro data according to the TSE framework (Zhang, 2012). This thesis focuses on the type of error indicated in blue.

Measurement error occurs when the measured response differs from the real value (Groves et al., 2004; Zhang, 2012). In surveys, this may be caused by factors related to the respondent, the interviewer, the questionnaire or the collection method. In administrative data, this may be caused by factors related to the processing, linking and editing of the data (Oberski, 2017). Measurement error affects the reliability and validity of statistical analyses, and should be identified and reduced as much as possible (Biemer, 2010; Wikman & Wärneryd, 1990).

However, measurement error is often well-hidden in the data. One possible way to detect measurement error is by looking for impossible combinations of response values. Another way is to repeat the measurement process and examine inconsistencies between both sets of responses (Boeschoten et al., 2017). However, even then, systematic errors may still remain hidden (Biemer, 2009). A better way to detect measurement error is therefore to compare a data source to a gold standard that is assumed to be error-free. This is called validation (Wikman & Wärneryd, 1990). Interestingly enough, administrative registers are often used to validate surveys, while surveys are used to validate administrative registers (Oberski, 2017). Nevertheless, both sources are not error-free (see Section 2.1). Different ways are therefore needed to estimate the amount of measurement error in both sources.

One approach for which a gold standard is not required is the application of latent variable models (LVMs) to linked data from multiple sources (Oberski, 2017). In LVMs, the true value of some measured characteristic is considered as an unobserved or latent variable that is measured by several indicators that represent the available data sources. By analysing the degree of (in)consistency between the indicators, the amount of measurement error conditional on the latent variable can be estimated and corrected estimates can be obtained[1].

[1]That is, on assumptions such as mentioned in Section 3.1.1.

Chapter 3

Three latent variable models (LVMs)

In this chapter, latent class (LC) analysis (see Section 3.1), latent class tree (LCT) analysis (see Section 3.2) and tree-multiple imputation of latent classes (tree-MILC) analysis (see Section 3.3) are described.

3.1 Latent class (LC) analysis

LC analysis (Lazarsfeld, 1950) is a method that is widely used for identifying subgroups based on responses to a set of indicator variables (Magidson et al., 2020). The method is based on the idea that the parameters of some statistical model differ across a number of unobserved subgroups or *latent classes* (Vermunt & Magidson, 2004). Together, these latent classes form the categories of a categorical latent variable. LC analysis is often used for various applications, such as clustering, scaling and random-effects modelling.

In the current thesis, LC analysis is used to estimate and correct for measurement error (ME) in the ER and LFS. The latent variable, in this case, represents an observation's true contract type and consists of the latent classes 'permanent', 'flexible' and 'other'. The indicator variables, in turn, represent the observed contract types as measured by the ER and LFS.

3.1.1 Model assumptions

Let X denote a categorical latent variable that is assumed to be the true value of some unobserved population characteristic that is measured by a set of indicators. Any LC model for X has the following assumptions (Biemer, 2011):

1. The sample consists of n units sampled without replacement from a large population of N units using simple random sampling.
2. The probability that a unit responds is the same for any two units in the population.
3. The indicators are all indicators of the same latent variable X.
4. The indicators are mutually independent within each latent class. Hence, X explains all associations between the indicators.
5. The amount of measurement error is equal for each respondent within each latent class.

3.1.2 The model

Let K denote the number of latent classes, k a particular latent class, $\mathbf{Y}$ a complete response vector for the indicators $Y_1, ..., Y_J$, and $\mathbf{y}$ a complete response vector of the same length for one observation. The basic LC model can be represented as

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{k=1}^{K} P(X = k) P(\mathbf{Y} = \mathbf{y} | X = k), \tag{3.1}$$

where $P(X = k)$ is the probability of belonging to class k, and $P(\mathbf{Y} = \mathbf{y}|X = k)$ the probability of having response $\mathbf{y}$ conditional on belonging to class k. Since the indicators are assumed to be mutually independent within each latent class, the latter probability can be simplified to

$$P(\mathbf{Y} = \mathbf{y}|X = k) = \prod_{j=1}^{J} P(Y_j = y_j|X = k). \tag{3.2}$$

Combining Equations (3.1) and (3.2) yields the following model:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{k=1}^{K} P(X = k) \prod_{j=1}^{J} P(Y_j = y_j|X = k). \tag{3.3}$$

3.1.3 Estimation

The parameters in Equation (3.3) are unknown and need to be estimated. Estimation is typically done using maximum likelihood (ML) (Vermunt & Magidson, 2004). For this purpose, the following log-likelihood function is maximised:

$$\log L(\theta, \mathbf{y}) = \sum_{i=1}^{n} \log P(\mathbf{Y} = \mathbf{y}_i). \tag{3.4}$$

To find the ML estimates, a combination of the Expectation-Maximization (EM) algorithm and the Newton-Raphson (NR) algorithm is often used. Both algorithms require starting values for the parameters to be estimated. These starting values are adjusted iteratively until the fit of the model does not improve anymore beyond a threshold value.

Some problems may, however, occur during estimation. One problem is that local maxima in the log-likelihood function may cause the EM and NR algorithms to converge to different maxima. To prevent this, different sets of random starting values are often used. In this case, the solution is chosen that yields the highest value of the log-likelihood function.

A second problem is that a model may be non-identified. A model is non-identified when multiple sets of parameter values yield the same log-likelihood maximum. For non-identified models, estimating the model parameters is not possible. Model identification can be verified by checking if the information matrix of the model is invertible. In general, at least three indicators are needed to identify an LC model, although two indicators may also be combined with a covariate (Boeschoten et al., 2017) (see Section 3.1.4).

A final problem is that some model parameter estimates may lie on the boundary of the parameter space (Vermunt & Magidson, 2004). This occurs when some of the probabilities in Equation (3.3) are equal to zero or one. Boundary solutions may lead to various issues, including

numerical problems in estimation, and can be prevented by imposing constraints on the model parameters, or by taking other a priori information about possible parameter values into account (Galindo Garre & Vermunt, 2006).

3.1.4 Covariates

In the basic LC model described so far, only indicator variables have been included. The latent variable X may, however, also be related to one or more explanatory variables that measure something other than the attribute of interest (Vermunt, 2017). These variables are typically referred to as covariates. Including one or more covariates can help improve an LC model in terms of goodness of fit or identification (Boeschoten et al., 2017).

Let $\mathbf{Z}$ denote a complete response vector for the categorical covariates $Z_1, \ldots, Z_C$, and $\mathbf{z}$ the complete response vector for one observation. If $\mathbf{Y}$ is assumed to be independent of $\mathbf{Z}$ conditional on X, $\mathbf{Z}$ can be added to Equation (3.3) as follows

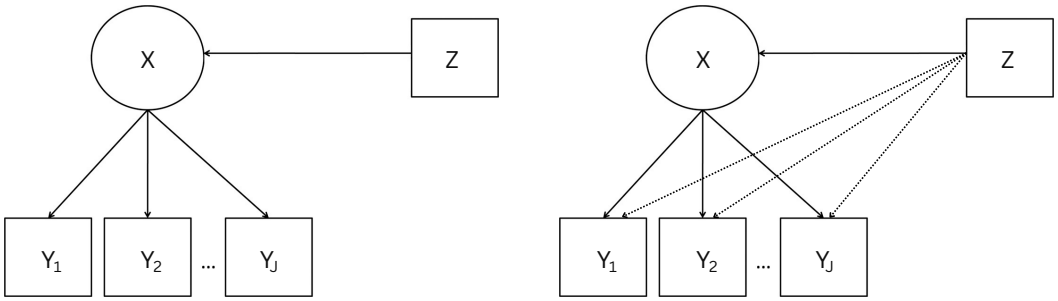
$$P(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z}) = \sum_{k=1}^{K} P(X = k|\mathbf{Z} = \mathbf{z}) \prod_{j=1}^{J} P(Y_j = y_j|X = k). \tag{3.5}$$

If $\mathbf{Y}$ is assumed to be dependent on both X and $\mathbf{Z}$, Equation (3.5) can be expanded by including direct effects (DEs) of $\mathbf{Z}$ on $\mathbf{Y}$ (Masyn, 2017):

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z}) = \sum_{k=1}^{K} P(X = k|\mathbf{Z} = \mathbf{z}) \prod_{j=1}^{J} P(Y_j = y_j|X = k, \mathbf{Z} = \mathbf{z}). \tag{3.6}$$

In Figure 3.1, a schematic overview is displayed of both situations. Note that in both situations, the assumption is made that the covariates $Z_1, \ldots, Z_C$ are measured without error.

Figure 3.1
Schematic overview of two LC models that include the latent variable X, the indicators $Y_1, \ldots, Y_J$, and the covariate Z



A. Without DEs **B.** With DEs

Note. Figure A shows an LC model where Z is independent of $\mathbf{Y}$ conditional on X. Figure B shows an LC model where direct effects (DEs) are included of Z on $\mathbf{Y}$.

3.1.5 Posterior class membership probabilities

After estimating an LC model, the posterior class membership probability for a particular observation to belong to latent class k can be obtained using Bayes' rule. If $\mathbf{Y}$ is assumed to be independent of $\mathbf{Z}$ conditional on X, this probability follows as

$$P(X = k|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = \frac{P(X = k|\mathbf{Z} = \mathbf{z}) \prod_{j=1}^{J} P(Y_j = y_j|X = k)}{P(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z})}. \tag{3.7}$$

If $\mathbf{Y}$ is assumed to be dependent on both X and $\mathbf{Z}$, the term $P(Y_j = y_j|X = k)$ can simply be replaced by the term $P(Y_j = y_j|X = k, \mathbf{Z} = \mathbf{z})$.

3.1.6 Latent class assignment

The posterior class membership probabilities can be used to assign observations to latent classes. This is often done using modal or proportional assignment (Dias & Vermunt, 2008).

In modal assignment, each observation is assigned to the latent class with the largest posterior class membership probability. Let W denote an assigned class for some observation with response pattern $\mathbf{y}$, and let s and k denote particular latent classes. The probability of being assigned to latent class s follows as

$$P(W = s|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = \begin{cases} 1, & \text{if } P(X = s|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) > P(X = k|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) \quad \forall s \neq k \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

In proportional assignment, each observation is assigned to a latent class with a weight that is equal to their posterior class membership probability for that class:

$$P(W = s|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = P(X = s|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}). \tag{3.9}$$

3.1.7 The population proportion estimator (PPE)

In addition, the posterior class membership probabilities can be used to estimate the proportion of elements in the population that belong to the true class k. Let $I(.)$ denote an indicator that takes on the value of 1 if the argument is true and 0 otherwise. The proportion of elements in the population that belong to the true class k is given by

$$P_k = \frac{1}{N} \sum_{i=1}^{N} I(X_i = k). \tag{3.10}$$

From this, the model-based population proportion estimator (PPE) $\hat{P}_{k_{LC}}$ follows as

$$\hat{P}_{k_{LC}} = \frac{1}{n} \sum_{i=1}^{n} P(X = k|\mathbf{Y} = \mathbf{y}_i, \mathbf{Z} = \mathbf{z}_i). \tag{3.11}$$

3.1.8 The ME probability estimator (MEPE)

Finally, the posterior class membership probabilities can be used to estimate the proportion of elements in the population that belong to the true class k, but for which class l is observed by indicator Y_j. This proportion is given by

$$P_{l|k}^{Y_j} = \frac{\sum_{i=1}^{N} I(X_i = k, Y_{ij} = l)}{\sum_{i=1}^{N} I(X_i = k)} = \frac{\sum_{i=1}^{N} I(X_i = k) I(Y_{ij} = l)}{\sum_{i=1}^{N} I(X_i = k)}. \tag{3.12}$$

Note that this proportion also reflects the probability for indicator Y_j to measure the true class k as the observed class l for any element in the population. If this probability is non-zero for any $l \neq k$, ME may occur. $P_{l|k}^{Y_j}$ can therefore also be considered as a ME probability. Consequently, the model-based ME probability estimator (MEPE) follows as

$$\hat{P}_{l|k_{LC}}^{Y_j} = \frac{\sum_{i=1}^{n} P(X = k|\mathbf{Y} = \mathbf{y}_i, \mathbf{Z} = \mathbf{z}_i) I(Y_{ij} = l)}{\sum_{i=1}^{n} P(X = k|\mathbf{Y} = \mathbf{y}_i, \mathbf{Z} = \mathbf{z}_i)}. \tag{3.13}$$

Both $P_{l|k}^{Y_j}$ and $\hat{P}_{l|k_{LC}}^{Y_j}$ with $l, k \in \{1, \ldots, K\}$ can be represented in a matrix (see Table 3.1). In this matrix, the diagonal elements represent the probability that a true class is measured correctly, whereas the non-diagonal elements represent the probability that a true class is (erroneously) measured as a different class. Consequently, an increase in one diagonal element leads to a decrease in one or more non-diagonal elements in the same row (and vice versa).

Table 3.1
Measurement error (ME) probability matrix for an indicator Y_j with ME probabilities $P_{l|k}^{Y_j}$ of .05 for any $l \neq k$ with $l, k \in \{1, 2, 3\}$

	Permanent	**Flexible**	**Other**
Permanent	.90	.05	.05
Flexible	.05	.90	.05
Other	.05	.05	.90

Note. The rows represent the true values as estimated by the model. The columns represent the observed values as measured by indicator Y_j.

3.2 Latent class tree (LCT) analysis

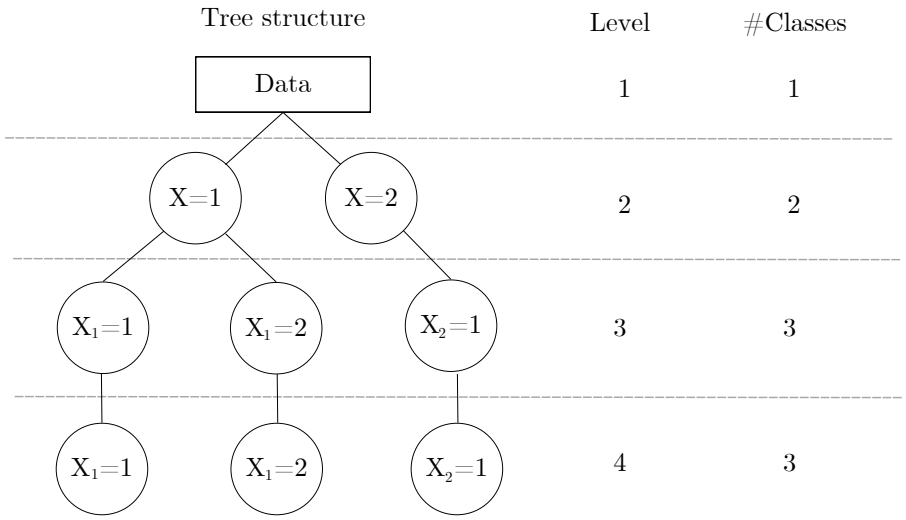
An extension of LC analysis, which is referred to as latent class trees (LCT), was developed by Van Den Bergh (2018). In this extended method, a hierarchical tree structure is imposed on the latent classes. As a result, a clear insight can be obtained into how the latent classes are formed and how models with different number of classes are related to each other.

In this section, an adaptation of the original algorithm is developed. In this adaptation, a two-level tree structure is used to distinguish between the latent classes 'permanent', 'flexible' and 'other'. In the following subsections, both the original algorithm (see Section 3.2.1) and the adapted algorithm (see Section 3.2.3) are described. In addition, the reasons for adapting the original algorithm are described (see Section 3.2.2). The remaining subsections apply to models that result from both algorithms.

3.2.1 The original algorithm

The original LCT algorithm starts by estimating both a 1- and a 2-class LC model on a parent node that encapsulates the entire data set. If a 2-class model is preferred according to some fit measure, such as the Bayesian information criterion (BIC), two new nodes are created and the data set is split in two subsamples using proportional assignment (see Section 3.1.6). For each subsample, 1- and 2-class models are then estimated and compared again. This time, however, each observation is weighted using the previously obtained posterior class membership probabilities. These steps are repeated until only 1-class models are preferred. Figure 3.2 shows an example of the structure of an LCT model with three latent classes.

Figure 3.2
Schematic overview of an LCT model with three latent classes



Note. Adapted from Van Den Bergh (2018). At the beginning, the data is split into two subsamples (i.e. $X = 1$ and $X = 2$) (level 2). For $X = 1$, a 2-class model is preferred, whereas for $X = 2$, a 1-class model is preferred. Note that the fit measures are not displayed in this image. $X = 1$ is therefore split again into two subsamples ($X_1 = 1$ and $X_1 = 2$), whereas $X = 2$ is not (level 3). After this split, only 1-class models are preferred for each node (level 4). Hence, level 3 and 4 show the final solution.

3.2.2 Motivation for adapting the original algorithm

The reasons for adapting the original algorithm are twofold. The first is that the original algorithm does not allow the user to specify the number of child nodes per parent node; it simply chooses the number of child nodes based on some fit measure (see Section 3.2.1). Consequently, if both classes at the second level of the tree prefer 2-class models, the model automatically chooses four latent classes instead of three (see Figure 3.2). Nevertheless, when correcting for measurement error in the ER and the LFS, the latent variable should always be trichotomous to represent the three different contract types.[1]

[1] An initial analysis on a simulated data set using the built-in implementation of LCT in Latent GOLD® 6.0 showed that in some cases, four latent classes were indeed distinguished.

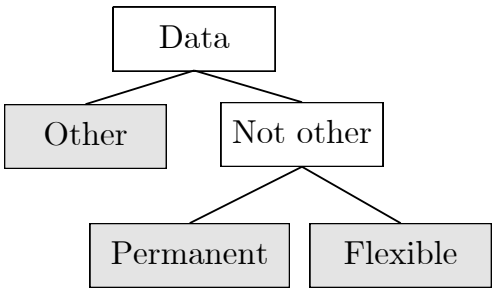
The second reason to adapt the original algorithm is to exert influence on which latent class is distinguished first. In the linked data, some covariates are missing for many observations for whom the contract type 'other' was observed by the ER (see Chapter 1). By distinguishing the class 'other' first, it is therefore possible to only include covariates after this class is already distinguished. In the adapted algorithm, both of these features are incorporated.

3.2.3 The adapted algorithm

The adapted algorithm starts by estimating a 3-class LC model on the entire data set to distinguish between the latent classes 'permanent', 'flexible' and 'other'. However, to distinguish the latent class 'other' first, the latent classes 'permanent' and 'flexible' are merged again by combining the posterior probabilities for both classes. The resulting latent class 'not other' therefore contains all observations that do not belong to the class 'other' (see Figure 3.3).

In the next step, a 2-class LC model is estimated on the entire data set to split the latent class 'not other' into the classes 'permanent' and 'flexible'. However, similar as in the original algorithm, the posterior probabilities of belonging to the latent class 'not other' are now included as weights. Note that as a result, observations with a weight of zero are excluded from the model. Furthermore, note that at this point, it is possible to include covariates that are only available for the latent classes 'permanent' and 'flexible'. In the end, the final posterior probabilities are computed using Equation (3.19).

Figure 3.3
Schematic overview of how the latent classes 'permanent', 'flexible', and 'other' are distinguished in LCT and tree-MILC analysis in the current thesis



3.2.4 The model

Let $\mathbf{Y}$ denote the complete vector of responses, X_{parent} a particular parent node, k a particular child node, J the number of indicators, and Y_j a particular indicator. The 2-class LC model at a particular parent node follows as

$$P(\mathbf{Y} = \mathbf{y}|X_{parent}) = \sum_{k=1}^{2} P(X_{child} = k|X_{parent}) \prod_{j=1}^{J} P(Y_j = y_j|X_{child} = k, X_{parent}). \quad (3.14)$$

Adding covariates without direct effects yields

$$P(\mathbf{Y}=\mathbf{y}|X_{parent},\mathbf{Z}=\mathbf{z})=\sum_{k=1}^{2}P(X_{child}=k|X_{parent},\mathbf{Z}=\mathbf{z})\prod_{j=1}^{J}P(Y_j=y_j|X_{child}=k,X_{parent}). \tag{3.15}$$

Note that the term $P(Y_j = y_j|X_{child} = k, X_{parent})$ can simply be replaced by the term $P(Y_j = y_j|X_{child} = k, X_{parent}, \mathbf{Z} = \mathbf{z})$ when adding covariates with direct effects. This applies to all future equations.

3.2.5 Estimation

To estimate the parameters of an LCT model at the node X_{parent}, the following log-likelihood function is maximised:

$$\log L(\theta,\mathbf{y},X_{parent},\mathbf{z})=\sum_{i=1}^{n}w_{i|X_{parent}}P(\mathbf{Y}=\mathbf{y}_i|X_{parent},\mathbf{Z}=\mathbf{z}_i). \tag{3.16}$$

Here, $w_{i|X_{parent}}$ denotes the weight for observation i at the parent class. This weight equals the posterior probability of belonging to the parent class for that particular observation.

After performing a split, the weights for the two newly formed child classes are obtained by

$$w_{i|X_{child}=1}=w_{i|X_{parent}}P(X_{child}=1|\mathbf{Y}=\mathbf{y},X_{parent},\mathbf{Z}=\mathbf{z}) \tag{3.17}$$

and

$$w_{i|X_{child}=2}=w_{i|X_{parent}}P(X_{child}=2|\mathbf{Y}=\mathbf{y},X_{parent},\mathbf{Z}=\mathbf{z}). \tag{3.18}$$

3.2.6 Posterior class membership probabilities

The posterior class membership probability of belonging to a particular node k follows as

$$P(X_{child}=k|\mathbf{Y}=\mathbf{y},X_{parent},\mathbf{Z}=\mathbf{z})= P(X_{child}=k|X_{parent},\mathbf{Z}=\mathbf{z})\frac{\prod_{j=1}^{J}P(Y_j=y_j|X_{child}=k,X_{parent})}{P(\mathbf{Y}=\mathbf{y}|X_{parent},\mathbf{Z}=\mathbf{z})}. \tag{3.19}$$

Similar as in Section 3.1, these probabilities can be used to estimate the population proportions and the ME probabilities. Consequently, the PPEs $\hat{P}_{k_{LCT}}$ and the MEPEs $\hat{P}^{Y_j}_{l|k_{LCT}}$ are given by Equations (3.11) and (3.13), respectively.

3.3 Tree-multiple imputation of latent classes (tree-MILC) analysis

A second extension of LC analysis − and at the same time an adaptation of multiple imputation of latent classes (MILC) (Boeschoten et al., 2017) − is tree-multiple imputation of latent classes (tree-MILC) (Remmerswaal, 2022). In tree-MILC analysis, LC analysis is combined with multiple imputation (MI) in a hierarchical tree structure that consists of two levels. Similar to LCT analysis, tree-MILC analysis allows for including covariates at different levels of the tree.

3.3.1 The algorithm

The tree-MILC algorithm starts by taking M bootstrap samples from the original data set. These bootstrap samples are taken to account for the uncertainty of the imputations that are created at a later step. To each bootstrap sample, a 3-class LC model is applied to make a distinction between the latent classes 'permanent', 'flexible' and 'other'. However, to distinguish the latent class 'other' first (see Section 3.2.2), the latent classes 'permanent' and 'flexible' are merged again by combining the posterior probabilities for both classes. The posterior class membership probabilities that result from this step are used to impute latent classes for all observations in the original data set. Note that the latent classes are imputed using proportional assignment.

In the next step, the imputed values are used to create two subsets (see Figure 3.3). The first subset contains all observations for which the latent class 'other' is imputed. The other subset contains all remaining observations. To the latter subset, a 2-class LC model is applied to distinguish between the classes 'permanent' and 'flexible'. The posterior probabilities that result from this model are used to impute latent classes again. For the observations in this subset, the existing imputations are replaced by these newly obtained imputations.

Finally, for each set of imputations, estimates of interest are calculated and pooled using Rubin's pooling rules (Rubin, 1987). Let $\hat{\theta}_m$ denote an estimate of interest based on a particular bootstrap sample m. The pooled estimates are obtained by

$$\overline{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{3.20}$$

The total variance of these estimates can be calculated using

$$V_T = \overline{V}_W + \Big(1 + \frac{1}{M}\Big) V_B, \tag{3.21}$$

where $\overline{V}_W$ is the pooled within variance and V_B the between variance of the estimates. The pooled within variance is given by

$$\overline{V}_W = \frac{1}{M} \sum_{m=1}^{M} \frac{\hat{\theta}_m(1 - \hat{\theta}_m)}{n}. \tag{3.22}$$

The between variance is, in turn, given by

$$V_B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \overline{\theta})^2. \tag{3.23}$$

Figure 3.4 shows a schematic overview of the tree-MILC algorithm.

3.3.2 The model and estimation

The 2-class tree-MILC model at a parent node X_{parent} is equal to the 2-class LCT model as described in Equation (3.14). To estimate the parameters, the same log-likelihood function is maximised. This function is given by

$$\log L(\theta, \mathbf{y}, X_{parent}, \mathbf{z}) = \sum_{i=1}^{n} w_{i|X_{parent}} P(\mathbf{Y} = \mathbf{y}_i | X_{parent}, \mathbf{Z} = \mathbf{z}_i). \tag{3.24}$$

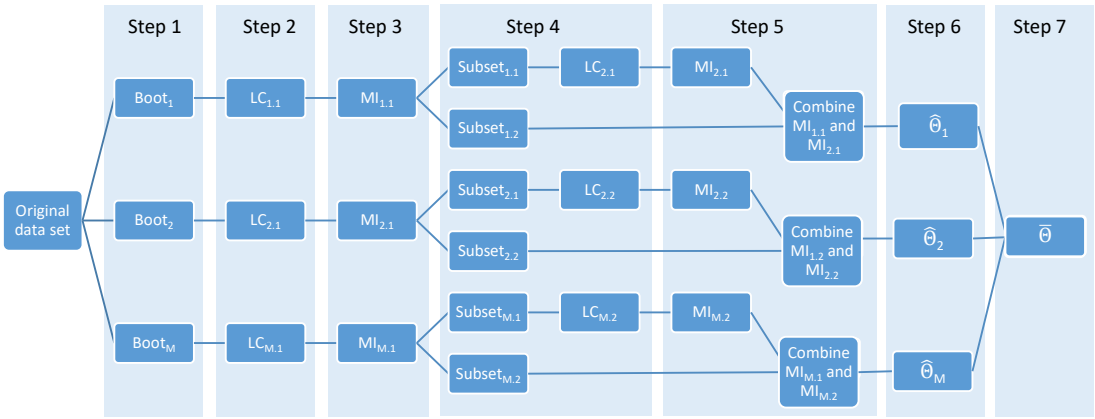
In tree-MILC, however, the weights are either equal to 0 or 1. For all observations with a weight of 1, latent classes are imputed. Let s denote a particular imputed latent class for observation i. New weights for the two child classes are obtained by

$$w_{i|X_{child}=1} \begin{cases} 0, & \text{if } w_{i|X_{parent}} = 0 \text{ or } s \neq 1, \\ 1, & \text{if } s = 1, \end{cases} \tag{3.25}$$

and

$$w_{i|X_{child}=2} \begin{cases} 0, & \text{if } w_{i|X_{parent}} = 0 \text{ or } s \neq 2, \\ 1, & \text{if } s = 2. \end{cases} \tag{3.26}$$

Figure 3.4
Schematic overview of the tree-MILC algorithm



Note. M bootstrap samples are taken from the original data set (step 1). To each bootstrap sample, a 3-class LC model is applied (step 2). The posterior class membership probabilities for the classes 'permanent' and 'flexible' are combined (not shown). Latent classes are imputed for all observations in the original data set (step 3). Subsets are created to which a 2-class LC model is applied (step 4). Imputations are obtained using the posterior probabilities from the second LC model; they replace the imputations from step 3 where necessary (step 5). Estimates of interest ($\hat{\theta}$) are calculated (step 6) and pooled ($\bar{\theta}$) (step 7).

3.3.3 Posterior class membership probabilities

Halfway through tree-MILC analysis, the data is split into two subsets (see Section 3.3.1). The 2-class LC model that distinguishes between the latent classes 'permanent' and 'flexible' is only applied to one of these subsets. For observations in the 'other' subset, computing the posterior probabilities for the classes 'permanent' and 'flexible' is therefore not possible.

3.3.4 The PPE and the MEPE

Let S_i^m denote an imputed latent class for observation i based on bootstrap sample m. With these imputations, the proportion of elements in the population that belong to the true class k can be estimated:

$$\hat{P}_{k_m} = \frac{1}{n} \sum_{i=1}^{n} I(S_i^m = k). \tag{3.27}$$

Using Equation (3.20), the PPE follows as

$$\hat{P}_{k_{TM}} = \frac{1}{M} \sum_{m=1}^{M} \hat{P}_{k_m}. \tag{3.28}$$

Similarly, the imputations obtained with bootstrap sample m can be used to estimate the ME probability $P_{l|k_m}^{Y_j}$:

$$\hat{P}_{l|k_m}^{Y_j} = \frac{\sum_{i=1}^{n} I(S_i^m = k) I(Y_{ij} = l)}{\sum_{i=1}^{n} I(S_i^m = k)}. \tag{3.29}$$

Using Equation (3.20), the MEPE follows as

$$\hat{P}_{l|k_{TM}}^{Y_j} = \frac{1}{M} \sum_{m=1}^{M} \hat{P}_{l|k_m}^{Y_j}. \tag{3.30}$$

Chapter 4

Simulation study 1: Comparing LC, LCT and tree-MILC analysis without missing covariates

In this chapter, a simulation study is conducted to compare the performance of LC, LCT and tree-MILC analysis in conditions without missing covariates. Note that by conducting a simulation study, it is possible to compare the model estimates against a number of known values. In addition, it is possible to study the performance of each method in various conditions.

4.1 Data generation

Latent GOLD® 6.0 (Vermunt & Magidson, 2021) was used to simulate 200 data sets ($n = 10,000$). Each data set consisted of four trichotomous categorical indicators ($Y_1, \ldots, Y_4$) and one dichotomous categorical covariate (Q). The indicators represented four fictitious data sources in which observed contract types (i.e. 'permanent', 'flexible' or 'other') were recorded for all observations. The covariate Q was only included to identify an LC model with two indicators.

To simulate data in Latent GOLD, four theoretical population models were constructed (see Vermunt and Magidson (2021) for more information). Each model was constructed by providing values for all logit parameters associated with an LC model. Realistic values for $P(X = k)$ and $P(X = k|Q = q)$ with $k \in \{1, 2, 3\}$ and $q \in \{1, 2\}$ were obtained by performing an LC analysis on linked data from the ER and LFS for Dutch citizens in the age of 25 to 55 from 2016 to 2018. Since there was no particular requirement for the strength of association between X and Q, the variable *gender* was included as a readily available covariate. This variable was weakly associated (V $\approx$ .06) with the observed contract types in the ER and LFS. The parameters of the resulting LC model yielded true proportions of respectively .60, .14 and .26 for the contract types 'permanent' (P), 'flexible' (F), and 'other' (O).

The amount of ME was, however, made to vary per population model. For this purpose, different values were provided for the parameters $P(Y_j = y_j|X = k)$ with $j \in \{1, \ldots, 4\}$. To find suitable values, a ME probability matrix was constructed for each indicator. The parameters that corresponded to these matrices were subsequently computed as described in Appendix A.

The indicators in the first three population models were constructed using ME probabilities that corresponded to respectively 10%, 20% and 30% ME. The same ME probability matrix was used to construct all four indicators in each model (see Table 4.1). In addition, in each model, the ME probabilities were equal for all combinations of observed and true contract types.

The indicators in the fourth population model were constructed using ME probabilities that were equal to the estimated proportions of inaccuracies in the ER and the LFS for respondents in the age of 25 to 55 in the first quarter of 2018 (see Table 4.2) (Bakker et al., 2021). The indicators Y_1 and Y_3 were constructed using the estimated proportions of inaccuracies for the LFS, whereas the indicators Y_2 and Y_4 were constructed using the estimated proportions of inaccuracies for the ER. Consequently, the indicators Y_1 and Y_3 represented the LFS, whereas the indicators Y_2 and Y_4 represented the ER. The total estimated proportion of inaccuracies in the fourth population model was approximately 7%. Note that most inaccuracies occurred when Y_1 and Y_3 measured the contract type 'flexible' as either 'permanent' ($P_{P|F}^{Y_1,Y_3} = .30$) or 'other' ($P_{O|F}^{Y_1,Y_3} = .07$).

In the end, 50 data sets of size $n = 10,000$ were simulated from each population model. This resulted in 50 data sets with 10% ME, 50 data sets with 20% ME, 50 data sets with 30% ME, and 50 data sets with a realistic 7% ME. Appendix B.1 shows an example of a Latent GOLD syntax file that was used to simulate a data set with 30% ME. A different seed was used to generate each data set.

Table 4.1

Measurement error (ME) probability matrices used to construct the indicators in the population models with 10%, 20% and 30% ME

	10% ME			**20% ME**			**30% ME**		
	P	**F**	**O**	**P**	**F**	**O**	**P**	**F**	**O**
P	.90	.05	.05	.80	.10	.10	.70	.15	.15
F	.05	.90	.05	.10	.80	.10	.15	.70	.15
O	.05	.05	.90	.10	.10	.80	.15	.15	.70

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types.

Table 4.2

The proportion of estimated inaccuracies in the Labour Force Survey (LFS) and the Employment Register (ER) in the first quarter of 2018 for people in the age of 25 to 55 according to hidden Markov models (HMMs)

	LFS			**ER**		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	.982	.007	.011	.931	.067	.002
F	.302	.623	.074	.083	.916	.001
O	.040	.027	.933	.004	.000	.996

Note. From Bakker et al. (2021). P = Permanent, F = Flexible, O = Other.

4.2 Simulation conditions

This simulation study had 24 conditions. These conditions differed with respect to:

- The amount of ME (a realistic 7%; 10%; 20%; 30%),
- The number of indicators (two and a covariate: Y_1, Y_2, Q; three: Y_1, Y_2, Y_3; four: Y_1, Y_2, Y_3, Y_4)
- The sample size (small: $n = 1,000$; large: $n = 10,000$[1]).

The simulation conditions were replicated $N_{\text{Sim}} = 50$ times. For each condition in replication r with $r \in \{1, \ldots, N_{\text{Sim}}\}$, the rth simulated data set with the right amount of ME was selected (i.e. a realistic 7%, 10%, 20% or 30% ME) (see Section 4.1). Subsequently, for conditions with a small sample size, a subset was created that contained a random sample from this data set ($n = 1,000$). This sample was drawn using the seed r. Finally, LC, LCT and tree-MILC analysis was performed using the right number of indicators.

4.3 Model implementation

LC, LCT and tree-MILC analysis were implemented in R version 4.1.3 using RStudio 2022.02.01. Each method involved the estimation of one or more LC models. Estimation was done using Latent GOLD® 6.0. Appendix G provides a link to an online repository that contains the R code that was used for this thesis.

4.3.1 LC analysis

To estimate an LC model in R using Latent GOLD, a text file was generated that contained the appropriate Latent GOLD syntax (Vermunt & Magidson, 2021) (see Appendix B.3). A shell command was used in R to execute this file in Latent GOLD. After estimating the model, the output provided by Latent GOLD was imported back to R for further analysis.

For a detailed description of how LC models are estimated in Latent GOLD, see Vermunt and Magidson (2016). For each model, the same parameters were used as in Appendix B.3. The only difference was that to prevent local minima, 3200 random starting sets were used for models with a small sample size, whereas 100 sets were used for models with a large sample size.

One general problem with LC models, however, is that the class labels of any estimated LC model are arbitrary (Tueller et al., 2011). Nevertheless, to correctly average the results over all simulation replications, the class labels had to be consistent across replications. For every class label permutation, a ME probability matrix was therefore computed. Class labels were consequently adjusted to match the permutation for which the sum of the diagonal elements in the ME probability matrix was highest.

4.3.2 LCT analysis

LCT analysis was implemented in R using the algorithm described in Section 3.2. The 3-class and 2-class LC models were estimated as described in Section 4.3.1.

[1] Initially, an extra condition with $n = 20,000$ observations was included. However, a first analysis showed that there were virtually no differences between the conditions with $n = 10,000$ and $n = 20,000$ observations. This condition was therefore omitted from further analyses.

4.3.3 Tree-MILC analysis

Tree-MILC analysis was implemented in R using the algorithm described in Section 3.3. The number of bootstrap samples taken was 5. This number is usually sufficient for obtaining point estimates with a satisfactory relative efficiency (Rubin, 1987). The 3-class and 2-class LC models were estimated as described in Section 4.3.1.

4.4 Performance measures

The performance of LC, LCT and tree-MILC analysis was evaluated in several ways. In this section, these different ways are described.

4.4.1 Bias, variance and RMSE of the PPEs

The first way to evaluate the performance of the three methods was to compare the bias, the variance and the RMSE of the PPEs (see Section 3.1.7).

An estimator is considered to perform well if it is (approximately) unbiased and precise (Bethlehem, 2009). Let $\hat{\theta}$ denote an estimator (e.g. a PPE) and θ the population parameter. The bias of $\hat{\theta}$ is defined as the difference between the expected value of $\hat{\theta}$ and the true value:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta). \tag{4.1}$$

If the bias is equal to zero, $\hat{\theta}$ is considered to be unbiased. If $\hat{\theta}$ is unbiased, estimates are produced that are on average equal to the population parameter (i.e. $\mathbb{E}(\hat{\theta}) = \theta$).

The estimator $\hat{\theta}$ is furthermore considered to be precise if its variance is small. The variance of $\hat{\theta}$ is given by

$$\text{Var}(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) = \mathbb{E}(\hat{\theta}^2) - \mathbb{E}(\hat{\theta})^2. \tag{4.2}$$

To measure the bias and the variance simultaneously, the RMSE was used. Formally, the RMSE is defined as

$$\text{RMSE}(\hat{\theta}) = \sqrt{\mathbb{E}((\hat{\theta} - \theta)^2)}. \tag{4.3}$$

Using the property $\mathbb{E}(\hat{\theta}^2) = \text{Var}(\hat{\theta}) + \mathbb{E}(\hat{\theta})^2$ from Equation (4.2), this equals

$$\begin{aligned} \text{RMSE}(\hat{\theta}) &= \sqrt{(\text{Var}(\hat{\theta} - \theta) + \mathbb{E}((\hat{\theta} - \theta))^2} \\ &= \sqrt{\text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2}. \end{aligned} \tag{4.4}$$

In the current simulation study, the RMSE was approximated using

$$\text{RMSE}(\hat{\theta}) \approx \sqrt{\frac{\sum_{t=1}^{N_{\text{Sim}}} \left(\hat{\theta} - \theta\right)^2}{N_{\text{Sim}}}}. \tag{4.5}$$

4.4.2 (Mean) bias, variance and RMSE of the MEPEs

The second way to evaluate the performance of LC, LCT and tree-MILC analysis was to compare the (mean) bias, the (mean) variance and the (mean) RMSE of the MEPEs. In order to do this, the expected values and the RMSE values were firstly computed as described in Section 4.4.1. This resulted in $J \cdot K^2$ expected values and $J \cdot K^2$ RMSE values per condition.

The expected values and the RMSE values of MEPEs for indicators with the same amount of ME were, however, expected to be rather similar. To summarise the results, the expected values and the RMSE values were therefore averaged over all indicators in conditions with 10%, 20% and 30% ME. Likewise, in conditions with a realistic 7% ME, the expected values and RMSE values were averaged (separately) over the indicators Y_1 and Y_3 (if $J > 2$), and over the indicators Y_2 and Y_4 (if $J = 4$) (see Section 4.1).

Furthermore, to get an impression of how accurately the total amount of ME was estimated, the sum was computed of the biases of estimators of the probability that a contract type was observed correctly (i.e. $P^{Y_j}_{\mathrm{P|P}}, P^{Y_j}_{\mathrm{F|F}}, P^{Y_j}_{\mathrm{O|O}}$). The result was averaged over all indicators:

$$\text{Mean summed bias} = \frac{1}{J}\sum_{j=1}^{J}\Big(\text{Bias}(\hat{P}^{Y_j}_{\mathrm{P|P}}) + \text{Bias}(\hat{P}^{Y_j}_{\mathrm{F|F}}) + \text{Bias}(\hat{P}^{Y_j}_{\mathrm{O|O}})\Big). \tag{4.6}$$

If the resulting value was negative, the total amount of ME was overestimated (and vice versa).

4.4.3 Mean entropy R^2

The third way to evaluate the performance of LC, LCT and tree-MILC analysis was to compare the mean entropy R^2 of the different models. The entropy R^2 measures the degree of the dispersion of the posterior distributions, and indicates how well latent class membership can be predicted (Boeschoten et al., 2017). To compute the entropy R^2, the entropy was firstly computed:

$$\text{Entropy } (X|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} P(X = k|\mathbf{Y} = \mathbf{y}_i, \mathbf{Z} = \mathbf{z}_i) \log P(X = k|\mathbf{Y} = \mathbf{y}_i, \mathbf{Z} = \mathbf{z}_i). \tag{4.7}$$

The entropy was then rescaled to a value between 0 and 1, where 1 means perfect prediction (Dias & Vermunt, 2008):

$$\text{Entropy } R^2 = \frac{\text{Entropy } (X|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})}{n \cdot \log K}. \tag{4.8}$$

Finally, the mean entropy R^2 was computed by averaging the entropy R^2 over all $N_{Sim} = 50$ replications. Note that the mean entropy R^2 was only computed for LC and LCT models, since for tree-MILC models, computing the posterior class membership probabilities was not possible (see Section 3.3).

4.5 Results

In this section, the results of the simulation study are described. Note that in conditions with two indicators, the covariate Q was included to achieve model identification ($V \approx .06$).

4.5.1 Results for the PPEs and the MEPEs

The results for the PPEs and the MEPEs are shown in Figures 4.1 and 4.2, and C.1-C.4 in Appendix C.[2] Figure 4.1 shows the results for estimators of the proportion of permanent contracts in conditions with a small sample size, whereas Figures C.1- C.2 show the results for all PPEs. Similarly, Figure 4.2 shows the results for estimators of the probability that a permanent contract was observed correctly in conditions with a small sample size, whereas Figures C.3-C.8 show the results for all estimators of the probabilities that a contract type was observed correctly.

The results for the PPEs and the MEPEs are described in three parts. Firstly, a general overview is given of the biases of the PPEs and the MEPEs. Secondly, the performance of the PPEs and the MEPEs is compared across the different simulation conditions. Lastly, the performance of LC, LCT and tree-MILC is compared within the simulation conditions. Note that to summarise the results for the MEPEs, only the performance of estimators of the probability that a contract type was observed correctly (i.e. $P^{Y_j}_{\mathrm{P|P}}$, $P^{Y_j}_{\mathrm{F|F}}$ and $P^{Y_j}_{\mathrm{O|O}}$) is discussed. For the bias and variance of the other MEPEs in biased conditions, see Figures C.5-C.8.

4.5.1.1 Inspecting the bias

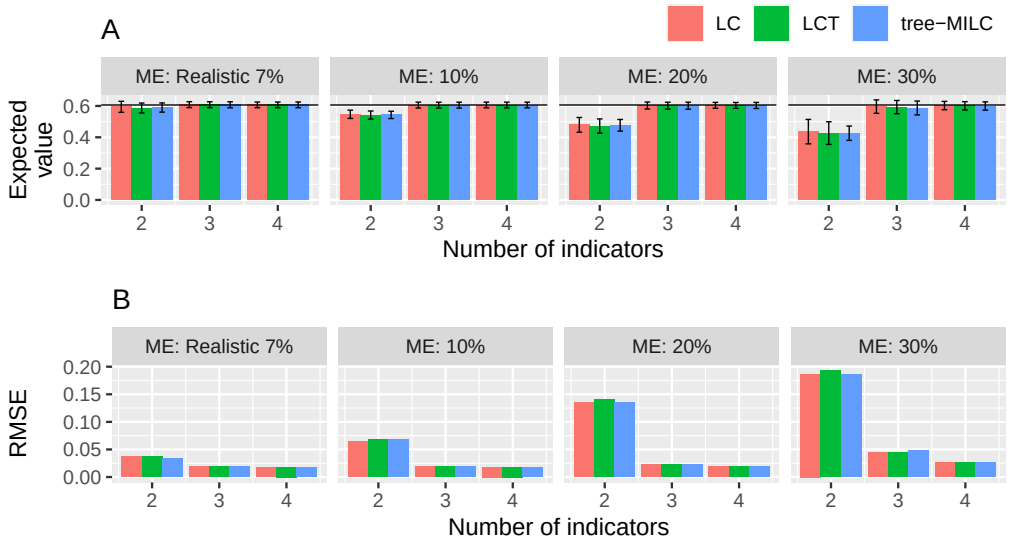
A considerable amount of bias was observed for PPEs and MEPEs in conditions with two indicators and 10%, 20%, or 30% ME; Figures 4.1A and 4.2A show that the expected values of these estimators deviated noticeably from the true proportions. Similarly, a small bias was observed for PPEs and MEPEs in conditions with two indicators, a small sample size, and a realistic 7% ME, as well as in conditions with three indicators, a small sample size, and 30% ME. In all other conditions, the PPEs and the MEPEs were unbiased.

Figure 4.1 shows that whenever bias was present, the direction of the bias differed per estimator. For instance, in biased conditions, the PPEs underestimated the proportion of elements with a permanent contract (see Figure 4.1A), but overestimated the proportion of elements with a flexible or an 'other' contract (see Figure C.1). Likewise, the MEPEs overestimated the probability that a permanent contract was observed correctly (see Figure 4.2A), but underestimated the probability that a flexible or an 'other' contract was observed correctly (see Figure C.3).

Furthermore, the relative size of the bias differed per estimator. For instance, of the three PPEs, estimators of the proportion of permanent contracts had the largest bias, whereas estimators of the proportion of 'other' contracts had the smallest bias (see Figure C.1). In contrast, of the three MEPEs, estimators of the probability that a permanent contract was observed correctly had the largest bias, whereas estimators of the probability that a flexible contract was observed correctly had the smallest bias (see Figure C.3). Overall, the total amount of ME was overestimated by the MEPEs, as shown by the negative values in Figures 4.3 and C.9.

[2] This section summarises the results in all figures in Appendix C. However, to avoid repeatedly mentioning all figures in the text, only the in-text figures (Figures 4.1 and 4.2) are referred to as much as possible. The figures in Appendix C are only referred to when the results are more clearly shown in these figures.

Figure 4.1
Expected value and RMSE of population proportion estimators (PPEs) for permanent contracts in conditions with a small sample size (n=1,000)



Note. ME = measurement error. The black lines show the true proportions of observations with a permanent contract. The bias is given by the difference between the expected values and the true proportions. The error bars show the standard deviations of the expected values.

4.5.1.2 Comparing the PPEs and the MEPEs across simulation conditions

Number of indicators

When comparing the PPEs and the MEPEs across simulation conditions, it was found that the bias of the PPEs and the MEPEs disappeared when a third indicator was added (see Figures 4.1A and 4.2A). The only exception, however, was the condition with 30% and a small sample size (see Figure 4.1A). Nevertheless, in this condition, the bias did also disappear when a fourth indicator was added. Likewise, in all conditions, the variance and the RMSE of the PPEs (see Figures 4.1A and 4.1B) and the MEPEs (see Figures 4.2A and 4.2B) decreased when a third indicator was added. In contrast, when a fourth indicator was added, the variance and the RMSE only decreased further of PPEs and MEPEs in conditions with a small sample size and 20% ME, and in all conditions with 30% ME (see Figures 4.1A, 4.1B, 4.2A and 4.2B).

Sample size

Figures C.1 and C.3 show that overall, the PPEs and the MEPEs had a larger bias and variance in conditions with a small sample size as compared to a large sample size. The opposite was, however, true for PPEs and MEPEs in conditions with two indicators and 20% or 30% ME, where a (slightly) larger variance was observed in conditions with a large sample size as compared to a small sample size (see Figures C.1 and C.3).

Amount of ME

The bias and the variance of PPEs and MEPEs in conditions with two indicators increased as the amount of ME increased (see Figure 4.1A and 4.2A). In contrast, in conditions with three or four indicators, an increase in ME was mostly only associated with an increase in variance.

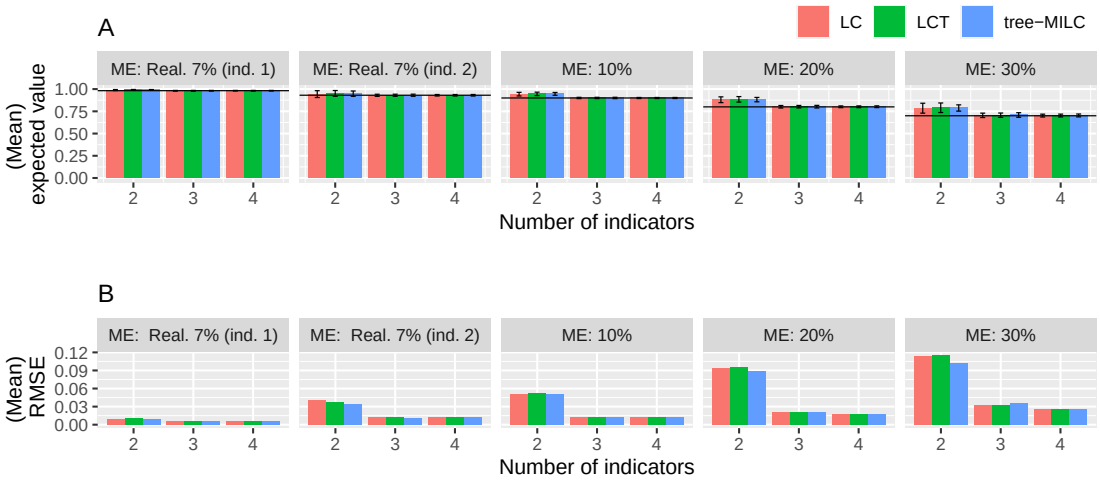
4.5.1.3 Comparing LC, LCT and tree-MILC within simulation conditions

When comparing LC, LCT and tree-MILC analysis within the simulation conditions, no method was found to consistently outperform the other methods in terms of RMSE (see Figures C.2 and C.4). Instead, the performance of LC, LCT and tree-MILC analysis differed slightly for the PPEs and MEPEs. For example, one particular method could perform best for the PPEs, but perform averagely for the MEPEs (or vice versa).

Overall, however, it was concluded that tree-MILC analysis performed best in conditions with a realistic 7% ME, and in conditions with two indicators. In these conditions, LCT analysis often had the relatively worst performance. In contrast, LC and LCT analysis performed best in conditions with three or four indicators and a small sample size. In all other conditions, all methods performed equally well. In Appendix C.4, a more detailed description is provided of the performance of the three methods.

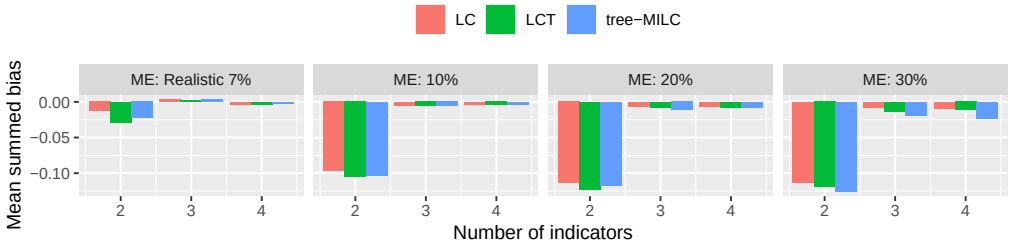
Figure 4.2

(Mean) expected value and (mean) RMSE of estimators of the probability that a permanent contract was observed correctly ($P^{Y_j}_{P|P}$) in conditions with a small sample size (n=1,000)



Note. ME = measurement error. For estimators in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. For estimators in conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (if $J = 4$) (denoted by 'ind. 2'). The black lines show the true ME probabilities in the simulated data. The bias is given by the difference between the (mean) expected values and the true ME probabilities. The error bars show the standard deviations of the (mean) expected values.

Figure 4.3
Mean summed bias of estimators of the probability that a contract type was observed correctly ($P^{Y_j}_{P|P}$, $P^{Y_j}_{F|F}$ and $P^{Y_j}_{O|O}$) in conditions with a small sample size (n=1,000)



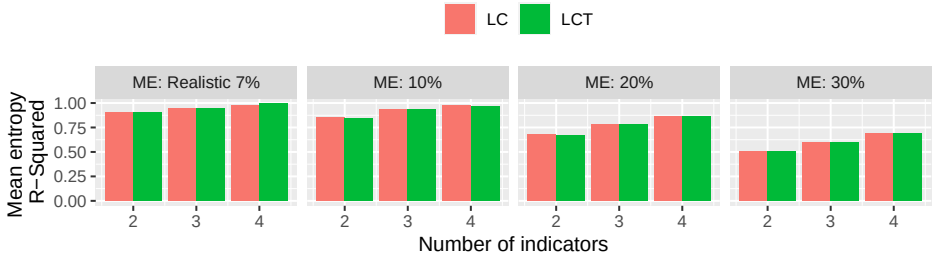
Note. ME = measurement error, P = permanent, F = flexible, O = other. The summed bias was averaged over the number of indicators.

4.5.2 Mean entropy R^2

Figure 4.4 shows the mean entropy R^2 for LC and LCT models in conditions with a small sample size. In addition, Figure C.10 shows the mean entropy R^2 for LC and LCT models in all conditions. The results showed that the mean entropy R^2 increased as the number of indicators increased (see Figure 4.4). The mean entropy R^2, however, also decreased as the amount of ME increased. No differences were found between LC and LCT analysis (see Figure 4.4). Similarly, no differences were found between models with a small and a large sample size (see Figure C.10).

Note that the results also showed that the entropy R^2 is not necessarily a measure of accuracy. For example, the mean entropy R^2 of models in conditions with two indicators and 10% ME was approximately equal to that of models in conditions with four indicators and 20% ME (see Figure 4.4). Nevertheless, Figure 4.1A shows that the PPEs and MEPEs of models in the first condition were biased, whereas the PPEs and MEPEs of models in the second condition were unbiased. From this, it can be concluded that models with equally dispersed posterior distributions may still perform differently in terms of accuracy.

Figure 4.4
Mean entropy R^2 for LC and LCT models in conditions with a small sample size (n=1,000)



Note. ME = measurement error.

4.5.3 Conclusion

In conclusion, the PPEs and MEPEs were often considerably biased in conditions with two indicators and a covariate. The exact amount of the bias depended on the sample size and the amount of ME. When a third indicator was added, the bias of the PPEs and MEPEs disappeared. In addition, the variance of the PPEs and MEPEs decreased. When a fourth indicator was added, the variance of PPEs and MEPEs in conditions with 20% or 30% ME decreased even further.

Overall, tree-MILC analysis performed best in conditions with a realistic 7% ME, and in conditions with two indicators. In these conditions, LCT analysis had the relatively worst performance. In contrast, LC and LCT analysis performed best in conditions with three or four indicators and a small sample size. In all other conditions, all methods performed equally well. Note that in all conditions, the differences between the methods were, however, rather small.

Chapter 5

Simulation study 2: Comparing LC, LCT and tree-MILC analysis with missing covariates

In the previous chapter, the performance of LC, LCT and tree-MILC analysis was compared in a simulation study. In this simulation study, a covariate was only included where necessary to achieve model identification. Nevertheless, the fit of an LVM can be improved further when covariates are included that are (more) strongly associated with the latent variable.

To enhance the accuracy of the HMM estimates, Bakker et al. (2021) identified several covariates that were strongly associated with the inconsistencies between the ER and LFS. One problem, however, was that some covariates were missing for approximately 90% of the observations for which the contract type 'other' was observed by the ER. To understand why, note that the ER only contains information on Dutch insured employees. As a result, some respondents, such as those who were self-employed or unemployed according to the ER, were only present in the LFS and not in the ER. To solve this problem, the contract type 'other' was recorded for the ER for these respondents. However, covariates that were only recorded by the ER were also missing for these respondents. To avoid multicollinearity in the HMMs, all respondents with missing covariates were assigned to existing covariate categories that seemed most suitable. Nevertheless, inaccurate assignments may have affected the accuracy of the estimates.

To overcome this problem, this chapter develops a new approach to include missing covariates in LC, LCT and tree-MILC analysis. The best approach is subsequently used in a simulation study that aims to study to which extent including missing covariates using this approach reduces the bias and the variance of the PPEs and the MEPEs as compared to not including missing covariates. In addition, it is studied whether using this approach to include missing covariates yields differences between LC, LCT and tree-MILC analysis.

5.1 Solving the problem of multicollinearity

Similar to the HMMs (Bakker et al., 2021), multicollinearity occurs when multiple missing covariates are included in an LC model. Let Z_1 and Z_2 be two dichotomous covariates that are strongly associated with inconsistencies between the trichotomous indicators Y_1 and Y_2. In addition, let Z_1 and Z_2 be missing for most observations for which the value 2 is observed by Y_1 (i.e. $Y_1 = 2$). Furthermore, let these observations be assigned to an extra category for Z_1 and Z_2 (i.e.

$Z_1 = 3$ and $Z_2 = 3$), such that this category does not provide any meaningful information, but serves as a residual category for these observations. Multicollinearity occurs because when both Z_1 and Z_2 are included, it becomes impossible to distinguish the individual effects of $Z_1 = 3$ and $Z_2 = 3$ on the latent class X. In other words, it becomes impossible to estimate the parameters $P(X = k|\mathbf{Z} = \mathbf{z})$ in Equation (3.5).

5.1.1 Approach 1: Including covariates at different levels of the tree in LCT and tree-MILC analysis

One potential approach to overcome the problem of missing covariates is to make use of the tree structures of LCT and tree-MILC analysis. Both methods namely provide the opportunity to include missing covariates at the second level of the tree, while excluding them at the first level of the tree (see Sections 3.2 and 3.3). To see whether this approach was suitable for including missing covariates, an initial analysis was conducted. In this analysis, data was simulated as described in Section 5.2. In addition, LCT and tree-MILC analysis was performed on a data set with 10% ME (seed 1), while including three indicators and two missing covariates.

The results showed that multicollinearity did, however, still occur in both models. In LCT analysis, multicollinearity occurred, because observations with missing covariates had a weight that was larger than zero at the second level of the tree (see Section 3.2). Consequently, the model at the second level of the tree was estimated on a data set that still included observations with missing covariates. In tree-MILC analysis, multicollinearity occurred, because some observations with missing covariates were assigned to the subset that contained the observations for which the latent class 'not other' was imputed (see Section 3.3). Similar as in LCT analysis, the model at the second level of the tree was therefore estimated on a data set that included observations with missing covariates.

5.1.2 Approach 2: Adding parameter restrictions

To avoid multicollinearity, the absence of any effects of $Z_1 = 3$ and $Z_2 = 3$ on X was explicitly specified in a second approach. For this purpose, the following restriction was added

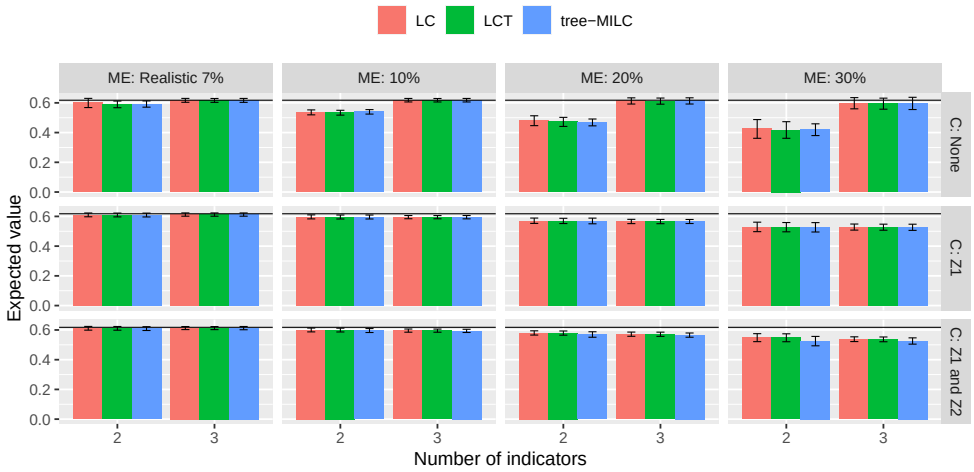
$$P(X = k|\mathbf{Y} = \mathbf{y}, Z_1 = z_1, Z_2 = 3) = P(X = k|\mathbf{Y} = \mathbf{y}, Z_1 = z_1).$$

Note that this restriction was also applied to LC analysis. Consequently, missing covariates were also included at the first level of the tree in LCT and tree-MILC analysis.

An initial analysis showed that this approach did, however, also not yield the desired results. For example, when simulating data as described in Section 5.2 and when replicating the conditions with $n = 1,000$ in Section 5.3 $N_{Sim} = 25$ times, larger biases were found for PPEs in conditions with three indicators and two missing covariates as compared to conditions with three indicators and no missing covariates (see Figure 5.1). Furthermore, in all conditions with one or two missing covariates, the ME probabilities for indicator Y_1 to observe a contract type correctly (i.e. $P^{Y_1}_{P|P}$, $P^{Y_1}_{F|F}$, and $P^{Y_1}_{O|O}$, but especially $P^{Y_1}_{O|O}$) were estimated as larger than the true values (see Table 5.1). In contrast, the ME probabilities for the other indicators to observe a contract type correctly were estimated as smaller than the true values. Nevertheless, Table 5.2 shows that models in conditions without missing covariates did correctly estimate the ME probabilities as approximately equal for all indicators. These results suggest that the models with missing covariates were unable to accurately capture the relationship between the indicator Y_1 and the missing covariates.

Figure 5.1

Expected value of population proportion estimators (PPEs) for permanent contracts in conditions with a small sample size (n=1,000) when adding parameter restrictions



Note. ME = measurement error; C = missing covariates.

Table 5.1

Measurement error (ME) probabilities as estimated by one LC model on a data set with 30% ME, while including three indicators and two missing covariates

	Indicator Y_1			**Indicator Y_2**			**Indicator Y_3**		
Model	**P**	**F**	**O**	**P**	**F**	**O**	**P**	**F**	**O**
P	.847	.137	.016	.717	.114	.169	.680	.140	.179
F	.176	.803	.021	.158	.494	.349	.170	.479	.351
O	.000	.000	1.00	.330	.224	.447	.427	.148	.425

Note. P = Permanent, F = Flexible, O = Other. The data set was generated using the seed 1.

Table 5.2

Measurement error (ME) probabilities as estimated by one LC model on a data set with 30% ME, while including three indicators

	Indicator Y_1			**Indicator Y_2**			**Indicator Y_3**		
Model	**P**	**F**	**O**	**P**	**F**	**O**	**P**	**F**	**O**
P	.702	.122	.176	.762	.116	.122	.692	.156	.152
F	.057	.777	.167	.132	.722	.146	.180	.674	.147
O	.193	.194	.613	.099	.207	.694	.272	.102	.626

Note. P = Permanent, F = Flexible, O = Other. The data set was generated using the seed 1.

5.1.3 Approach 3: Adding direct effects and parameter restrictions

To ensure that the models accurately captured the relationship between the indicator Y_1 and the missing covariates, a third approach was developed. In this approach, direct effects of Z_1 and Z_2 were added on Y_1, while specifying the restriction that whenever Z_1 or Z_2 equals 3, the indicator Y_1 has the value 2:

$$P(Y_1 = 2|X = k, Z_1 = 3) = 1, \;\; P(Y_1 \neq 2|X = k, Z_1 = 3) = 0,$$

and

$$P(Y_1 = 2|X = k, Z_2 = 3) = 1, \;\; P(Y_1 \neq 2|X = k, Z_2 = 3) = 0.$$

Note that in all other cases, Z_1 and Z_2 did not have a direct influence on Y_1. In addition, the absence of any effects of $Z_1 = 3$ and $Z_2 = 3$ on X was specified using

$$P(X = k|\mathbf{Y} = \mathbf{y}, Z_1 = z_1, Z_2 = 3) = P(X = k|\mathbf{Y} = \mathbf{y}, Z_1 = z_1).$$

Consequently, missing covariates were included in all steps of LCT and tree-MILC analysis. An initial analysis showed that this approach did, in fact, yield the desired results. In the remainder of this chapter, a simulation study is therefore conducted in which this approach is used.

5.2 Data generation

The data for this simulation study consisted of four trichotomous indicators $(Y_1, \ldots, Y_4)$, one dichotomous covariate (Q), and two trichotomous covariates $(Z_1,\ Z_2)$. The indicators and the covariate Q were as described in Section 4.1. The covariates Z_1 and Z_2 represented the missing covariates that were strongly associated with the inconsistencies between the ER and LFS.

The data was generated in two steps. In the first step, Latent GOLD® 6.0 was used to simulate 100 data sets of size $n = 1,000$, and 100 data sets of size $n = 10,000$. The data sets contained the variables as mentioned above. One difference, however, was that Z_1 and Z_2 were simulated as dichotomous instead of trichotomous covariates. The covariates Q, Z_1, and Z_2 were based on respectively the variables *gender*, *job duration*, and *main economic activity of the employer* in the linked data from ER and the LFS. Appendix E.2 provides a description of how the latter two variables were recoded to dichotomous variables. To simulate the data, similar steps were taken as described in Section 4.1. Appendix B.2 shows an example of a Latent GOLD syntax file that was used to simulate a data set with 30% ME. In the end, the contract types 'permanent', 'flexible' and 'other' had true proportions of respectively .62, .25, and .13.

In the second step, RStudio 2022.02.01 (R version 4.1.3) was used to transform the covariates Z_1 and Z_2 into missing covariates. For this purpose, a random sample was drawn in each data set that consisted of 90% of the observations for which the contract type 'other' was observed by the indicator Y_1. The observations in this sample were subsequently assigned to a new category of the covariates Z_1 and Z_2 (i.e. $Z_1 = 3$ and $Z_2 = 3$). Table 5.3 shows the strengths of association between the covariates Q, Z_1, and Z_2, and the indicators $Y_1, \ldots, Y_4$ in data sets with different amounts of ME.

Table 5.3
Cramer's V between the covariates Q, Z_1 and Z_2 and the indicators $Y_1, \ldots, Y_4$ in simulated data sets with different amounts of measurement error (ME)

	Realistic 7% ME		**10% ME**		**20% ME**		**30% ME**	
	Y_1/Y_3	Y_2/Y_4	Y_1/Y_3	Y_2/Y_4	Y_1/Y_3	Y_2/Y_4	Y_1/Y_3	Y_2/Y_4
Q	.06	.03	.05	.03	.06	.04	.05	.03
Z_1	.72	.65	.73	.55	.72	.39	.70	.25
Z_2	.69	.62	.69	.51	.68	.34	.68	.20

Note. To compute these values, data sets were generated with a seed of 1.

5.3 Simulation conditions

This simulation study had 48 conditions. These conditions differed with respect to:

- The number of indicators (two: Y_1, Y_2 and the covariate Q; three: Y_1, Y_2, Y_3)[1],
- The number of missing covariates (none; only Z_1; both Z_1 and Z_2),
- The sample size (small: $n = 1,000$; large: $n = 10,000$),
- The amount of ME (a realistic 7%; $10\%; 20\%; 30\%$).

The simulation conditions were replicated $N_{\text{Sim}} = 50$ times. The conditions were implemented as described in Section 4.2.

5.4 Model implementation

LC, LCT and tree-MILC analysis was performed using the same model parameters as described in Section 4.3. The only difference was that an increased maximum number of ER and NR iterations (i.e. each 10 million) was used to achieve model convergence. Direct effects and parameter restrictions were added as described in Section 5.1.3. Appendix B.3 shows an example of a Latent GOLD syntax file for an LC model that included the two missing covariates Z_1 and Z_2.

5.5 Performance measures

The performance measures in Section 4.4 were used to compare LC, LCT and tree-MILC analysis.

5.6 Results

In this section, the results of the second simulation study are described. Note (again) that in conditions with two indicators, the covariate Q was included to achieve model identification.

[1]The results in Section 4.5 showed that there were very little differences between conditions with three and four indicators. The condition with four indicators was therefore omitted in this simulation study.

5.6.1 Results for the PPEs and the MEPEs

The results for the PPEs and the MEPEs are shown in Figures 5.2 and 5.3, and D.1-D.8 in Appendix C.4.2.[2] Figure 5.2 shows the results for estimators of the proportion of elements with a permanent contract in conditions with a small sample size, whereas Figures D.1-D.4 show the results for all PPEs in all conditions. Similarly, Figure 5.3 shows the results for estimators of the probability that a flexible contract was observed correctly, whereas Figures D.5-D.8 show the results for MEPEs in all conditions.

The results in this section are divided into two parts. In the first part, the performance of the PPEs and MEPEs is compared across the different simulation conditions. In the second part, the performance of LC, LCT and tree-MILC analysis is compared within the simulation conditions.

Note that since the results for conditions with no missing covariates are already described in Section 4.5.1, they are not described again. In addition, the direction and the relative size of the biases were found to be equal as in Section 4.5.1.1, and are not described again. Thirdly, similar as in Section 4.5.1, the current section only describes the performance of estimators of the probability that a contract type was observed correctly. For the bias and variance of the other MEPEs in biased conditions, see Figures D.9-D.13 in Appendix C.4.2.

5.6.1.1. Comparing the PPEs and the MEPEs across simulation conditions

Number of missing covariates

The results showed that the bias of most PPEs (see Figure 5.2A) and MEPEs (see Figures D.3 and D.5) disappeared when one missing covariate (Z_1) was added. However, in none of the conditions was adding Z_1 enough to eliminate the bias of all PPEs (see Figures D.1 and D.3) and MEPEs (see Figure 5.3A).[3] In contrast, the bias of (almost) all PPEs and MEPEs disappeared when a second missing covariate (Z_2) was added (see Figures 5.2A and 5.3A). Note that the only exceptions were conditions with two or three indicators, a small sample size, and 30% ME, in which small biases were still present after Z_2 was added.

Likewise, when one missing covariate (Z_1) was added, the variance of the PPEs and the MEPEs decreased in all conditions with two indicators, as well as in conditions with three indicators and 30% ME (see Figures 5.2A and 5.3A). In contrast, including a second missing covariate (Z_2) only resulted in an additional decrease in variance of PPEs (see Figures D.1 and D.3) and MEPEs (see Figure 5.3A) in conditions with two indicators.

[2] This section summarises the results in all figures in Appendix C.4.2. However, to avoid repeatedly mentioning all figures in the text, only the in-text figures (Figures 5.2 and 5.3) are referred to as much as possible. The figures in Appendix C.4.2 are only referred to when the results are more clearly shown in these figures.

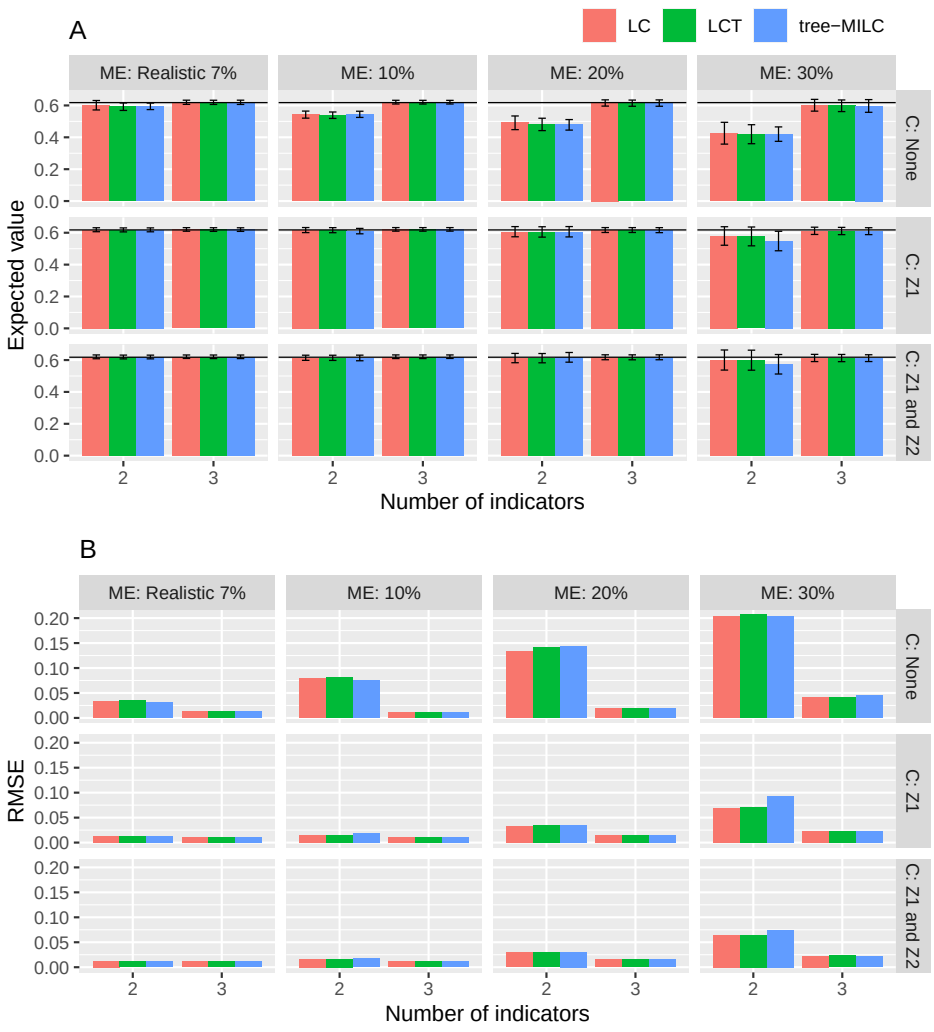
[3] The bias disappeared of PPEs in conditions with a realistic 7% and 10% ME, as well as ME probability estimators of the probability that a permanent contract was observed correctly in conditions with realistic 7%, 10% and 20% ME. The remaining estimators were, however, still biased.

Number of indicators

Similarly, Figures 5.2A and 5.3A show that the bias of all PPEs and MEPEs in conditions with two indicators disappeared when a third indicator was added. In addition, the variance of all PPEs and MEPEs decreased. Interestingly enough, a larger decrease in RMSE was observed when a third indicator was added as compared to when one missing covariate (Z_1) or two missing covariates (Z_1 and Z_2) were added (see Figures 5.2B and 5.3B).

Figure 5.2

Expected value and RMSE of population proportion estimators (PPEs) for permanent contracts in conditions with a small sample size (n=1,000)



Note. ME = measurement error; C = missing covariates. The black lines show the true proportions of observations with a permanent contract. The bias is given by the difference between the expected values and the true proportions. The error bars show the standard deviations of the expected values.

Sample size

Furthermore, in conditions with one (Z_1) or two missing covariates (Z_1 and Z_2), a larger bias and variance was observed for PPEs (see Figures D.1 and D.3) and MEPEs (see Figures D.5 and D.7) in conditions with a small sample size as compared to a large sample size.

Amount of ME

Finally, the bias and the variance of the PPEs and the MEPE increased as the amount of ME increased in conditions with two indicators, and one (Z_1) or two missing covariates (Z_1 and Z_2) (see Figures 5.2A and 5.3A). In contrast, in conditions with three indicators, and one (Z_1) or two missing covariates (Z_1 and Z_2), an increase in ME was (mostly) only associated with an increase in variance of the PPEs and the MEPEs (see Figures 5.2A and 5.3A).

5.6.1.2. Comparing LC, LCT and tree-MILC analysis within simulation conditions

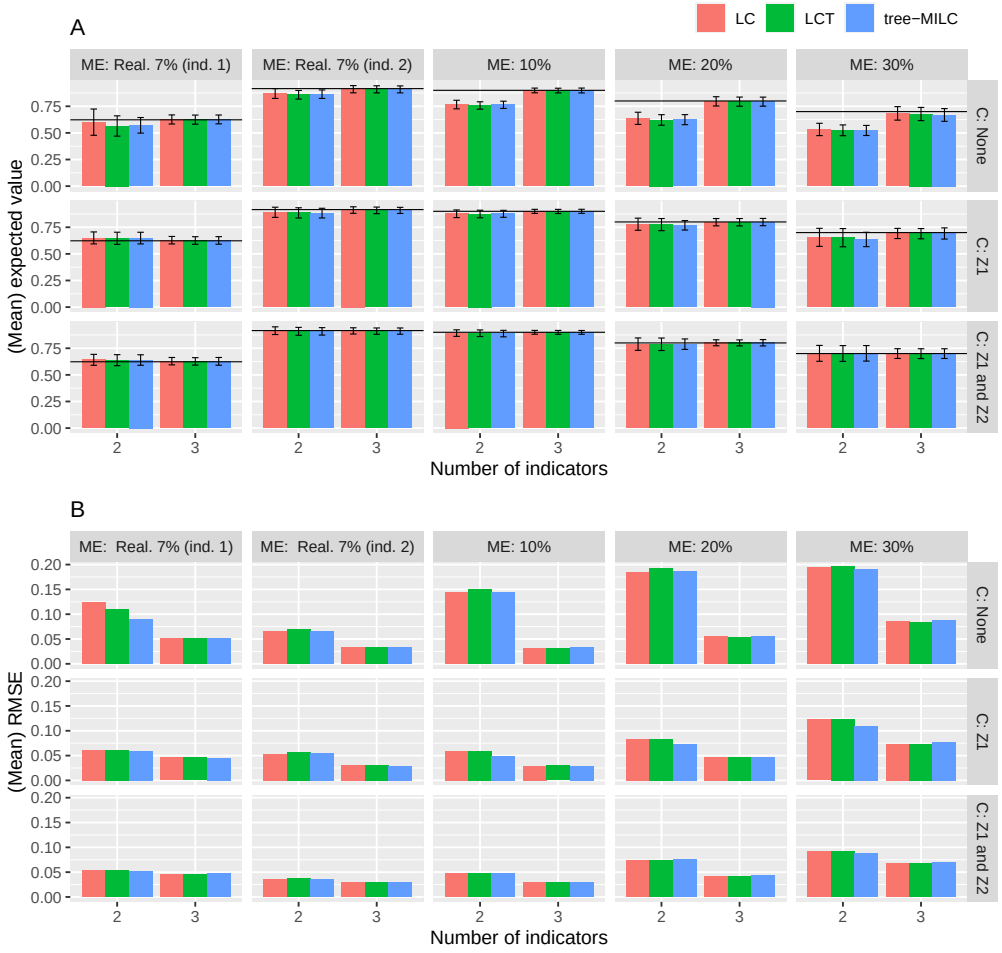
When comparing LC, LCT and tree-MILC within the simulation conditions, the methods were again found to differ slightly for the PPEs and MEPEs (see Figures D.2, D.4,D.6, and D.8). For example, one particular method could perform best in terms of the PPEs, but at the same time perform averagely in terms of the MEPEs (or vice versa). Furthermore, in one specific condition (i.e. two indicators, one missing covariate, a small sample size and 30% ME), LC performed best in terms of the PPEs, whereas tree-MILC performed best in terms of the MEPEs.

Overall, tree-MILC performed best in all conditions with one missing covariate, whereas LC and LCT performed best in all conditions with two missing covariates and a small sample size. In all other conditions, the three methods performed equally well. Appendix D.6 provides a more detailed description of the performance of the three methods.

5.6.2 Mean entropy R^2

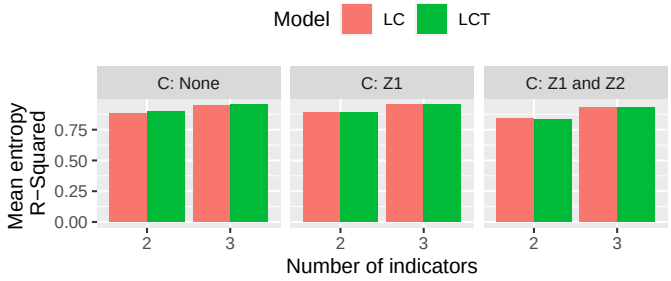
Figure 5.4 shows the mean entropy R^2 for LC and LCT models in conditions with a small sample size and 30% ME. In addition, Figure D.15 in Appendix C.4.2 shows the mean entropy R^2 for LC and LCT models in all conditions. Overall, no differences in mean entropy R^2 were observed between models with no missing covariates and one missing covariate (Z_1) (see Figures 5.4 and Figure D.15). Interestingly enough, however, the mean entropy R^2 decreased when a second missing covariate (Z_2) was added. In addition, similar to Chapter 4, the mean entropy R^2 decreased when the amount of ME increased (see Figure D.15), and increased when a third indicator was added (see Figures 5.4 and Figure D.15). No differences were observed between LC and LCT analysis.

Figure 5.3
(Mean) expected value and (mean) RMSE of estimators of the probability that a permanent contract was observed correctly ($P^{Y_j}_{P|P}$) in conditions with a small sample size (n=1,000)



Note. ME = measurement error; C = missing covariates. Note that for estimators in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. For estimators in conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (if $J = 4$) (denoted by 'ind. 2'). The black lines show the true ME probabilities in the simulated data. The bias is given by the difference between the (mean) expected values and the true ME probabilities. The error bars show the standard deviations of the (mean) expected values.

Figure 5.4
Mean entropy R^2 in conditions with a small sample size (n=1,000) and 30% measurement error



Note. C = missing covariates.

5.6.3 Conclusion

In conclusion, the results showed that the bias and the variance of the PPEs and the MEPEs decreased when one missing covariate was included. When a second missing covariate was included, the bias and the variance of the PPEs and the MEPEs decreased further. Nevertheless, the largest decreases in bias and variance were observed when a third indicator was added.

When comparing the performance of LC, LCT and tree-MILC analysis, only small differences were found. Overall, tree-MILC analysis performed best in conditions with one missing covariate. In contrast, LC and LCT analysis performed best in conditions with two missing covariates and a small sample size. In all other conditions, all methods performed equally well.

Chapter 6

Analyses of real linked data from the ER and the LFS

From the notion in Chapters 1 and 5 that LC, LCT and tree-MILC analysis could potentially overcome the limitations of the HMMs, the question arises to which extent these methods yield different estimates as compared to the HMMs when applied to real data from the ER and the LFS. In this chapter, this question is addressed. Note that since the simulation studies in Chapters 4 and 5 showed that LCT analysis is a more complex, but not better performing method than LC analysis, the decision was made to exclude LCT analysis in this chapter.

6.1 Data

In the analyses in this chapter, real linked data from the ER and the LFS for respondents in the age of 25 to 55 between 2016 and 2018 was used. The choice for age group was made, because Bakker et al. (2021) found that these respondents had the fewest estimated inconsistencies between the ER and the LFS. Thus, by focusing on this age group, the most accurate estimates could be obtained. Note that the data in this study was equal to the data used for the HMMs (Bakker et al., 2021).

The data consisted of several observations per respondent (see Chapter 1). However, to minimise the impact of changes in the observed and true contract types over time, a subset was created that only included the first observations of respondents in which both the ER and the LFS were recorded. This subset only included the first observations in the first quarters of 2016, 2017 and 2018. In addition, to conduct separate analyses for each year, three different subsets were created for 2016 ($n = 12,843$), 2017 ($n = 11,910$), and 2018 ($n = 14,682$).

Each subset consisted of two indicators that contained the contract types as observed by the ER and the LFS. For a detailed description on how the contract types were operationalised, see Bakker et al. (2021). Table 6.1 shows how the observed contract types differed for the ER and the LFS. Furthermore, the data consisted of nine covariates that were strongly associated with the inconsistencies between the ER and the LFS. Of these covariates, the covariates *gender, education level, interview manner*, and *migration background* were available for all respondents. In contrast, the covariates *company size, contract hours, economic activity, job duration*, and *software cluster* were missing for approximately 90% of the respondents for which the contract type 'other' was observed in the ER. Appendix E.1 provides a short description of all covariates.

Table 6.1
Contingency table for the contract types as observed by the Employment Register (ER) and the Labour Force Survey (LFS) for respondents in the age of 25 to 55 in the first quarters of 2016, 2017 and 2018

	LFS								
	2016			2017			2018		
ER	**P**	**F**	**O**	**P**	**F**	**O**	**P**	**F**	**O**
P	6370	202	124	5809	185	116	7502	202	154
F	994	1263	152	898	1328	145	1100	1633	173
O	121	108	3509	119	98	3212	141	114	3633

Note. P = Permanent, F = Flexible, O = Other.

6.2 Model specification

To minimise the impact of changes in the observed and true contract types over time, LC and tree-MILC analyses were conducted separately for 2016, 2017 and 2018. In each analysis, the two indicators and the nine covariates were included as described in Section 6.1. To avoid multicollinearity, the missing covariates *company size, contract hours, economic activity, job duration*, and *software cluster* were included by adding direct effects and parameter restrictions as described in Section 5.1.3. The analyses were performed as described in Section 4.3. The same model parameters were used as in the simulation study described in Section 5.4.

6.2.1 Performance evaluation

To obtain insights into the fit of the LC models, the entropy R^2 was computed (see Section 4.4). In addition, to obtain insights into the precision of the tree-MILC estimates, the standard deviation of the pooled estimates was computed. Note that this was done by taking the square root of the variance as described in Section 3.3.

6.3 Results

Figure 6.1 shows the population proportions as estimated by the LC and the tree-MILC models. In addition, Tables F.1-F.6 in Appendix F show the ME probabilities as estimated by the LC and the tree-MILC models. For the population proportions and the ME probabilities as estimated by the HMMs, see the discussion paper by Bakker et al. (2021, p. 33-34). Finally, to facilitate comparison of the ME probability estimates, Figures 6.2, F.2, and F.3 show the differences in the ME probability estimates between the LC models and the HMMs, and between the tree-MILC models and the HMMs.

6.3.1 Entropy R^2 of the LC models

For the first quarters of 2016, 2017 and 2018, entropy R^2 values of respectively $0.94, 0.95$, and 0.94 were found. These results show that the LC models had a high degree of certainty when

classifying individuals into latent classes. For the corresponding HMMs, similar entropy R^2 values of 0.95 were found (Bakker et al., 2021, p. 69). Do note, however, that Section 4.5.2 showed that entropy R^2 is not necessarily a measure of accuracy.

6.3.2 Standard deviation of the tree-MILC estimates

The standard deviance of the population proportion estimates of tree-MILC analysis ranged from 0.0051 to 0.0246 in the first quarters of 2016, 2017 and 2018 (see Table F.8). These results show that the tree-MILC estimates had a high level of precision.

6.3.3 Population proportion estimates

Figure 6.1 shows that the population proportion estimates differed for the three methods. For example, the estimated proportions of permanent contracts were consistently higher for the LC and the tree-MILC models than for the HMMs. In contrast, the estimated proportions of flexible and 'other' contracts were consistently lower for the LC and the tree-MILC models than for the HMMs.

In addition, the results showed that the tree-MILC estimates were more similar to the estimates from the LFS, while the estimates from the LC models and the HMMs were more similar to the estimates from the ER. The tree-MILC models, as a result, showed different trends for the estimated proportions of permanent contracts as compared to the other two methods. For example, although the LC models and the HMMs showed respective increases of 0.18 and 0.14 from 2016 to 2018, the tree-MILC models showed an overall decrease of 0.003 from 2016 to 2018. Moreover, tree-MILC was the only method that showed a decrease in 2017, but an increase in 2018.

Furthermore, when looking at the estimated proportions of flexible contracts, the LC and the tree-MILC models showed respective increases of 0.009 and 0.028 between 2016 and 2018. The HMMS did, however, not show any increases or decreases at all. In addition, due to tree-MILC's resemblance to the LFS as mentioned earlier, there was a considerable difference of 0.04 in the estimated proportions of flexible contracts between tree-MILC and the HMM in 2016.

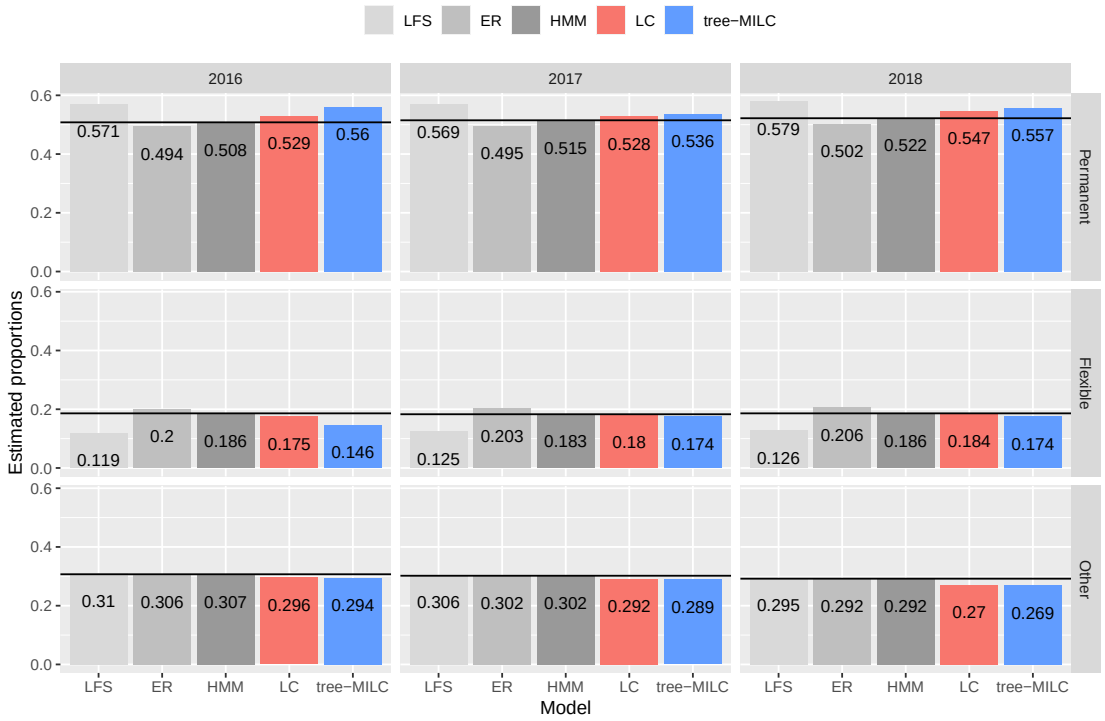
Finally, the LC and the tree-MILC models showed a larger decrease in the estimated proportions of 'other' contracts from 2016 to 2018 than the HMMs. Note, however, that the differences between the three methods were smaller than the original differences between the ER and the LFS.

6.3.4 ME probability estimates

Furthermore, small differences were found in the ME probability estimates between the LC models and the HMMs, and between the tree-MILC models and the HMMs (see Figures 6.2, F.2, and F.3). For example, both the LC and the tree-MILC models estimated the probability that the LFS observed a flexible contract as a permanent contract ($P^{\text{LFS}}_{P|F}$) as lower than the HMMs. Consequently, the LC and the tree-MILC models estimated the probability that the LFS measured a flexible contract correctly ($P^{\text{LFS}}_{F|F}$) as higher than the HMMs.

Overall, the differences between the LC and the HMM estimates (see Figure 6.2A) were smaller than those between the tree-MILC and the HMM estimates (see Figure 6.2B). In addition, the differences between the LC and the HMM estimates from 2016 to 2018 were more

Figure 6.1
Proportions per contract type as estimated by the Labour Force Survey (LFS), the Employment Register (ER), the hidden Markov models (HMMs), the LC models and the tree-MILC models for respondents in the age of 25 to 55 in the first quarters of 2016, 2017 and 2018



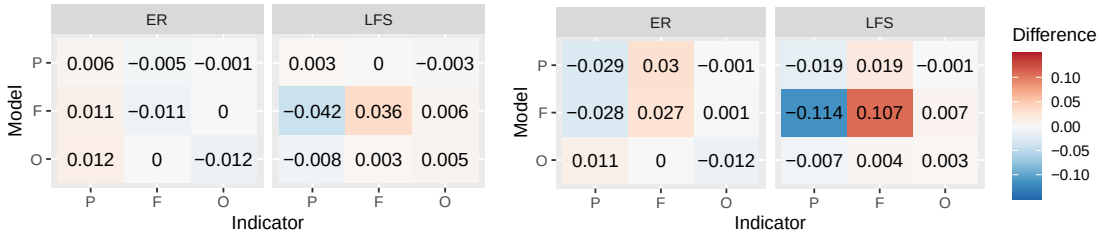
Note. The HMM estimates are denoted by the black lines.

consistent than the differences between the tree-MILC and the HMM estimates (see Figures 6.2, F.2, and F.3). For example, the latter differences were considerably larger in 2016 than in 2017 and 2018. Note that this finding is in line with the considerable difference of 0.04 that was found in Section 6.3.3 for the estimated proportions of flexible contracts between the tree-MILC model and the HMM in 2016.

6.4 Additional analyses

The results in the previous section showed that the LC estimates, the tree-MILC estimates, and the original HMM estimates differed from each other. However, it is still unknown what caused these differences. For example, one explanation could be that in LC and tree-MILC analysis, the missing covariates were included in a different way than in the HMMs. Nevertheless, it could also be that the differences were caused by inherent differences between the methods. To investigate the plausibility of both explanations, additional LC and tree-MILC analyses were performed.

Figure 6.2
Differences in measurement error (ME) probability estimates for respondents in the age of 25 to 55 in the first quarter of 2016



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a measurement error (ME) probability as lower than the HMM (and vice versa).

6.4.1 Two different approaches

In the additional analyses, two different approaches were used. In the first approach, no missing covariates were included, whereas in the second approach, observations with missing covariates were assigned to the same categories as in the HMMs. In Appendix E.1, it is described how this was done. In the end, the results were compared to the results in Section 6.3.

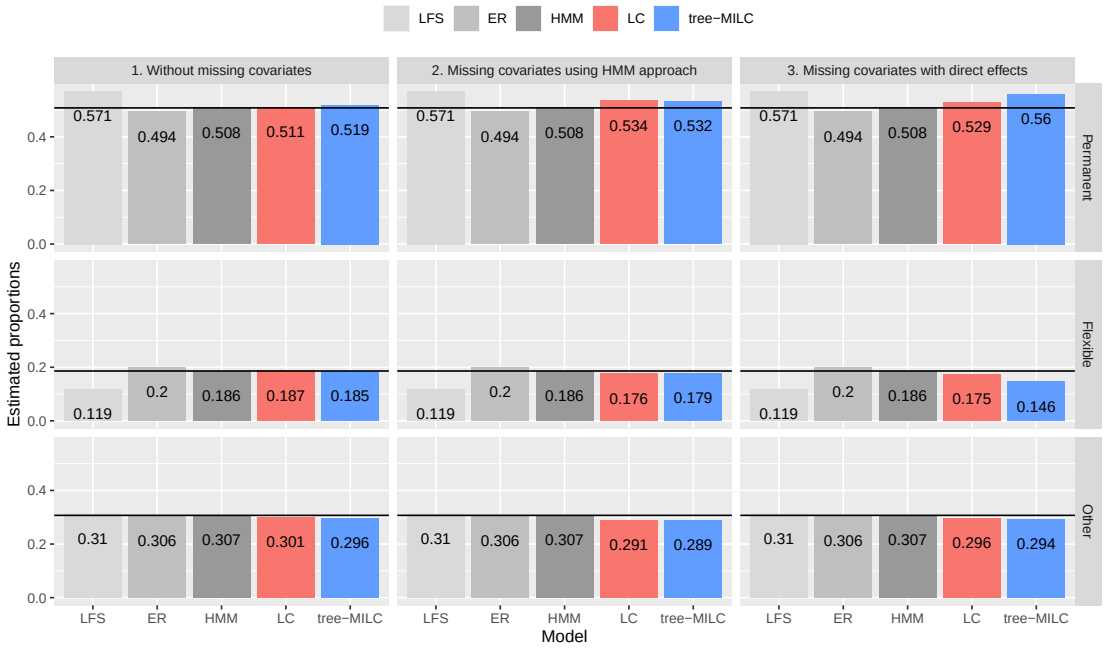
6.4.2 Results of the additional analyses

The population proportion estimates for the additional analyses are shown in Figures 6.3, F.4, and F.5 for respectively 2016, 2017, and 2018. The differences in the ME probability estimates between the LC models and the HMMs, and between the tree-MILC models and the HMMs are shown in Figures F.6-F.11.

The results showed that when observations with missing covariates were assigned to the same categories as in the HMMs, the LC and the tree-MILC estimates were still different from the HMM estimates (see Figure F.4). For example, the estimated proportion of permanent contracts was still higher for the LC and the tree-MILC models than for the HMMs. These results suggest that the differences in the results were most likely due to inherent differences between the methods rather than the fact that covariates were included in different ways. One difference between the methods is, for instance, that the LC and the tree-MILC analyses only included one observation per respondent, whereas the HMMs also modelled the development of ME over time.

Furthermore, the results showed that there were little to no differences between the LC and the tree-MILC estimates when not including missing covariates at all, or when assigning observations with missing covariates to the same categories as in the HMMs (see Figure 6.3). In contrast, larger differences were found when missing covariates were included with direct effects and parameter restrictions (see Chapter 5). These results suggest that the differences between the LC and the tree-MILC models may, in fact, have been caused by including missing covariates using direct effects and parameter restrictions.

Figure 6.3
Proportions per contract type as estimated by the Labour Force Survey (LFS), the Employment Register (ER), the hidden Markov models (HMMs), the LC models and the tree-MILC models for respondents in the age of 25 to 55 in the first quarter of 2016



Note. The HMM estimates are denoted by the black lines.

Chapter 7

Discussion

This thesis aimed to compare the performance of LC, LCT and tree-MILC analysis to correct for measurement error in the ER and the LFS. For this purpose, two simulation studies were conducted: one without missing covariates (Chapter 4) and one with missing covariates (Chapter 5). For the second simulation study, a new approach was developed in which missing covariates were included in LC, LCT and tree-MILC analysis without having to assign observations with missing covariates to existing categories. In the end, LC and tree-MILC analysis was performed on real linked data for respondents in the age of 25 to 55 in the first quarters of 2016, 2017, and 2018. In this way, the estimates obtained with LC and tree-MILC analysis could be compared to the estimates obtained with HMMs in previous research (Bakker et al., 2021) (Chapter 6).

In this final section, the results of this study are summarised and interpreted (see Section 7.1). In addition, the limitations of this study are described (see Section 7.2), and suggestions for future research are given (see Section 7.3). Finally, the use of model-based estimation methods for the production of official statistics is discussed (see Section 7.4).

7.1 Summary and interpretation of the results

The results in Chapter 5 showed that the best approach to include missing covariates was by using direct effects and parameter restrictions. Furthermore, the results of the simulation studies in Chapters 4 and 5 showed that estimators in conditions with two indicators and a covariate were often considerably biased. The bias and the variance decreased when one or two missing covariates were included using direct effects and parameter restrictions. Nevertheless, the largest decreases in bias and variance were observed when a third indicator was added. Overall, little differences were found between LC, LCT and tree-MILC analysis. However, since LCT analysis was a more complex but not better performing method than LC analysis, LCT analysis was excluded from the analyses on real linked data in Chapter 6. These analyses showed that the LC, the tree-MILC, and the original HMM estimates were different from each other. Additional analyses showed that this could be due to inherent differences between the methods. Nevertheless, these differences were smaller than the original differences between the ER and the LFS.

The finding that estimators had the largest bias in conditions with two indicators and one covariate is in line with previous research. For example, Wurpts and Geiser (2014) found that when using LC analysis for clustering purposes, data sets with larger sample sizes, more indicators, and larger covariate effects yielded less biased model parameters. Both results can be explained by the fact that the parameters of an LC model can be more accurately estimated

when more information is available.

Furthermore, possible explanations can be provided for the remaining results. For example, the fact that only small differences were found between LC, LCT, and tree-MILC analysis in the simulation studies can be explained by the fact that all three methods used LC analysis to distinguish between the latent classes. In addition, each method included missing covariates using the same direct effects and parameter restrictions. Interestingly enough, larger differences were found when applying LC and tree-MILC analysis to the real data. One potential explanation for this finding is that the simulated data may not have been able to fully capture the complexity and underlying structure of the real data (see also Section 7.2). Alternatively, the estimates obtained for the real data may have been influenced by random variability, since the models were only estimated once for the real data, but multiple times in the simulation studies. Finally, the results showed that the best approach to include missing covariates was by using direct effects and parameter restrictions. This finding can be explained by the fact that this approach provided the models with information about how the missing covariates were related to the indicators. In the other approach, in which this type of information was not provided, the models were unable to accurately capture the relationship between the indicators and the missing covariates.

7.2 Limitations of the current study

This study did have some limitations. Firstly, in the analyses of the real data, a number of assumptions were made that may not have been met by the real data. For example, the assumption was made that covariates were measured without error. However, in the real data, covariates may have been measured with error in the same way as the indicators. Furthermore, the assumption was made that the amount of measurement error was equal for each respondent within each latent class. Nevertheless, some respondents may have been more likely to provide incorrect responses as compared to other respondents. Future research should therefore investigate in which ways these assumptions can be relaxed. One suggestion to relax the second assumption would for instance be to add an additional latent variable that models unobserved heterogeneity in a similar way as done by Bakker et al. (2021) and Pavlopoulos and Vermunt (2015).

Secondly, the simulated data may not have fully captured the complexity and the underlying structure of the real data. For example, unlike the real data, the covariates were simulated without measurement error. In addition, in the simulated data, the amount of measurement error was equal for each respondent within each latent class. Nevertheless, these differences may have limited the degree to which the results of the simulation studies are generalisable to analyses of the real data.

Finally, the limited number of replications in the simulation studies may have introduced simulation noise in the results. For example, the finding that some estimators had a larger variance in conditions with a large sample size as compared to a small sample size suggests that simulation noise may have been present. When interpreting the results, it should therefore be taken into account that the biases and variances observed in the simulation studies are only approximations of the true values.

7.3 Implications and suggestions for future research

Although several estimates have been obtained for the contract types of Dutch citizens, it is still unknown which estimates − whether it be the LC's, the tree-MILC's, or the HMM's − are most accurate. One suggestion for further research would therefore be to extend the simulation study by also including the HMMs. In this way, the performance of LC and tree-MILC analysis can be compared to the performance of the HMMs. In addition, future research that aims to correct for measurement error in the ER and the LFS could use the approach proposed in this study to include missing covariates. Moreover, a similar approach for including missing covariates could potentially be applied to the HMMs. Future research could therefore investigate how this can be done and include this approach in the extended simulation study as mentioned earlier. Furthermore, since the accuracy of the estimates improved considerably when a third indicator was added, the current findings suggest that it may be beneficial for CBS to find a third indicator that also measures the contract types of Dutch citizens. For this purpose, big data sources can potentially be used. Note, however, that such a third indicator may contain more measurement error than the ER and the LFS. Future research should therefore also investigate to which extent adding a third indicator of a lower quality will influence the accuracy of the estimates. Lastly, an important question that still needs to be addressed is whether the ER and the LFS measure the same concept. So far, previous research has only found inconclusive results (Restrepo Estrada, 2023). Nevertheless, if the ER and the LFS do measure somewhat different concepts, there would be no longer be a need to correct for measurement error in both sources. Future research should therefore continue to address this question.

7.4 Using model-based estimation methods in practice

Finally, LVMs, such as the LC models, the tree-MILC models, and the hidden Markov models, could potentially be used for the production of official statistics. Although CBS currently still relies more on traditional design-based approaches due to their robustness against model-misspecification, methodologists argue that NSIs should not be afraid to use model-based estimation methods for the production of official statistics (Buelens, 2016).[1] By using model-based estimation methods, it would, for instance, not only be possible to improve the accuracy of the estimates, but also to gain insights into changes in the data that would go unnoticed by using design-based approaches alone. However, in order for CBS to move into this direction, further research should be conducted on both a methodological and an organisational level.

Future methodological research should investigate which model is most suitable for the production of official statistics. Note that choosing such a model may involve making a trade-off between the performance of the model and the complexity of implementing the model in practice. To investigate which model is most suitable, the issues mentioned in Sections 7.2-7.3 should be addressed. In addition, in order to align with the current statistics, future research should investigate whether these models can also be used to distinguish between subcategories of the contract types 'flexible' (e.g. temporary, agency, and on-call contracts) and 'other' (e.g. non-employed, self-employed, or director-major shareholders). Similarly, future research should investigate whether these models can be used to obtain estimates on a regional level (e.g. by potentially combining the results of these models with small area estimation methods).

[1] See for example Buelens (2016) for a number of guidelines that can be used to evaluate model-based estimation methods based on the principles of objectivity and reliability of the European Code of Practice.

The way in which official statistics are produced can, however, not be changed without gaining support from the employees who produce the statistics. Future research should therefore also be conducted on an organisational level. For example, future research could focus on how the results of an LVM should be communicated to employees, such as researchers, subject-matter specialists, or managers with various degrees of statistical training, in order to improve their understanding of these models. Furthermore, exploratory interviews could be conducted to investigate whether these employees understand the added value of measuring one concept with multiple variables, trust the results of an LVM, and recognize the necessity of using such models. By conducting these interviews, the current level of support can be assessed, and potential barriers can be identified that need to be addressed to increase the level of support. Finally, it is important to engage in open and transparent communication with the employees who produce the statistics. This includes being clear about the limitations of these models at all times. Nevertheless, being clear about the limitations of these models may also raise the question of whether more transparency is required with regard to the statistics that are currently produced. One final suggestion for future research would therefore be to explore the potential benefits of including information on measurement uncertainty in official statistics. Including this type of information could potentially help users of the data understand and compare the reliability of both the design-based and model-based estimates.

Disclaimer

The views expressed in this thesis are those of the author and do not necessarily reflect the views of Statistics Netherlands.

Acknowledgements

I would like to thank my supervisors, Dr. S. Scholtus, Dr. Reinoud Stoel and Dr. S. J. W. Willems, for their excellent guidance throughout my internship. In addition, I would like to thank my fellow EMOS students for their friendship and support in the past months. It was a pleasure to occupy the room next to the coffee machine on the third floor together.

Bibliography

Bakker, B. F. M. (2011a). *Micro integration*. Statistics Netherlands.

Bakker, B. F. M. (2011b). Micro-integration. State of the art. In *ESSnet on Data Integration WP1 State of the Art on Statistical Methodologies for Data Integration* (pp. 77–107). (ESSnet). Eurostat.

Bakker, B. F. M., Gringhuis, G., Hoogland, J., Van Der Linden, F., Michiels, J., Pannekoek, J., Scholtus, S., & Smits, W. (2021). *Tijdelijke en vaste contracten. Verschillen tussen de schattingen uit de Polisadministratie en de Enquête beroepsbevolking verklaard?* CBS. https://www.cbs.nl/-/media/_pdf/2021/33/tijdelijke-en-vaste-contracten.pdf

Bethlehem, J. (2009). *Applied survey methods*. John Wiley & Sons, Inc.

Biemer, P. P. (2009). Chapter 12 - Measurement errors in sample surveys. In C. R. Rao (Ed.), *Handbook of statistics* (pp. 281–315). Elsevier. https://doi.org/10.1016/S0169-7161(08)00012-6

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. P. (2011). *Latent class analysis of survey error*. John Wiley & Sons, Inc.

Boeschoten, L., Oberski, D., & De Waal, T. (2017). Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent class Modelling (MILC). *Journal of Official Statistics*, *33*(4), 921–962. http://dx.doi.org/10.1515/JOS-2017-0044

Buelens, B. (2016). Model based estimation at Statistics Netherlands. *European Conference on Quality in Official Statistics (Q2016)*.

De Waal, T., Van Delden, A., & Scholtus, S. (2019). Multi-source statistics: Basic situations and methods. *International Statistical Review*, *88*(1), 1–26. https://doi.org/10.1111/insr.12352

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*, 643–659. https://doi.org/10.1007/s00180-007-0103-7

Galindo Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics*, *28*(2), 173–198.

Groves, R. M., Fowler, F. J., Couper, M. P., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. John Wiley & Sons, Inc.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (Ed.), *Measurement and Prediction* (pp. 362–472). Princeton University Press.

Magidson, J., Vermunt, J. K., & Madura, J. P. (2020). Latent class analysis. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *Sage research methods foundations*. Sage Publications. https://doi.org/10.4135/9781526421036883636

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling*, *24*(2), 180–197. https://doi.org/10.1080/10705511.2016.1254049

Oberski, D. L. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. In *Total survey error in practice* (pp. 339–358). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119041702.ch16

Pankowska, P., Bakker, B. F. M., Oberski, D., & Pavlopoulos, D. (2018). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. *Statistical Journal of the IAOS*, *34*(3), 317–329. https://doi.org/10.3233/SJI-170368

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology*, *41*(1), 197–214. https://doi.org/10.1177/102425890401000206

Remmerswaal, D. (2022). *Comparing the performance of MILC and tree-MILC to estimate and correct for multiple sources of errors in combined datasets: A simulation study* (Master's thesis). Utrecht University.

Restrepo Estrada, F. G. (2023). *Do Employment Register and Labour Force Survey measure the same employment contract type concept?* (Master's thesis). Leiden University.

Rubery, J., Keizer, A., & Grimshaw, D. (2016). Flexibility bites back: The multiple and hidden costs of flexible employment policies. *Human Resource Management Journal*, *26*(3), 235–251. https://doi.org/10.1111/1748-8583.12092

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc.

Smits, W., Gringhuis, G., Bakker, B. F. M., Hoogland, J., van der Linden, F., Michiels, J., Pannekoek, J., & Scholtus, S. (2021). *Verschillen tussen schattingen van flexibele en vaste arbeidsrelaties*. https://www.cbs.nl/nl-nl/longread/rapportages/2021/verschillen-tussen-schattingen-van-flexibele-en-vaste-arbeidsrelaties?onepage=true

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(1), 110–131. https://doi.org/10.1080/10705511.2011.534695

Van Den Bergh, M. (2018). *Latent class trees* (Doctoral dissertation). Proefschriftenmaken.nl.

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2021). *LG-Syntax user's guide: Manual for Latent GOLD syntax module version 6.0*. Arlington, MA: Statistical Innovations Inc.

Vermunt, J. K. (2017). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2004). *Latent class analysis in the Sage encyclopedia of social science research methods*.

Wikman, A., & Wärneryd, B. (1990). Measurement errors in survey questions: Explaining response variability. *Social Indicators Research*, *22*(2), 199–212.

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a monte-carlo study. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00920

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

Computing the logit parameters

This section describes how the logit parameters were computed in Section 4.1. These parameters were used to construct population models with different amounts of measurement error (ME).

A.1 Multinomial logistic regression parameterisation

The probabilities in Equation (3.3) can be parameterised using multinomial logistic regression models. Let X denote a latent variable, K the number of latent classes, and k a particular latent class. The probability $P(X = k)$ can be parameterised by

$$P(X = k) = \frac{\exp(\alpha_k)}{\sum_{k'=1}^{K} \exp(\alpha_{k'})}, \tag{A.1}$$

where α_k is a logistic intercept that corresponds to the log-odds of belonging to class k. Furthermore, let Y_j denote an indicator with R_j categories, and r a particular category of indicator Y_j. The ME probability $P(Y_j = j|X = k)$ can be parameterised by

$$P(Y_j = r|X = k) = \frac{\exp(\alpha_r + \beta_{rk})}{\sum_{r'=1}^{R_j} \exp(\alpha_{r'} + \beta_{r'k})}, \tag{A.2}$$

where α_r is a logistic intercept that corresponds to the log-odds of having category r on indicator Y_j. In addition, β_{rk} is a logistic slope that corresponds to the log-odds of having category r on indicator Y_j conditional on belonging to the latent class k.

A.2 Computing the logit parameters

For a latent variable X and an indicator Y_j with three classes, Equation (A.2) follows as

$$P(Y_j = j|X = k) = \frac{\exp(\alpha_j + \beta_{jk})}{\exp(\alpha_1 + \beta_{1k}) + \exp(\alpha_2 + \beta_{2k}) + \exp(\alpha_3 + \beta_{3k})}. \tag{A.3}$$

Since α_1, β_{11}, β_{12}, β_{13}, β_{21}, and β_{31} correspond to the reference categories, they are equal to 0. The ME probabilities $P(Y_j = y_j|X = k)$ are therefore given by the equations in Table A.1.

Table A.1
Measurement error (ME) probabilities for the trichotomous indicator Y_j and the trichotomous latent variable X parameterised using multinomial regression models

	$Y_j = 1$	$Y_j = 2$	$Y_j = 3$
X=1	$P(Y_j = 1\|X = 1) = \frac{1}{1+\exp(\alpha_2)+\exp(\alpha_3)}$	$P(Y_j = 2\|X = 1) = \frac{\exp(\alpha_2)}{1+\exp(\alpha_2)+\exp(\alpha_3)}$	$P(Y_j = 3\|X = 1) = \frac{\exp(\alpha_3)}{1+\exp(\alpha_2)+\exp(\alpha_3)}$
X=2	$P(Y_j = 1\|X = 2) = \frac{1}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{32})}$	$P(Y_j = 2\|X = 2) = \frac{\exp(\alpha_2+\beta_{22})}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{32})}$	$P(Y_j = 3\|X = 2) = \frac{\exp(\alpha_3+\beta_{32})}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{32})}$
X=3	$P(Y_j = 1\|X = 3) = \frac{1}{1+\exp(\alpha_2+\beta_{23})+\exp(\alpha_3+\beta_{33})}$	$P(Y_j = 2\|X = 3) = \frac{\exp(\alpha_2+\beta_{23})}{1+\exp(\alpha_2+\beta_{23})+\exp(\alpha_3+\beta_{33})}$	$P(Y_j = 3\|X = 3) = \frac{\exp(\alpha_3+\beta_{33})}{1+\exp(\alpha_2+\beta_{23})+\exp(\alpha_3+\beta_{33})}$

To find the logit parameters that correspond to the ME probability matrix for the population model with 10% ME in Table 4.1, the following equations need to be solved:

$$\frac{1}{1+\exp(\alpha_2)+\exp(\alpha_3)} = 0.90, \tag{A.4}$$

$$\frac{\exp(\alpha_2)}{1+\exp(\alpha_2)+\exp(\alpha_3)} = 0.05, \tag{A.5}$$

$$\frac{1}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{32})} = 0.05, \tag{A.6}$$

$$\frac{\exp(\alpha_2+\beta_{22})}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{32})} = 0.90, \tag{A.7}$$

$$\frac{\exp(\alpha_2+\beta_{22})}{1+\exp(\alpha_2+\beta_{23})+\exp(\alpha_2+\beta_{23})} = 0.05, \tag{A.8}$$

$$\frac{\exp(\alpha_3+\beta_{33})}{1+\exp(\alpha_2+\beta_{22})+\exp(\alpha_3+\beta_{33})} = 0.90. \tag{A.9}$$

Solving these equations yields $\alpha_2 = \alpha_3 = -\log(18)$, $\beta_{22} = \beta_{32} = \log(324)$, and $\beta_{23} = \beta_{33} = \log(18)$. Note that the logit parameters that correspond to other ME probability matrices (e.g. in Tables 4.1 and 4.2) can be computed in a similar way.

Appendix B

Latent GOLD syntax

B.1 Latent GOLD syntax for generating a data set in simulation study 1

An example of a Latent GOLD syntax file for generating a data set (n=10,000) with 30% measurement error in the first simulation study.

```
version = 6.0
infile 'ExampleData.dat'

model
    title 'simulation3';
    options
    algorithm
        tolerance=1e-08 emtolerance=0.01 emiteration=500 nriterations=500;
    startvalues
        seed=1 sets=100 tolerance=1e-05 iterations=100;
    montecarlo
        seed=1 replicates=500 tolerance=1e-008;
    bayes
        categorical=1 variances=1 latent=1 poisson=1;
    missing includeall;
    output
     parameters=first standarderrors profile reorderclasses iterationdetails;
outfile 'simDat3_iteration7.dat' simulation=1 seed=7;
    variables
         caseid id;
         caseweight w20000;
         dependent Y1 nominal 3, Y2 nominal 3, Y3 nominal 3, Y4 nominal 3;
         independent Q nominal;
         latent cluster nominal 3;
     equations
         cluster <- 1 + Q;
         Y1      <- 1 + cluster;
         Y2      <- 1 + cluster;
```

```
         Y3      <- 1 + cluster;
         Y4      <- 1 + cluster;
{
 -0.9662 -1.626 0.22 0.2607
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 }
end model
```

B.2 Latent GOLD syntax for generating a data set in simulation study 2

An example of a Latent GOLD syntax file for generating a data set (n=1,000) with 30% measurement error in the second simulation study.

```
version = 6.0
infile 'exampleDat_1000.dat'

model
    title 'simulation3';
    options
    algorithm
        tolerance=1e-08 emtolerance=0.01 emiteration=500 nriterations=500;
    startvalues
        seed=1 sets=100 tolerance=1e-05 iterations=100;
    montecarlo
        seed=1 replicates=500 tolerance=1e-008;
    bayes
        categorical=1 variances=1 latent=1 poisson=1;
    missing includeall;
    output
     parameters=first standarderrors profile reorderclasses iterationdetails;
outfile 'simDat3_iteration45.dat' simulation=1 seed=45;
    variables
         caseid id;
         caseweight w;
         dependent Y1 nominal 3, Y2 nominal 3, Y3 nominal 3;
         independent q nominal, SBIgroep nominal, baanduur nominal;
         latent cluster nominal 3;
     equations
         cluster <- 1 + q + SBIgroep + baanduur;
         Y1      <- 1 + cluster;
         Y2      <- 1 + cluster;
         Y3      <- 1 + cluster;
{
 0.8252 -0.3192 -0.1506 -0.1584 -2.5392 2.2426 -4.6105 -3.4073
```

```
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 -1.54045 -1.54045 3.0809 1.54045 1.54045 3.0809
 }
end model
```

B.3 Latent GOLD syntax for estimating a standard LC model

An example of a Latent GOLD syntax file for estimating a standard LC model with three latent classes, two indicators (Y_1 and Y_2), and one covariate (Q).

```
version = 6.0
infile 'LC_1-2-1-1000-1_data.dat'

model title '1-2-1-1000-1';
    options
    algorithm
        tolerance=1e-08 emtolerance=0.01 emiterations=1000 nriterations=1000;
    startvalues
        seed=1 sets=3200 tolerance=1e-05 iterations=100;
    bayes
        categorical=1 variances=1 latent=1 poisson=1;
    missing includeall;
    output
     parameters=first standarderrors profile reorderclasses iterationdetails;
     outfile 'LC_1-2-1-1000-1_output.dat' classification keep=id;
    variables
dependent Y1 nominal, Y2 nominal;
independent Q nominal;
latent Cluster nominal 3;
    equations
Cluster <- 1 + Q;
Y1 <- 1 + Cluster;
Y2 <- 1 + Cluster;
end model
```

B.4 Latent GOLD syntax for estimating a standard LC model with two missing covariates

An example of a Latent GOLD syntax file for estimating a standard LC model with three latent classes, two indicators (Y_1 and Y_2), one non-missing covariate (Q), and two non-missing covariates (Z_1 and Z_2).

```
version = 6.0
infile 'LC_50-3-null-Z1-Z2-10000-4_data.dat'

model title '50-3-null-Z1-Z2-10000-4';
    options
    algorithm
        tolerance=1e-08 emtolerance=0.01 emiterations=10000000 nriterations=10000000;
    startvalues
        seed=1 sets=100 tolerance=1e-05 iterations=100;
    bayes
        categorical=1 variances=1 latent=1 poisson=1;
    missing includeall;
    output
     parameters=first standarderrors profile reorderclasses iterationdetails;
     outfile 'LC_50-3-null-Z1-Z2-10000-4_output.dat' classification keep=id;
    variables
dependent Y1 nominal, Y2 nominal, Y3 nominal;
independent  Z1 nominal, Z2 nominal;
latent Cluster nominal 3;
    equations
Cluster <- 1 + (kk) Z1 + (ll) Z2;
Y1 <- 1 | Cluster + (aa~ful) 1 | Z1 + (bb~ful) 1 | Z2;
Y2 <- 1 | Cluster;
Y3 <- 1 | Cluster;
ll[1,3] = 0; ll[1,4] = 0;
aa[1,] = 0;
aa[2,] = 0;
aa[3,1] = -100;
aa[3,2] = 0;
aa[3,3] = -100;
bb[1,] = 0;
bb[2,] = 0;
bb[3,1] = -100;
bb[3,2] = 0;
bb[3,3] = -100;
end model
```

Appendix C

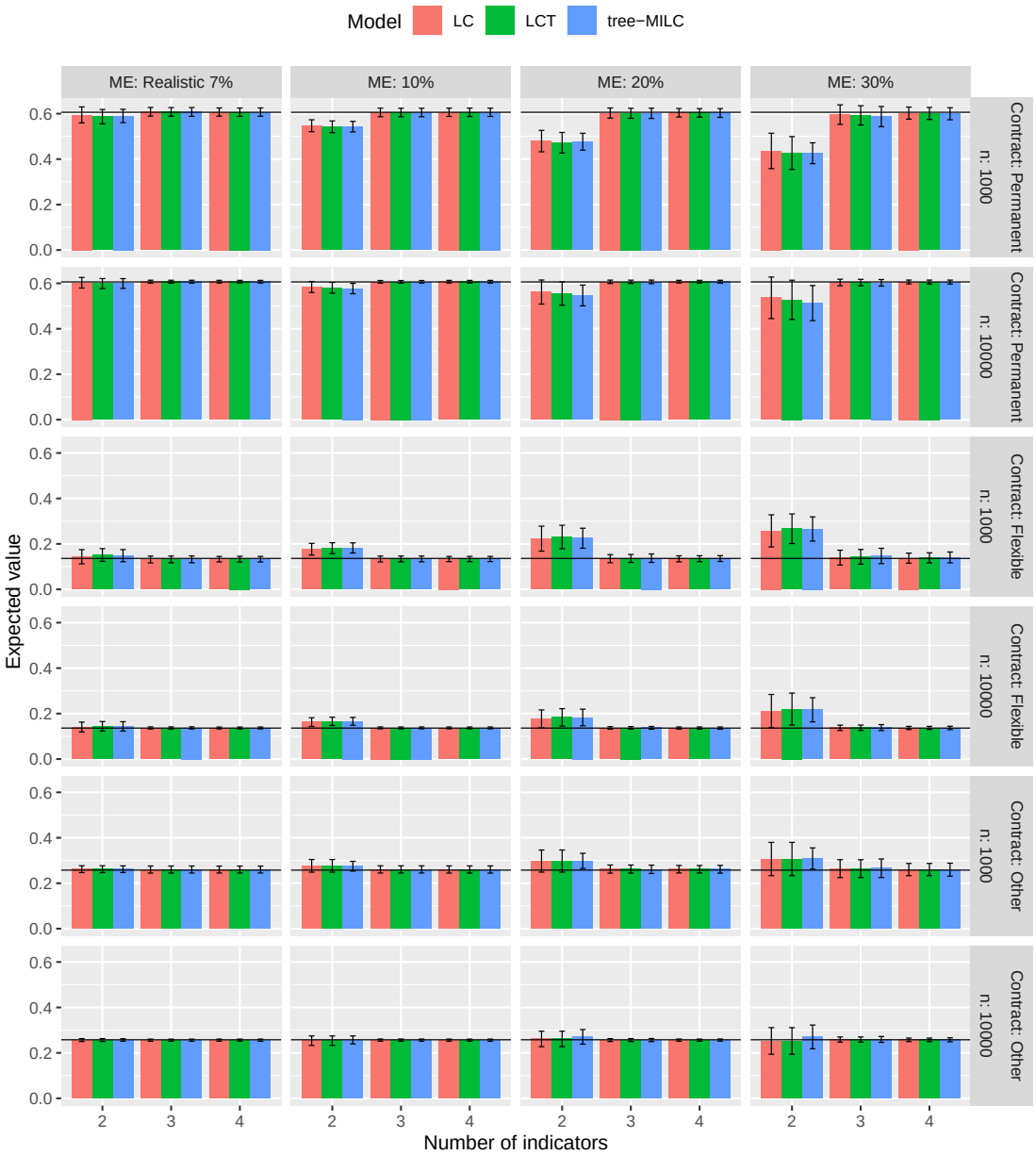
Results of simulation study 1

C.1 Population proportion estimators (PPEs)

See next page.

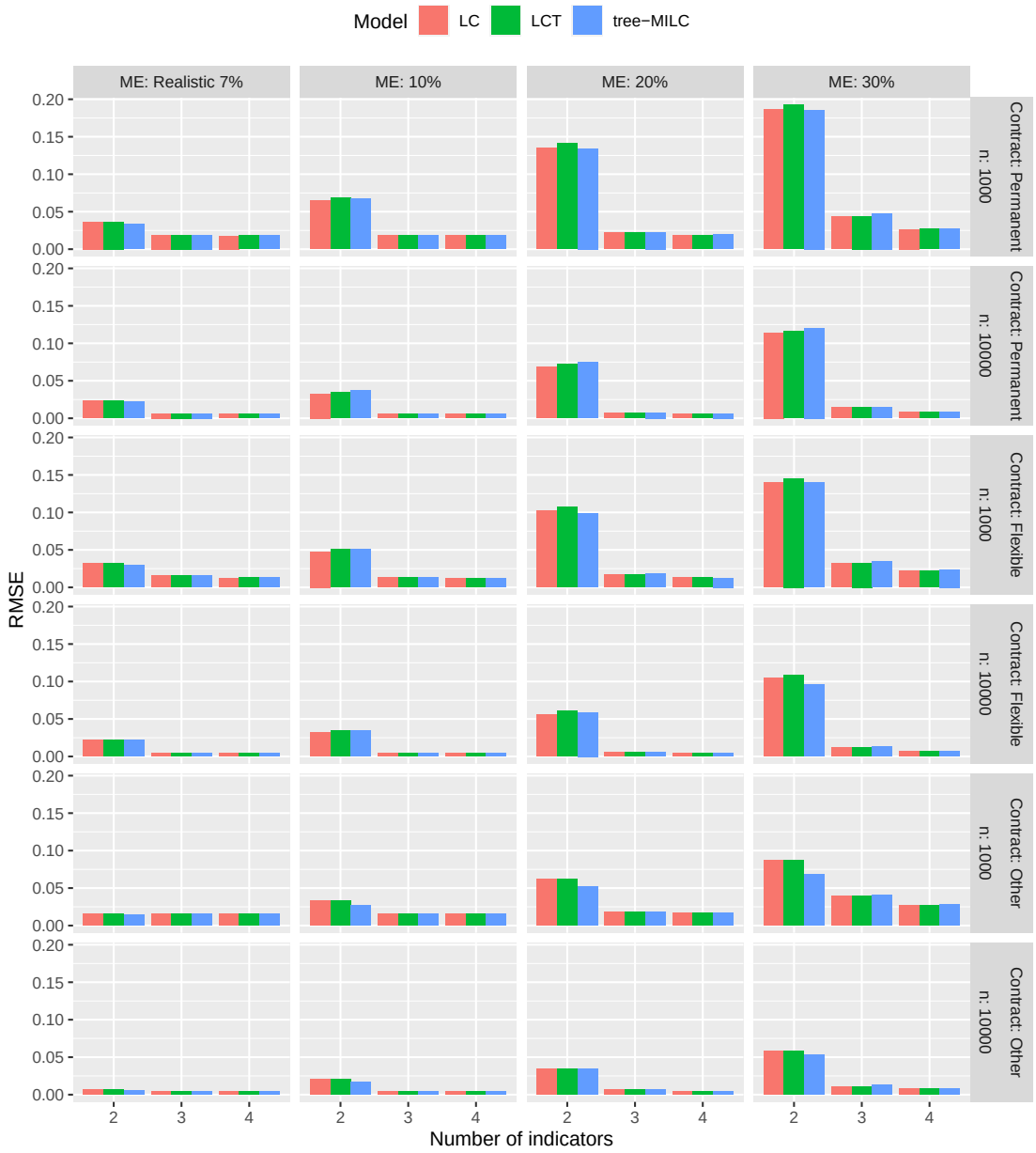
Figure C.1

Expected value of the population proportion estimators (PPEs) in the first simulation study



Note. ME = measurement error. The black lines show the true proportions in the simulated data. The bias is given by the difference between the expected values and the true proportions. The error bars show the standard deviations of the expected values.

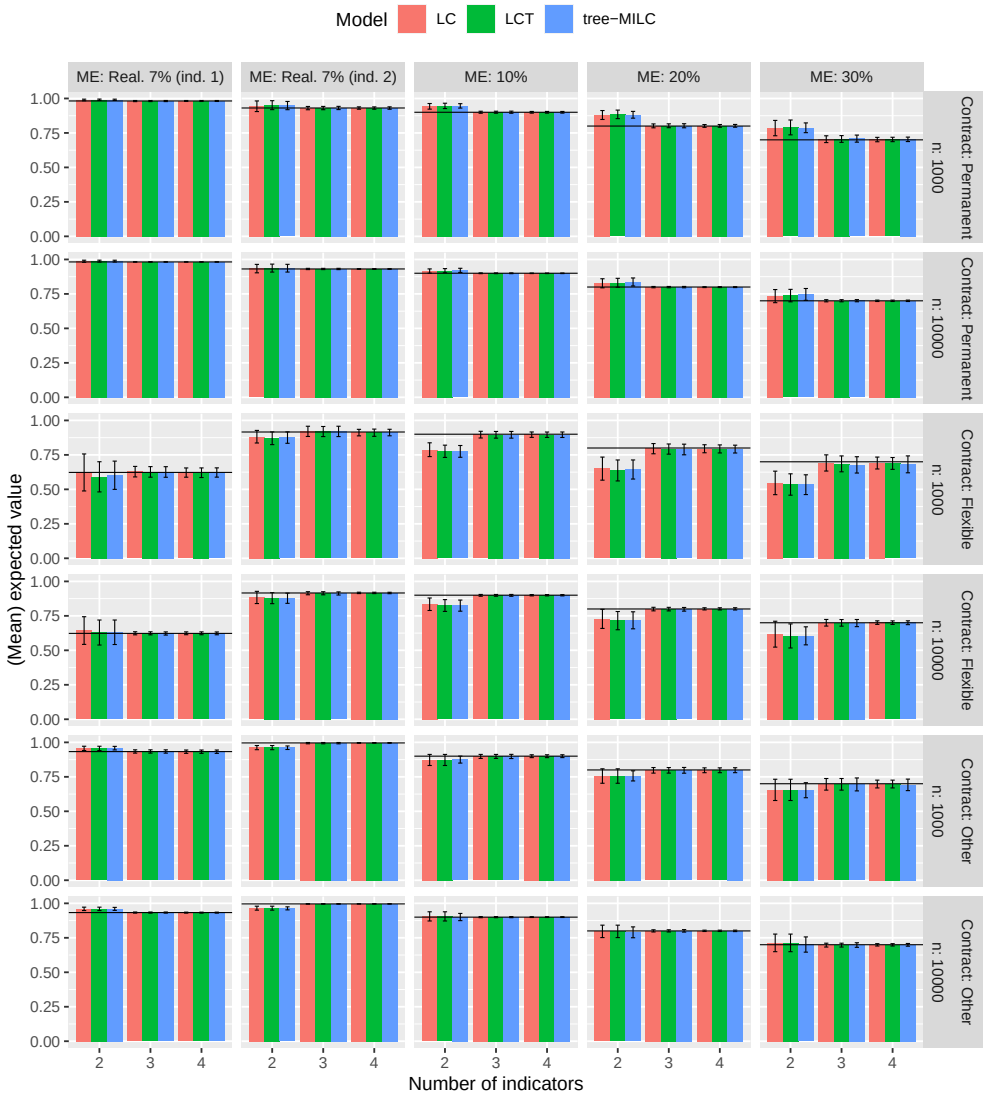
Figure C.2
RMSE of the population proportion estimators (PPEs) in the first simulation study



Note. ME = measurement error.

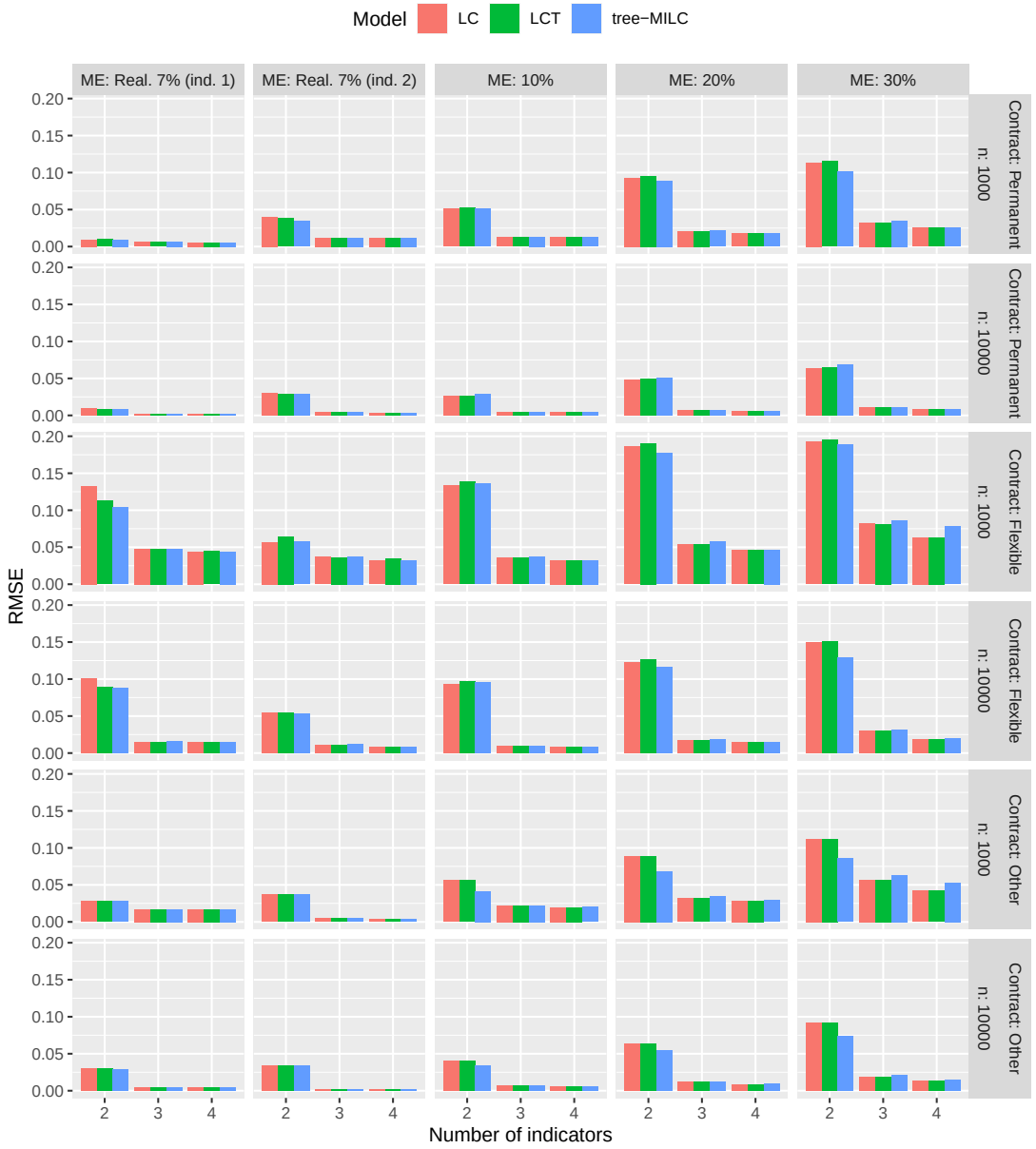
C.2 Measurement error probability estimators (MEPEs)

Figure C.3
Mean expected value of the measurement error (ME) probability estimators (MEPEs) in the first simulation study



Note. For estimators in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. For estimators in conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2'). The black lines show the true ME probabilities in the simulated data. The bias is given by the difference between the expected values and the true ME probabilities. The error bars show the standard deviations of the expected values.

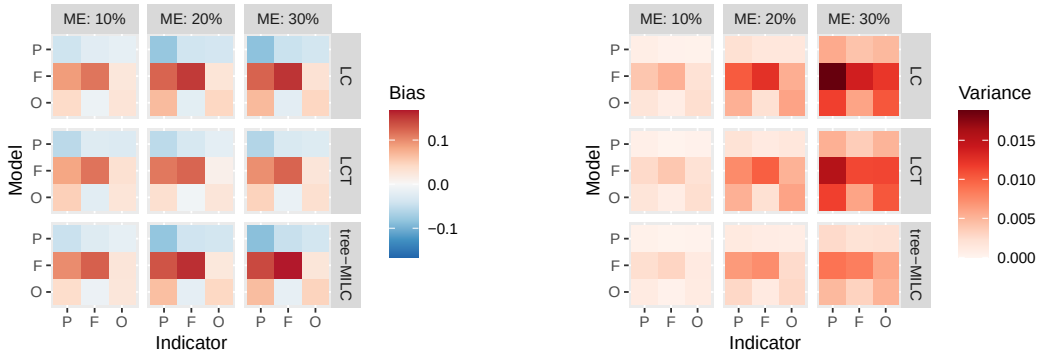
Figure C.4
Mean RMSE of the measurement error (ME) probability estimators (MEPEs) in the first simulation study



Note. For estimators in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. For estimators in conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2').

Figure C.5

Mean bias and mean variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a small sample size (n=1,000), 2 indicators and a covariate in the first simulation study

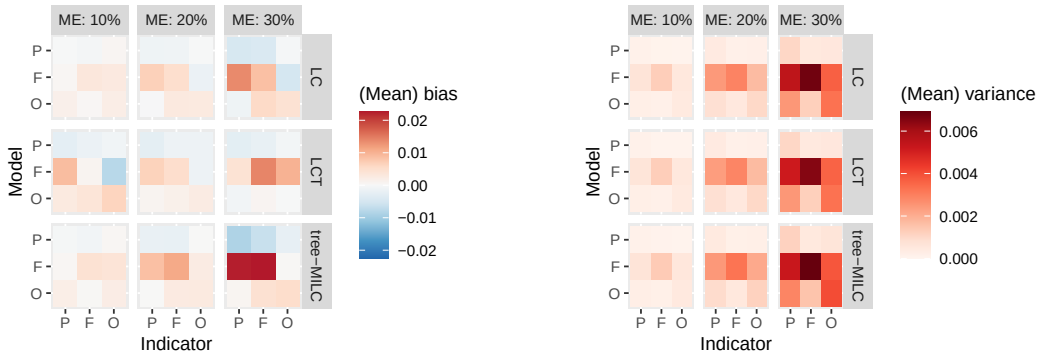


A. Bias **B**. Variance

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. Note that the bias (A) and the variance (B) were averaged over the two indicators. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure C.6

(Mean) bias and (mean) variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a small sample size (n=1,000), 3 indicators and a covariate in the first simulation study

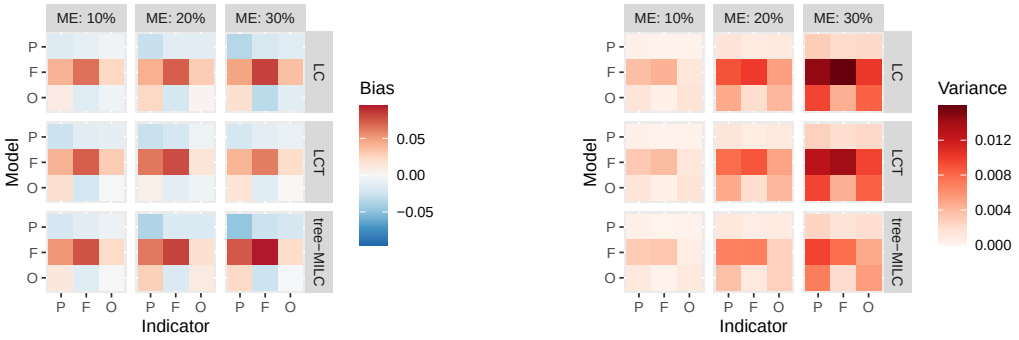


A. Bias **B**. Variance

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. Note that the bias (A) and the variance (B) were averaged over the three indicators. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure C.7

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a large sample size (n=10,000), 2 indicators and a covariate in the first simulation study

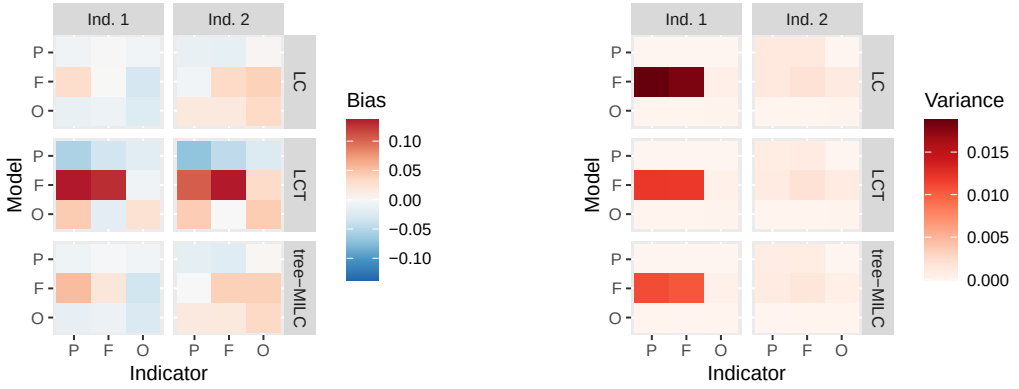


A. Bias **B**. Variance

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. Note that the bias (A) and the variance (B) were averaged over the two indicators. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure C.8

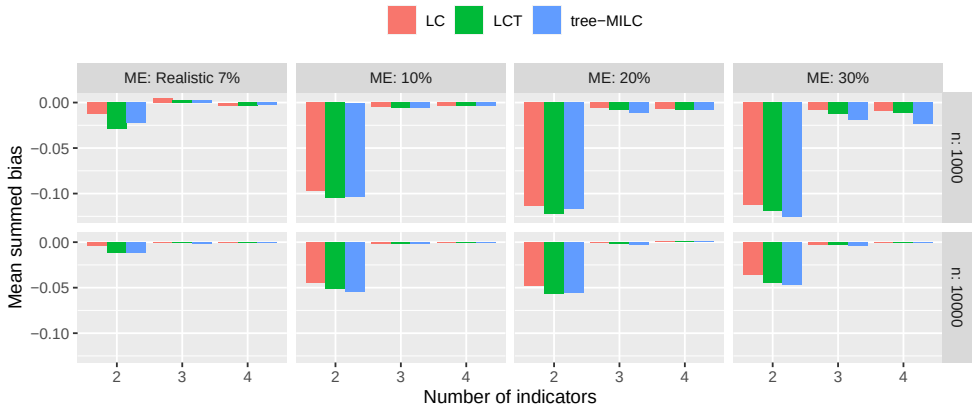
Mean bias and mean variance of measurement error (ME) probability estimators (MEPEs) in conditions with a realistic amount of ME, a small sample size (n=1,000), 2 indicators and a covariate in the first simulation study



A. Bias **B**. Variance

Note. Bias (A) and variance (B) for the indicators Y_1 (i.e. denoted by 'ind. 1') and Y_2 (i.e. denoted by 'ind. 2'). The rows represent the true contract types. The columns represent the observed contract types. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

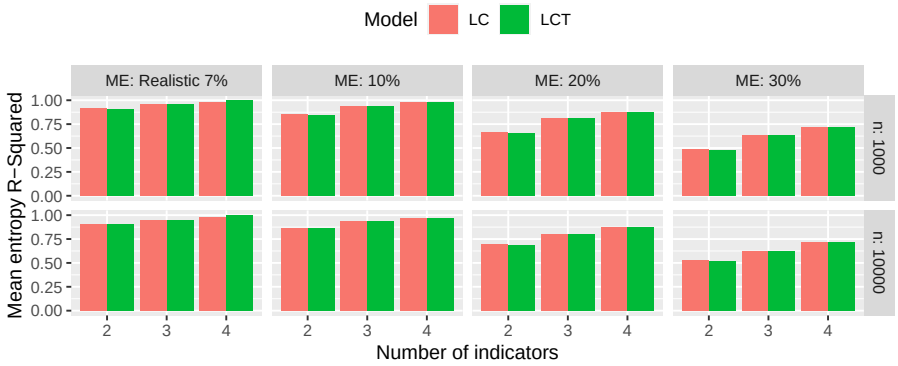
Figure C.9
Mean summed bias of estimators of the probability that a contract type was observed correctly in the first simulation study



Note. Mean summed bias of estimators of the probability that a permanent (P), flexible (F), or an 'other' (O) contract type was observed correctly (i.e. $P^{Y_j}_{\mathrm{P|P}}), P^{Y_j}_{\mathrm{F|F}}, P^{Y_j}_{\mathrm{O|O}}$) in conditions with different amounts of measurement error (ME), different numbers of indicators, and different sample sizes. The summed bias was averaged over the number of indicators. Negative values indicate that the amount of ME was overestimated (and vice versa).

C.3 Mean entropy R^2

Figure C.10
Mean entropy R^2 in the first simulation study



Note. ME = measurement error.

C.4 Supplement to Section 4.5.1: Comparing LC, LCT and tree-MILC analysis within simulation conditions

C.4.1 Comparing the PPEs

When comparing the PPEs of LC, LCT and tree-MILC analysis, only small differences were found. For example, overall, tree-MILC had the lowest RMSE in all conditions with a realistic 7% ME, and in all conditions with two indicators and a small sample size (see Figure C.1). Figure C.2 shows that this was because tree-MILC had the smallest variance. Of the other two methods, LCT had the largest bias and (thus) the largest RMSE (see Figures C.2 and C.1).

Likewise, LCT had the largest RMSE in conditions with two indicators and a large sample size (see Figure C.2). In these conditions, however, the performance of LC and tree-MILC differed per contract type. For example, in conditions with two indicators, a large sample size, and 30% ME, tree-MILC had the smallest RMSE for 'other' contracts, but the largest RMSE for permanent contracts (see Figure C.2). In the same conditions, however, LC had the largest RMSE for 'other' contracts, but the smallest RMSE for flexible contracts. Figure C.1 shows that these differences were mainly caused by differences in bias. Overall, it was concluded that in these conditions, LC and tree-MILC had the equally best performance.

Lastly, LC and LCT had the equally lowest RMSE in conditions with three or four indicators, a small sample size, and 30% ME (see Figure C.2). In these conditions, tree-MILC, on the other hand, had a slightly larger bias and (thus) a slightly larger RMSE (see Figures C.1 and C.2). In all other conditions, all three methods performed equally well in terms of RMSE (see Figure C.2).

C.4.2 Comparing the MEPEs

Slightly different results were found when comparing the MEPEs of LC, LCT and tree-MILC. For example, tree-MILC had the smallest variance and (thus) the lowest RMSE in all conditions with two indicators (see Figures C.3 and C.4). Nevertheless, in these conditions, LCT often had the largest RMSE (see Figure C.4).

Tree-MILC, on the other hand, had the largest RMSE in conditions with three or four indicators, a small sample size, and 20% or 30% ME (see Figure C.4). Figure C.3 shows that this was because tree-MILC often had a (slightly) larger variance. In contrast, in these conditions, LC and LCT performed equally well in terms of RMSE (see Figure C.4). No differences in RMSE were observed between the three methods in all other conditions (see Figure C.4).

Appendix D

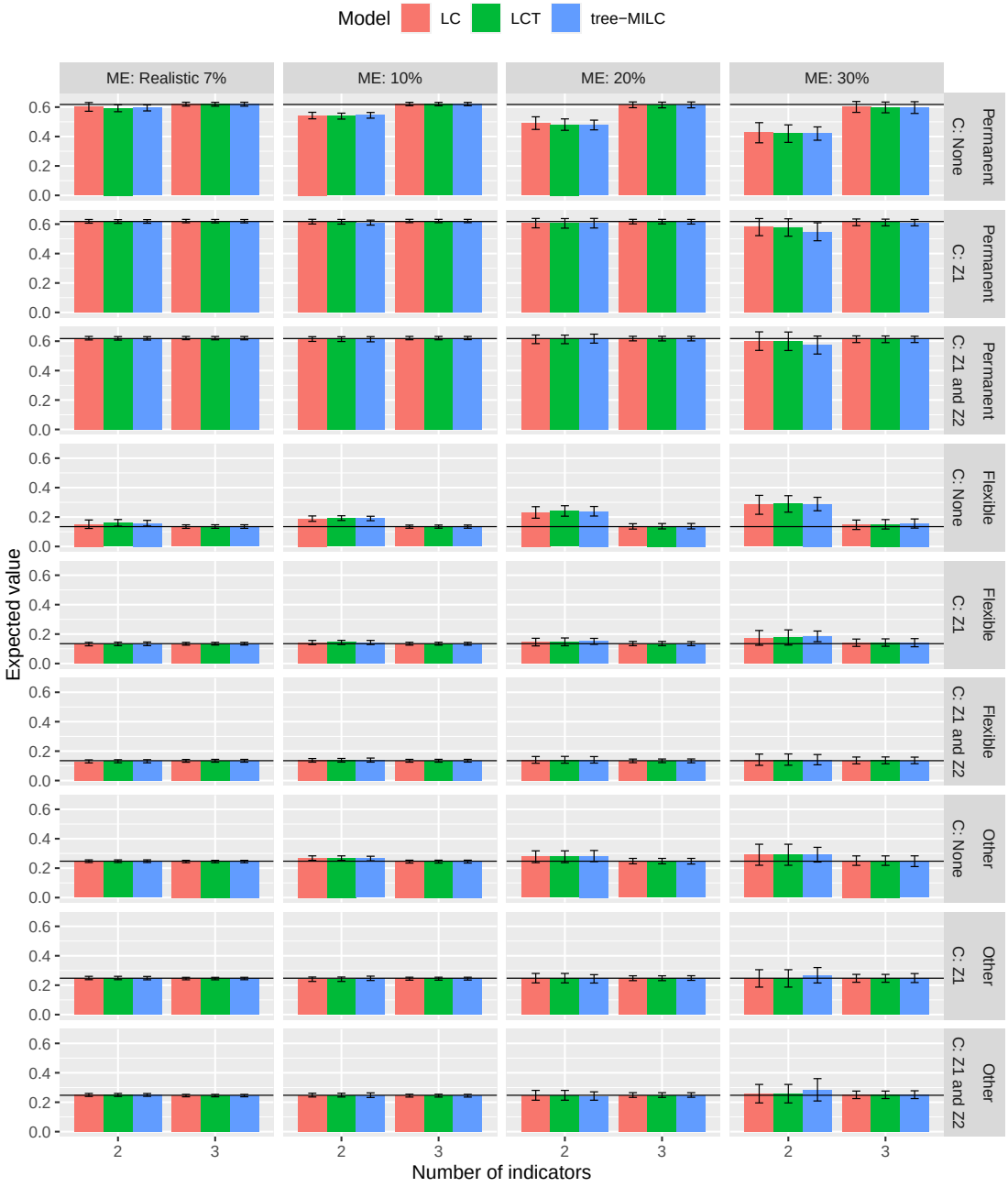
Results of simulation study 2

D.1 Population proportion estimators (PPEs) (n=1,000)

See next page.

Figure D.1

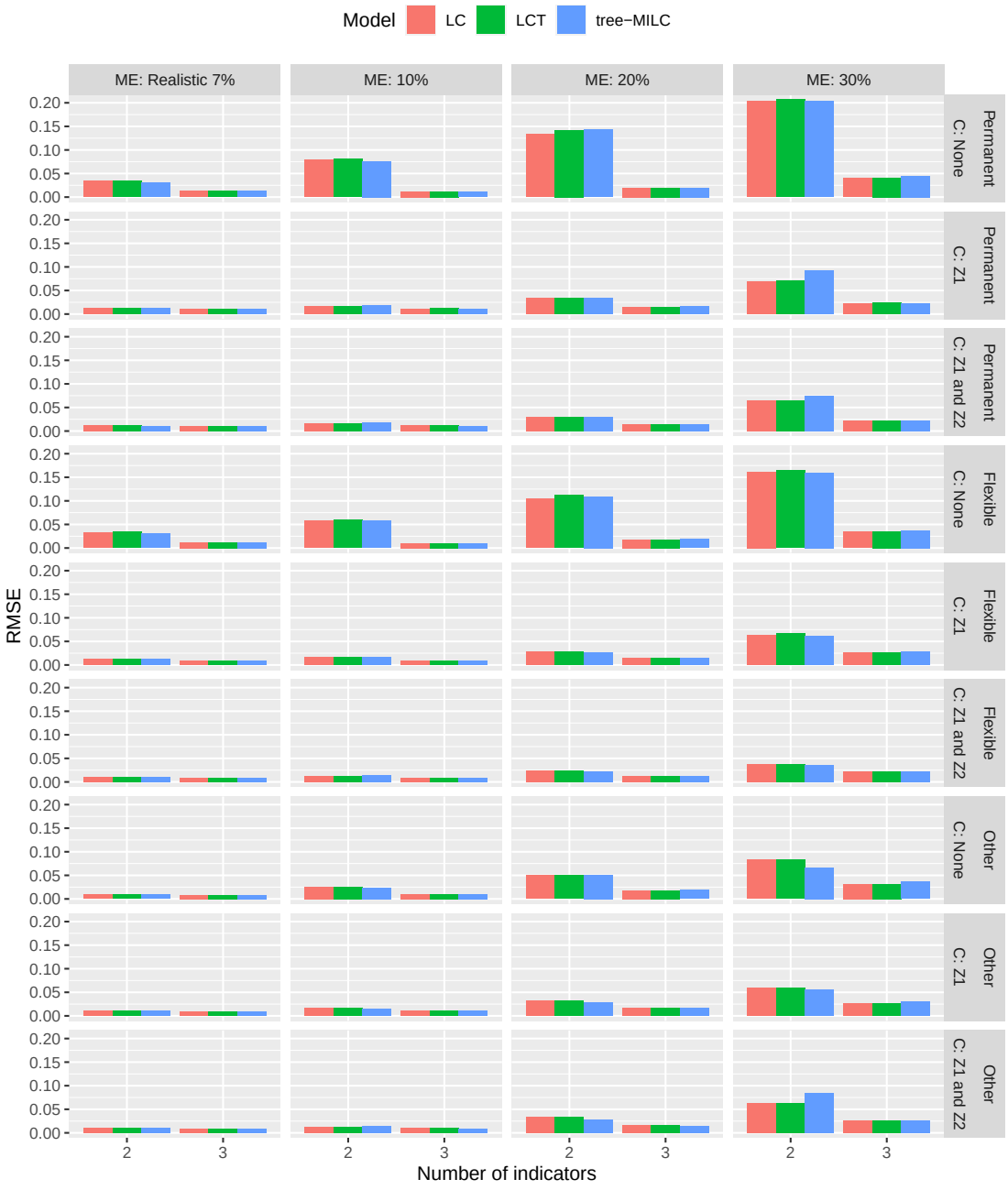
Expected value of population proportion estimators (PPEs) in conditions with a small sample size (n=1,000) in the second simulation study



Note. ME = measurement error; C = missing covariates. The black lines show the true proportions in the simulated data. The bias is given by the difference between the expected values and the true proportions. The error bars show the standard deviation of the expected values.

Figure D.2

RMSE of population proportion estimators (PPEs) in conditions with a small sample size (n=1,000) in the second simulation study

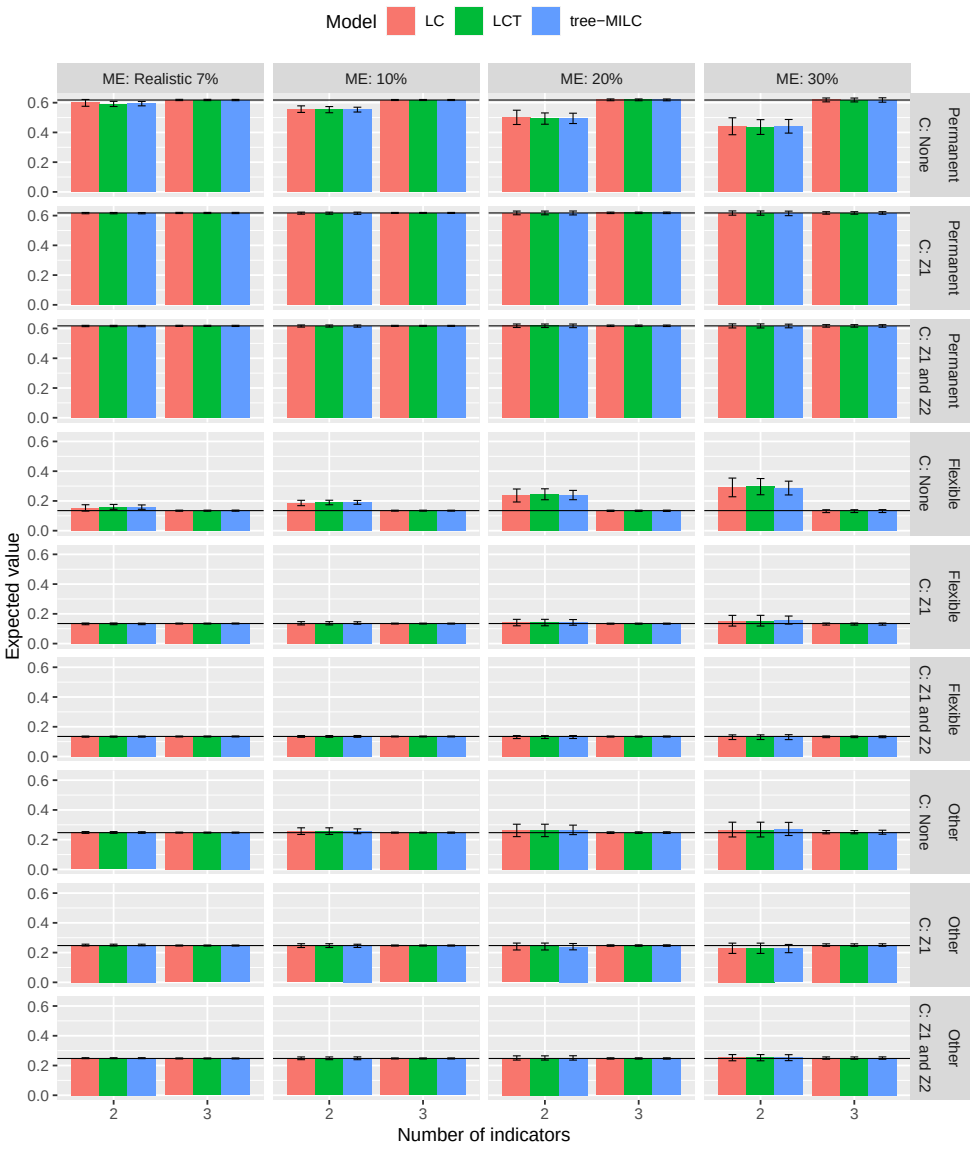


Note. ME = measurement error; C = missing covariates.

D.2 Population proportion estimators (PPEs) (n=10,000)

Figure D.3

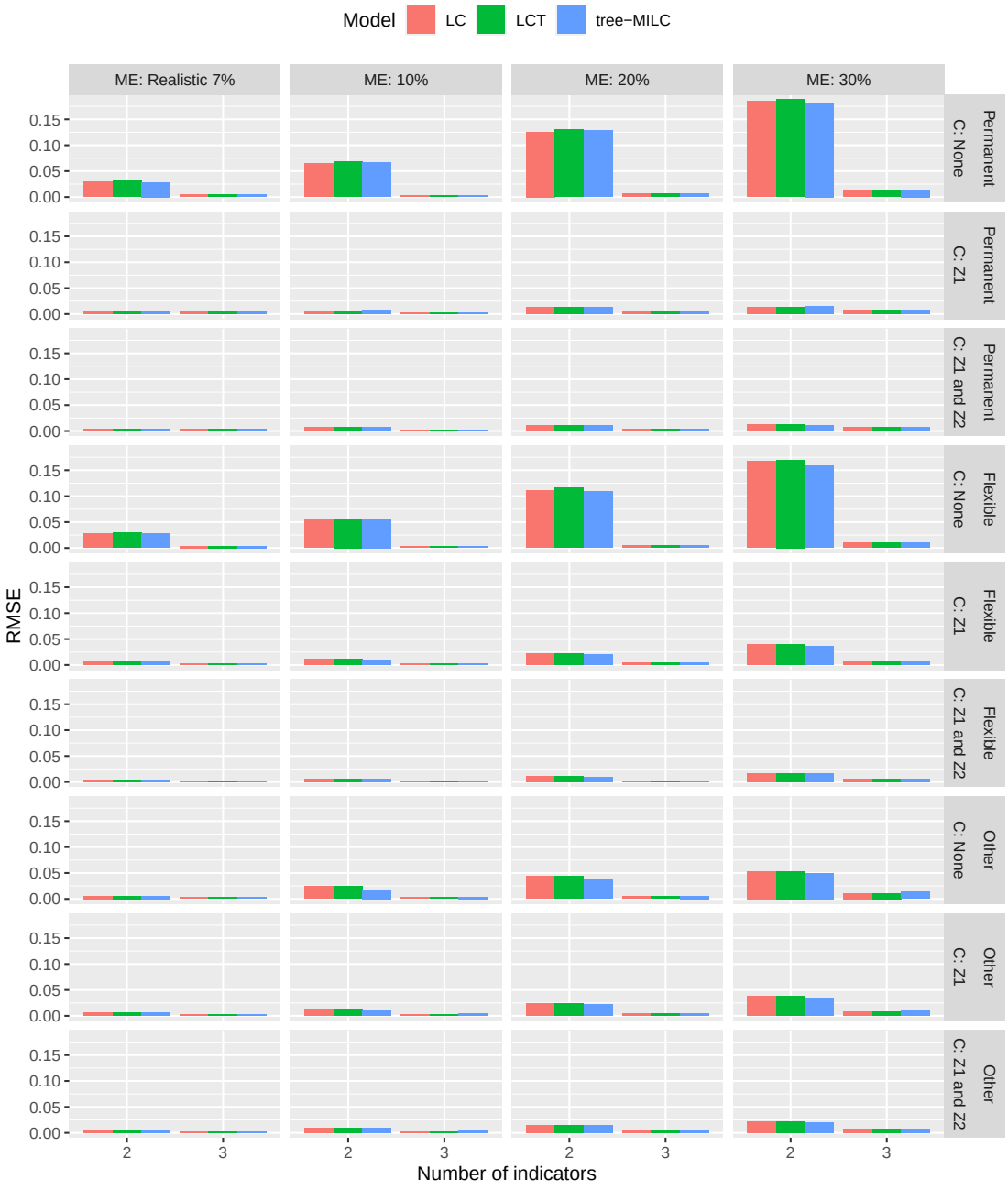
Expected value of population proportion estimators (PPEs) in conditions with a large sample size (n=10,000) in the second simulation study



Note. ME = measurement error; C = missing covariates. The black lines show the true proportions in the simulated data. The bias is given by the difference between the expected values and the true proportions. The error bars show the standard deviation of the expected values.

Figure D.4

RMSE of population proportion estimators (PPEs) in conditions with a large sample size (n=10,000) in the second simulation study



Note. ME = measurement error; C = missing covariates.

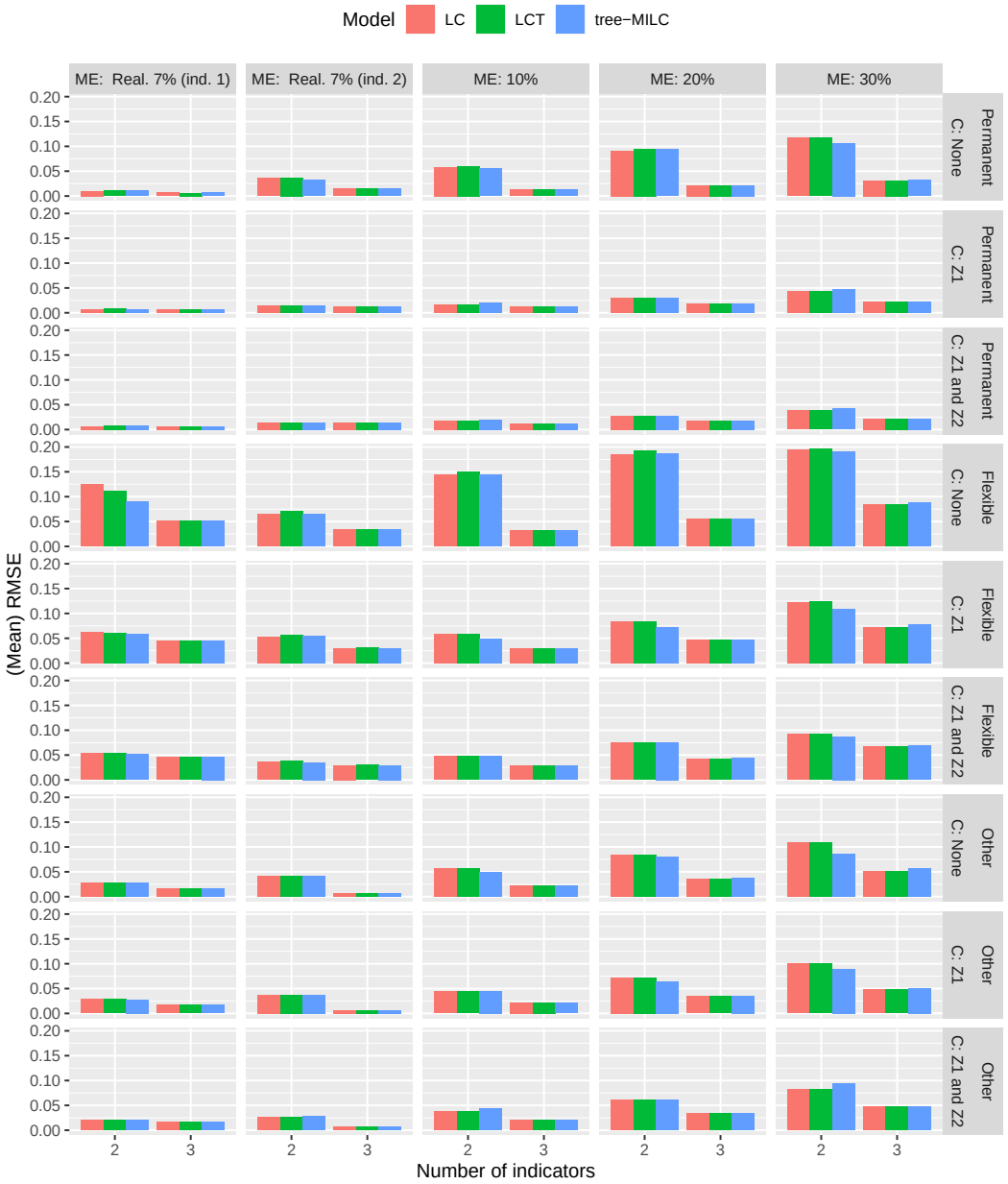
D.3 Measurement error probability estimators (MEPEs) (n=1,000)

Figure D.5

Expected value of measurement error (ME) probability estimators (MEPEs) in conditions with a small sample size (n=1,000) in the second simulation study

Note. ME = measurement error; C = missing covariates. Note that in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. In conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2'). The black lines show the true ME probabilities in the simulated data. The bias is given by the difference between the expected values and the true ME probabilities. The error bars show the standard deviation of the mean expected values.

Figure D.6
RMSE of measurement error (ME) probability estimators (MEPEs) in conditions with a small sample size (n=1,000) in the second simulation study



Note. ME = measurement error; C = missing covariates. Note that in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. In conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2').

D.4 Measurement error probability estimators (MEPEs) (n=10,000)

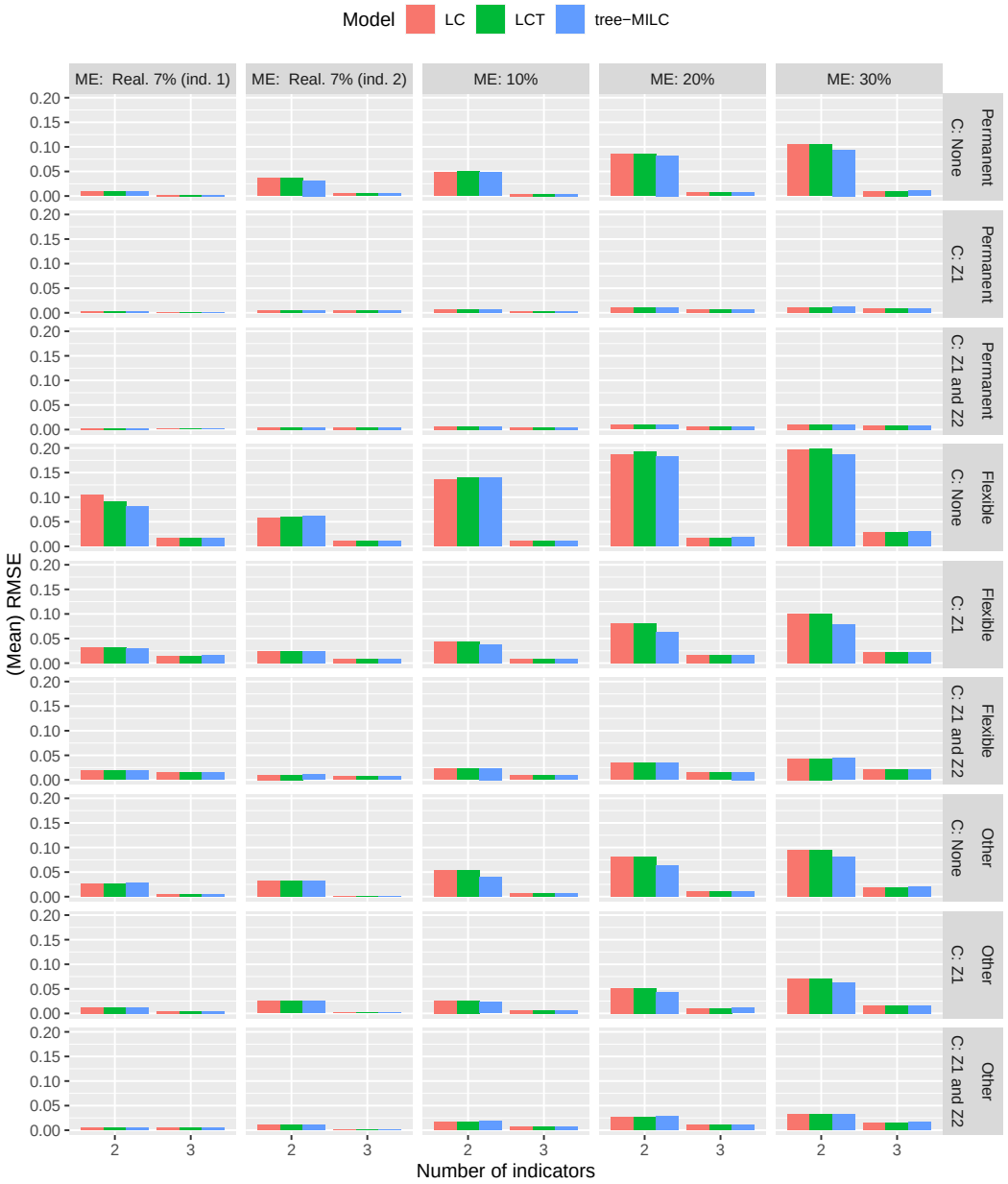
Figure D.7

Expected value of measurement error (ME) probability estimators (MEPEs) in conditions with a large sample size (n=10,000) in the second simulation study

Note. ME = measurement error; C = missing covariates. Note that in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. In conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2'). The black lines show the true ME probabilities in the simulated data. The bias is given by the difference between the expected values and the true ME probabilities. The error bars show the standard deviation of the mean expected values.

Figure D.8

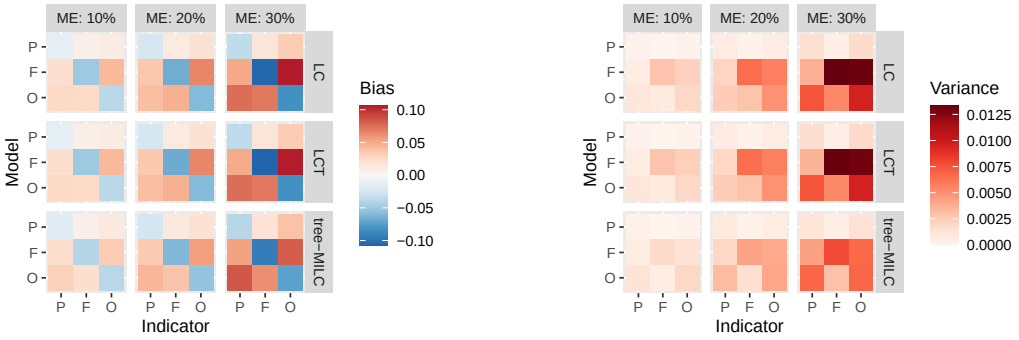
RMSE of measurement error (ME) probability estimators (MEPEs) in conditions with a small sample size (n=1,000) in the second simulation study



Note. ME = measurement error; C = missing covariates. Note that for estimators in conditions with 10%, 20% and 30% ME, the expected values and RMSE were averaged over the J indicators. For estimators in conditions with a realistic 7% amount of ME, the expected values and RMSE were averaged over the indicators Y_1 and Y_3 (i.e. if $J > 2$) (denoted by 'ind. 1'), and over the indicators Y_2 and Y_4 (i.e. if $J = 4$) (denoted by 'ind. 2').

Figure D.9

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a small sample size (n=1,000), 2 indicators and the covariates Q and Z_1 in the second simulation study

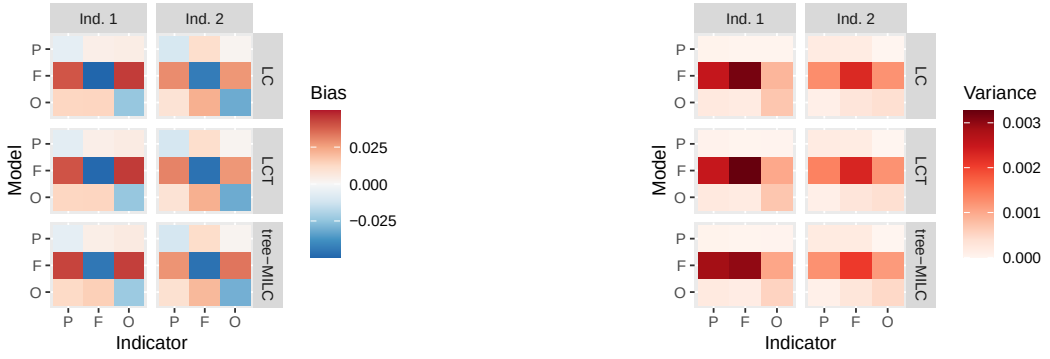


A. Bias **B**. Variance

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. Note that the bias (A) and the variance (B) were averaged over the two indicators. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure D.10

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with a realistic 7% ME, a small sample size (n=1,000), 2 indicators and the covariates Q and Z_1 in the second simulation study

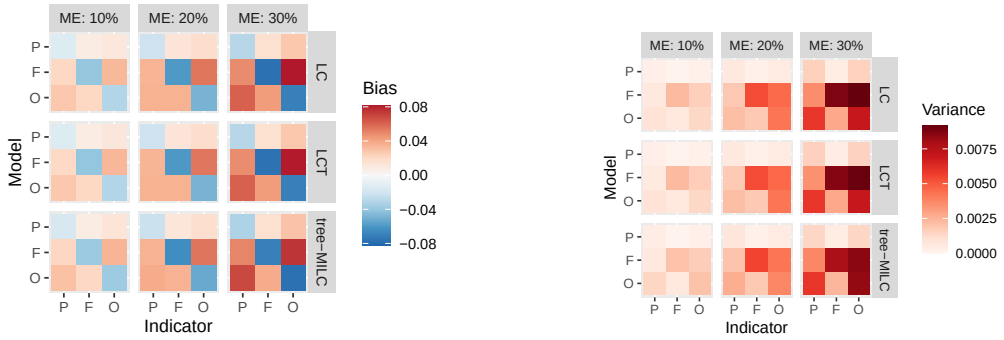


A. Bias **B**. Variance

Note. Bias (A) and variance (B) for the indicators Y_1 (i.e. denoted by 'ind. 1') and Y_2 (i.e. denoted by 'ind. 2'). The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure D.11

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a small sample size (n=1,000), 2 indicators and the covariates Q, Z_1 and Z_2 in the second simulation study

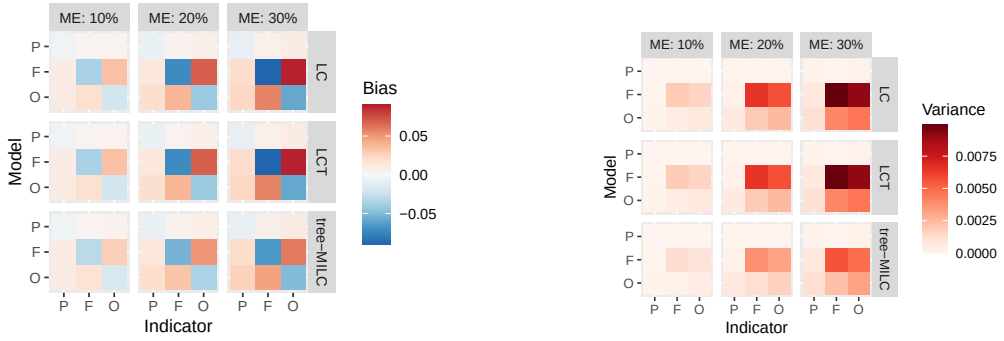


A. Bias **B**. Variance

Note. Bias (A) and variance (B) for the indicators Y_1 (i.e. denoted by 'ind. 1') and Y_2 (i.e. denoted by 'ind. 2'). The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure D.12

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with 10%, 20% and 30% ME, a large sample size (n=10,000), 2 indicators and the covariates Q and Z_1 in the second simulation study

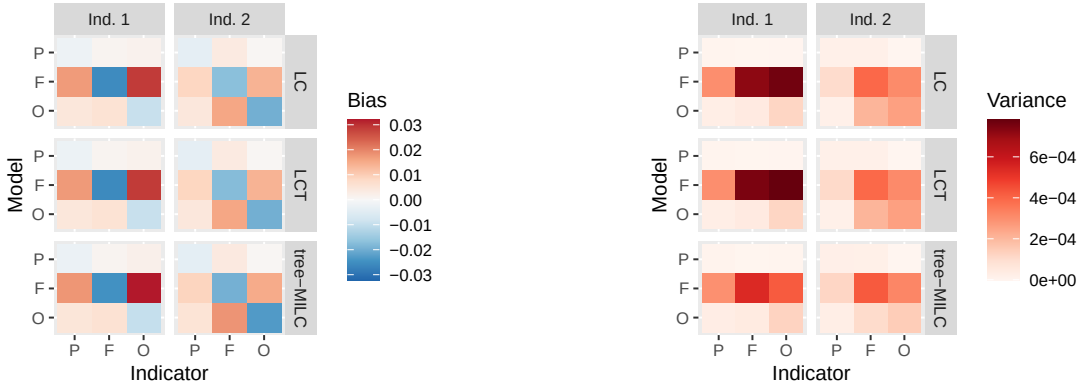


A. Bias **B**. Variance

Note. The rows represent the true contract types 'permanent' (P), 'flexible' (F) and 'other' (O). The columns represent the observed contract types. Note that the bias (A) and the variance (B) were averaged over the two indicators. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by −1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure D.13

Bias and variance of measurement error (ME) probability estimators (MEPEs) in conditions with a realistic 7% ME, a large sample size (n=10,000), 2 indicators and the covariates Q and Z_1 in the second simulation study

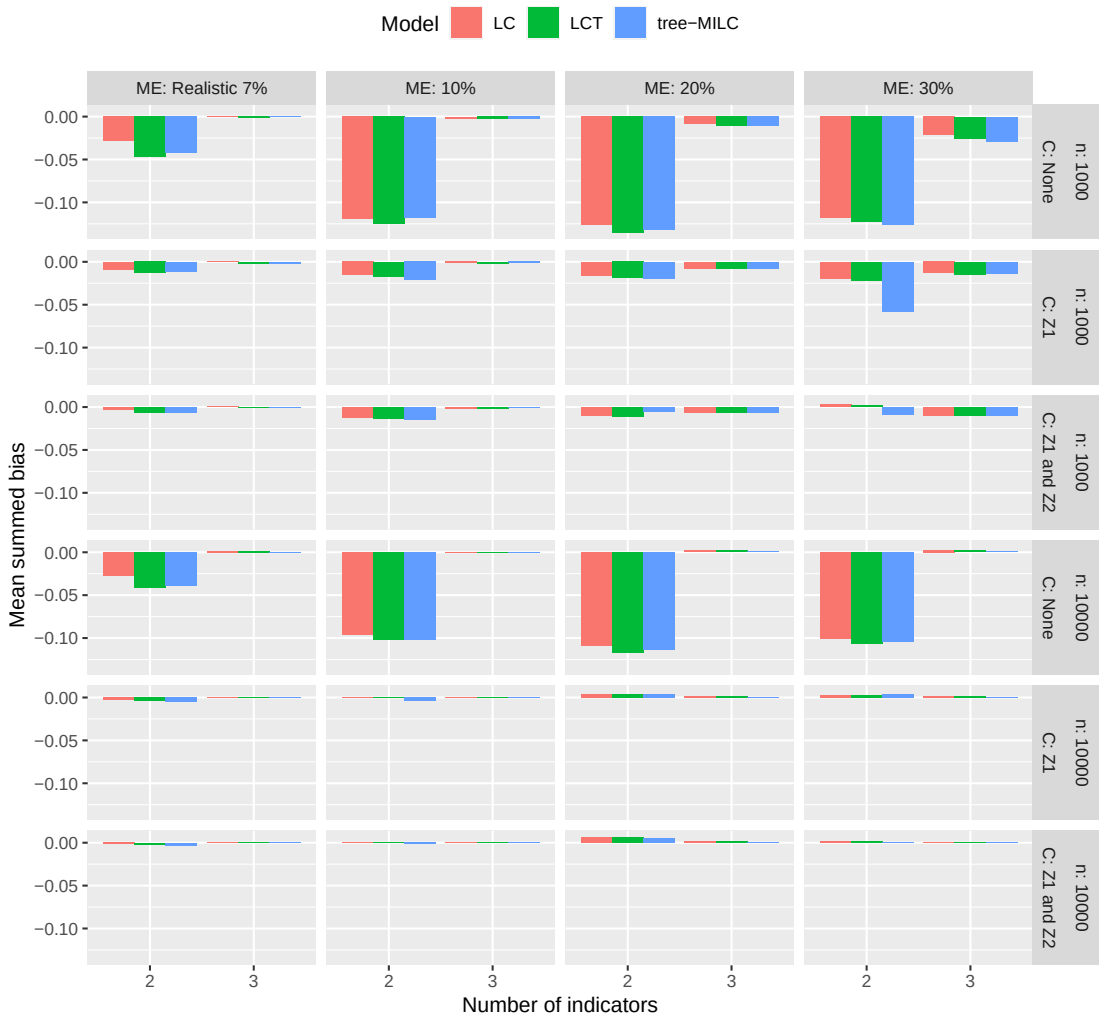


A. Bias **B**. Variance

Note. Bias (A) and variance (B) for the indicators Y_1 (i.e. denoted by 'ind. 1') and Y_2 (i.e. denoted by 'ind. 2'). The rows represent the true contract types. The columns represent the observed contract types. To ensure consistent interpretation, the bias of MEPEs on the diagonal was multiplied by -1. A positive bias means that the amount of ME was overestimated, whereas a negative bias means that the amount of ME was overestimated.

Figure D.14

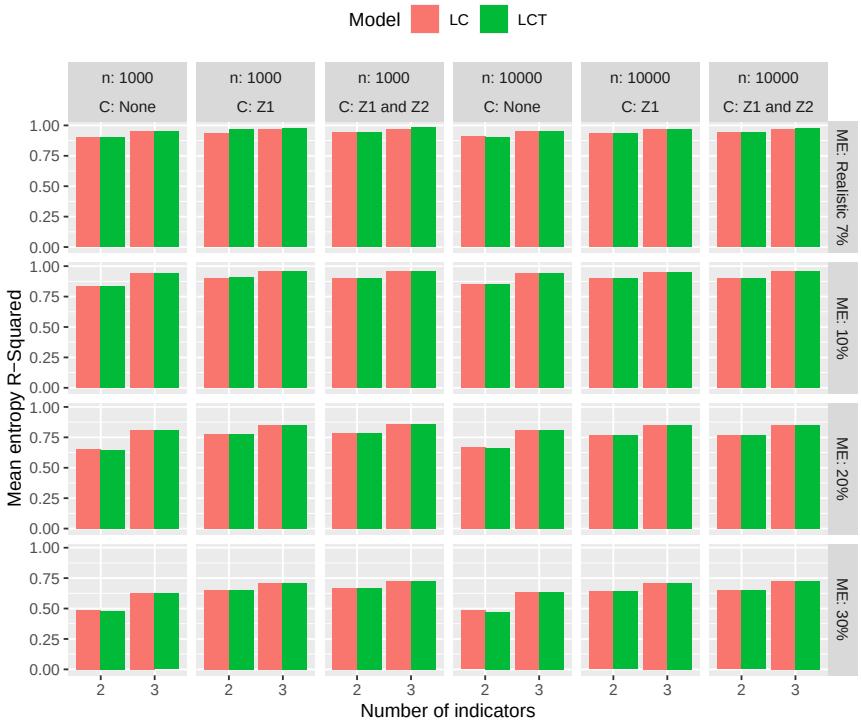
Mean summed bias of estimators of the probability that a contract type was observed correctly $(P^{Y_j}_{P|P}, P^{Y_j}_{F|F}, P^{Y_j}_{O|O})$ *in the second simulation study*



Note. ME = measurement error, C = missing covariates, P = permanent, F = flexible, O = other. The summed bias was averaged over the number of indicators. Negative values indicate that the amount of ME was overestimated.

D.5 Mean entropy R^2

Figure D.15
Mean entropy R^2 for LC and LCT models in the second simulation study



Note. ME = measurement error; C = missing covariates.

D.6 Supplement to Section 5.6.1: Comparing LC, LCT and tree-MILC analysis within simulation conditions

D.6.1 Comparing PPEs

When comparing the PPEs of LC, LCT and tree-MILC within the simulation conditions, only small differences were observed in conditions with two indicators and 30% ME. For example, in conditions with two indicators, a small sample size, and 30% ME, LC had the smallest RMSE when one (Z_1) or two missing covariates (Z_1 and Z_2) were included (see Figures D.2 and D.4). In contrast, tree-MILC and LCT had (respectively) a larger bias and a larger variance (see Figures D.1 and D.3). Of the latter two methods, tree-MILC had the largest RMSE. Nevertheless, in conditions with two indicators, a large sample size, and 30% ME, tree-MILC had the smallest RMSE when one missing covariate (Z_1) was included (see Figures D.1 and D.3). Figures D.2 and D.4 show that this was because tree-MILC had the smallest variance. The other two methods, on the other hand, performed equally well. In all other conditions, no differences between the three methods were observed.

D.6.2 Comparing the MEPEs

In contrast, slightly different results were found when comparing the MEPEs of LC, LCT and tree-MILC. For example, tree-MILC had the smallest RMSE in all conditions with one missing covariate (Z_1), and 10%, 20% or 30% ME (both sample sizes) (see Figures D.2 and D.4). Figures D.1 and D.3 show that this was because tree-MILC had a smaller variance. Nevertheless, in conditions with two missing covariates (Z_1 and Z_2) and a small sample size, tree-MILC had a larger RMSE as compared to the other two equally performing methods (see Figures D.2 and D.4). Figures D.1 and D.3 show that in these conditions, tree-MILC had a larger bias. In all other conditions, no differences were observed between the three methods.

Appendix E

Covariates in real linked data from the ER and the LFS

The real linked data from the ER and the LFS contained nine categorical covariates that were strongly associated with the inconsistencies between the ER and the LFS (see Chapter 6). A short description of all covariates and their categories is provided below. In addition, it is described how the covariates were recoded to simulate data in the second simulation study.

E.1 Description of all covariates

Table E.1 shows the covariates that were available for all respondents. Table E.2 shows the missing covariates as described in Section 5.1. The categories in bold in Table E.2 represent the extra categories to which respondents with missing covariates were assigned in the additional analyses in Section 6.4. Note that in these additional analyses, the covariate *number of contract hours* was added in its original form with five categories.

Table E.1
Overview of non-missing covariates in the real linked data from the ER and the LFS

Covariate	Description	Categories
Gender	Gender of the respondent	1: Male 2: Female
Education level	Highest education level of the respondent	1: Low 2: Middle 3: High 4: Unknown
Interview manner	The person who filled in the survey	1: The respondent 2: Someone else
Migration background	Migration background of the respondent	1: The Netherlands 2: Morrocco 3: Turkey 4: Suriname 5: Netherlands Antilles 6: Other non-Western countries 7: Other Western countries

Table E.2
Overview of missing covariates in the real linked data from the ER and the LFS

Covariate	Description	Categories	
Number of contract hours	Number of contract hours of the respondent	1: <12 hours 2: 12-20 hours 3: 20-30 hours	4: >30 hours 5: Missing
Company size	Size of the company the respondent is employed at	1: Large 2: Middle	**3: Small** 4: Missing
Main economic activity of the employer	Economic sector of the company at which the respondent is employed	**1: Other** 2: Agriculture, forestry and fishing 3: Retail 4: Water and air transportation 5: Hospitality	6: Finance 7: Job placement, employment agencies, and human resources management 8: Government and education 9: Health care 10: Missing
Job duration	Number of months for which the respondent has been employed	**1: < 3 months** 2: 3-6 months 3: 6-12 months 4: 12-24 months	5: 24-36 months 6: >36 months 7: Missing
Software cluster	Type of software used to register the respondent in the ER	1: Cluster 1 2: Cluster 2 3: Cluster 3	4: Cluster 4 **5: Cluster 5** 6: Missing

Note. The categories in bold represent the categories to which respondents with missing covariates were assigned in the additional analyses in Section 6.4.

E.2 Recoding the covariates in simulation study 2

The missing covariates *main economic activity of the employer* and *job duration* were represented by the covariates Z_1 and Z_2 in the second simulation study (see Chapter 5 and Section E.1). The covariate *main economic activity of the employer* originally consisted of ten categories (see Table E.2). However, in Section 5.2, the number of categories was reduced to two (see Table E.3). The covariate was recoded in such a way that the strength of association between the covariate and the observed contract types was maintained. Similarly, the covariate *job duration* was recoded to two categories. Of this covariate, the first category consisted of observations with missing observations and observations who were employed for less than 36 months, whereas the second category consisted of observations who were employed for more than 36 months.

Table E.3
Recoded categories of the covariate 'main economic activity of the employer'

Recoded categories			
1:	Other Finance Water and air transportation Government and education Health care Missing	2:	Agriculture, forestry, and fishing Retail Hospitality Job placement, employment agencies and human resources management

Appendix F

Results of analyses of real linked data from the ER and the LFS

F.1 Measurement error probability estimates

Table F.1
Measurement error probabilities in the first quarter of 2016 as estimated by LC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9376	0.0622	0.0002	0.9877	0.0055	0.0068
F	0.1179	0.8814	0.0008	0.2902	0.6330	0.0768
O	0.0160	0.0000	0.9839	0.0318	0.0288	0.9394

Note. P = Permanent, F = Flexible, O = Other.

Table F.2
Measurement error probabilities in the first quarter of 2016 as estimated by tree-MILC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9027	0.0973	0.0000	0.9664	0.0242	0.0094
F	0.0794	0.9189	0.0016	0.2180	0.7036	0.0784
O	0.0153	0.0002	0.9845	0.0328	0.0304	0.9367

Note. P = Permanent, F = Flexible, O = Other.

Table F.3
Measurement error probabilities in the first quarter of 2017 as estimated by LC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9339	0.0660	0.0001	0.9809	0.0096	0.0094
F	0.2896	0.9104	0.0000	0.2491	0.6762	0.0747
O	0.0121	0.0007	0.9871	0.0345	0.0287	0.9368

Note. P = Permanent, F = Flexible, O = Other.

Table F.4
Measurement error probabilities in the first quarter of 2017 as estimated by tree-MILC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9253	0.0742	0.0005	0.9746	0.0142	0.0112
F	0.0886	0.9074	0.0041	0.2300	0.6916	0.0784
O	0.0096	0.0007	0.9897	0.0326	0.0279	0.9396

Note. P = Permanent, F = Flexible, O = Other.

Table F.5
Measurement error probabilities in the first quarter of 2018 as estimated by LC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9442	0.0553	0.0005	0.9809	0.0100	0.0091
F	0.0883	0.9109	0.0008	0.2713	0.6604	0.0683
O	0.0170	0.0019	0.9810	0.0349	0.0293	0.9358

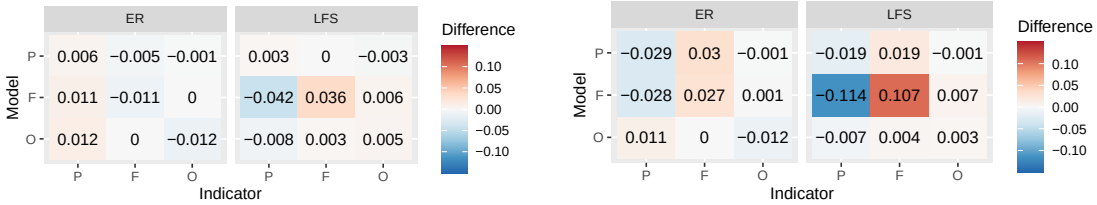
Note. P = Permanent, F = Flexible, O = Other.

Table F.6
Measurement error probabilities in the first quarter of 2018 as estimated by tree-MILC

	ER			LFS		
Model	**P**	**F**	**O**	**P**	**F**	**O**
P	0.9284	0.0707	0.0009	0.9734	0.0160	0.0106
F	0.0770	0.9210	0.0020	0.2403	0.6867	0.0731
O	0.0172	0.0028	0.9800	0.0348	0.0293	0.9359

Note. P = Permanent, F = Flexible, O = Other.

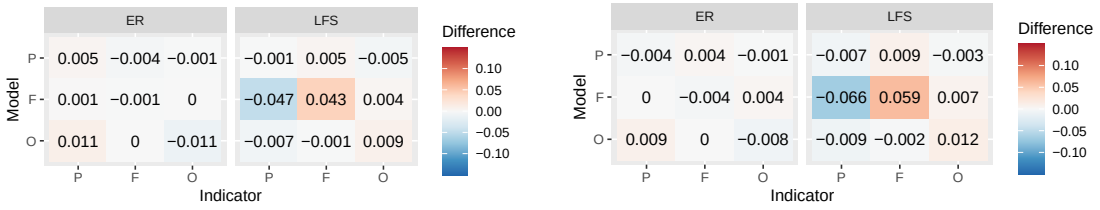
Figure F.1
Differences in ME probability estimates in the first quarter of 2016 for LC and tree-MILC models



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

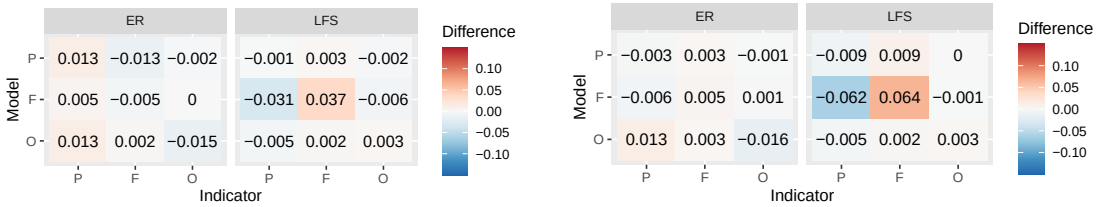
Figure F.2
Differences in ME probability estimates in the first quarter of 2017 for LC and tree-MILC models



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

Figure F.3
Differences in ME probability estimates in the first quarter of 2018 for LC and tree-MILC models



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

F.2 Results of additional analyses

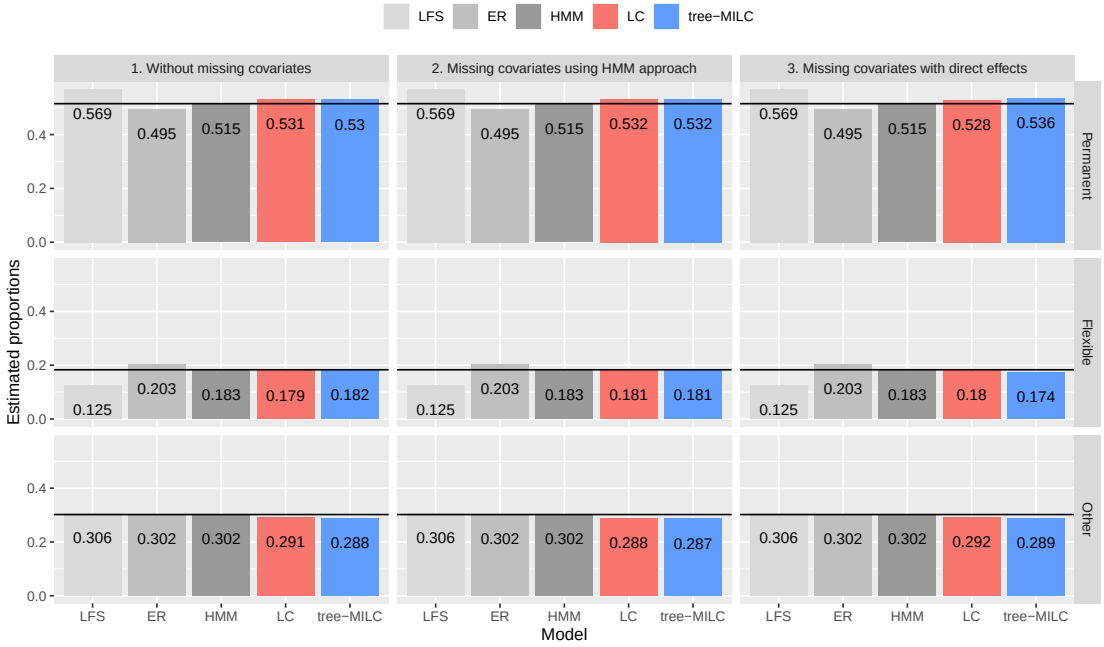
Table F.7
Entropy R^2 for LC models in the first quarters of 2016, 2017 and 2018

	Missing covariates with direct effects			Missing covariates using HMM approach			Without missing covariates		
	2016	**2017**	**2018**	**2016**	**2017**	**2018**	**2016**	**2017**	**2018**
Entropy R^2	0.9402	0.9455	0.9440	0.9428	0.9479	0.9473	0.8845	0.8926	0.8853

Table F.8
Variance of the pooled estimates for the population proportion estimates for tree-MILC models in the first quarters of 2016, 2017 and 2018

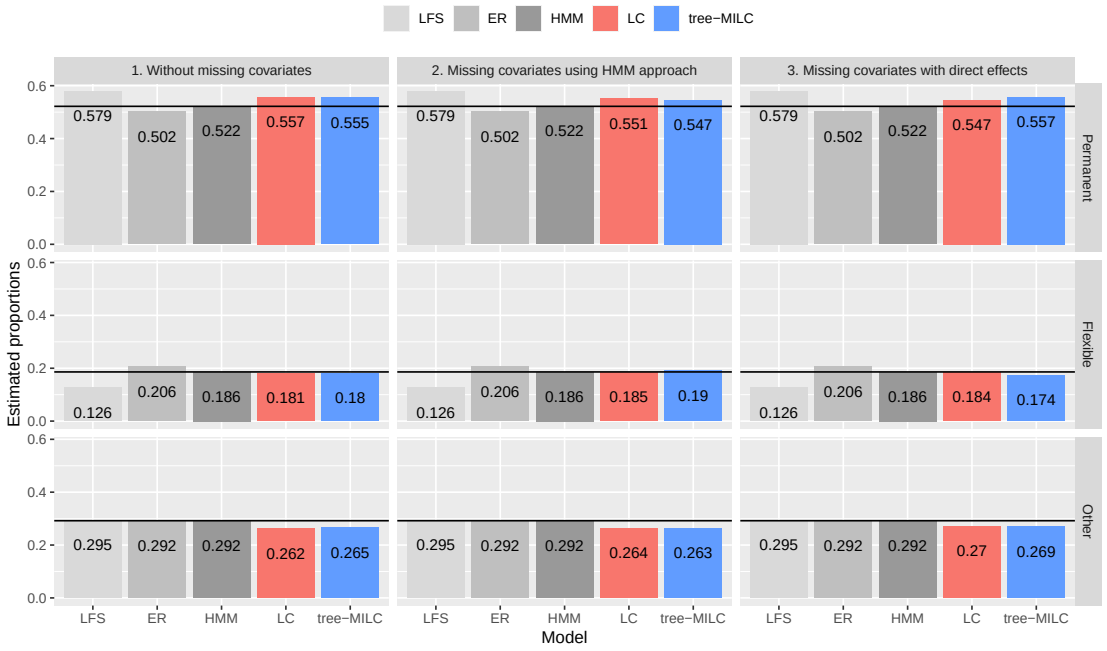
	Missing covariates with direct effects			Missing covariates using HMM approach			Without missing covariates		
	2016	**2017**	**2018**	**2016**	**2017**	**2018**	**2016**	**2017**	**2018**
Permanent	0.0200	0.0223	0.0216	0.0079	0.0095	0.0094	0.0246	0.0107	0.0117
Flexible	0.0198	0.0191	0.0238	0.0072	0.0061	0.0051	0.0213	0.0155	0.0093
Other	0.0074	0.0056	0.0082	0.0073	0.0061	0.0087	0.0094	0.0086	0.0114

Figure F.4
Estimated proportions per contract type in the first quarter of 2017



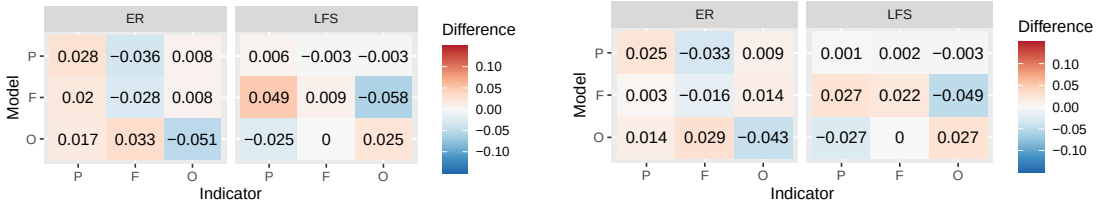
Note. This figure shows the LC, the tree-MILC and the HMM estimates. In addition, the original Labour Force Survey (LFS) and Employment Register (ER) estimates are shown. To enhance comparison, the HMM estimates are denoted the black lines.

Figure F.5
Estimated proportions per contract type in the first quarter of 2018



Note. This figure shows the LC, the tree-MILC and the HMM estimates. In addition, the original Labour Force Survey (LFS) and Employment Register (ER) estimates are shown. To enhance comparison, the HMM estimates are denoted the black lines.

Figure F.6
Differences in ME probability estimates in the first quarter of 2016 for LC and tree-MILC models without missing covariates

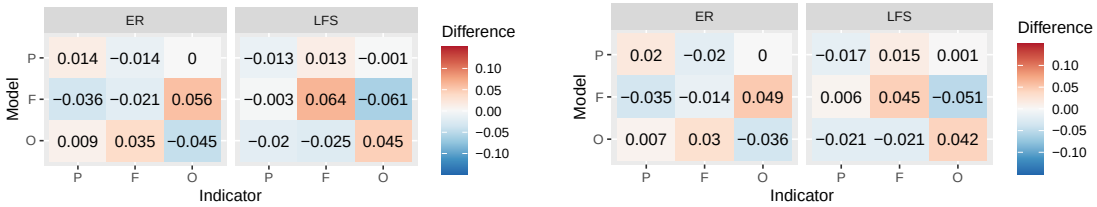


A. Differences between LC and the HMM

B. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

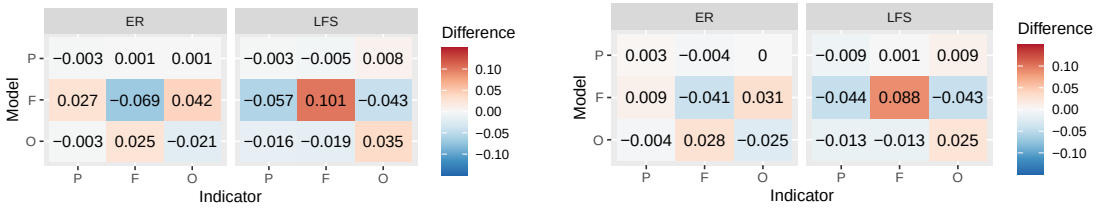
Figure F.7
Differences in ME probability estimates in the first quarter of 2017 for LC and tree-MILC models without missing covariates



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

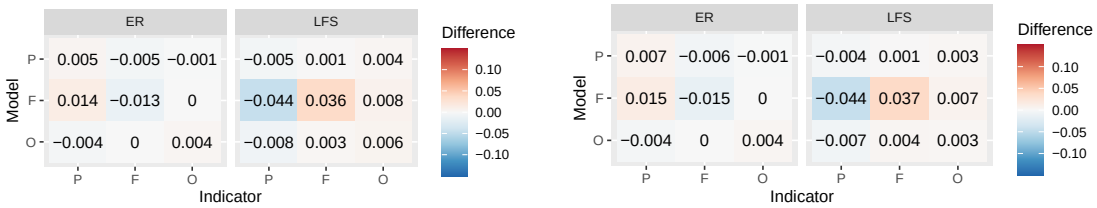
Figure F.8
Differences in ME probability estimates in the first quarter of 2018 for LC and tree-MILC models without missing covariates



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

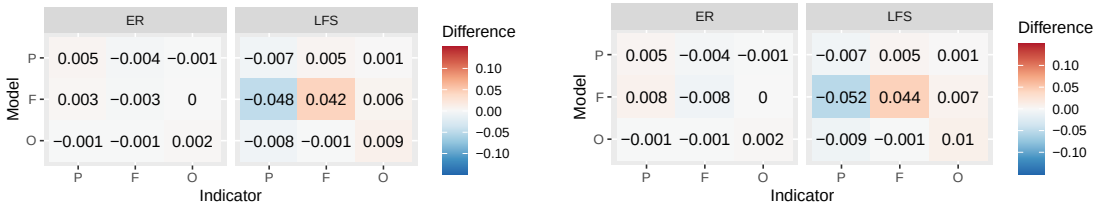
Figure F.9
Differences in ME probability estimates in the first quarter of 2016 for LC and tree-MILC models with missing covariates included using the HMM approach



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

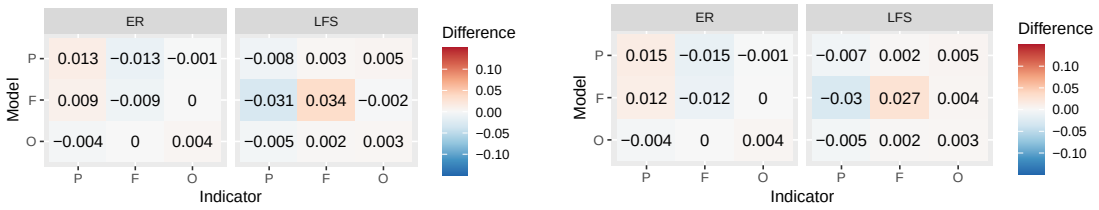
Figure F.10
Differences in ME probability estimates in the first quarter of 2017 for LC and tree-MILC models with missing covariates included using the HMM approach



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

Figure F.11
Differences in ME probability estimates in the first quarter of 2018 for LC and tree-MILC models with missing covariates included using the HMM approach



A. Differences between LC and the HMM **B**. Differences between tree-MILC and the HMM

Note. A negative value indicates that the LC or the tree-MILC model estimated a probability as lower than the HMM (and vice versa).

Appendix G

GitHub repository

The R code used for this thesis is available at https://github.com/eliscamastenbroek/master_thesis. In the README.md file of this repository, instructions to reproduce the analyses can be found. Note that the real data from the ER and the LFS (Chapter 6) is not publicly available.