Universiteit Leiden
The Netherlands

Do Employment Register and Labour Force Survey measure the same employment contract type concept?

Frederick Goldwyn Restrepo Estrada

Primary thesis advisors:
Dr. Sander Scholtus & Prof. Dr. Bart Bakker (CBS)

Secondary thesis advisor:
Dr. Zsuzsa Bakk (Universiteit Leiden)

Defended on January 18, 2023

MASTER THESIS

STATISTICS AND DATA SCIENCE

UNIVERSITEIT LEIDEN

Disclaimer

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

Acknowledgements

I would like to thank Sander Scholtus, Bart Bakker, and Zsuzsa Bakk for their insights, advice, and guidance throughout the project. Additionally, I would like to thank Sander Scholtus and Bart Bakker for their guidance at Statistics Netherlands. Furthermore, I would like to thank Sander Scholtus and Zsuzsa Bakk for their feedback on earlier versions of this report. Lastly, I would like to thank Zsuzsa Bakk for their guidance on working with Latent GOLD and I would like to thank Jeroen Vermunt for their comments on the inner workings of the program.

Contents

1 Abstract

Employment contract types are often distinguished in permanent, flexible, and other types of contracts. Accurate estimates of their frequencies are valuable for socio-economic research and legislative purposes. In member states of the European Union, the Labour Force Survey (LFS) is used to acquire such estimates. In the Netherlands, in addition to the LFS, the Employment Register (ER) is available with which employment contract type frequencies can be estimated. Estimates based on the two indicators are known to differ substantially and consistently. Studies have found several plausible contributing factors for the inconsistencies. However, when taken into account (excluding measurement error), a substantive part of the inconsistencies remains unexplained. The true employment contract type can be regarded as a latent variable of which the ER and the LFS are indicators. Potentially, the inconsistencies between the indicators result from a difference in measured concept. Aside from the true employment contract type, direct effects (DEs) may exist from external covariates on the recorded employment contract type in the ER or the LFS. If so, such covariates are a source of differential item functioning (DIF) for the indicators. This study focuses on potential DIF for the ER and the LFS. An attempt is made to deduce DIF using latent class (LC) analysis. LC models in which various types of DEs are included are compared with a stepwise likelihood-ratio test (LRT) method, based on Masyn (2017), and an exhaustive Bayesian information criterion (BIC) method. Over multiple datasets, the results for both methods were inconsistent. Additionally, there was little agreement between the methods. The exhaustive BIC method was more conservative as all best-fitting models were nested in the best-fitting models of the stepwise LRT method. For testing the performance of the assessed methods in scenarios with two indicators, an additional simulation study is included. It was found that when no DEs were present, both methods deduced the correct relationships in all cases. However, when DEs were present, both methods performed poorly in deducing the correct relationships. Correct relationships between covariates and indicators were more often found when DIF was relatively simple and effect sizes were relatively large. The moderate success of the stepwise LRT method with two indicators had not been described in any literature thus far. As the results for the real data were inconsistent and the simulation study showed poor performance overall for the assessed methods, no decisive evidence was found that a specific covariate is a source of DIF for the employment contract type as recorded in the ER or the LFS. However, as there are hints for DIF, a difference in measured concept cannot be ruled out. Follow-up research should consider other avenues to investigate the question at hand as the assessed methods gave unsatisfactory results.

Keywords and Phrases: employment contract type, inconsistency, latent class analysis, differential item functioning, stepwise likelihood-ratio tests

2 Introduction

Accurate reports of employment contract type frequencies and transition rates are useful for socio-economic research and legislative purposes. Three types of employment contracts are often distinguished: *permanent*, *flexible* and *other*. Member states of the European Union (EU) and some of their partners use a survey, the Labour Force Survey (LFS), to measure various aspects of their labour force, among which employment contract type frequencies. In each participating state, the survey is conducted by the state's respective national statistical institute (Eurostat, 2022). Often, the LFS is the only source available for estimating the employment contract type frequencies of an EU member state. For some states, however, additional sources are available from which estimates can be obtained. In the Netherlands, in addition to the LFS, Statistics Netherlands (CBS): the national statistical institute of the Netherlands, has access to data from the Employment Register (ER). Both sources give an independent estimate for employment contract type frequencies in the Netherlands. The LFS is completed by the employees themselves, while the information in the ER is provided by employers based on their personnel records.

Starting from 2006, data from the ER has been available to CBS. Soon after, it became apparent that estimates for employment contract types frequencies based on the ER differed from those based on the LFS. Throughout subsequent years, inconsistencies were observed. Even when correcting for differences in population and definitions of employment contract types, inconsistencies between the two sources persisted (Bakker et al., 2021). Figure 1 illustrates how proportions of different employment contract types developed between 2016 and 2018 based on the ER and the LFS. The results in Figure 1 have already been corrected for all systematic differences between the two sources that are known by CBS. Between the sources, a difference is noticeable for *permanent* and *flexible*, while *other* seems consistent. Which source is the most accurate and where the remaining inconsistencies stem from is unknown.

Measurement errors are known to occur in survey data (Biemer, 2011). Although an assumption exists that measurement errors do not occur in administrative data, considerable measurement errors may be present as shown by Oberski (2017), Bakker (2012), and Zhang (2012). Furthermore, the measured concept in registers may not correspond with the desired concept as registers are usually not kept primarily for statistical purposes, but for other administrative purposes. Therefore, it is possible that the register keepers use different definitions for their own administrative purposes than the definitions one would like to use when making statistics. In addition: it may be of interest to individuals, or other entities, to be registered in a certain way; there may exist a substantial administrative delay; or the way in which the register keeper conducts their work can lead to biased registrations (Bakker, 2012, p. 9; and references).

The true employment contract type could be considered as a latent (unobserved) variable. Pavlopoulos and Vermunt (2015) used an extended hidden Markov (HM) model to estimate the

latent employment contract type by combining the data from the ER and the LFS. (In the combined dataset, for each respondent, there is an observation from both the ER and the LFS.) With an estimate of the latent employment contract type, they estimated the measurement error probabilities of the ER and the LFS. For this, the inconsistency between the estimated employment contract type and the measured employment contract type was used. Their findings indicated that both sources contain measurement errors. Later on, Pankowska et al. (2018) replicated the model, and thereafter, Bakker et al. (2021) extended the model by allowing data from respondents that did not indicate to have changed jobs to be serially correlated in both sources. As in the prior studies, Bakker et al. (2021) found that the estimates for the latent employment contract type frequencies and transition rates differed from both the estimates of the ER and the LFS. In addition, Bakker et al. (2021) examined effects of covariates on the inconsistency between the estimated employment contract type and the measured employment contract type, for which some covariates appeared significant.

The aforementioned studies have a few caveats. It is unclear whether the found classes of the latent variable correspond well with the distinguished true employment contract types. This assumption is important for interpreting the model results as measurement error probabilities. If the assumption is incorrect, then the error probabilities found are not the probability that the measured employment contract type deviates from the true employment contract type, but the probability that the measured employment contract type deviates from a different concept. Also, the methods used assume that both sources measure the same underlying concept (latent variable) of employment relationships (Bakker et al., 2021). If this assumption is violated, then the models used may not be valid. It has been noted that the LFS does not necessarily measure the legal employment contract type. For example, some flexible employees may have an informal agreement with their employer that they will be appointed to a permanent position. In such a situation, a contract may be recorded as flexible in the ER, while recorded as permanent in the LFS. Whether the previously made assumptions are appropriate is unknown. Therefore, in this study, the focus will be on this topic.

This study aims to assess whether there are systematic differences in the underlying concept measured in the ER and the LFS with respect to employment contract type in the Netherlands. For this, latent class (LC) analysis will be used. This method has several applications, among which investigating measurement errors in survey data (Biemer, 2011). For investigating measurement errors, LC analysis is not limited to survey data only (Biemer, 2011; Masyn, 2017). Oberski (2017) showed how measurement errors can be estimated from survey data and administrative data in a combined dataset using an LC model. In this study, for investigating the inconsistency between the ER and the LFS, a similar approach will be used where the true employment contract type will be modelled as a latent variable with observations from the ER and the LFS as indicator variables. In contrast to previous research, instead of examining

potential effects of covariates on the inconsistency between the estimated employment contract type and the measured employment contract type afterwards, models used in this study allow covariates to have effects on the measured employment contract type when estimating the model.

In addition to the true employment contract type, the measured employment contract type of one of the indicators may also depend on one or more external covariates. If this is the case, then there is measurement non-invariance for the indicators. The covariate is then a source of differential item functioning (DIF) for the indicator. With DIF, there is an indicator specific systematic error due to the effect of a covariate. In addition, there may exist a random error that influences the measured value. One can investigate whether there is evidence for any error with respect to measuring the same latent variable for both sources. However, to attribute any error found to a systematic or a random part is beyond the scope of this study. (There is insufficient data available for such an inquiry.) If there is a systematic error, then the measured concept consists of the true employment contract type and the systematic error. If this error differs in both indicators, then the measured concepts for employment contract type are different. Note that if the measured concept is identical in both sources, it may still deviate from the desired concept: true employment contract type.

Assessed will be whether covariates can be identified that are a source of DIF. If one or more covariates are found that have a different effect on the indicators, this indicates that there is an underlying conceptual difference between the indicators used to measure employment contract type. Measurement invariance for the indicators in an estimated model for all examined covariates indicates absence of systematic differences in underlying concepts. However, there may still be unobserved covariates that are a source of DIF. Therefore, one can only detect DIF, but not completely rule it out.

The research question will mainly be investigated by applying a stepwise method for detecting DIF described by Masyn (2017). In this study, the applicability of this method will also be evaluated. If evidence for DIF is found, one should assess whether the effects are substantial or negligible. Additionally, a simulation study is included to examine the reliability of the methods used when there truly is a conceptual difference and when there is no conceptual difference. Previous studies showed that the inconsistencies between the sources and the relationships with other covariates differed for respondents of different age groups (Bakker et al., 2021). Respondents aged 15 to 25 showed the greatest inconsistencies. For this reason, in this exploratory study, the focus is on this age group.

The contribution of this study can help CBS to continue their research into the inconsistencies between the estimates from the ER and the LFS. If it is understood why the estimates differ between the two sources, then the method of estimating the employment contract type frequencies

may be improved to acquire more accurate estimates in the future. In the next sections of this report, the following is discussed: Section 3 gives an introduction to basic LC analysis and modifications when adding covariates; Section 4 gives a description of the real data, the way in which an identifiable model is obtained; and the methods used to test for DIF; Section 5 gives a description of the results for identifying an LC model and the methods used to test for DIF; Section 6 gives a description of a simulation study for testing the performance of the assessed methods in a situation with two indicators; and Section 7 gives a discussion of all results.

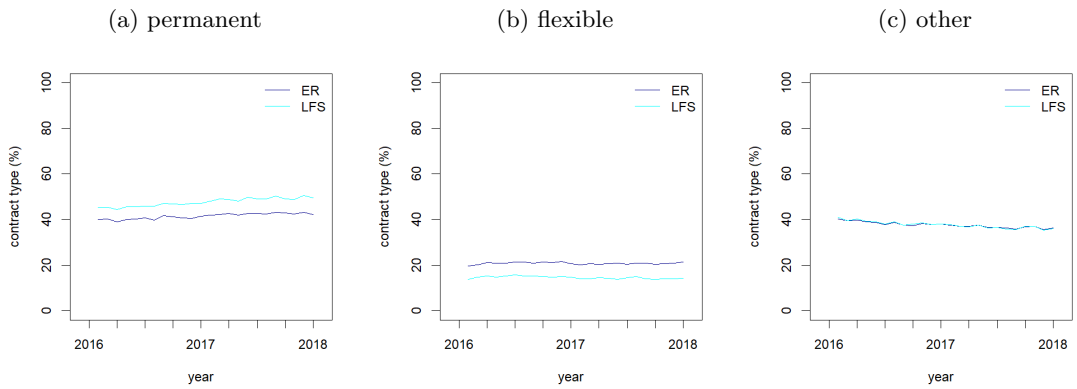

Figure 1: Development of the proportion of (a) *permanent*, (b) *flexible*, and (c) *other* employment contract types in the Netherlands between 2016 and 2018 based on the Employment Register (ER) and the Labour Force Survey (LFS). Represented are respondents who started with the LFS in 2016.

3 Latent class analysis

In literature, the terminology used for LC analysis is context dependent and therefore inconsistent. To aid readability when consulting other sources, commonly used synonyms for certain terms are provided. For example, originally LC analysis was referred to as latent structure analysis. The LC model is also referred to as a binomial (finite) mixture model (Van den Bergh, 2018, p. 1).

In its essence, LC analysis is a method for relating a set of indicators to a set of classes of a latent (unobserved), categorical variable. The indicators are also referred to as dependent variables, endogenous variables, items, outcome variables, observed variables, outputs, and responses (Hagenaars and McCutcheon, 2002, p. 91; Vermunt, 2010). The latent classes are also referred to as clusters. The assumption is that the cases, which have values on the indicators, belong to a set of latent classes (Hagenaars & McCutcheon, 2002, p. 89). The cases are also referred to as objects, observational units, observations, subjects, or respondents. The indicator values of a case are also referred to as observed scores or response. In addition to indicators, covariates can also be included in LC analysis. Such covariates may be referred to as concomitant variables, exogenous variables, explanatory variables, external variables, grouping variables, independent variables, inputs, or predictors (Hagenaars and McCutcheon, 2002, p. 96; Vermunt, 2010). A unique combination of indicator and covariate values for a case is referred to as a pattern. Separately, indicator patterns and covariate patterns may also be indicated. Frequencies are a measure of how often patterns occur.

LC analysis is a model-based approach where one assumes a statistical model for the population from which the sample originates (Hagenaars & McCutcheon, 2002, p. 90). For cases that belong to the same latent class, one assumes that the indicator values originate from the same probability distribution. The parameters from this distribution are unknown and need to be estimated (Hagenaars & McCutcheon, 2002, p. 89). In LC analysis, it is assumed that each case belongs to a single class. As it is a probabilistic classification approach, the uncertainty of the class membership can be taken into account (Hagenaars and McCutcheon, 2002, p. 91; Vermunt, 2010).

In LC analysis, the latent variable is categorical, while the indicators may be of any type. Similar analysis methods for other types of data go by different names. For example, factor analysis is an analogue of LC analysis in which the latent variable is continuous (Van den Bergh, 2018, p. 1). In this study, only categorical indicators will be used. LC analysis is used for various purposes. It may be used for classifying cases into latent classes or for relating class membership to one or more covariates. For example, one may do an exploratory study to investigate associations between latent classes and covariates or create a model for predicting class membership based on covariates (Vermunt, 2010). In this study, the focus is on the former.

Below, the basic LC model, estimation of parameters, and classification of cases is described. Further down, the addition of covariates to an LC model is described.

3.1 Basic latent class model

An LC model is estimated with logistic regression. Multinomial logistic regression is used for a latent variable with more than two classes and multiple logistic regression is used for indicators with more than two categories. If both are the case, then multiple multinomial logistic regression is used.

Let K represent the number of indicators, and k represent a single indicator. Let N represent the number of cases, and i represent a single case. Let $\mathbf{Y}_i$ represent the indicator pattern of case i; δ represent an indicator pattern; and Y_{ik} represent an indicator value for the case i on indicator k. Let X represent a discrete latent class variable and X_i its value for case i. Let T represent the number of classes in X, and t or s represent a single latent class. Let R_k represent the number of categories for indicator k and r represent a single category for indicator k. With this, one can define a basic LC model with a single latent variable for $P(\mathbf{Y}_i = \delta)$ as

$$P(\mathbf{Y}_i = \delta) = \sum_{t=1}^{T} P(X_i = t)P(\mathbf{Y}_i = \delta|X_i = t) \tag{1}$$

Note that in literature, similar to the terminology, there are some different ways in which equivalent models are defined (see for example, Hagenaars and McCutcheon, 2002, p. 94; Masyn, 2017; Vermunt, 2010).

In a basic LC model, given the LC, indicators are assumed to be independent. This is the assumption of local independence and can be defined as

$$P(\mathbf{Y}_i = \delta|X_i = t) = \prod_{k=1}^{K} P(Y_{ik} = r|X_i = t). \tag{2}$$

The parameters to estimate for an LC model are the class proportions $P(X_i = t)$ and the multinomial parameters $P(Y_{ik} = r|X_i = t)$.

In the case of this study, with two indicators and three latent classes, Equation (2) becomes

$$P(Y_{i1} = r, Y_{i2} = r'|X_i = t) = P(Y_{i1} = r|X_i = t)P(Y_{i2} = r'|X_i = t), \tag{3}$$

and Equation (1) becomes

$$P(Y_{i1} = r, Y_{i2} = r') = \sum_{t=1}^{3} P(X_i = t)P(Y_{i1} = r|X_i = t)P(Y_{i2} = r'|X_i = t). \tag{4}$$

There are two main methods for estimating parameter estimates in LC models. These is maximum likelihood (ML) estimation and maximum a posteriori (MAP) estimation. For this study, ML will be used. Generally, the estimates do not differ much. However, MAP avoids that probabilities or variances become zero. This problem can also be mitigated by adding a small amount of prior information so that the parameters stay within possible parameter values (parameter space; Hagenaars and McCutcheon, 2002, p. 97). When using ML estimation for estimating parameter values, the log-likelihood function is maximised. The log-likelihood function for a basic LC model for $P(\mathbf{Y}_i = \delta)$ can be defined as

$$\mathcal{L}_{(\text{BASIC})} = \sum_{i=1}^{N} \log P(\mathbf{Y}_i = \delta) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X_i = t) P(\mathbf{Y}_i = \delta | X_i = t) \right]. \tag{5}$$

In the case of this study, with two indicators and three latent classes, Equation (5) becomes

$$\begin{aligned} \mathcal{L}_{(\text{BASIC})} &= \sum_{i=1}^{N} \log P(Y_{i1} = r, Y_{i2} = r') \\ &= \sum_{i=1}^{N} \log \left[\sum_{t=1}^{3} P(Y_{i1} = r | X_i = t) P(Y_{i2} = r' | X_i = t) P(X_i = t) \right]. \end{aligned} \tag{6}$$

To estimate a model one can use dummy coding. The advantage of this is that estimated parameter values can be interpreted as a change in the log-odds with respect to specified reference categories. With dummy coding, the parameter values of the reference categories are zero.

The probabilities $P(X_i = t)$ and $P(Y_{ik} = r|X_i = t)$ seen in Equation (5) can be parameterised using multinomial logistic regression models. $P(X_i = t)$ can be defined as

$$P(X_i = t) = \frac{\exp(a_t)}{\sum_{t'=1}^{T} \exp(a_{t'})}, \tag{7}$$

where

a_t is a logistic intercept for the true contract type. It corresponds to the log-odds of the respondent having true contract type t overall.

Likewise, $P(Y_{ik} = r|X_i = t)$ can be defined as

$$P(Y_{ik} = r|X_i = t) = \frac{\exp(\alpha_r + \beta_{rt})}{\sum_{r'=1}^{R_k} \exp(\alpha_{r'} + \beta_{r't})}, \tag{8}$$

where

α_r is a logistic intercept for the registered contract type. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k overall; and

β_{rt} is a logistic slope for the true contract type. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k given that the respondent's true contract type is t.

3.2 Defining the model

Before a basic LC model (without covariates) can be constructed, one needs to decide what indicators to use and how many latent classes the latent variable will have. The number of latent classes and their sizes are unknown (Hagenaars & McCutcheon, 2002, p. 89). Usually, the number of latent classes is not determined a priori, but with an iterative process. A set of models is fitted, the best of which is chosen on the basis of fit measures (for example, AIC or BIC) and interpretability (Van den Bergh, 2018, p. 2). There also exist possibilities for relaxing the local independence assumption for specified indicator pairs or putting restrictions on certain model parameters to find a more parsimonious model.

Key is that the model is identifiable, meaning that the number of independent parameters to be estimated may not exceed the number of patterns in the data (Vermunt, 2010). For this study, the aim is to use three latent classes as these may correspond one-to-one with the true contract types. Note that it is not possible to identify a model with three latent classes and two indicators that each consist of three categories. A way to resolve this is by adding a covariate to

the model that is not related to the indicators but is related to the latent variable. This will be discussed in Section 4.2.

One can evaluate the fit of a basic LC model in various ways. Commonly, Bayesian information criterion (BIC) or likelihood ratio chi-square (L^2) are used to assess model fit. When comparing models with a different number of latent classes, the model with the lowest BIC or lowest, non-significant L^2 is preferred. For assessing the fit of a single model on its own, one may examine whether the L^2 indicates if the model fit is not statistically worse than that of the 'saturated' model. The 'saturated' model is a model with as many independent parameters as there are patterns in the data (Bishop et al., 1975, p. 9). The validity of any restricted parameters should also be assessed (Hagenaars and McCutcheon, 2002, p. 90). When models are nested they can be compared with one another using a likelihood-ratio test (LRT).

Another type of statistic that can be used for determining what number of latent classes is appropriate is bivariate residual (BVR). The distribution of the BVR statistic is undefined. Generally, however, a chi-square distribution with one degree of freedom is assumed (Janssen et al., 2018, 5, and references). With this, a BVR for a pair of indicators larger than 3.84 indicates that there is a strong dependency left. If this is the case, then the model does not sufficiently explain correlations between that pair of variables and the local independence assumption is violated (Vermunt and Magidson, 2005, p. 125; Hagenaars and McCutcheon, 2002). Note that extremely small BVRs may point towards overfitting. For more information about BVRs, see Vermunt and Magidson (2005).

3.3 Estimating parameters

Parameters may be estimated using full information maximum likelihood (FIML). Software capable of performing LC analysis with covariates may use an expectation-maximization (EM) or Newton-Raphson (NR) algorithm for finding optimal values for the parameters (Vermunt, 2010). EM is the more stable, but NR converges faster when it is close to the solution. Hagenaars and McCutcheon (2002) states that the ideal algorithm starts with a number of EM iterations and when the estimates are close to the final solution, switches to NR iterations. In this way, the strengths of both algorithms are combined. One gets the stability of EM when the optimum is still far away and the speed of NR when the optimum is close by. Well known in LC analysis is the problem of finding local solutions. This can be limited by using multiple start values for the algorithm (Hagenaars & McCutcheon, 2002).

3.4 Latent class membership

In LC analysis, a case's posterior class membership probability is determined with the estimated model parameters and the case's indicator values. Note that it is also possible to classify new

cases that belong to the same population as the sample (Hagenaars and McCutcheon, 2002, p. 91).

One may be interested in the parameters of the model and or the classification of cases. The latent class assigned to case i is represented with W_i. For assigning cases to classes, the main classification method is modal assignment, which means that each case is assigned to the class for which it has the highest probability (Hagenaars & McCutcheon, 2002). Also widely used is proportional assignment. Modal assignment estimates W_i as the value of t for which $P(X_i = t|\mathbf{Y}_i = \delta)$ is largest. Meaning that it yields a hard separation in which case i is treated as belonging to class t with weight $P(W_i = t|\mathbf{Y}_i = \delta) = 1$ if $P(X_i = t|\mathbf{Y}_i = \delta)$ is largest and with weight $P(W_i = t|\mathbf{Y}_i = \delta) = 0$ otherwise. Proportional assignment treats cases as belonging to latent class t with probability $P(X_i = t|\mathbf{Y}_i = \delta)$. Meaning that it yields a soft separation with weights $P(W_i = t|\mathbf{Y}_i = \delta) = P(X_i = t|\mathbf{Y}_i = \delta)$. Lastly, there exists random assignment in which W_i is acquired by randomly drawing from $P(X_i = t|\mathbf{Y}_i = \delta)$, yielding a hard separation (Vermunt, 2010, 8, and references). With Bayes rule, this posterior class membership probability of belonging to latent class t given the data can be defined as

$$P(X_i = t|\mathbf{Y}_i = \delta) = \frac{P(\mathbf{Y}_i = \delta|X_i = t)P(X_i = t)}{P(\mathbf{Y}_i = \delta)}. \tag{9}$$

Note that $P(X_i = t)$ is the prior probability of belonging to latent class t.

If one substitutes Equation (1) and (2) back into Equation (9), one gets

$$P(X_i = t|\mathbf{Y}_i = \delta) = \frac{\prod_{k=1}^{K} P(Y_{ik} = r|X_i = t)P(X_i = t)}{\sum_{t'=1}^{T}\prod_{k=1}^{K} P(Y_{ik} = r|X_i = t')P(X_i = t')}. \tag{10}$$

(The logistic parameterisations of $P(Y_{ik} = r|X_i = t)$ and $P(X_i = t)$ were given in Equation (8) and (7), respectively.) Classification error may also be assessed with the conditional probability $P(W_i = s|X_i = t)$ that represents the probability of the estimated value conditional on the true value. The probability can be defined as

$$P(W_i = s|X_i = t) = \sum_{\delta\in\Delta} P(\mathbf{Y}_i = \delta|X_i = t)P(W_i = s|\mathbf{Y}_i = \delta), \tag{11}$$

where δ is an indicator pattern of the set of all possible indicator patterns Δ. (One sums over all possible indicator patterns.) As noted, with proportional assignment, $P(W_i = s|\mathbf{Y}_i = \delta) = P(X_i = t|\mathbf{Y}_i = \delta)$; and with modal assignment, $P(W_i = s|\mathbf{Y}_i = \delta) = 0 \vee P(W_i = s|\mathbf{Y}_i = \delta) = 1$. $P(\mathbf{Y}_i = \delta|X_i = t)$ is defined in Equation (2) and parameterised in Equation (8). For a more detailed description, see Vermunt (2010).

3.5 Latent class model with covariates

Covariates can be added to an LC model. There are a few different ways described in which an LC model with covariates can be modelled. Well established methods are the one-step approach and the three-step approach. These approaches are discussed in Section 3.5.1, and 3.5.2, respectively. A two-step approach has also been developed (Bakk & Kuha, 2018). This approach will not be discussed, however, as it was not used in this study.

Normally, given the LC membership, indicators are assumed to be independent of covariates if present in the model. This is the assumption of measurement invariance or the absence of DIF (Janssen et al., 2018; Masyn, 2017). For a covariate vector $\mathbf{Z}_i$, this can be defined as

$$P(\mathbf{Y}_i = \delta | X_i = t, \mathbf{Z}_i = \lambda) = P(\mathbf{Y}_i = \delta | X_i = t), \tag{12}$$

where λ is a covariate pattern. If the assumption is violated, then the parameter estimates of an LC model can have a strong bias (Janssen et al., 2018; Masyn, 2017). This results from residual associations between indicators and covariates which have not been accounted for in the model (Masyn, 2017). One can relax the assumption of measurement invariance by adding a direct effect (DE) from one or more covariates on one or more indicators (Janssen et al., 2018; Masyn, 2017; Vermunt, 2010). With this, one allows for the value of indicators to depend on the covariates. In Section 3.5.1 and Section 3.5.2, a short overview is given of the one-step and three-step approach respectively with a formal definition for when DEs are present.

For more information on LC models with covariates without DEs on indicators, see Vermunt (2010). For more information on LC model with covariates with DEs on indicators, see Janssen et al. (2018).

3.5.1 One-step approach

Without direct effects

If one adds covariates that affect the latent variable without directly affecting the indicators (no DEs; an assumption that given the latent class, the indicators and the covariate are independent), then Equation (1) becomes

$$P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda) = \sum_{t=1}^{T} P(X_i = t|\mathbf{Z}_i = \lambda)P(\mathbf{Y}_i = \delta|X_i = t), \tag{13}$$

(Vermunt, 2010). Note that via the latent class, the covariate does have an indirect effect (IE) on the indicator.

In Equation (13), one again assumes local independence for the indicators values in $\mathbf{Y}_i$ as in Equation (2). Note that here, one also assumed that given X, $\mathbf{Y}_i$ is independent of $\mathbf{Z}_i$ (Vermunt, 2010). In other words, no DIF is assumed.

For categorical covariates, the probability $P(X_i = t|\mathbf{Z}_i = \lambda)$, in which $\mathbf{Z}_i$ for Q covariates is denoted as $Z_{i1}, \ldots, Z_{iQ}$ to distinguish the individual covariates, is usually parameterised with a multinomial logistic regression model that can be defined as

$$P(X_i = t|Z_{i1} = \theta_1, \ldots, Z_{iQ} = \theta_Q) = \frac{\exp\left(\alpha_t + \beta^{(1)}_{t\theta_1} + \cdots + \beta^{(Q)}_{t\theta_Q}\right)}{\sum_{t'=1}^{T} \exp\left(\alpha_{t'} + \beta^{(1)}_{t'\theta_1} + \cdots + \beta^{(Q)}_{t'\theta_Q}\right)}, \tag{14}$$

where

α_t is a logistic intercept for the true contract type. It corresponds to the log-odds of the respondent having true contract type t overall; and

$\beta^{(q)}_{t\theta_q}$ is a logistic slope for the covariate category. It corresponds to the log-odds of the respondent having contract type t given that the respondent has covariate category θ_q for covariate q.

Note that parameters for interaction effects are not included in Equation (14). They may be included with additional γ parameters for specific covariate pair interactions on the right hand side. However, one should indicate the specific interactions in some way on the left hand side as well (which becomes unclear, rather fast). For a variant of Equation (14) for numeric covariates, see Vermunt (2010).

With covariates, the overall probability $P(X_i = t)$ can be defined as

$$P(X_i = t) = \sum_{\lambda=1}^{\Lambda} P(\mathbf{Z}_i = \lambda)P(X_i = t|\mathbf{Z}_i = \lambda). \tag{15}$$

The multinomial parameters defining $P(\mathbf{Y}_i = \delta|X_i = t)$ (present in Equation (13), which use the same parameterisation as defined in Equation (8), as $\mathbf{Y}_i$ is independent of $\mathbf{Z}_i$, given the LC) and the estimates of the α and β parameters (present in Equation (14)) can be obtained by maximising a log-likelihood function based on $P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda)$. This can be defined as

$$\mathcal{L}_{(\text{1-STEP: IE})} = \sum_{i=1}^{N} \log P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda) = \sum_{i=1}^{N} \log \sum_{t=1}^{T} P(X_i = t|\mathbf{Z}_i = \lambda) P(\mathbf{Y}_i = \delta|X_i = t) \quad (16)$$

With direct effects

By adding DEs from the covariates on the indicators, one relaxes the conditional probabilities $P(\mathbf{Y}_i = \delta|X_i = t)$ as $\mathbf{Y}_i$ does not only depend on the latent variable X but also on the covariate vector $\mathbf{Z}_i$. If one allows the covariates to have a DE on the indicators, then one modifies Equation (13) (which was already a modified version of Equation (1)) to get

$$P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda) = \sum_{t=1}^{T} P(X_i = t|\mathbf{Z}_i = \lambda) P(\mathbf{Y}_i = \delta|X_i = t, \mathbf{Z}_i = \lambda), \quad (17)$$

which holds for both uniform and non-uniform DIF (Janssen et al., 2018). For a formal definition of DIF in general, uniform DIF, and non-uniform DIF, see Masyn (2017).

In Equation (17), just like Equation (2) when no covariates were added, the second term at the right hand side (assuming local independence) can be defined as

$$P(\mathbf{Y}_i = \delta|X_i = t, \mathbf{Z}_i = \lambda) = \prod_{k=1}^{K} P(Y_{ik} = r|X_i = t, \mathbf{Z}_i = \lambda). \quad (18)$$

When DEs are included, the term at the right hand side of Equation (18) can be parameterised for a single categorical covariate as

$$P(Y_{ik} = r|X_i = t, Z_{iq} = \theta) = \frac{\exp\left(\alpha_r + \beta_{rt} + \gamma_{r\theta} + \zeta_{rt\theta}\right)}{\sum_{r'=1}^{R_k} \exp\left(\alpha_{r'} + \beta_{r't} + \gamma_{r'\theta} + \zeta_{r't\theta}\right)}, \quad (19)$$

where

α_r is a logistic intercept for the registered contract type. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k overall;

β_{rt} is a logistic slope for the true contract type. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k given that respondent's true contract type is t;

$\gamma_{r\theta}$ is a logistic slope for the covariate. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k given that the respondent has covariate category θ; and

$\zeta_{rt\theta}$ is a logistic slope for the interaction effect for the true contract type and covariate. It corresponds to the log-odds of the respondent being registered as having contract type r on indicator k given that the respondent has true contract type t and covariate category θ.

In Equation (19), the ζ parameter is only included for non-uniform DIF. For uniform DIF, it is left out. It is of course possible to include multiple covariates, both with and without interaction effects. If desired, one has to simply add the parameters for the specific covariate categories and interactions within the exponential function of the numerator and denominator. As this study deals with categorical covariates, this is difficult to express in a single formula. For a concise way to parameterise Equation (19) for numeric covariates, see Janssen et al. (2018).

When using a one-step approach with covariates that have DEs on the indicators, one estimates the LC model by maximising the log-likelihood function $\mathcal{L}_{(\text{1-STEP: DE})}$ for $P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda)$ as

$$\begin{aligned}\mathcal{L}_{(\text{1-STEP: DE})} &= \sum_{i=1}^{N} \log P(\mathbf{Y}_i = \delta|\mathbf{Z}_i = \lambda) \\ &= \sum_{i=1}^{N} \log \sum_{t=1}^{T} P(X_i = t|\mathbf{Z}_i = \lambda) P(\mathbf{Y}_i = \delta|X_i = t, \mathbf{Z}_i = \lambda).\end{aligned} \tag{20}$$

For Equation (20), the term $P(\mathbf{Y}_i = \delta|X_i = t, \mathbf{Z}_i = \lambda)$ is defined in Equation (18) and (19). Note that by adding DEs in the one-step approach, it requires the complete re-estimation of the model. This may change the definitions of the latent classes. Depending on the size of the dataset, re-estimating the model may take some time.

3.5.2 Three-step approach

The three-step approach consist of the following steps: (1) Estimating a simple LC model without covariates (see Section 3.1). (2) Assigning cases to a latent class on the basis of their indicator values $\mathbf{Y}_i$ and the estimated parameters $P(X_i = t)$ and $P(Y_{ik} = r|X_i = t)$ (see section 3.4). (3) Investigating the association between the assigned class memberships and covariates with, for

example, a logistic regression model (Vermunt, 2010).

A shortcoming of the three-step approach is that classification errors can be made in the second step (see section 3.4). The classification error is the probability that a case belongs to class $X_i = t$ while it is assigned to class $W = s$ where $t \neq s$. Bolck et al. (2004) showed that the three-step approach underestimates the relationships between covariates and class membership, irrespective of whether a modal, random, or proportional assignment was used. They demonstrate that the estimates of the parameters that relate to the effect of $\mathbf{Z}_i$ are biased towards 0. Bolck et al. (2004) demonstrate an approach to mitigate this bias, in the third step, taking into account classification error found in the second step; which was later named the BCH approach by Vermunt (2010). However, as discussed by Vermunt (2010), the method still has some limitations. Vermunt (2010) proposed a modified BCH approach that removes some limitations of the original BCH approach: the bias-adjusted BCH approach. (For a description of the original and the bias-adjusted BCH approach, see Bolck et al. (2004) and Vermunt (2010)) In addition to this, Vermunt (2010) proposed an alternative to the (bias-adjusted) BCH approach: the ML approach. The ML approach has a similar logic to the BCH approach, however, Vermunt (2010) deemed it to be more direct. In the ML approach, a LC model is defined in which the class assignment in step two serves as a single response variable with known measurement error probabilities. In the LC model the relevant predictors can be introduced while keeping the measurement model fixed. Vermunt (2010) states that the ML approach is more elegant, easier to use, and easier to extend to more complex situations. (For example, having multiple separately constructed latent variables or measurement models which differ across categories.) In their simulation study, they showed that it is more efficient and that it yields smaller standard errors for the covariate effects compared to the BCH approach. In the next section, an overview is given of the ML approach.

Without direct effects

Standard approach

In the third step of the three-step approach, using a multinomial logistic regression model, one estimates the effect of the covariates on the estimated class membership W_i. Note that for performing the third step, the dataset has to be expanded in the second step. When using modal, or random assignment, a new column needs to be added with the assigned class membership. When using proportional assignment, a column needs to be added with the posterior class membership probabilities for each latent class. For categorical covariates, if covariates are added without DEs, a multinomial logistic regression model can be parameterised as

$$P(W = t|Z_{iq} = \theta) = \frac{\exp(\alpha_t + \beta_{t\theta})}{\sum_{t'=1}^{T} \exp(\alpha_{t'} + \beta_{t'\theta})}, \tag{21}$$

where

α_t is the logistic intercept for assigned class. It corresponds to the log-odds of the respondent being assigned to class t overall; and

$\beta_{t\theta}$ is the logistic slope for the covariate. It corresponds to the log-odds of the respondent being assigned to class t given that the respondent has covariate category θ.

As with aforementioned equations, one can add multiple covariates by adding additional terms in the numerator and denominator.

Of interest are the β parameters. Estimates can be found by maximising the following log-likelihood function

$$\mathcal{L}_{(\text{3-STEP: IE})} = \sum_{i=1}^{N}\sum_{t=1}^{T} \underbrace{P(W=t|\mathbf{Y}_i=\delta)}_{\text{fixed}} \log P(W=t|\mathbf{Z}_i=\lambda) \tag{22}$$

In (22), $P(W=t|\mathbf{Y}_i=\delta)$ is a fixed weight that has already been defined in step two of the three-step approach. For modal and random assignment, it is 1 or 0, and for proportional assignment it is the posterior class membership probability $P(X=t|\mathbf{Y}_i=\delta)$. (Note that no DEs are included; only IEs of the covariate.)

ML approach

The probability $P(W=s|\mathbf{Z}_i=\lambda)$ is related to the probability $P(X=t|\mathbf{Z}_i=\lambda)$ as follows:

$$P(W=s|\mathbf{Z}_i=\lambda) = \sum_{t=1}^{T} P(X=t|\mathbf{Z}_i=\lambda)P(W=s|X=t) \tag{23}$$

(Bolck et al., 2004; Vermunt, 2010). One can see that the form of Equation (23) is similar to that of Equation (13); depicting a one-step approach for an LC model with covariates. There are two differences, however. The assigned class W_i replaces the vector with indicator values $\mathbf{Y}_i$ and the error probabilities $P(W_i=s|X_i=t)$ are fixed (known after step two). Equation (23) can be viewed as an LC model with one indicator and fixed error probabilities (Vermunt, 2010). With this realization, Vermunt (2010) proposed an alternative correction for the third step of the three-step approach. One includes the covariates in a LC model with one nominal indicator: the assigned class membership; and in which the classification error probabilities $P(W_i=s|X_i=t)$ are fixed based on estimates obtained in step two. (In reality, one estimates a one-step approach LC model for the third step in the three-step LC model approach). For finding the parameters of interest, one maximised the following log-likelihood function:

$$\mathcal{L}_{(\text{3-STEP ML IE})} = \sum_{i=1}^{N} \log \sum_{t=1}^{T} \underbrace{P(X=t|\mathbf{Z}_i=\lambda)}_{\text{free}} \underbrace{P(W=s|X=t)}_{\text{fixed}} \tag{24}$$

With the ML estimate, one obtains the probabilities $P(X = t|\mathbf{Z}_i = \lambda)$ and the accompanying logistic parameters.

Vermunt (2010) does note that the standard errors may still be underestimated as the classification error probabilities $P(W = s|X = t)$ are fixed when finding the ML estimate, while they are estimated themselves in step two and hence have a degree of uncertainty that should be accounted for.

With direct effects

Adding DEs with the three-step approach is often difficult, if not impossible. For estimating the structural model in the third step with the standard approach, only information from the assigned class is used. The information from the separate indicators is lost. Because of this, in the third step of the standard approach, one cannot add DEs from covariates on indicators. It is possible to add DEs in the first step. For this, however, the exact structure of the relations needs to be known beforehand. This is very rarely the case (Janssen et al., 2018). Therefore, it is not considered in this study. (How to add DEs in the first step is described by Vermunt and Magidson (2021).)

In contrast to the third step of the standard approach, it is possible to add DEs in the third step of the ML approach. For this, one uses the same method as described in Section 3.5.1 for a one-step approach, with one indicator. The only difference being that the parameters in Equation (19) are not estimated, but fixed. Only the parameters for $P(X_i = t|\mathbf{Z}_i = \lambda)$ are estimated. This probability can be parameterised with a multinomial logistic regression model as in Equation (14).

4 Methods

4.1 Data description

For this study, real pseudonymised data collected by CBS was used. The data originates from the ER and the LFS collected between 2016 and 2020.

The ER was provided by the Institute for Employee Insurance (UWV): an independent governing body that falls under the responsibility of the ministry of Social Affairs and Employment (SZW). The ER was based on the total of all wage declarations to the taxation authority of the Netherlands (Belastingdienst) and was recorded each month. The declarations were filled out by the employers.

The LFS was a rotating panel survey with five polls conducted by CBS. In this survey, respondents were interviewed on five separate occasions over a fifteen-month period with three months between interviews. The interviews were not conducted at a fixed time, but somewhere within the three-month interval. Thus, not all respondents started at the same calendar time.

Prior to this study, datasets were generated by deterministically linking the ER and the LFS, retaining only the data of respondents who were present in both datasets. The combined datasets consisted of a case identifier, two indicators (one for each source), and ten covariates. For each panel of the LFS, a separate dataset was created, whereby individual datasets were denoted by their starting year. In this way, there was a 2016, 2017 and 2018 dataset. See Section 4.1.1 and 4.1.2 for a description of the indicators and covariates, respectively.

A time series consisting of one observation per month for a period of fifteen months was available for each respondent. This time series always contained missing data due to delayed entry, panel attrition, and the fact that the LFS was not being held monthly. In this exploratory study, only used were the first observations of each respondent where both the ER and the LFS were recorded. In this selection there was no missing data for any of the variables. Previous studies showed that the inconsistencies between the indicators for employment contract type and the association with external covariates differed for different age groups (Bakker et al., 2021). Respondents aged 15 to 25 showed the greatest inconsistencies. Therefore, the focus of this exploratory study will be on this age group only. With the aforementioned selection, the datasets for 2016, 2017, and 2018 consisted of 16,258, 16,052, and 18,362 complete cases, respectively.

4.1.1 Indicators

There were two indicators for the respondent's employment contract type: the contract as recorded in the ER (*conER*), and the contract as recorded in the LFS (*conLFS*). Definitions

used for employment contract types were those defined by Bakker et al. (2021).

The type of employment contract was determined based on the number of contract hours and the duration of the contract. For employment contract type, three category levels were distinguished with coding: *permanent* = 1, *flexible* = 2, and *other* = 3. A *permanent* contract was defined as an employment contract for an indefinite period of time with a fixed or varying number of contract hours per week. A *flexible* contract was defined as an employment contract for a definite period of time with any number of contract hours. It included on-call contracts and employment agency contracts. The *other* category included respondents who did not have a payed job, who were self-employed (freelancers) or who were director major-shareholder (DGA). The latter are individuals who own a large portion of the shares in a private or public company and have the highest, or a relatively high managerial position in said company. For respondents who had multiple jobs at the same time, only the job with the highest income was taken into account. For additional information, see the appendix of Bakker et al. (2021).

A contingency table for the two indicators is given in Table 1. As can be seen, there are a considerable amount of inconsistencies between the two.

Table 1: Contingency table of the two indicators for employment contract type in the Netherlands: employment contract type as recorded in the Employment Register (*conER*) and employment contract type as recorded in the Labour Force Survey (*conLFS*). Shown are the first complete observations of respondents aged 15 to 25 of the 2016 dataset. Employment contract types are indicated with their corresponding coding.

	conLFS		
conER	1	2	3
1	1026	662	159
2	1664	5185	1117
3	188	668	5589

4.1.2 Covariates

Ten covariates were available that may be of importance in explaining the inconsistencies between the ER and the LFS. All covariates were treated as categorical without ordering. The categories used for the covariates were similar to those of Bakker et al. (2021), but not identical. A short description of all covariates and their corresponding categories is given below.

Age group (*ageGro*): age group to which the respondent belongs. Two category levels with coding: *age 15 to 19* = 1 and *age 19 to 25* = 2.

Company size (*comSiz*): size of the company where the respondent is employed, based on the total number of employees of the company. Four category levels with coding: *large company* = 1; *middle-size company* = 2; *small company* = 3; and *missing/not applicable* = 4. A small, middle-size, and large company was defined as having less than ten, ten to one hundred, or one hundred or more employees, respectively.

Contract hours (*conHou*): number of contract hours per week of the respondent. Five category levels with coding: *less than 12 hours* = 1; *12 to 20 hours* = 2; *20 to 30 hours* = 3; *more than 30 hours* = 4; and *missing/not applicable* = 5.

Economic activity (*ecoAct*): economic sector in which the respondent is employed. Ten category levels with coding: *other* = 1; *agriculture, forestry and fishing* = 2; *retail* = 3; *water and aviation transport* = 4; *hotel, restaurants and cafe* = 5; *finance* = 6; *job placement, employment agencies and human resources management* = 7; *government and education* = 8; *healthcare* = 9; and *missing/not applicable* = 10.

Education level (*eduLev*): highest education level attained by the respondent. Four category levels with coding: *low* = 1; *middle* = 2; *high* = 3; and *unknown* = 4.

Gender: recorded gender of the respondent. Two category levels with coding: *male* = 1 and *female* = 2.

Interview manner (*intMan*): whether the survey was filled out by the respondent of interest or someone else (proxy interview). Two category levels with coding: *respondent of interest* = 1 and *someone else* = 2. Although this covariate only directly relates to the LFS, it may indirectly relate to the ER via a relationship with another (unobserved) covariate.

Job duration (*jobDur*): number of months that the respondent has been employed in their current job. Seven category levels with coding: *less than 3 months* = 1; *3 to 6 months* = 2; *6 to 12 months* = 3; *12 to 24 months* = 4; *24 to 36 months* = 5; *more than 36 months* = 6; and *missing/not applicable* = 7.

Migration background (*migBac*): migration background of the respondent. Seven category levels with coding: *Netherlands* = 1; *Morocco* = 2; *Turkey* = 3; *Suriname* = 4; *Netherlands Antilles* = 5; *other non-western countries* = 6; and *other western countries* = 7. Note that the categories represent common migration backgrounds for Dutch citizens (CBS, 2022).

Software cluster (*sofClu*): group to which the software belongs that was used by the employer

to register the respondent in the ER. Six category levels with coding: *cluster 1* = 1; *cluster 2* = 2; *cluster 3* = 3; *cluster 4* = 4, *cluster 5* = 5; and *missing/not applicable* = 6. Note that each software cluster consists of different types of software that are similar in some way. Although this covariate only directly relates to the ER, it may indirectly relate to the LFS via a relationship with another (unobserved) covariate.

Each covariate was derived from one of the two sources. Directly recorded in or linked to the ER were: *comSiz*, *conHou*, *ecoAct*, *jobDur*, and *sofClu*. Directly recorded in or linked to the LFS were: *ageGro*, *eduLev*, *gender*, *intMan*, and *migBac*.

For a more detailed description of the covariates, similar category levels, and a justification for those category levels used, see Bakker et al. (2021). The study by Bakker et al. (2021) found, for different age groups, that some covariates appear to be related to the observed inconsistencies between the indicators.

Note that the covariates *comSiz*, *ecoAct*, *jobDur*, *sofClu*, and *conHou* could only be applicable for respondents who were perceived to have a paid job according to the ER at the time the ER was recorded. For these covariates, respondents who did not appear in the ER were directly assigned to the covariate category *missing/not applicable*. As a result, patterns where *conER* had the value *permanent* or *flexible* and the covariate had the value *missing/not applicable* could not occur. The aforementioned patterns always had a frequency of zero. As noted, all patterns where the covariate had the value *missing/not applicable* were not directly observed but derived. To indicate this, the frequencies of these patterns are referred to as derived frequencies.

4.2 Latent class model identification

For this study, available were two trichotomous categorical indicators: *conER* and *conLFS*. The aim was to model a categorical latent variable that represented the true employment contract type. Ideally, one would construct a latent variable consisting of three classes. In this way, there could be a one-to-one correspondence between the LCs and the true employment contract types. However, it is not possible to identify a basic three-class LC model with only two trichotomous indicators. In this situation, there are more parameters to estimate than there are degrees of freedom available. (For how to find the number of parameters to estimate and the degrees of freedom of an LC model with only categorical indicators, see Appendix 9.1. Note that an LC model can only be estimated if the degrees of freedom are non-negative). A way in which this issue can be resolved is by adding a covariate to the basic LC model that is related to the latent variable, but not directly related to the indicators. In this way, the covariate helps identify the model by making more degrees of freedom available for estimating parameters. Such a model, with an covariate added for identification, will be referred to as a semi-basic LC model. A drawback of this approach is that it is no longer possible to investigate potential DEs of the

identifying covariate on the indicators. Furthermore, by adding a covariate that has an effect on the latent variable, the interpretation of the LCs changes. The way in which a covariate for identifying the model was found, is described in Section 4.2.1 below.

4.2.1 Covariate for identification

For finding a suitable covariate with which the semi-basic LC model could be identified, findings of Bakker et al. (2021) for respondents aged 15 to 25 were used. Bakker et al. (2021) used an HM model to estimate a latent employment contract type and assessed which covariates were related to the twelve possible inconsistencies between their HM model estimates and the employment contract type as recorded in either the ER or the LFS. The covariates available in this study were similar to those of Bakker et al. (2021). The only difference being that for some covariates, more category levels are distinguished in this study. Therefore, effects found by Bakker et al. (2021) may also be reflected in the datasets at hand. A covariate was searched that was not (or least) related to any inconsistency between the HM model estimate and the employment contract type as recorded in the ER or the LFS. Once a candidate covariate had been found, for the 2016 dataset, it was assessed whether the candidate covariate fitted the aforementioned criteria. (That is, having an effect on the latent variable and having no effect on any of the indicators). For this, a Wald test was used. This statistic gives an estimate for the decrease of the log-likelihood value in the unrestricted model given the constraint that parameters of interest equal zero (Vermunt & Magidson, 2016, p. 83). The identifying covariate should add enough degrees of freedom to make the model identifiable. Otherwise, a different identifying covariate has to be used. Once an identified model was acquired by adding a covariate, potential DEs of all remaining covariates could be examined.

4.2.2 Latent class model fit

Once an identified semi-basic three-class LC model had been found, it was assessed if the model adequately fitted the data by comparing it with a one- and two-class model. If the identifying covariate added enough degrees of freedom, a comparison could also have been made with a four-class model. The models with a different number of classes were compared on BIC, L^2 and BVRs. Note that with two indicators, there is only one BVR (between the two indicators).

4.3 Parameter restrictions

As indicated in Section 4.1.2, certain patterns could not occur due to the way the data was collected and processed. Since the frequencies of these patterns were not directly observed but derived, the estimated effects of the associated covariate category levels were not considered of interest. Therefore, for covariates with derived frequencies, an attempt was made to determine whether evidence for DIF was associated with the derived frequencies only. For these covariates, in addition to models with DEs in which all parameters were free, models were tested with two

different sets of parameter restrictions. The specific parameters that could be restricted were the γ and ζ parameters in Equation (18).

For the first set parameters restriction, denoted by R1, all parameters relating to derived frequencies were set equal to zero. This set of parameter restrictions indicated that the effect of the covariate category in question on the indicator was equal to that of the reference category. If a model with this set of parameter restrictions fitted the data worse than a model without, then the DIF would appear to result mainly from the derived frequencies.

For the second set of parameter restriction, denoted by R2, only parameters that related to the derived frequencies were left free. This set of parameter restrictions indicated that only the effect of the covariate category in question on the indicator was different from that of the reference category. If a model with this set of parameter restrictions fitted the data equally well as a model without, then the DIF would appear to result mainly from the derived frequencies.

4.4 Testing for differential item functioning

The question is whether any of the covariates available are a source of DIF on one of the indicators: *conER* and *conLFS*. To answer this question, it was assessed for each covariate what kind of model best fitted the data. For this, two different methods were used: a stepwise LRT method and an exhaustive BIC method. For each method, models that allowed for various forms of DIF were constructed and compared. (See Appendix 9.6, Figure 2 for a schematic representation of the tested LC models.) The different methods are described in the next sections.

4.4.1 Stepwise likelihood-ratio test method

An approach was used based on the method described by Masyn (2017). They used a stepwise procedure to determine if a particular covariate is a source of DIF for any indicator in an LC model. In this stepwise approach, the one-step approach is mainly used for constructing LC models with covariates (see Section 3.5.1). However, for one of the steps, the three-step approach is also used (see Section 3.5.2). For each covariate of interest, one can perform the stepwise LRT method. However, it is not possible to test for effects of multiple covariates at the same time. Therefore, interactions between different covariates cannot be ruled out. In the stepwise procedure, one starts with an overall test for any evidence of DIF. If any evidence for DIF is found, then with stepwise iterations, one tests for non-uniform DIF and uniform DIF for each indicator. The adapted stepwise LRT method used for this study is described below. For a detailed description of the original method, see Masyn (2017). The original method has been slightly modified. It is indicated at the steps where this is the case.

Step 0: test for the number of LCs. One tests basic LC models with a different number of

LCs without adding covariates. The best-fitting model is selected based on some criteria. For this step, this study deviated from the original method as the aim was to use a three-class LC model as described in Section 4.2. However, it was tested how well the model fits the data as described in Section 4.2.2.

Step 1: overall test for measurement non-invariance. With the one-step approach, one estimates two models with T classes, as determined in *step 0*. One model (M1.0) is the null model in which covariate Z is no source for DIF. In this model, Z has an effect on the latent variable, but no DE on any of the indicators. In the other model (M1.1), Z has an effect on the latent variable and additionally a class-varying DE on all the included indicators. That is, Z is a source for non-uniform DIF for all indicators. The models M1.0 and M1.1 are compared using an LRT.

If the result is non-significant ($p > 0.05$), then one continues with *step 7*. With a non-significant result, one does not reject the null model. There is no sufficient evidence that Z is a source of measurement non-invariance. If the result is significant ($p \leq 0.05$), then one continues with *step 2*. With a significant result, one rejects the null model. There is sufficient evidence that Z is a source of DIF for one or more indicators in one or more LCs.

Step 2: test for non-uniform DIF per indicator. One starts with the unconditional T-class LC model from *step 0*. One saves the modal class assignment W_i for all respondents. (This is the classification for each respondent for which the respondent has the highest posterior class probability.) Based on the estimates from the model selected in *step 0*, one calculates the average modal class assignment error rates, which can be defined as

$$P(W_i = s|X = t) = \frac{P(X_i = t|W_i = s)P(W_i = s)}{P(X_i = t)}. \tag{25}$$

With the three-step approach, one estimates a new T-class model, with W_i as a nominal indicator in the third step. In this model, the class specific multinomial intercepts for W_i are fixed at the estimated values for the error rates as defined in Equation (25). Also included in this model in the third step is the first indicator k and the tested covariate Z as having an effect on the latent variable. This is the null model for *step 2* (M2.0.1), which represents no DIF for indicator k. One also estimates an alternative model in which Z has a class-varying DE on k (M2.1.1). With an LRT, one compares the models M2.0.1 and M2.1.1. This is repeated for all K indicators where M2.0.k is compared with M2.1.k successively.

The results of all LRTs of *step 2* are collected. Based on the results, one then identifies indicators with significant evidence for non-uniform DIF. (Note that this procedure is similar to the ML approach of the three-step approach for a LC model with covariates.)

Step 3: Verify the joint results of *step 2*. With the one-step approach, an LC model is esti-

mated in which all non-uniform DIF effects found in *step 2* are included. The new model (M3.0) is compared using an LRT with model M1.0 and M1.1 to verify that that the new model has a better fit than M1.0 (no DIF) and no worse fit than M1.1 (non-uniform DIF for all indicators). Note that this step is not foolproof. The original method, as described by Masyn (2017), does not indicate what steps to take if M3.0 fits significantly worse than M1.1. If this would be the case, it was chosen to continue with *step 3* with a caveat. Furthermore, this step can be skipped if evidence is found in *step 2* that Z is a source of DIF for all indicators as in this case M3.0 equals M1.1.

Step 4: Test for uniform DIF. For all indicators for which evidence for non-uniform DIF was found in *step 2*, one compares M3.0 using an LRT with a model in which the DE of Z on the indicator is constrained to be class-invariant. That is, Z is a source of uniform DIF. All other DEs of Z on the other indicators are kept as class-varying. The one-step approach is used for estimating all new models. For each comparison, if the result is non-significant, then there is insufficient evidence that there is non-uniform DIF for that indicator. If significant, then the opposite holds true.

Step 5: verify the joint results from *step 4*. With the one-step approach, one estimates an LC model in which all class-varying (non-uniform) DEs from model M3.0 that were found to be non-significant in *step 4* are constrained to be class-invariant (uniform). With an LRT, this new model (M5.0) is then compared with M3.0 to verify that it does not have a significantly worse fit. Note that like *step 3*, *step 5* is not foolproof. The original method does not indicate what steps to take if M5.0 fits significantly worse than M3.0. If this would be the case, it was chosen to continue with a caveat.

Step 5R: test for DIF with restrictions. After *step 5*, the original method as described by Masyn (2017) accepts all found DEs as part of the best-fitting model. For covariates with derived frequencies, this study adds an additional step. In this step, using an LRT, a comparison is made between M5.0 and a model in which parameters that relate to derived frequencies are fixed at zero (M5.R1); and a comparison is made between M5.0 and a model in which parameters that relate to non-derived frequencies are fixed at zero (M5.R2). (See Section 4.3 for details.) The new models are estimated with a one-step approach. The purpose of this step is to deduce if the evidence for DIF found in the previous steps originates from the extreme values of the derived frequencies.

If M5.R1 fits the data as well as M5.0, then the evidence for DIF does not appear to originate from the derived frequencies. If M5.R2 fits the data as well as M5.0, then the evidence for DIF seems to originate from the derived frequencies. If both M5.R1 and M5.R2 fit the data as well as M5.0, then one could join the results in a new model (M5.R3) in which there are no DEs of Z on the indicator for which derived frequencies exist. One could then compare M5.R3 with M5.0

using an LRT. If there is no significant difference, then there is no evidence for DIF for that indicator after all, contradicting findings of *step 2*. This last part indicates that this step is not foolproof either.

Step 6: Evaluate the practical and substantive impact of perceived DIF. One examines the final model of *step 5*. One considers the number of indicators with (uniform or non-uniform) DIF and the magnitude of the DEs of Z. One may determine whether one can substantively compare the LC profile for each value of Z. Only if one deems the profiles to be comparable, then one continues to *step 7*. If one deems that at different values of Z, the LC profiles are not comparable, then one needs to try to interpret the LC formation and corresponding class proportions for different values (or ranges) of Z. Note that in this study, any DIF found is considered to be relevant if the exponent for any of the multinomial regression coefficients of Z on the latent variable is greater than 1.25 or its inverse.

Step 7: Evaluate the association between the covariate and LC membership. One can only do a valid comparison if there is sufficient measurement invariance. One conducts an overall test for an association between Z and the latent variable by comparing two models with an LRT. For the null model of this step (M7.0), all multinomial regression coefficients for Z on the latent variables are fixed at zero. This model is compared with a model where all multinomial regression coefficients of Z on the latent variable are freely estimated (M7.1). Note that if there is any evidence of DIF, M7.1 is the same as M5.0. If there is no evidence for DIF, M7.1 is the same as M1.0. All models are estimated with the one-step approach. If there is no significant difference (M7.0 fits the data as well as M7.1), then there is no evidence for an association between the LC membership and Z, after one accounts for measurement non-invariance due to Z. If there is a significant difference, then one can evaluate the specific parameters and differences between the parameters to identify how membership in each LC is differentiated by values of Z.

In this study, this last step is not of any particular interest. The aim is to solely detect DIF for any indicator and not how the covariate effects class membership. Therefore, this step is left out.

Note that although the entire procedure successively tests simplified models, not all conceivable models are tested. As noted, it is therefore possible that after a certain step, the best found model fits the data worse than a model from a previous step. Also, as the original method by Masyn (2017) does not include a correction for multiple testing, no multiple testing correction is included when applying the method in this study as well.

4.4.2 Exhaustive Bayesian information criterion method

An exhaustive search was used where all possibly considered models (with different DEs from the covariates on the indicators) were constructed for each covariate. As with the stepwise LRT method, for covariates with derived frequencies, additional models with parameter restrictions (as described in Section 4.3) were tested. For covariates without derived frequencies, nine models were created. (For both *conER* and *conLFS*, *no DIF*, *uniform DIF*, and *non-uniform DIF* were possible. This resulted in $3 \times 3 = 9$ possible models.) For covariates with derived frequencies, 21 models were created. (For *conER*, *no DIF*, *uniform DIF (R1)*, *uniform DIF (R2)*, *uniform DIF*, *non-uniform DIF (R1)*, *non-uniform DIF (R2)*, and *non-uniform DIF* were possible. For *conLFS*, *no DIF*, *uniform DIF*, and *non-uniform DIF* are possible. This resulted in $7 \times 3 = 21$ possible models.) All models were estimated with the one-step approach. For each covariate, the model with the lowest BIC was selected as the best-fitting model. Note that when using BIC, non-nested models can be compared and multiple testing corrections are not required.

4.5 Method validation

For the individual datasets of 2016, 2017, and 2018, the stepwise LRT method and the exhaustive BIC method were applied. It was examined whether the results per method were consistent across datasets and whether the results of the different methods were consistent for the same dataset. The former was used to validate the methods. The aim was to unravel whether structural properties of the datasets were observed or whether noise was being modelled. As the inconsistencies between *conER* and *conLFS* appeared to be consistent over time, and the method of observation for the ER and the LFS was unchanged between the starting years of 2016, 2017, and 2018, the same measurement model was expected to apply (Bakker et al., 2021). That is, any covariate effect found should be present in all three datasets. Note that all three datasets have the same recorded covariates and are similar in size (see Section 4.1.2). Ideally, the two methods would find the same best-fitting model for each covariate. This would strengthen the reliability of the results. If the results differ, then similar trends may still be of interest. In Section 5, for illustrative purposes, the results for the 2016 dataset are described in detail. The results for the other years are concisely described thereafter.

4.6 Performing analyses

With RStudio (RStudio Team, 2021, version 1.4.1103) as development environment, a script was made in R (R Core Team, 2022, version 4.1.3) that created a Latent GOLD syntax (LGS) file with which all models of interest for a given covariate could be estimated in Latent GOLD (Vermunt & Magidson, 2016, version 5.1.0.20301). The R script created an LGS file by modifying template text files with the function *brew* from the package *brew* (Horner & Hunt, 2022). The

template files contained all the necessary information for latent GOLD with placeholder fields for the dataset and the specific model properties. The LGS files were opened in Latent GOLD with which all specified models were estimated. In the syntax, the *write* and *append* options were used to instruct Latent GOLD to print results to a comma-separated values (CSV) file. R scripts were also made for finding the best-fitting model with the stepwise LRT method and the exhaustive BIC method. For comparing models, both methods used the already estimated models stored in CSV files. (For a link to an online repository containing the code, see Appendix 9.7.)

Latent GOLD uses the preferred algorithm of Hagenaars and McCutcheon (2002), starting with an EM algorithm and followed with an NR algorithm (see Section 3.3). Note that the user themselves decides the number of iterations and the convergence criteria. In Latent GOLD, one can specify the number of random start values. (In some preliminary analysis, it was found that with the default settings, the ML estimates differed when using different starting values. Therefore, a large number of starting values was used in hopes of finding the global maximum.)

For estimating all models, 3,200 sets with 100 EM iterations and a (default) tolerance of 1e−5 were chosen. (The EM algorithm stops if the tolerance or the maximum number of iterations is reached.) The tolerance criterion used by Latent GOLD is the sum of the absolute value of the relative changes in the parameters (Vermunt & Magidson, 2016, p. 60). (The number of sets indicate how many times this is repeated with random start values.) Thereafter, the best 320 sets (10% of sets with the lowest log-posterior or L^2) are selected with which another 200 EM iterations (twice the chosen number of EM iterations) are performed. The best of these models is then iterated to convergence. For this, another 500 EM iterations were chosen with a (default) tolerance of 1e−2. Lastly 10,000 NR iterations with a (default) tolerance of 1e−8 were chosen for finding parameter estimates. For details on the algorithms used in Latent GOLD, see Vermunt and Magidson (2016). (For an example of how the settings are indicated in an LGS file, see Appendix 9.2.)

To estimate an LC model with categorical covariates, the observed frequencies of all possible patterns in the data are used. If patterns do not occur in the dataset, then those patterns have a frequency of zero which makes estimating the associated parameters difficult. This is because with an iterative process the estimation procedure tries to find parameter combinations that result in probabilities equal to zero for these frequencies. This may result in the model not converging, because the logit parameters will tend to minus infinity to indicate such probabilities. (In addition, the standard errors will become very large.) This problem is also present for the datasets at hand as not all possible patterns occurred (see Section 4.1.2). The problem was mitigated by making use of an option in Latent GOLD that adds a small prior when estimating a model. The addition of this prior can be interpreted as adding a small number of observations for every possible pattern. With it, the model converges more quickly and the parameter esti-

mates become less extreme. For detailed information on the use of a prior in LC modelling, see Vermunt and Magidson (2016). For how to restrict parameters in Latent GOLD as discussed in Section 4.3, see Vermunt and Magidson (2013, pp. 43–46).

5 Results

The results are described in the following order. First the identification and the fit of the semi-basic LC model is described. Thereafter, the results of the application of the stepwise LRT method and the exhaustive BIC method for the 2016, 2017, and 2018 dataset is described. Lastly, a comparison between the results of the different methods for each year is given.

5.1 Latent class model identification

5.1.1 Covariate for identification

The study of Bakker et al. (2021) examined effects of external covariates on inconsistencies between their estimated latent employment contract type, according to an HM model (*conHM*), and the recorded employment contract type for *conER* or *conLFS*. All three distinguished between a *permanent*, *flexible*, and *other* employment contract type. This allowed for six possible inconsistencies between estimate and indicator. With two indicators, this makes for twelve possible inconsistencies. (For example, *conHM: permanent* & *conER: flexible* or *conHM: flexible* & *conLFS: other*.)

From the results of Bakker et al. (2021), a covariate was only considered to have a relevant effect if 1) at least one of the covariate category levels had a significant effect on the inconsistency ($p \leq 0.05$); and 2) the exponent of the coefficient (which represents an odds ratio (OR) with respect to the reference category) of at least one of the covariate category levels was greater than 1.25 or less than its inverse. The covariate was considered irrelevant if one of the conditions was not met. The results from this assessment are in Table 2. Only covariates that were related to all inconsistencies were taken into account. Note that the covariate category levels used for the HM model do not all match the covariate category levels used in this study, but they are similar. The difference being that in this study, more category levels were often distinguished. As can be seen in Table 2, with the criteria used, the most irrelevant covariate appeared to be *gender*. Therefore, this covariate was considered for identifying the model.

Table 2: Covariates examined by Bakker et al. (2021) and the number of times they were considered irrelevant for explaining the inconsistencies between the hidden Markov (HM) model estimates and the recorded employment contract type in the Employment Register or the Labour Force Survey. Distinguished were *permanent*, *flexible*, and *other* employment contracts. With it, there are twelve possible inconsistencies. The first column indicates the assessed covariate (Z); the second column indicates the number of covariate category levels (Ncat); and the third column indicates the number of times the covariate was considered irrelevant for the twelve inconsistencies (irrelevant).

Z	Ncat	irrelevant
ageGro	2	8/12
gender	2	12/12
comSiz	3	8/12
conHou	4	6/12
eduLev	4	9/12
jobDur	6	9/12
migBac	7	5/12
ecoAct	9	3/12

To check whether the potential identifying covariate *gender* itself was not a source of DIF on any of the indicators in the datasets at hand, a model was constructed in which it was allowed to have a DE on both indicators with the one-step approach. By assessing the Wald statistics, it was deduced whether the covariate meets the requirements of having an effect of the latent variable while not having a significant effect on the indicators. Unfortunately, adding the covariate *gender* (with two category levels) to the three-class model at hand without DEs, leaves the model with 0 degrees of freedom. Therefore, it was not possible to test for additional DEs without adding additional covariates to the model. For this, the covariate *ageGro* was also added. This covariate was chosen for several reasons: it used identical category levels in the study by Bakker et al., 2021 as in this study; it appeared to be relatively irrelevant; and it had the fewest number of category levels. In Table 3, one can see that the covariate *gender* had no significant effect on any of the indicators, but it does have a significant effect on the latent variable. The covariate *gender*, therefore, seemed to be suitable for identifying the model. The covariate was added to the LC model, in which it had an effect on the latent variable and no effect on the indicators.

Table 3: Wald test statistics for the overall effects of the covariates *gender* and *ageGro* on the latent variable (*conLat*) and the indicators (*conER* and *conLFS*). Indicated are the terms, the Wald statistic (W), the degrees of freedom (df), and the corresponding p-values (p).

term			W	df	p
conLat	←	gender	20.22	2	< 0.001
conLat	←	ageGro	321.22	2	< 0.001
conER	←	gender	2.77	2	0.250
conLFS	←	gender	4.22	2	0.121

5.1.2 Latent class model fit

Adding the covariate *gender* (with two category levels) made it only just possible to identify a semi-basic LC model with three classes. The resulting model then had 0 degrees of freedom left. Therefore, such a model could only be compared with a one- or two-class model. For a comparison with a model with four of more classes, too few degrees of freedom were available. Since this study was mainly interested in a three-class model, it was left at that.

For assessing the model fit, the 2016 dataset was used. In Table 4, fit measures are given for the one-, two- and three-class model. Of the models examined, the three-class model fits the data best. It has the lowest BIC and the BVR indicates that there is no residual association left. The BVRs of the one- and two-class model indicate that there is a strong dependency left. Of the models examined, the three-class model also has the lowest L^2. Usually, with a positive number of degrees of freedom, an associated significant p-value (for example, $p \leq 0.05$) indicates that the model is rejected in favour of the 'saturated' model. However, for the three-class model, there are 0 degrees of freedom. With it, there is no way to affirm or reject the model as the data cannot freely vary in any way (Eisenhauer, 2008). Therefore, this fit measure was not considered. Lastly, in the three-class model, there appears to be no residual association between the identifying covariate *gender* and any of the indicators.

Table 4: Results for the one-, two- and three-class latent class (LC) model. Shown for each LC model are the number of latent classes (Ncla); the likelihood-ratio chi-square statistic (L^2) with its corresponding degrees of freedom (df) and p-value (p); Bayesian information criterion based on the L^2 (BIC[L^2]); and bivariate residual (BVR).

Ncla	L^2	df	p	BIC(L^2)	BVR
1	10407.90	12	< 0.001	10291.54	2505.96
2	795.59	6	< 0.001	737.41	206.94
3	0.22	0	< 0.001	0.22	0.0001

5.2 Testing for differential item functioning

For all best-fitting models for all years, the relevance of the DEs were assessed based on the parameter estimates. Parameters were considered relevant if they were significant ($p \leq 0.05$) and indicated a substantial effect ($coef \notin [-\log(1.25), \log(1.25)]$). A model was considered overall relevant if at least one parameter of the DEs met these criteria. It was found that almost all models with DIF showed a relevant effect of the tested covariate. The only exceptions were for the best-fitting model according to the stepwise LRT method for *ageGro* with the 2016 dataset; and the best-fitting model according to the stepwise LRT method for *intMan* with the 2017 dataset. Note that these covariates only had two category levels and no derived frequencies.

5.2.1 2016 dataset

For the 2016 dataset, with the Latent GOLD settings as described in Section 3.1, all estimated models converged.

The results for the stepwise LRT method are in Table 5. (For detailed results of the stepwise LRT method for all tested covariates, see Appendix 9.4.)

With the stepwise LRT method, a dead end was reached for *eduLev*. Additionally, when restrictions were included, a dead end was also reached for *jobDur*. For *eduLev*, during *step 2*, no evidence for DIF was found. This contradicts *step 1* in which evidence for some DIF was found. For *jobDur* with restrictions, in *step 5R*, for *conER* both R1 and R2 were found to fit the data as well as a model with uniform DIF for *conER* with all parameters free. Combining R1 and R2 gives a model with no DIF for *conER*. This contradicts *step 2* in which evidence for some DIF for *conER* was found.

For the stepwise LRT method results without considering restrictions, for six of the nine covariates, some evidence for DIF was found. DIF was mainly found for the covariates with derived frequencies. If any DIF was found, it was mainly of the form *uniform DIF* for *conER* and *uniform DIF* for *conLFS*. The only exception being for *ageGro*. For this covariate, DIF was only found for *conER*. Additionally, the type of DIF was non-uniform.

When restrictions were included, for two of the five covariates with derived frequencies (*sofClu* and *ecoAct*), it was found that a model in which parameters were set to zero that related to the derived frequencies (R1) fit the data as well as an unrestricted model. For these covariates, the evidence for DIF seems to originate mainly from the observed frequencies. Therefore, the evidence for DIF does not seem to be an artefact of the manner in which the data was processed. For the covariate *jobDur*, when including restrictions, no substantial evidence for DIF was found. For the remaining covariates, it seems that the DIF stems from both the derived frequencies and

observed frequencies as neither tested restrictions fit the dataset. If these results prove consistent, then all DIF found seems interpretable to some extent.

The results for the exhaustive BIC method are in Table 6. (For detailed results of the exhaustive BIC method for all tested covariates, see Appendix 9.4.)

For the 2016 dataset (without restrictions), the exhaustive BIC method found DIF for four of the nine covariates. All covariates that showed some DIF had derived frequencies. Additionally, DIF was mainly found for *conER*. Only for *comSiz*, DIF was found solely for *conLFS*. When considering restrictions, there were no contradicting results. That is, the best-fitting model with restrictions is always nested in the best-fitting model without restrictions. For *sofClu*, it was found that DIF originates from the observed frequencies (R1) and for *jobDur*, it was found that DIF mainly originated from the derived frequencies (R2).

As a demonstration, for the results of the both the stepwise LRT and the exhaustive BIC method (without restrictions), for the best-fitting models that included DIF, the profiles per category levels are respectively in Tables 14 and 15 in the Appendix. Ideally, for all LCs, the probability $P(Y|X, Z)$ would show a similar trend, that is, the greatest values would be for a single category of the indicator. This would give a consistent interpretation to the classes. Furthermore, each LC would ideally be related to a different category level of the indicator. This would be the one-to-one relationship of the latent employment contract type and the observed employment contract type. This was not always found to be the case. (Note that the LCs are arbitrarily assigned when estimating the model. Therefore, one should not consider the numeric names of the LCs.) It can be seen that DIF is mainly related to the category *other* of *conER* and *conLFS* and with the category *missing/not applicable* for the covariates. However, when trying to account for these covariate and indicator category levels by adding the aforementioned restrictions, substantial evidence for DIF seems to remain.

Method comparison

Agreements between the methods are shown in Table 7. As can be seen, only for three of the nine covariates (*intMan*, *eduLev*, and *migBac*), there is agreement. The agreement specifically being that there is no DIF. Assessing additional restrictions does not lead to more agreement. For all disagreements, the exhaustive BIC method always found less complex DIF than the stepwise LRT method. There was no instance in which the exhaustive BIC method found DIF for an indicator that was not found with the stepwise LRT method. This indicates that the exhaustive BIC method is more conservative than the stepwise LRT method. When including restrictions, there are a few exceptions. For *jobDur*, the BIC method found that the DIF originates from the

derived frequencies (as R2 fits the data best). This contrasts with the results of the stepwise LRT method that ended up contradicting a previous step, as previously mentioned. Additionally, for *ecoAct*, the exhaustive BIC method found evidence for DIF that originated from both the observed and derived frequencies. In contrast, the stepwise LRT method only found substantial evidence for DIF originating from the observed frequencies.

Table 5: Results for the stepwise likelihood-ratio test method (LRT-M). For each covariate (Z), the number of category levels (Ncat) and the best-fitting model is shown. Indicated are the corresponding differential item functioning (DIF) effects of the covariate on the employment contract type as recorded in the Employment Register (*conER*) and the Labour Force Survey (*conLFS*); the log-likelihood (LL); the number of parameters (Npar); and the Bayesian information criterion based on the LL (BIC[LL]). Shown are results when restrictions are (a) excluded and (b) included. *No DIF*, *uniform DIF*, and *non-uniform DIF* are denoted with 0, 1, and 2, respectively. DIF with parameter restrictions is denoted with a superscript. Covariates with derived frequencies are indicated with an asterisk and covariates for which the method reached a dead end are indicated with a superscript *F*.

(a) LRT-M without restrictions.

Z	Ncat	conER	conLFS	LL	Npar	BIC(LL)
ageGro	2	2	0	-26711.24	24	53655.19
intMan	2	0	0	-27275.89	18	54726.31
comSiz*	4	1	1	-16248.80	34	32827.28
eduLev^{F}	4	0	0	-26909.03	22	54031.39
conHou*	5	1	1	-16084.53	40	32556.91
sofClu*	6	1	1	-16132.96	46	32711.96
jobDur*	7	1	1	-15547.86	52	31599.93
migBac	7	0	0	-27128.27	28	54528.05
ecoAct*	10	1	1	-15880.40	70	32439.53

(b) LRT-M with restrictions.

Z	Ncat	ER	LFS	LL	Npar	BIC(LL)
comSiz*	4	1	1	-16248.80	34	32827.28
conHou*	5	1	1	-16084.53	40	32556.91
sofClu*	6	1^{R1}	1	-16133.58	44	32693.81
jobDur*^{F}	7	0	1	-15554.24	40	31496.34
ecoAct*	10	1^{R1}	1	-15880.64	68	32420.63

Table 6: Results for the exhaustive Bayesian information criterion method (exhaustive BIC method) in which all possible models were examined. For each covariate (Z), the number of category levels (Ncat) and the best-fitting model is shown. Indicated are the corresponding differential item functioning (DIF) effects of the covariate the employment contract type as recorded in the Employment Register (*conER*) and the Labour Force Survey (*conLFS*), the log-likelihood (LL), the number of parameters (Npar), and the BIC based on the LL (BIC[LL]). Shown are results when restrictions are (a) excluded and (b) included. *No Dif*, *uniform DIF*, and *non-uniform DIF* are denoted with 0, 1, and 2, respectively. DIF with parameter restrictions is denoted with a superscript. Covariates with derived frequencies are indicated with an asterisk.

(a) BIC-M without restrictions.

Z	Ncat	conER	conLFS	LL	Npar	BIC(LL)
ageGro	2	0	0	-26720.95	18	53616.43
intMan	2	0	0	-27275.89	18	54726.31
comSiz*	4	0	1	-16261.10	28	32793.70
eduLev	4	0	0	-26909.03	22	54031.39
conHou*	5	0	0	-16119.61	24	32471.94
sofClu*	6	1	0	-16144.08	36	32637.22
jobDur*	7	1	0	-15552.96	40	31493.77
migBac	7	0	0	-27128.27	28	54528.05
ecoAct*	10	1	0	-15937.06	52	32378.34

(b) BIC-M with restrictions.

Z	Ncat	conER	conLFS	LL	Npar	BIC(LL)
comSiz*	4	0	1	-16261.10	28	32793.70
conHou*	5	0	0	-16119.61	24	32471.94
sofClu*	6	1^{R1}	0	-16143.42	34	32616.53
jobDur*	7	1^{R2}	0	-15557.85	30	31406.59
ecoAct*	10	1	0	-15937.06	52	32378.34

Table 7: Agreement between the stepwise likelihood-ratio test method (LRT-M) and the exhaustive Bayesian information criterion method (BIC-M). Shown are results when restrictions are (a) excluded and (b) included. Indicated are the tested covariates (Z); the number of covariate category levels (Ncat); the best-fitting models according to the assessed method (LRT-M and BIC-M); and the number of agreements as a fraction of the total number of covariates tested (agreement). Best-fitting models are indicated by A-B, where A and B denote the type of differential item functioning (DIF) for the employment contract type as recorded in the Employment Register and the Labour Force Survey, respectively. For A and B, 0 denotes *no DIF*, 1 denotes *uniform DIF*, and 2 denotes *non-uniform DIF*. An asterisk denotes covariates with derived frequencies. DIF with parameter restrictions are denoted with a superscript with the specific restriction. The data is of respondents aged 15 to 25 that started with the Labour Force Survey in 2016.

(a) Agreement without restrictions

Z	Ncat	LRT-M	BIC-M
ageGro	2	2-0	0-0
intMan	2	0-0	0-0
comSiz*	4	1-1	0-1
eduLev	4	0-0	0-0
conHou*	5	1-1	0-0
sofClu*	6	1-1	1-0
jobDur*	7	1-1	1-0
migBac	7	0-0	0-0
ecoAct*	10	1-1	1-0
agreement			3/9

(b) Agreement with restrictions.

Z	Ncat	LRT-M	BIC-M
comSiz*	4	1-1	0-1
conHou*	5	1-1	0-0
sofClu*	6	1^{R1}-1	1^{R1}-0
jobDur*	7	0-1	1^{R2}-0
ecoAct*	10	1^{R1}-1	1-0
agreement			0/5

5.3 Method validation

With the settings used, all constructed models for the 2017 and 2018 dataset converged. (For the exact settings used in Latent GOLD, see Section 3.1.)

As with the 2016 dataset, the 2017 and 2018 dataset reached a dead end or ended up with a contradiction at some points.

For the 2017 dataset, the stepwise LRT method reached a dead end for the covariate *sofClu* with restrictions. In *step 5R*, both R1 and R2 were found to fit the data as well as a model without restrictions. By combining R1 and R2, one gets a model without DIF, which contradicts *step 2* where some DIF was found. In this case, for the final model, no DIF was selected.

For the 2018 dataset, for *eduLev*, the stepwise LRT method reached a dead end. In *step*

2, a model was found with non-uniform DIF for *conER* only. However, in *step 3*, this model performed significantly worse than a model with uniform DIF for both indicators, contradicting *step 2*. In this case, the steps were continued and a final model with uniform *DIF* for *conER* was found.

The results of the validation analysis examining agreement between the results of different years can be seen in Table 8. The main focus was on agreement between the results of different years using the same method.

For the stepwise LRT method, without restrictions considered, there was agreement for six of the nine covariates over the three years. Disagreement only occurred for covariates without derived frequencies. For the exhaustive BIC method, without restrictions considered, there was agreement for seven of the nine covariates over the years. Disagreement only occurred for covariates with derived frequencies. There is no overlap between covariates with inconsistent results for the stepwise LRT method and the exhaustive BIC method. It can be seen that without considering restrictions, there is less agreement between years for the stepwise LRT method than for the exhaustive BIC method. Although the latter more often finds a model without DIF. Some agreement was found with no DIF and uniform DIF, but none for non-uniform DIF. Although, as seen in Table 5 and 6, non-uniform DIF was rarely found to begin with.

When restrictions are considered, agreement is lower for both methods. For both the stepwise LRT method and the exhaustive BIC method, for covariates that had consistent results without considering restrictions, results with restrictions considered are far less consistent. For both methods, there was only one non-overlapping covariate that was consistent over the three years. This was *comSiz* for stepwise LRT method and *jobDur* for exhaustive BIC method.

There is little agreement between the stepwise LRT method and exhaustive BIC method of the same year. Only for the covariate *migBac*, where no DIF was found, there is full agreement between the methods and the years.

For the 2017 dataset, when restrictions were not considered, there was agreement between methods for three of the nine covariates (*ageGro*, *eduLev*, and *migBac*). As for the 2016 dataset, there was agreement only for a model with no DIF. For all disagreements, all best-fitting models according to the exhaustive BIC method were nested in the best-fitting models according to the stepwise LRT method. Non-uniform DIF was never found.

When restrictions were considered, there was agreement between the methods in only one instance. Only for the covariate *sofClu*, there was agreement where a best-fitting model was found with no restrictions. For best-fitting models that included restrictions, the best-fitting

models of the exhaustive BIC method were not always nested in the best-fitting models of the stepwise LRT method.

Note that with the stepwise LRT method, it was found that DIF caused by *conHou* and *jobDur* seemed to originate from the observed frequencies (R1). In contrast, with the exhaustive BIC method it seemed to originate from the derived frequencies (R2). For *ecoAct*, only with the exhaustive BIC method the best-fitting model included restrictions.

For the 2018 dataset, without considering restrictions, there was only one agreement between methods for the covariate *migBac*, where no DIF was found. As with the 2016 and 2017 dataset, best-fitting models of the exhaustive BIC method were always nested in best-fitting models of the stepwise LRT method. When restrictions were considered, all models were nested also, except for *jobDur*. For this covariate, the BIC indicated that the DIF originated from the derived frequencies (R2) while the stepwise LRT method indicated that the DIF originated from the observed frequencies (R1).

Table 8: Agreement between the 2016, 2017, and 2018 results for the stepwise likelihood-ratio test method (LRT-M) and the exhaustive Bayesian information criterion method (BIC-M). Shown are the results for (a) the LRT-M without restrictions; (b) the BIC-M without restrictions; (c) the LRT-M with restrictions; and (d) the BIC-M with restrictions. Indicated are the tested covariates (Z), the number of covariate category levels (Ncat), the best-fitting model for each year (best-fitting model), and the number of agreements as a fraction of the total number of covariates tested (agreement *year*). The best-fitting models are indicated by A-B, where A and B denote the type of differential item functioning (DIF) for the employment contract type as recorded in the Employment Register and the Labour Force Survey, respectively. For A and B, 0 denotes *no DIF*, 1 denotes *uniform DIF*, and 2 denotes *non-uniform DIF*. Covariates with derived frequencies are denoted with an asterisk. DIF with parameter restrictions is denoted with a superscript. Disagreements are indicated with bold text.

(a) LRT-M agreement without restrictions.

		best-fitting model		
Z	Ncat	2016	2017	2018
ageGro	2	**2-0**	**0-0**	**1-1**
intMan	2	**0-0**	**0-1**	**2-0**
comSiz*	4	1-1	1-1	1-1
eduLev	4	**0-0**	**0-0**	**1-0**
conHou*	5	1-1	1-1	1-1
sofClu*	6	1-1	1-1	1-1
jobDur*	7	1-1	1-1	1-1
migBac	7	0-0	0-0	0-0
ecoAct*	10	1-1	1-1	1-1
agreement 2016		.	.	.
agreement 2017		7/9	.	.
agreement 2018		6/9	7/9	.

(b) BIC-M agreement without restrictions.

		best-fitting model		
Z	Ncat	2016	2017	2018
ageGro	2	0-0	0-0	0-0
intMan	2	0-0	0-0	0-0
comSiz*	4	**0-1**	**0-1**	**1-1**
eduLev	4	0-0	0-0	0-0
conHou*	5	0-0	0-0	0-0
sofClu*	6	**1-0**	**0-1**	**0-1**
jobDur*	7	1-0	1-0	1-0
migBac	7	0-0	0-0	0-0
ecoAct*	10	1-0	1-0	1-0
agreement 2016		.	.	.
agreement 2017		8/9	.	.
agreement 2018		7/9	8/9	.

(c) LRT-M agreement with restrictions.

		best-fitting model		
Z	Ncat	2016	2017	2018
comSiz*	4	1-1	1-1	1-1
conHou*	5	**1-1**	**1^{R1}-1**	**1-1**
sofClu*	6	**1^{R1}-1**	**0-1**	**1^{R2}-1**
jobDur*	7	**0-1**	**1^{R1}-1**	**1^{R1}-1**
ecoAct*	10	**1^{R1}-1**	**1-1**	**1-1**
agreement 2016		.	.	.
agreement 2017		1/5	.	.
agreement 2018		2/5	3/5	.

(d) BIC-M agreement with restrictions.

		best-fitting model		
Z	Ncat	2016	2017	2018
comSiz*	4	**0-1**	**0-1**	**1^{R2}-1**
conHou*	5	**0-0**	**1^{R2}-0**	**1^{R2}-0**
sofClu*	6	**1^{R1}-0**	**0-1**	**1^{R2}-1**
jobDur*	7	1^{R2}-0	1^{R2}-0	1^{R2}-0
ecoAct*	10	**1-0**	**1^{R2}-1**	**1^{R2}-1**
agreement 2016		.	.	.
agreement 2017		2/5	.	.
agreement 2018		1/5	3/5	.

For disagreements with the exhaustive BIC method, examined was whether the best-fitting models from different years were substantially different when estimated on the same data. The difference in BIC was evaluated with the criteria of Raftery (1995, p. 139). They consider a difference in BIC of 0–2, 2–6, 6–10, and > 10 to be *weak*, *positive*, *strong*, and *very strong* evidence, respectively. For *weak* evidence, the models are considered to fit the data equally well. A similar comparison was not made for the stepwise LRT method as the method did not use BIC to select a model. Also, the best-fitting models could not be compared with an LRT as they did not always meet the requirement of being nested. In Table 9, one can see that without considering restrictions, the best-fitting models always had *positive* or greater evidence of having a better fit on the data. When considering the assessed restrictions, it was nearly always the case as well.

Table 9: Difference in Bayesian information criterion (BIC) value of the best-fitting model of one year and the best-fitting model of another year, where the models are estimated on the same data. The best fitting models were found with the exhaustive BIC method. Indicated are the tested covariates (Z); the number of covariate category levels (Ncat); the data with which the comparison was made (*year* data); the data with which the model was found against which the best-fitting model is compared (*year* mod); and the difference in BIC value between the models (BIC difference). Where the same best-fitting model was found for the compared years, a BIC value of zero is shown.

(a) BIC difference without restrictions.

		BIC difference					
		2016 data		2017 data		2018 data	
Z	Ncat	2017 mod	2018 mod	2016 mod	2018 mod	2016 mod	2017 mod
ageGro	2	0	0	0	0	0	0
intMan	2	0	0	0	0	0	0
comSiz	4	0	33.58	0	25.01	9.52	9.52
eduLev	4	0	0	0	0	0	0
conHou	5	0	0	0	0	0	0
sofClu	6	3.63	3.63	20.59	0	22.40	0
jobDur	7	0	0	0	0	0	0
migBac	7	0	0	0	0	0	0
ecoAct	10	0	0	0	0	0	0

(b) BIC difference with restrictions.

		BIC difference					
		2016 data		2017 data		2018 data	
Z	Ncat	2017 mod	2018 mod	2016 mod	2018 mod	2016 mod	2017 mod
comSiz	4	0	10.26	0	2.72	26.20	26.20
conHou	5	7.80	7.83	14.41	0	19.37	0
sofClu	6	24.30	35.35	1.10	13.16	7.33	1.87
jobDur	7	0	0	0	0	0	0
ecoAct	10	3.18	3.20	59.53	0	86.22	0

6 Simulation Study

In addition to the real data, the performance of the methods was assessed with a simulation study. For this, simulated datasets with specified DIF relationships were generated. The examined methods were used to find the true DIF relationships of these datasets. The reliability of the methods was assessed on the basis of the extent to which the pre-specified DIF relations were found. Tested were various forms of DIF (uniform and non-uniform), sample sizes, and effect sizes. First, the design of the simulation study is described. Thereafter, the results of the simulation study are presented.

6.1 Simulation design

Simulated data was generated with Latent GOLD, with which one can generate simulated data by specifying an LC model with all associated logit parameter values. (For how to generate simulated data using Latent GOLD, see Vermunt and Magidson [2013, pp. 16–17, 31–32, 82–86].) In simulated data from Latent GOLD, all covariate category levels occur equally often.

The simulated data was given the same structure as the real data. The semi-basic LC model included two nominal indicators: Y_1 and Y_2, each with three category levels. In addition, the model included a nominal identifying covariate I with two category levels. (Note that the number of category levels was chosen to be the same as used for the real data.) As with the real data, the identifying covariate was added to make the model identifiable (see Section 4.2 for details). A nominal covariate Z with two category levels was added to the semi-basic model. This covariate could have a DE on one or more indicators. (For example, a uniform DIF effect on Y_1 and a non-uniform DIF effect on Y_2. A particular combination of DEs is also referred to as a DIF relationship.) As with the real data, dummy coding was used in which the logit parameters were relative to the indicated reference category levels.

For all logit parameters related to $P(X_i = t)$ and $P(X_i = t|I_i)$, the parameters of the semi-basic LC model estimated from the real data were used. (See Equation (7) and Equation (14) respectively for the equations in which these logit parameters are related to the aforementioned probabilities. See Appendix 9.2 for the logit parameter values.) For the other parameters, a small, medium, and large effect was considered. For these effects, logit parameters equal to $\log(1.25)$, $\log(2)$, and $\log(5)$ were chosen respectively. These logit parameters give an OR of 1.25, 2, and 5 respectively with respect to the reference category. Lastly, no effect was represented with a parameter value of 0, which gives an OR of 1 with respect to the reference category.

The covariate Z, representing a covariate that could be a source of DIF, was chosen with two category levels to limit the number of parameters to be specified. With three latent classes,

three category levels of an indicator Y, two category levels of covariate Z, and dummy coding, for a main effect of Z on Y (as with uniform DIF), there are a total of $3 \times 2 = 6$ parameters. Two are free and four refer to reference category levels (equal to 0). To specify these parameter values in a simple way, one can repeat the parameter value with an inverted sign. For example, for a small effect, one gets $\log(1.25)$ and $-\log(1.25)$, which gives an OR of 1.25 and 0.8 relative to the reference category respectively. For a main and interaction effect of Z on Y (as with non-uniform DIF), there are a total of $3 \times 2 \times 2 = 18$ parameters. Six are free and twelve refer to reference category levels. To specify these parameter values in a simple way, one can repeat the parameter value, followed by two 0 parameter values, and then repeat the parameter values two more times with an inverted sign. For example, for a medium effect, one gets $\log(2)$, $\log(2)$, 0, 0, $-\log(2)$, and $-\log(2)$. This gives an OR of 2, 2, 1, 1, 0.5, and 0.5 with respect to the reference category respectively.

Although many other parameter combinations are conceivable, investigating combination-specific effects is beyond the scope of this study. With more category levels for covariate Z, the number of parameters to choose from increases rapidly, raising the question of whether different parameter combinations yield different results. For this reason, this study limits itself to the simplest situation in which covariate Z has only two category levels.

The covariate Z was chosen to have a small effect on the latent variable X. That is, the OR of $P(X = 2|Z = 2)$ with respect to $P(X = 1|Z = 2)$ and $P(X = 2|Z = 1)$ with respect to $P(X = 1|Z = 1)$ is 1.25. Likewise, the OR of $P(X = 3|Z = 2)$ with respect to $P(X = 1|Z = 2)$ and $P(X = 3|Z = 1)$ with respect to $P(X = 1|Z = 1)$ is 0.8. (Note that there are six parameters in total, of which two are free when using dummy coding.)

The classification error probabilities (or class separation level) for both indicators was chosen to be 0.1. Meaning that the probability to be correctly classified was a 0.8 for both indicators. For the specific parameter values used to achieve this, see Appendix 9.2. The idea is to have a low classification error probability, similar to that of the real data. (See Appendix 9.3 for the classification error probabilities of the indicators *conER* and *conLFS* of the real data.)

For the simulation study, sample size (*samSiz*), covariate effect (*covEff*), and effect size (*effSize*) were varied. These are referred to as simulation covariates. For the sample size, $N = 2{,}000$ (*2k*) and $N = 20{,}000$ (*20k*) were used. The former was considered to be small and the latter was considered to be moderate in size (in the context of data handled by CBS). Also, the latter is of similar size to the real data. For the covariate effect, all unique combinations of *no*, *uniform*, and *non-uniform DIF* were used. These were 0-0, 1-0, 2-0, 1-1, 1-2 and 2-2, where 0 denotes *no DIF*, 1 denotes *uniform DIF*, and *2* denotes *non-uniform DIF*. In the representation *A-B*, *A* denotes the first indicator (Y_1) and *B* denotes the second indicator (Y_2). For the effect size, *none*

(*N*), *small* (*S*), *medium* (*M*), and *large* (*L*) were used, for which the aforementioned parameter values are used.

Per condition, ten replicate simulated datasets were generated. More were considered, but due to run time constrains, not feasible. For each replication, the stepwise conditional LRT and the exhaustive BIC approach were used for finding the best-fitting model. As the data was readily available, an additional exhaustive Akaike information criterion (AIC) approach was used, using the same procedure as the exhaustive BIC approach. The outcome variable was the best-fitting model according to one of the three methods. The goal was to discover to what extent the applied methods could deduce the true DIF relationships.

Finally, an additional analysis was performed on the results of the simulation study. For this analysis, the method used for finding a best-fitting model was also included as a simulation covariate (*method*). The aim was to discover which simulation covariates (*samSiz*, *covEff*, *effSiz* and *method*) were significant in explaining the results. For *effSiz*, the category *N* only existed for simulated datasets that did not contain DEs, while the category levels *S*, *M*, and *L* only existed for simulated datasets that did contain DEs. As this complicates comparison, cases where the simulated dataset did not contain DEs were omitted from this analysis. Thus, in all remaining cases, DEs were present in the simulated datasets. For the analysis, an outcome variable was defined that indicated what kind of best-fitting model was found (*found*). A distinction was made between three category levels: the best-fitting model includes no DEs (*none*); the best-fitting model includes DEs that do not all match the DEs of the simulated data (*incorrect*); and the best-fitting model includes DEs that all match the simulated data (*correct*). Only the presence of the relationships found was important, not the effect sizes. The analysis was performed using a multinomial logistic regression model.

For the analysis, a model was created in R with the function *multinom* from the package *nnet*. The best-fitting model was found using both forward and backward selection with both AIC and BIC as selection criteria. The stepwise model selection with AIC as selection criterion was performed with the *step* function from the *stats* package. The stepwise model selection with BIC as selection criterion was performed manually. (Although the *step* package has options for performing stepwise model selection with BIC as selection criterion for *lm* and *glm* objects from the *stats* package, the option is missing for *multinom* objects.)

6.2 Simulation results

The results of the simulation study for finding the best-fitting model using the stepwise LRT, exhaustive BIC, and exhaustive AIC method are shown in Tables 10a, 10b, and 10c, respectively. One can see that the relatively simple DIF relations were occasionally found. The more complex ones, on the other hand, were rarely if ever found. For datasets in which the covariate was not a

source of DIF on any of the indicators, the stepwise LRT and the exhaustive BIC method always found the correct model. For the exhaustive AIC method, out of ten cases, this was only the case in eight and seven for a dataset with 2,000 and 20,000 cases, respectively. In the remaining cases, the exhaustive AIC method found non-existent uniform DIF for one of the indicators. However, of the three methods, the exhaustive AIC method was best able to retrieve the most complex DIF relationships, although rarely.

Table 10: Best-fitting models for simulated datasets according to (a) the stepwise likelihood-ratio test (LRT) method; (b) the exhaustive Bayesian information criterion (BIC) method; and (c) the exhaustive Akaike information criterion (AIC) method. Data was simulated with different parameter combinations and sample sizes from a model with two indicators (Y_1 and Y_2), one identifying covariate (I), and one covariate of interest (Z). The sample size (*samSiz*); direct effects (DEs) of the covariate on the indicators (*covEff*); and effect size (*effSiz*) are indicated for each simulated dataset in the first three columns. Examined sample sizes were 2,000 (*2k*) and 20,000 (*20k*). For the abbreviations of the form *A*-*B*, *A* denotes the DEs of Z on Y_1 and *B* denotes the DEs of Z on Y_2. Z could have no DE, a uniform DE, or a non-uniform DE on Y_1 or Y_2. These DEs are represented by 0, 1, and 2, respectively. Examined effect sizes were small (S), medium (M), large (L), and none (N). For each unique combination of *samSiz*, *covEff*, and *effSiz*, there were ten replications. The last eight columns indicate for each combination how often a particular best-fitting model was found. Cells where the best-fitting model matched the model that generated the simulated data are highlighted with a grey background. For each combination, models that were never found to be the best fit are indicated with grey text.

(a) Best-fitting models for simulated datasets according to the stepwise LRT method.

simulated data			best-fitting model								
samSiz	covEff	effSiz	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2
2k	0-0	N	10	0	0	0	0	0	0	0	0
2k	1-0	S	10	0	0	0	0	0	0	0	0
2k	1-0	M	9	0	0	1	0	0	0	0	0
2k	1-0	L	0	0	0	1	9	0	0	0	0
2k	1-1	S	10	0	0	0	0	0	0	0	0
2k	1-1	M	10	0	0	0	0	0	0	0	0
2k	1-1	L	10	0	0	0	0	0	0	0	0
2k	2-0	S	10	0	0	0	0	0	0	0	0
2k	2-0	M	10	0	0	0	0	0	0	0	0
2k	2-0	L	0	2	0	2	2	0	4	0	0
2k	2-1	S	10	0	0	0	0	0	0	0	0
2k	2-1	M	9	0	0	0	1	0	0	0	0
2k	2-1	L	0	1	0	1	8	0	0	0	0
2k	2-2	S	10	0	0	0	0	0	0	0	0
2k	2-2	M	10	0	0	0	0	0	0	0	0
2k	2-2	L	2	0	2	0	5	0	1	0	0
20k	0-0	N	10	0	0	0	0	0	0	0	0
20k	1-0	S	10	0	0	0	0	0	0	0	0
20k	1-0	M	0	0	0	0	10	0	0	0	0
20k	1-0	L	0	0	0	0	10	0	0	0	0
20k	1-1	S	10	0	0	0	0	0	0	0	0
20k	1-1	M	8	0	0	0	2	0	0	0	0
20k	1-1	L	1	0	0	0	9	0	0	0	0
20k	2-0	S	9	0	0	1	0	0	0	0	0
20k	2-0	M	0	1	1	0	7	0	1	0	0
20k	2-0	L	0	0	2	0	5	0	0	3	0
20k	2-1	S	9	0	0	1	0	0	0	0	0
20k	2-1	M	0	0	0	0	10	0	0	0	0
20k	2-1	L	0	0	0	0	10	0	0	0	0
20k	2-2	S	6	1	0	0	0	0	3	0	0
20k	2-2	M	0	0	1	0	8	0	0	1	0
20k	2-2	L	0	0	0	0	10	0	0	0	0

(b) Best-fitting models for simulated datasets according to the exhaustive BIC method.

simulated data			best-fitting model								
samSiz	covEff	effSiz	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2
2k	0-0	N	10	0	0	0	0	0	0	0	0
2k	1-0	S	10	0	0	0	0	0	0	0	0
2k	1-0	M	9	0	0	1	0	0	0	0	0
2k	1-0	L	0	4	0	6	0	0	0	0	0
2k	1-1	S	10	0	0	0	0	0	0	0	0
2k	1-1	M	10	0	0	0	0	0	0	0	0
2k	1-1	L	10	0	0	0	0	0	0	0	0
2k	2-0	S	10	0	0	0	0	0	0	0	0
2k	2-0	M	10	0	0	0	0	0	0	0	0
2k	2-0	L	0	9	0	0	1	0	0	0	0
2k	2-1	S	10	0	0	0	0	0	0	0	0
2k	2-1	M	8	1	0	1	0	0	0	0	0
2k	2-1	L	0	3	0	7	0	0	0	0	0
2k	2-2	S	10	0	0	0	0	0	0	0	0
2k	2-2	M	10	0	0	0	0	0	0	0	0
2k	2-2	L	7	1	0	0	2	0	0	0	0
20k	0-0	N	10	0	0	0	0	0	0	0	0
20k	1-0	S	10	0	0	0	0	0	0	0	0
20k	1-0	M	0	2	0	8	0	0	0	0	0
20k	1-0	L	0	0	0	10	0	0	0	0	0
20k	1-1	S	10	0	0	0	0	0	0	0	0
20k	1-1	M	10	0	0	0	0	0	0	0	0
20k	1-1	L	10	0	0	0	0	0	0	0	0
20k	2-0	S	10	0	0	0	0	0	0	0	0
20k	2-0	M	0	8	0	1	1	0	0	0	0
20k	2-0	L	0	0	0	0	10	0	0	0	0
20k	2-1	S	9	0	0	1	0	0	0	0	0
20k	2-1	M	0	7	0	3	0	0	0	0	0
20k	2-1	L	0	0	0	9	1	0	0	0	0
20k	2-2	S	10	0	0	0	0	0	0	0	0
20k	2-2	M	0	0	0	0	10	0	0	0	0
20k	2-2	L	0	0	0	0	10	0	0	0	0

(c) Best-fitting models for simulated datasets according to the exhaustive AIC method.

simulated data			best-fitting model								
samSiz	covEff	effSiz	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2
2k	0-0	N	8	0	0	2	0	0	0	0	0
2k	1-0	S	6	2	0	2	0	0	0	0	0
2k	1-0	M	1	4	0	5	0	0	0	0	0
2k	1-0	L	0	3	0	6	1	0	0	0	0
2k	1-1	S	7	2	0	1	0	0	0	0	0
2k	1-1	M	7	0	0	0	3	0	0	0	0
2k	1-1	L	5	0	0	0	5	0	0	0	0
2k	2-0	S	9	1	0	0	0	0	0	0	0
2k	2-0	M	2	2	0	4	2	0	0	0	0
2k	2-0	L	0	3	0	0	7	0	0	0	0
2k	2-1	S	7	0	0	2	1	0	0	0	0
2k	2-1	M	1	3	0	6	0	0	0	0	0
2k	2-1	L	0	3	0	7	0	0	0	0	0
2k	2-2	S	9	0	0	1	0	0	0	0	0
2k	2-2	M	1	0	0	2	7	0	0	0	0
2k	2-2	L	0	1	0	0	9	0	0	0	0
20k	0-0	N	7	2	0	0	1	0	0	0	0
20k	1-0	S	0	5	0	5	0	0	0	0	0
20k	1-0	M	0	1	0	7	2	0	0	0	0
20k	1-0	L	0	0	0	7	3	0	0	0	0
20k	1-1	S	8	1	0	0	1	0	0	0	0
20k	1-1	M	3	0	0	0	7	0	0	0	0
20k	1-1	L	0	0	0	0	10	0	0	0	0
20k	2-0	S	2	2	0	3	3	0	0	0	0
20k	2-0	M	0	0	1	0	8	0	1	0	0
20k	2-0	L	0	0	1	0	4	0	4	1	0
20k	2-1	S	1	4	0	4	1	0	0	0	0
20k	2-1	M	0	3	1	2	4	0	0	0	0
20k	2-1	L	0	0	0	1	5	0	0	4	0
20k	2-2	S	0	0	0	2	8	0	0	0	0
20k	2-2	M	0	0	0	0	10	0	0	0	0
20k	2-2	L	0	0	0	0	9	0	0	1	0

6.2.1 Simulation covariate analysis

With forward and backward selection and with AIC and BIC as selection criteria, it was found that all explanatory simulation covariates were important for finding the best-fitting model. (See Appendix 9.5 for the detailed results of the stepwise analyses.) The results of the analysis in which all simulation covariates were included are shown in Table 11. Note that some DIF was present in all simulated datasets used in this analysis. It can be seen that as the sample size increases, the odds of finding incorrect or correct DIF increases, with the odds of finding correct DIF increasing the most. (In detail, the ratio of the odds of finding incorrect DIF with respect to finding no DIF for a sample size of 20,000 with respect to a sample size of 2,000 was $\exp(4.33) = 75.94$; and the ratio of the odds of finding correct DIF with respect to finding no DIF for a sample size of 20,000 with respect to a sample size of 2,000 was $\exp(4.47) = 114.43$.) Furthermore, it can be seen that as the complexity of DIF increases, the odds of finding incorrect and correct DIF generally decrease. (Complexity here refers to the number of parameters used to specify the DEs.) In general, the odds of finding correct DIF decreases the most. It can also be seen that as the effect size increases, the odds of finding incorrect DIF and correct DIF increase, with the latter increasing the most. Finally, for finding both incorrect and correct DIF with respect to finding no DIF, the exhaustive AIC method had the highest odds. This was followed by the stepwise LRT method, and the exhaustive BIC method had the lowest odds. (In detail, the ratio of the odds for finding incorrect DIF with respect to finding no DIF for the exhaustive BIC method with respect to the stepwise LRT method was $\exp(-1.12) = 0.33$; and the ratio of the odds for finding incorrect DIF with respect to finding no DIF for the exhaustive AIC method with respect to the stepwise LRT method was $\exp(4.11) = 60.94$. Furthermore, the ratio of the odds for finding correct DIF with respect to finding no DIF for the exhaustive BIC method with respect to the stepwise LRT method was $\exp(-0.24) = 0.79$; and the ratio of the odds for finding correct DIF with respect to finding no DIF for the exhaustive AIC method with respect to the LRT method was $\exp(5.59) = 267.74$.)

Table 11: Summary of the model in which all simulation covariates are included that explain which covariates are related to the type of best-fitting model found. Indicated are the terms (term), coefficient values (coef), standard errors (SE), and p-values (p). With dummy coding, the reference category levels have coefficient values of 0.

term			coef	SE	p
found = none	←	intercept	0	.	.
found = incorrect	←	intercept	-5.74	0.53	< 0.001
found = correct	←	intercept	-7.69	0.73	< 0.001
found = none	←	samSiz = 2k	0	.	.
found = incorrect	←	samSiz = 2k	0	.	.
found = correct	←	samSiz = 2k	0	.	.
found = none	←	samSiz = 20k	0	.	.
found = incorrect	←	samSiz = 20k	4.33	0.40	< 0.001
found = correct	←	samSiz = 20k	4.74	0.47	< 0.001
found = none	←	covEff = 1-0	0	.	.
found = incorrect	←	covEff = 1-0	0	.	.
found = correct	←	covEff = 1-0	0	.	.
found = none	←	covEff = 1-1	0	.	.
found = incorrect	←	covEff = 1-1	-7.39	0.77	< 0.001
found = correct	←	covEff = 1-1	-5.59	0.63	< 0.001
found = none	←	covEff = 2-0	0	.	.
found = incorrect	←	covEff = 2-0	-0.17	0.43	0.684
found = correct	←	covEff = 2-0	-2.69	0.56	< 0.001
found = none	←	covEff = 2-1	0	.	.
found = incorrect	←	covEff = 2-1	0.53	0.43	0.216
found = correct	←	covEff = 2-1	-3.02	0.68	< 0.001
found = none	←	covEff = 2-2	0		
found = incorrect	←	covEff = 2-2	-0.37	0.43	0.398
found = correct	←	covEff = 2-2	-11.48	22.01	0.602
found = none	←	effSiz = S	0	.	.
found = incorrect	←	effSiz = S	0	.	.
found = correct	←	effSiz = S	0	.	.
found = none	←	effSiz = M	0	.	.
found = incorrect	←	effSiz = M	4.25	0.41	< 0.001
found = correct	←	effSiz = M	5.06	0.61	< 0.001
found = none	←	effSiz = L	0	.	.

Continued on next page

Table 11 – continued from previous page

term			coef	SE	p
found = incorrect	←	effSiz = L	7.77	0.57	< 0.001
found = correct	←	effSiz = L	8.89	0.72	< 0.001
found = none	←	method = LRT	0	.	.
found = incorrect	←	method = LRT	0	.	.
found = correct	←	method = LRT	0	.	.
found = none	←	method = BIC	0	.	.
found = incorrect	←	method = BIC	-1.12	0.33	< 0.001
found = correct	←	method = BIC	-0.24	0.45	0.598
found = none	←	method = AIC	0	.	.
found = incorrect	←	method = AIC	4.11	0.45	< 0.001
found = correct	←	method = AIC	5.59	0.55	< 0.001

7 Discussion

There exists an unexplained inconsistency between estimates for employment contract type frequencies based on the ER and the LFS. The ER and the LFS are used, among other things, to determine a respondent's employment contract type, but they do not necessarily reflect their true employment contract type. The true employment contract type is a latent variable and the ER and the LFS are indicators of this latent variable. The purpose of this study was to assess if there is a difference in measured concept for employment contract type in the ER and the LFS. Hypothesised is that the measured concept differs which results in different estimates for employment contract type frequencies. In addition to the latent true employment contract type, there may exist external covariates that affect the measured employment contract type in one or more sources. Effects of such covariates lead to systematic errors in the measured employment contract type. If so, there is measurement non-invariance for the indicators and the covariate is a source of DIF. Evidence for DIF indicates a difference in measured concept. However, a lack of evidence cannot rule out DIF.

Using LC models, the aim was to test the two indicators for DIF. For this, ten covariates were available of which nine could be assessed. The assessed covariates were: *ageGro*, *intMan*, *comSiz*, *eduLev*, *conHou*, *sofClu*, *jobDur*, *migBac*, and *ecoAct*. For each individual covariate, DIF was tested for using two methods: a stepwise LRT method and an exhaustive BIC method. In each method, models were compared that allowed for different types of DEs of the tested covariate on the indicators. A best-fitting model was selected according to each method. The methods were validated by applying them on three datasets and comparing agreement between the results. These was a 2016, 2017, and 2018 dataset, in which the name denoted the starting year of the LFS. Assumed was that the data for the different years had the same structure. Some combinations of covariate and indicator category levels (patterns) were impossible to observe due to the way the data was processed. For such patterns the related frequencies were not observed but derived. This resulted in a skewed distribution in which certain patterns did not occur while others occurred with a high frequency. To assess whether any DIF found related to derived frequencies, additional models were assessed with parameter restrictions.

On a few occasions, the stepwise LRT method reached a dead end in which the algorithm could not advance or gave contradictory results. This questions the applicability of the method as it was found to not be foolproof. Dead ends were occasionally reached in *step 2*, *step 3* and *step 5R*. Although *step 5R* was inspired by the approach of Masyn (2017) by doing separate tests and pooling the results (as in *step 2* and *step 4*), it was not described in the original method. Thus, these dead ends cannot be attributed to the original method. In all other cases, the original approach did not include a way for dealing with dead ends.

For the three years, the results of the stepwise LRT method and the exhaustive BIC method

were not entirely consistent. Using the stepwise LRT method, three of the nine covariates had inconsistent results for the three years. Using the exhaustive BIC method, two of the nine covariates had inconsistent results for the years. Best-fitting models for covariates with consistent results using the same method for all years tested included no DIF and uniform DIF. Non-uniform DIF was never found to consistently fit the datasets. Nearly all best-fitting models with DIF were considered relevant based on the parameters of the DEs of the covariate on the latent variable. For this assessment, among other things, a threshold value of an OR of 1.25 was used. (This translated to relevant logit parameter values outside the interval $[-\log(1.25), \log(1.25)]$.) Only in two cases was the found DIF considered irrelevant. These cases involved covariates with only two category levels. This hints at there being a relationship between the number of category levels and the relevance of the DEs as defined in this study. For covariates with inconsistent best-fitting models over the years according to the exhaustive BIC method, the best-fitting models were considered to differ substantially in fit based on the difference in BIC. This questions the reliability of the results for the different covariates. There was little agreement between the results of the two methods for the same year. Noticeable was that the best-fitting model according to the exhaustive BIC method was either the same as or nested in the best-fitting model according to the stepwise LRT method. This indicates that the exhaustive BIC method is more conservative, overall. For both the 2016 and 2017 dataset, there was agreement between the methods for three of the nine covariates. The only agreements being that there was no DIF. For the 2018 dataset, there was agreement between the methods for two of the nine covariates. One covariate was found to be no source of DIF, while another was found to be a source of uniform DIF on both indicators. When simultaneously assessing the results of the different methods for the different years, there was only agreement for one covariate. This was *migBac* that was found to be no source of DIF.

For best-fitting models, profiles per category level showed that in some cases, based on the value of the covariate, there was a strong difference in the probability that a certain value for an indicator occurred. This could provide insight into the effects of DIF. However, as the covariates effects found were not fully consistent and had little agreement, such interpretations were left out. Ideally, in each profile, there would be an obvious relationship between a certain employment contract type according to the indicators and a single LC. Such a relationship would indicate that the LC represents that specific true employment contract type. For some best-fitting models that included DIF, it was found that some LCs did not strongly relate to a single employment contract type. In these cases, the one-to-one relationship between the different employment contract types and the LCs did not seem to hold. This makes interpreting the LCs as real employment contract types difficult.

When adding restrictions to the model to assess if the derived frequencies were a source of DIF, the results were inconsistent as well. For the results over the years, there was very little

agreement on the fit of the restrictions used. Therefore, the results are inconclusive in showing where potential DIF stems from. This seems to indicate that the restrictions are not well able to indicate the cause of DIF. If the restrictions are omitted, there is far more agreement for each method. All in all, no satisfactory way has been found to deal with the derived frequencies and it remains an open question. As there is less agreement when considering restrictions, it implies that the more models are considered, the less likely the methods assessed agree.

As the analysis of the real data showed inconsistent results for the different years and little agreement using different methods, a simulation study was performed to assess the reliability of the results found. Simulated datasets were generated with various sample sizes, types of DIF, and effect sizes. As with the real datasets, the stepwise LRT method and the exhaustive BIC method were used for finding the best-fitting model for each simulated dataset. In addition, an exhaustive AIC method was also tested. The performance of the methods was assessed and the effect of the different simulation conditions. For cases in which DIF was present, a multinomial logistic regression was used to assess the effect on finding no DIF, incorrect DIF, and correct DIF.

The results of the simulation study showed that when no DIF was present, both the stepwise LRT method and exhaustive BIC methods were able to find the correct model in all cases tested. In contrast, the exhaustive AIC method did find on a few occasions non-existing DIF. When DIF was present, none of the assessed methods performed well in finding the correct model. Overall, correct models with relatively simple DIF were occasionally found while correct models with relatively complex DIF were rarely if ever found. Both the stepwise LRT and exhaustive BIC method were unable to find that DIF was present if there was uniform DIF on a single covariate with a small effect size. However, for both methods, if any DIF was found, there truly was some DIF present in the model. The found relations however, were often incorrect. This indicates that complex relationships between covariates and indicators are difficult to determine correctly. For finding correct DIF or incorrect DIF with respect to finding no DIF, the exhaustive AIC method had the highest OR followed by the stepwise LRT method and the exhaustive BIC method. The OR of finding correct DIF or incorrect DIF with respect to finding no DIF increased as the sample size increased and as the effect size increased, with the OR of finding correct DIF increasing the most. Furthermore, the OR of finding correct DIF or incorrect DIF with respect to finding no DIF generally decreased as the complexity of DIF increased, with the OR of finding correct DIF decreasing the most.

Of the three methods assessed in the simulation study, the exhaustive BIC method was found to be most conservative, followed by the stepwise LRT method. Although the exhaustive AIC method proved to be the best able to retrieve the correct model, it also was found to give false positives. Most important in this study was whether there is convincing evidence for DIF. Therefore, the exhaustive AIC method was not considered for applications on the real dataset. The

stepwise LRT method and the exhaustive BIC method were both equally effective in preventing false positives in the simulation study. On the real dataset, the stepwise LRT method occasionally reached a dead end. For applications on real datasets, there is therefore no strong preference for either method.

Overall, there is an indication that the hypothesis of DIF existing for the ER or the LFS is true. This is due to DIF being found with one of the methods for all but one covariate. In the simulation study the assessed methods never found any DIF if no DIF was present. However, based on the results, the exact relationships between the tested covariates and the indicators cannot be deduced with any certainty. This is due to there being little agreement for what DEs are present for a covariate using different methods. Additionally, the results for different years are not entirely consistent. In the simulation study, it was found that the assessed methods performed poorly in finding the correct relationship. The results of the simulation study suggest that the findings for the real datasets should be interpreted with caution. It seems likely that any DIF detected in the real datasets does not accurately represent the true relationship between covariates and indicators. This may explain the inconsistency for some covariates. Furthermore, the results also suggest that if any DIF was found for one of the years for one of the methods, some type of DIF is likely present. This is due to there being no false positives found using the stepwise LRT method and exhaustive BIC method in the simulation study. For covariates for which no DIF was found, DIF cannot be ruled out as DIF with a relatively small effect size was often found to be undetected in the simulation study. Note that as only respondents aged 15 to 25 were included in this study, all statements only apply to this age group.

The findings in this study are in line with Bakker et al. (2021), who found that covariates are related to the inconsistencies between indicators.

For the analysis of the real data, there were some limitations to note. In this study, covariates were assessed individually for their effect. Therefore, interactions between covariates cannot be ruled out. Contrary to the employment contract type, measurement errors are not considered for the covariates. Whether the measured concept of any of the covariates correspond with the desired concepts cannot be assessed as all the covariates are only measured in one of the sources (Bakker, 2012). For the stepwise LRT method, multiple tests are performed without a multiple testing correcting. A Bonferroni correction or the Holm method cannot be used as the number of tests are unknown beforehand and the continuation of the method depends on the significance found in the previous step. Furthermore, with the available data, it was not possible to identify a basic LC model. By adding a covariate that was thought to not be a source of DIF, a semi-basic LC model was identified. With this approach, it was no longer possible to assess the identifying covariate as a potential source of DIF. Additionally, the definitions of the LCs may have changed. In this study, a relevant effect was considered to be a category level effect resulting in an OR of 1.25 or greater. No study in literature could be found that described how

the relevance of an effect was considered in a context similar to this study. However, in the context of epidemiological studies, Chen et al. (2010) considered an OR of 1.68, 3.47, and 6.71 to be a small, medium, and large effect, respectively. If the same interpretation holds in the context of this study, then this study was too liberal in considering the relevance of an effect.

An attempt was made to remove the effect of the covariate category levels that were related to derived frequencies by means of parameter restrictions. The results suggest that in some cases these category levels were the source of DIF. However, the restrictions used are not without question. Ideally, the parameters related to the derived frequencies would be completely omitted when estimating a model. However, as used, the restrictions implied that the restricted parameters are equal to those of the reference category. This could unintentionally lead to a worse fitting model as the derived frequencies may show a very different structure compared to frequencies related to the reference category.

For the simulation study, there were some limitations as well. The simulated covariate had two category levels. In the real data, seven of the nine covariates had a higher number of category levels. The results found in the simulation study may not hold true for situations with more category levels. Therefore, the extrapolation of the results should be taken with caution. Furthermore, due to run time constraints, for all simulation conditions, only ten replicates were tested. Although the results showed a clear pattern, the resolution of the patterns left something to be desired.

For future research into the topic of potential DIF as an explanation for the inconsistency between employment contract type frequencies as recorded in the ER and the LFS, there are a few recommendations. Research could be done into the application of a generalised multitrait-multimethod (GMTMM) model, as described by Oberski et al. (2017). This method has been proposed for estimating measurement error in both survey and administrative data. The feature of interest of the GMTMM model is that with a proper application, systematic and random measurement errors can be distinguished from one another. However, for estimating a GMTMM model, there must be several related concepts per respondent, each measured with the same (≥ 2) measurement methods. In concrete terms, for an application such as in this study, this means that, in addition to the employment contract type, another variable would have to be found that is measured in both the ER and the LFS, preferably in such a way that the systematic measurement errors per source in those different concepts resemble each other.

Furthermore, future research could focus on the *permanent* and *flexible* category levels only. All derived frequencies were related to the *other* category level. If the LC for *other* can be omitted, covariates where such derived frequencies occurred can be included without difficulty. Also, the frequencies of *other* contracts were most consistent between the ER and the LFS.

Furthermore, effects associated with the *other* category level are difficult to interpret because it concerns a residual group that includes various employment relationships.

For focusing only on the *permanent* and *flexible* class, one could try to split off all respondents estimated to belong to the *other* class. For this, a method described by Remmerswaal (2022) could be considered. The method combines the use of a decision tree and multiple imputation of latent classes (MILC; Van den Bergh, 2018). This adapted method is referred to as tree-MILC. In this application, instead of including employment contract types in a single three-class LC model, one would subdivide (partition) the employment contract types using a decision tree. At the first node of the tree, a two-class LC model could be created with a class for *permanent or flexible* and a class for *other*. At each branch, the LC for the respondents would be imputed. Subsequently, at the node for *permanent or flexible*, a new two-class LC model could then be created for the *permanent* and *flexible* class. In this way, in the latter model, effects of covariates on only the *permanent* and *flexible* classes could be investigated.

Finally, as noted by Bakker et al. (2021), audits can be considered for determining how respondents decide to answer the question of what their employment contract type is in the LFS. Likewise, audits can be considered for determining how employers decide to register their employees in the ER. Such audits are most informative where there is an inconsistency between the two indicators. Note that such audits are labour-intensive and expensive.

On another note, the stepwise LRT method proposed by Masyn (2017) may not work well for a model with two indicators. In the simulation study, it was found that the method performed poorly overall. On the real datasets, the results were inconsistent and the method reached a dead end occasionally. For additional studies into the application of this particular method, a more extensive simulation study could be carried out in which a different number of indicators are tested.

It may be possible to modify or extend the stepwise LRT method in some way so that its performance is (also) improved in situations with two indicators. A start can be made for this by examining in what situations the stepwise LRT method works well and when it does not, based on the results of the simulation study. For example, when the p-value is close to a critical value, an incorrect decision may more likely be made. Situations may be discovered in which the correct model is more frequently found.

8 References

Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*(4), 871–892.

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8–17.

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?* (Tech. rep.). CBS. https://www.cbs.nl/-/media/%5C_pdf/2021/33/tijdelijke-en-vaste-contracten.pdf

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

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* MIT Press.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political analysis*, *12*(1), 3–27.

CBS. (2022). *Hoeveel mensen met een migratieachtergrond wonen in nederland?* Retrieved July 14, 2022, from https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoeveel-mensen-met-een-migratieachtergrond-wonen-in-nederland-

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation®*, *39*(4), 860–864.

Eisenhauer, J. G. (2008). Degrees of freedom. *Teaching Statistics*.

Eurostat. (2022). *European union labour force survey (eu lfs)*. Retrieved June 27, 2022, from https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Horner, J., & Hunt, G. (2022). *Brew: Templating framework for report generation* [R package version 1.0-7]. https://CRAN.R-project.org/package=brew

Janssen, J. H., Van Laar, S., De Rooij, M. J., Kuha, J., & Bakk, Z. (2018). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 280–290.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180–197.

Oberski, D. L. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. *Total Survey Error in Practice*, 341–358.

Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, *112*(520), 1477–1489.

Pankowska, P., Bakker, B. F. M., Oberski, D. L., & 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.

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. do survey or register data tell the truth. *Survey Methodology*, *41*(1), 197–214.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Remmerswaal, D. (2022). *Comparing milc and tree-milc for estimating and correcting for multiple sources of errors in combined datasets* [MSc Thesis].

RStudio Team. (2021). *Rstudio: Integrated development environment for r*. Bosten, MA. http://www.rstudio.com/

Van den Bergh, M. (2018). *Latent class trees* [PhD Thesis]. Tilburg University.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2005). *Latent gold 4.0 user's guide*. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013). *Lg-syntax user's guide: Manual for latent gold 5.0 syntax module*. Statistical Innovations Inc. Belmont, MA.

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

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 356–364.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.

9 Appendix

9.1 Parameters to estimate and degrees of freedom

For an LC model with one identifying covariate and one tested covariate, the number of (independent or free) parameters to estimate can be found with:

$$\begin{aligned} Npar = (C-1)\cdot[1+(A-1)+(Z-1)] + \sum_{j=1}^{J}\big[&(C-1)\cdot[1+(M_j-1)]+\ldots \\ &I_{(1)}(j)\cdot(M_j-1)\cdot(Z-1)+\ldots \\ &I_{(2)}(j)\cdot(M_j-1)\cdot(Z-1)\cdot(C-1)\big], \end{aligned} \tag{26}$$

where

$Npar$	is the number of parameters to estimate;
C	is the number of latent classes;
A	is the number of categories of the identifying covariate;
Z	is the number of categories of the tested covariate;
J	is the number of indicators;
M_j	is the number of categories of indicator j;
$I_{(1)}(j)$	is an indicator function for a main effect of covariate Z on indicator j. It is 1 if there is a main effect and 0 otherwise; and
$I_{(2)}(j)$	is an indicator function for an interaction effect of covariate Z and the latent variable on indicator j. It is 1 if there is an interaction effect and 0 otherwise.

For the indicator functions in Equation (26):

$I_{(1)}(j) = 0 \wedge I_{(2)}(j) = 0$ if there is no DIF for indicator j;
$I_{(1)}(j) = 1 \wedge I_{(2)}(j) = 0$ if there is uniform DIF for indicator j; and
$I_{(1)}(j) = 1 \wedge I_{(2)}(j) = 1$ if there is non-uniform DIF for indicator j.

For an LC model, the number of degrees of freedom used by Latent GOLD can be found with:

$$df = P \cdot \left(\prod_{j=1}^{J} M_j - 1 \right) - Npar \tag{27}$$

where

df	is the degrees of freedom;
P	is the number of covariate patterns;
J	is the number of indicators;
M_j	is the number of categories of indicator j; and
$Npar$	is the number of parameters to estimate.

Note that Equations (26) and (27) hold if there are no missing values for indicators and covariates. In Equation (27), if there are no external covariates added, $P = 1$.

9.2 Example syntax simulating data

An example syntax for generating simulated data in Latent GOLD is shown in Listing 1. Here, simulated data is generated for a model with uniform DIF for indicator `Y1` and non-uniform DIF for indicator `Y2`. Comments are indicated with `//`. The model title has the following meaning: <covariate effect on `Y1` - covariate effect on `Y2` - effect size - sample size - replication number>.

For a three-class LC model with two categorical indicators and dummy first coding, to obtain a classification error probability of 0.1 for an indicator, the following parameters values are needed: 0, $-3\log(2)$, $-3\log(2)$, 0, 0, 0, 0, $6\log(2)$, $3\log(2)$, 0, $3\log(2)$, and $6\log(2)$. (The 0 parameters refer to all the reference categories.) These parameters were used for both indicator `Y1` and `Y2`. (Note that the specific aforementioned classification error probabilities apply to a model without covariates that are a source of DIF. When a covariate is added to the model that is a source of DIF, the classification error probabilities change.)

The '`Cluster <- Z`' parameters were chosen as $\log(1.25)$ and $-\log(1.25)$. This results in an OR of 1.25 and 0.8 with respect to the reference category respectively. The '`Y <- 1`' parameters were chosen as $-3\log(2)$ and $3\log(2)$; and the '`Y <- Cluster`' parameters were chosen as $6\log(2)$, $3\log(2)$, $3\log(2)$, and $6\log(2)$. Without covariates added, this result in a class separation with a classification error probability of 0.1. (This can be viewed as a matrix with 0.8 on the diagonal and 0.1 on the off-diagonal elements.) The '`Y1 <- Z`' parameters were chosen as $\log(2)$ and $-\log(2)$. This results in an OR of 2 and 0.5 with respect to the reference category respectively. Likewise, the '`Y2 <- Z|Cluster`' parameters were chosen as $\log(2)$, $\log(2)$, 0, 0, $-\log(2)$, and $-\log(2)$.

Listing 1: Latent GOLD simulate data syntax

```
//LG5.1//
version = 5.1
infile = 'exampleData.dat' quote = single

model
title '1-2-M-w2k-r1';
options
   output
      parameters=first profile;
      outfile 'simulatedData.dat' simulation=1 seed=2013;
variables
   caseid id;
   caseweight w2k;
```

```
   dependent Y1 nominal 3, Y2 nominal 3;
   independent I nominal, Z nominal;
   latent
      Cluster nominal 3;
equations
   Cluster <- 1 + I + Z;
   Y1 <- 1 + Cluster + Z;
   Y2 <- 1 + Cluster + Z|Cluster;
{ // defined parameters
-0.1604 -0.8187 // 'Cluster <- 1' estimates from real data
 0.2375 -0.0436 // 'Cluster <- I' estimates from real data
 0.2231 -0.2231 // 'Cluster <- Z'

-2.0794 2.0794               // 'Y1 <- 1'
 4.1589 2.0794 2.0794 4.1589 // 'Y1 <- Cluster'
 0.6931 -0.6931              // 'Y1 <- Z'

-2.0794 2.0794                              // 'Y2 <- 1'
 4.1589 2.0794 2.0794 4.1589                // 'Y2 <- Cluster'
 0.6931 0.6931 0      0     -0.6931 -0.6931 // 'Y2 <- Z|Cluster'
}
end model
```

9.3 Profile and ProbMeans-Posterior

Table 12: *Profile* of the three-class latent class model with the identifying covariate *gender* for the real data from the Employment Register (ER) and the Labour Force Survey (LFS) of 2016. The indicators are the employment contract type as recorded in the ER (*conER*) and the LFS (*conLFS*).

		conER			conLFS		
class	size	1	2	3	1	2	3
1	0.418	0.004	0.106	0.891	0.008	0.074	0.918
2	0.401	0.050	0.916	0.034	0.138	0.803	0.060
3	0.181	0.510	0.434	0.056	0.654	0.267	0.079
	overall	0.114	0.490	0.396	0.177	0.401	0.422

Table 13: *ProbMeans-Posterior* of the three-class latent class model with the identifying covariate *gender* for the real data from the Employment Register (ER) and the Labour Force Survey (LFS) of 2016. The indicators are the employment contract type as recorded in the ER (*conER*) and the LFS (*conLFS*).

		conER			conLFS			gender	
class	overall	1	2	3	1	2	3	1	2
1	0.418	0.013	0.090	0.940	0.020	0.077	0.910	0.436	0.400
2	0.401	0.175	0.749	0.035	0.312	0.803	0.057	0.372	0.432
3	0.181	0.812	0.160	0.025	0.668	0.121	0.034	0.192	0.168

Table 14: Profiles by covariate category level of the best-fitting models according to the stepwise likelihood-ratio test (LRT) method for the 2016 dataset. Three-class latent class models are estimated with employment contract type as recorded in the Employment Register (*conER*) and employement contract type as recorded in the Labour Force Survey (*conLFS*) as indicators. The covariate *gender* was included to identify the model. Shown are covariates for which the stepwise LRT method found a best-fitting model with differential item functioning (DIF). The highest probabilities $P(Y|X,Z)$ are in bold.

(a) Covariate: *ageGro*. Best-fitting model: *conER* with non-uniform DIF; *conLFS* with no DIF.

			conER			conLFS		
class	ageGro	size	1	2	3	1	2	3
1	1	0.437	0.093	**0.900**	0.007	0.147	**0.705**	0.149
1	2	0.432	0.066	**0.889**	0.044	0.147	**0.705**	0.149
2	1	0.537	0.011	0.073	**0.916**	0.015	0.041	**0.945**
2	2	0.339	0.002	0.155	**0.843**	0.015	0.041	**0.945**
3	1	0.026	0.364	**0.371**	0.265	**0.869**	0.097	0.034
3	2	0.230	**0.573**	0.404	0.023	**0.869**	0.097	0.034
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(b) Covariate: *comSiz*. Best-fitting model: *conER* with uniform DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	comSiz	size	1	2	3	1	2	3
1	1	0.638	0.005	**0.995**	< 0.001	0.141	**0.660**	0.199
1	2	0.437	0.144	**0.856**	< 0.001	0.321	**0.448**	0.230
1	3	0.038	0.191	**0.709**	0.100	0.128	0.068	**0.805**
1	4	0.620	< 0.001	< 0.001	**1.000**	0.003	0.081	**0.915**
2	1	0.219	**0.604**	0.396	< 0.001	**0.730**	0.227	0.043
2	2	0.096	**0.982**	0.018	< 0.001	**0.891**	0.083	0.027
2	3	0.076	**0.986**	0.011	0.003	**0.770**	0.027	0.203
2	4	0.367	< 0.001	< 0.001	**1.000**	0.073	0.115	**0.812**
3	1	0.143	0.005	**0.995**	< 0.001	0.046	**0.949**	0.005
3	2	0.467	0.152	**0.848**	< 0.001	0.139	**0.854**	0.007
3	3	0.886	0.222	**0.777**	< 0.001	0.266	**0.618**	0.116
3	4	0.013	< 0.001	0.001	**0.999**	0.008	**0.842**	0.150
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(c) Covariate: *conHou*. Best-fitting model: *conER* with uniform DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	conHou	size	1	2	3	1	2	3
1	1	0.733	0.154	**0.845**	< 0.001	0.134	**0.674**	0.192
1	2	0.802	0.134	**0.866**	< 0.001	0.188	**0.656**	0.156
1	3	0.667	0.077	**0.923**	< 0.001	0.035	**0.756**	0.210
1	4	0.048	0.087	**0.876**	0.037	0.002	0.055	**0.943**
1	5	0.988	< 0.001	< 0.001	**1.000**	0.025	0.098	**0.877**
2	1	0.220	0.010	**0.990**	< 0.001	0.165	**0.833**	0.002
2	2	0.133	0.008	**0.992**	< 0.001	0.222	**0.776**	0.002
2	3	0.013	0.004	**0.996**	< 0.001	0.044	**0.954**	0.002
2	4	0.432	0.005	**0.995**	< 0.001	0.038	**0.828**	0.135
2	5	0.009	< 0.001	0.001	**0.999**	0.192	**0.745**	0.063
3	1	0.047	**0.659**	0.341	< 0.001	**0.955**	0.045	< 0.001
3	2	0.065	**0.620**	0.380	< 0.001	**0.968**	0.032	< 0.001
3	3	0.320	0.468	**0.532**	< 0.001	**0.831**	0.169	< 0.001
3	4	0.520	**0.512**	0.487	< 0.001	**0.816**	0.168	0.016
3	5	0.002	< 0.001	< 0.001	**0.999**	**0.963**	0.035	0.002
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(d) Covariate: *sofClu*. Best-fitting model: *conER* with uniform DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	sofClu	size	1	2	3	1	2	3
1	1	0.813	0.015	**0.985**	< 0.001	0.176	**0.691**	0.133
1	2	0.846	0.232	**0.767**	< 0.001	0.205	**0.663**	0.132
1	3	0.860	0.030	**0.970**	< 0.001	0.251	**0.696**	0.052
1	4	0.956	0.034	**0.965**	0.001	0.267	**0.590**	0.143
1	5	0.250	0.025	**0.975**	< 0.001	0.106	**0.892**	0.002
1	6	0.908	< 0.001	< 0.001	**1.000**	0.023	0.110	**0.867**
2	1	0.181	**0.832**	0.167	< 0.001	**0.692**	0.250	0.058
2	2	0.136	**0.990**	0.010	< 0.001	**0.730**	0.218	0.052
2	3	0.054	**0.912**	0.088	< 0.001	**0.782**	0.200	0.018
2	4	0.033	**0.917**	0.077	0.006	**0.792**	0.161	0.047
2	5	0.030	**0.897**	0.103	< 0.001	**0.562**	0.437	0.001
2	6	0.046	< 0.001	< 0.001	**1.000**	0.177	0.079	**0.744**
3	1	0.006	0.025	**0.950**	0.024	0.052	0.001	**0.947**
3	2	0.018	0.338	**0.615**	0.047	0.060	< 0.001	**0.939**
3	3	0.086	0.053	**0.944**	0.003	0.166	0.002	**0.832**
3	4	0.011	0.041	**0.639**	0.319	0.072	< 0.001	**0.927**
3	5	0.720	0.045	**0.955**	< 0.001	**0.639**	0.027	0.334
3	6	0.046	< 0.001	< 0.001	**1.000**	0.001	< 0.001	**0.999**
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(e) Covariate: *jobDur*. Best fitting model: *conER* with uniform DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	jobDur	size	1	2	3	1	2	3
1	1	0.100	0.078	**0.915**	0.006	0.010	0.067	**0.924**
1	2	0.022	0.115	**0.856**	0.029	0.021	0.087	**0.892**
1	3	0.065	0.156	**0.835**	0.008	0.062	0.179	**0.759**
1	4	0.018	0.291	**0.641**	0.068	0.091	0.125	**0.784**
1	5	0.014	**0.564**	0.433	0.003	0.149	0.097	**0.755**
1	6	0.061	**0.534**	0.446	0.020	0.208	0.152	**0.640**
1	7	0.987	< 0.001	< 0.001	**1.000**	0.027	0.096	**0.877**
2	1	0.379	0.007	**0.993**	< 0.001	0.019	**0.753**	0.228
2	2	0.508	0.011	**0.989**	< 0.001	0.034	**0.790**	0.176
2	3	0.559	0.015	**0.985**	< 0.001	0.055	**0.865**	0.080
2	4	0.683	0.036	**0.963**	< 0.001	0.104	**0.788**	0.108
2	5	0.622	0.098	**0.902**	< 0.001	0.193	**0.690**	0.118
2	6	0.315	0.091	**0.909**	< 0.001	0.185	**0.746**	0.069
2	7	0.008	< 0.001	0.001	**0.999**	0.041	**0.799**	0.160
3	1	0.521	0.116	**0.884**	< 0.001	0.220	**0.630**	0.150
3	2	0.471	0.171	**0.829**	< 0.001	0.331	**0.569**	0.100
3	3	0.376	0.223	**0.777**	< 0.001	0.443	**0.519**	0.038
3	4	0.299	0.410	**0.590**	< 0.001	**0.617**	0.346	0.037
3	5	0.364	**0.666**	0.334	< 0.001	**0.769**	0.204	0.027
3	6	0.625	**0.647**	0.353	< 0.001	**0.757**	0.227	0.016
3	7	0.006	< 0.001	< 0.001	**1.000**	0.374	**0.541**	0.085
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(f) Covariate: *ecoAct*. Best-fitting model: *conER* with uniform DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	ecoAct	size	1	2	3	1	2	3
1	1	0.822	0.157	**0.841**	0.002	0.239	**0.596**	0.164
1	2	0.888	0.356	**0.644**	< 0.001	0.169	**0.606**	0.225
1	3	0.501	< 0.001	**0.999**	< 0.001	0.002	**0.896**	0.102
1	4	0.100	< 0.001	**0.995**	0.004	0.001	**0.734**	0.265
1	5	0.264	< 0.001	**0.999**	< 0.001	< 0.001	**0.834**	0.165
1	6	0.469	0.002	**0.931**	0.066	0.012	**0.872**	0.117
1	7	0.660	< 0.001	**1.000**	< 0.001	< 0.001	**0.804**	0.195
1	8	0.243	0.008	**0.992**	< 0.001	0.014	**0.855**	0.131
1	9	0.581	0.003	**0.997**	< 0.001	0.006	**0.733**	0.261
1	10	0.885	< 0.001	< 0.001	**1.000**	< 0.001	0.109	**0.890**
2	1	0.097	**0.987**	0.013	< 0.001	**0.994**	0.005	< 0.001
2	2	0.006	**0.996**	0.004	< 0.001	**0.992**	0.007	0.002
2	3	0.498	0.274	**0.726**	< 0.001	**0.554**	0.409	0.037
2	4	0.807	0.280	**0.719**	< 0.001	**0.445**	0.432	0.123
2	5	0.721	0.263	**0.737**	< 0.001	0.274	**0.628**	0.098
2	6	0.245	0.488	**0.500**	0.012	**0.868**	0.120	0.013
2	7	0.282	0.146	**0.854**	< 0.001	0.312	**0.577**	0.111
2	8	0.112	**0.771**	0.229	< 0.001	**0.886**	0.102	0.012
2	9	0.394	**0.553**	0.447	< 0.001	**0.776**	0.175	0.049
2	10	0.076	< 0.001	< 0.001	**1.000**	0.365	0.085	**0.550**
3	1	0.081	**0.728**	0.271	0.001	**0.925**	0.032	0.043
3	2	0.106	**0.888**	0.112	< 0.001	**0.876**	0.044	0.080
3	3	< 0.001	0.013	**0.987**	< 0.001	0.102	**0.577**	0.321
3	4	0.093	0.013	**0.980**	0.007	0.047	0.346	**0.607**
3	5	0.015	0.012	**0.988**	< 0.001	0.028	**0.495**	0.476
3	6	0.286	0.029	**0.863**	0.108	0.364	**0.385**	0.251
3	7	0.058	0.006	**0.994**	< 0.001	0.031	0.444	**0.524**
3	8	0.646	0.104	**0.896**	< 0.001	**0.395**	0.347	0.258
3	9	0.025	0.041	**0.959**	< 0.001	0.175	0.302	0.523
3	10	0.039	< 0.001	< 0.001	**1.000**	0.014	0.024	**0.962**
		overall	0.114	0.490	0.396	0.177	0.401	0.422

Table 15: Profiles by covariate category level of the best-fitting models according to the exhaustive Bayesian information criterion (BIC) method for the 2016 dataset. Three-class latent class models are estimated with employment contract type as recorded in the Employment Register (*conER*) and employment contract type as recorded in the Labour Force Survey (*conLFS*) as indicators. The covariate *gender* was included to identify the model. Shown are covariates for which the exhaustive BIC method found a best-fitting model with differential item functioning (DIF). The highest probabilities $P(Y|X, Z)$ are in bold.

(a) Covariate: *comSiz*. Best-fitting model: *conER* with no DIF; *conLFS* with uniform DIF.

			conER			conLFS		
class	comSiz	size	1	2	3	1	2	3
1	1	< 0.001	< 0.001	< 0.001	**1.000**	0.374	0.106	**0.520**
1	2	< 0.001	< 0.001	< 0.001	**1.000**	0.201	0.170	**0.629**
1	3	0.004	< 0.001	< 0.001	**1.000**	0.120	0.137	**0.743**
1	4	1.000	< 0.001	< 0.001	**1.000**	0.029	0.104	**0.867**
2	1	0.663	0.008	**0.992**	< 0.001	0.011	**0.809**	0.180
2	2	0.425	0.008	**0.992**	< 0.001	0.004	**0.852**	0.144
2	3	0.273	0.008	**0.992**	< 0.001	0.003	**0.800**	0.198
2	4	< 0.001	0.008	**0.992**	< 0.001	< 0.001	**0.724**	0.276
3	1	0.337	0.388	**0.612**	< 0.001	**0.741**	0.208	0.052
3	2	0.575	0.388	**0.612**	< 0.001	**0.503**	0.418	0.079
3	3	0.723	0.388	**0.612**	< 0.001	0.412	**0.461**	0.127
3	4	< 0.001	0.388	**0.612**	< 0.001	0.166	**0.585**	0.249
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(b) Covariate: *sofClu*. Best-fitting model: *conER* with uniform DIF; *conLFS* with no DIF.

			conER			conLFS		
class	sofClu	size	1	2	3	1	2	3
1	1	0.805	0.002	**0.998**	< 0.001	0.201	**0.691**	0.108
1	2	0.792	0.206	**0.794**	< 0.001	0.201	**0.691**	0.108
1	3	0.792	< 0.001	**1.000**	< 0.001	0.201	**0.691**	0.108
1	4	0.713	< 0.001	**0.999**	< 0.001	0.201	**0.691**	0.108
1	5	0.072	< 0.001	**1.000**	< 0.001	0.201	**0.691**	0.108
1	6	0.019	< 0.001	< 0.001	**1.000**	0.201	**0.691**	0.108
2	1	0.030	0.002	**0.990**	0.008	0.024	0.091	**0.885**
2	2	0.046	0.266	**0.717**	0.018	0.024	0.091	**0.885**
2	3	0.026	< 0.001	**0.993**	0.007	0.024	0.091	**0.885**
2	4	0.061	< 0.001	**0.933**	0.067	0.024	0.091	**0.885**
2	5	0.203	< 0.001	**0.998**	0.002	0.024	0.091	**0.885**
2	6	0.978	< 0.001	< 0.001	**1.000**	0.024	0.091	**0.885**
3	1	0.164	**0.979**	0.021	< 0.001	**0.651**	0.282	0.067
3	2	0.163	**1.000**	< 0.001	< 0.001	**0.651**	0.282	0.067
3	3	0.182	0.437	**0.563**	< 0.001	**0.651**	0.282	0.067
3	4	0.226	0.280	**0.719**	0.001	**0.651**	0.282	0.067
3	5	0.724	0.091	**0.909**	< 0.001	**0.651**	0.282	0.067
3	6	0.003	< 0.001	< 0.001	**0.999**	**0.651**	0.282	0.067
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(c) Covariate: *jobDur*. Best-fitting model: *conER* with uniform DIF; *conLFS* with no DIF.

			conER			conLFS		
class	jobDur	size	1	2	3	1	2	3
1	1	0.237	0.065	**0.932**	0.003	0.006	0.076	**0.918**
1	2	0.120	0.066	**0.928**	0.006	0.006	0.076	**0.918**
1	3	0.065	0.067	**0.925**	0.007	0.006	0.076	**0.918**
1	4	0.053	0.075	**0.904**	0.021	0.006	0.076	**0.918**
1	5	0.046	0.111	**0.888**	< 0.001	0.006	0.076	**0.918**
1	6	0.017	0.132	**0.840**	0.029	0.006	0.076	**0.918**
1	7	0.942	< 0.001	< 0.001	**1.000**	0.006	0.076	**0.918**
2	1	0.745	0.064	**0.936**	< 0.001	0.151	**0.797**	0.052
2	2	0.818	0.065	**0.935**	< 0.001	0.151	**0.797**	0.052
2	3	0.840	0.067	**0.933**	< 0.001	0.151	**0.797**	0.052
2	4	0.754	0.075	**0.925**	< 0.001	0.151	**0.797**	0.052
2	5	0.523	0.109	**0.891**	< 0.001	0.151	**0.797**	0.052
2	6	0.317	0.133	**0.867**	< 0.001	0.151	**0.797**	0.052
2	7	0.034	< 0.001	< 0.001	**1.000**	0.151	**0.797**	0.052
3	1	0.018	0.433	**0.567**	< 0.001	**0.745**	0.196	0.059
3	2	0.062	0.438	**0.562**	< 0.001	**0.745**	0.196	0.059
3	3	0.095	0.444	**0.556**	< 0.001	**0.745**	0.196	0.059
3	4	0.194	0.475	**0.524**	0.001	**0.745**	0.196	0.059
3	5	0.432	**0.578**	0.422	< 0.001	**0.745**	0.196	0.059
3	6	0.666	**0.631**	0.368	0.001	**0.745**	0.196	0.059
3	7	0.025	< 0.001	< 0.001	**1.000**	**0.745**	0.196	0.059
		overall	0.114	0.490	0.396	0.177	0.401	0.422

(d) Covariate: *ecoAct*. Best-fitting model: *conER* with uniform DIF; *conLFS* with no DIF.

			conER			conLFS		
class	ecoAct	size	1	2	3	1	2	3
1	1	0.603	0.155	**0.845**	< 0.001	0.119	**0.801**	0.079
1	2	0.679	0.350	**0.650**	< 0.001	0.119	**0.801**	0.079
1	3	0.816	0.079	**0.921**	< 0.001	0.119	**0.801**	0.079
1	4	0.569	0.118	**0.882**	< 0.001	0.119	**0.801**	0.079
1	5	0.863	0.156	**0.844**	< 0.001	0.119	**0.801**	0.079
1	6	0.666	0.064	**0.930**	0.006	0.119	**0.801**	0.079
1	7	0.851	0.040	**0.960**	< 0.001	0.119	**0.801**	0.079
1	8	0.528	0.068	**0.932**	< 0.001	0.119	**0.801**	0.079
1	9	0.591	0.120	**0.880**	< 0.001	0.119	**0.801**	0.079
1	10	0.015	< 0.001	< 0.001	**1.000**	0.119	**0.801**	0.079
2	1	0.091	0.131	**0.853**	0.016	0.002	0.096	**0.902**
2	2	0.162	0.312	**0.688**	< 0.001	0.002	0.096	**0.902**
2	3	0.001	0.067	**0.932**	0.001	0.002	0.096	**0.902**
2	4	0.146	0.100	**0.889**	0.011	0.002	0.096	**0.902**
2	5	0.056	0.135	**0.865**	< 0.001	0.002	0.096	**0.902**
2	6	0.079	0.021	0.367	**0.612**	0.002	0.096	**0.902**
2	7	0.137	0.034	**0.966**	< 0.001	0.002	0.096	**0.902**
2	8	0.168	0.058	**0.942**	< 0.001	0.002	0.096	**0.902**
2	9	0.148	0.103	**0.896**	< 0.001	0.002	0.096	**0.902**
2	10	0.959	< 0.001	< 0.001	**1.000**	0.002	0.096	**0.902**
3	1	0.306	**0.586**	0.414	< 0.001	**0.969**	0.008	0.023
3	2	0.159	**0.806**	0.194	< 0.001	**0.969**	0.008	0.023
3	3	0.183	0.398	**0.602**	< 0.001	**0.969**	0.008	0.023
3	4	0.285	**0.508**	0.492	< 0.001	**0.969**	0.008	0.023
3	5	0.081	**0.589**	0.411	< 0.001	**0.969**	0.008	0.023
3	6	0.255	0.331	**0.619**	0.049	**0.969**	0.008	0.023
3	7	0.012	0.244	**0.756**	< 0.001	**0.969**	0.008	0.023
3	8	0.304	0.360	**0.640**	< 0.001	**0.969**	0.008	0.023
3	9	0.261	**0.514**	0.486	< 0.001	**0.969**	0.008	0.023
3	10	0.026	< 0.001	< 0.001	**1.000**	**0.969**	0.008	0.023
		overall	0.114	0.490	0.396	0.177	0.401	0.422

9.4 Detailed results 2016 dataset

For all the tested covariates for the 2016 dataset, the detailed results of the stepwise LRT method and the exhaustive BIC method are shown in Tables 16 and 17, respectively. The stepwise LRT method is based on the method described by Masyn (2017) with an additional step in which parameter restrictions are assessed for covariates with derived frequencies. For the exhaustive BIC method, parameter restrictions are also assessed for covariates with derived frequencies. For *eduLev*, at *step 2*, no DIF is found, contradicting *step 1* (see Table 16d). For *jobDur*, at *step 5R*, both M5.R1 and M5.R2 do not perform significantly worse than M5.0. This indicates that a model with no DIF (M5.R3) for ER should suffice, contradicting *step 2* (see Table 16g).

Table 16: Results of all steps for the stepwise likelihood-ratio test (LRT) method of all covariates tested for the 2016 dataset. Indicated are the step numbers (step); the model names (model); the type of differential item functioning (DIF) for the indicators: employment contract type as recorded in the Employment Register (*conER*) and employment contract type as recorded in the Labour Force Survey (*conLFS*); the log-likelihood (LL); the number of estimated parameters (Npar); the models compared (comparison); the LRT statistic (LRTS) with the associated the degrees of freedom (df) and p-values (p). For the type of DIF, 0, 1, and 2 represent *no DIF*, *uniform DIF*, and *non-uniform DIF*, respectively. For covariates with derived frequencies, models with parameter restrictions have been tested. The models with parameter restrictions are indicated with grey text. Steps where no new comparison is made have been left out.

(a) Covariate *ageGro*. Note M5.0 = M3.0.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-26720.95	18	M1.0 vs. M1.1	27.22	12	0.007
.	M1.1	2	2	-26707.34	30	.	.	.	.
2	M2.0.1	0	.	-38779.54	12	M2.0.1 vs. M2.1.1	68.20	6	< 0.001
.	M2.1.1	2	.	-38745.44	18	.	.	.	.
.	M2.0.2	.	0	-39503.93	12	M2.0.2 vs. M2.1.2	8.87	6	0.181
.	M2.1.2	.	2	-39499.50	18	.	.	.	.
3	M.3.0	2	0	-26711.24	24	M1.0 vs. M3.0	19.42	6	0.004
.	.	.	.	.	.	M3.0 vs. M1.1	7.80	6	0.253
4	M4.1	1	0	-26719.19	20	M4.1 vs. M3.0	15.90	4	0.003

(b) Covariate *comSiz*. Note M3.0 = M1.1.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-16337.63	22	M1.0 vs. M1.1	183.57	36	< 0.001
.	M1.1	2	2	-16245.84	58	.	.	.	.
2	M2.0.1	0	.	-21684.46	16	M2.0.1 vs. M2.1.1	387.30	18	< 0.001
.	M2.1.1	2	.	-21490.81	34	.	.	.	.
.	M2.0.2	.	0	-26709.26	16	M2.0.2 vs. M2.1.2	455.88	18	< 0.001
.	M2.1.2	.	2	-26481.32	34	.	.	.	.
4	M4.1	1	2	-16247.43	46	M4.1 vs. M3.0	3.17	12	0.994
.	M4.2	2	1	-16248.11	46	M4.2 vs. M3.0	4.54	12	0.971
5	M5.0	1	1	-16248.80	34	M5.0 vs. M3.0	5.92	24	1.000
5R	M5.R1	1^{R1}	1	-16253.48	32	M5.R1 vs. M5.0	9.35	2	0.009
.	M5.R2	1^{R2}	1	-16256.54	30	M5.R2 vs. M5.0	15.47	4	0.004

(c) Covariate *conHou*. Note M3.0 = M1.1.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-16119.61	24	M1.0 vs. M1.1	81.46	48	0.002
.	M1.1	2	2	-16078.88	72	.	.	.	.
2	M2.0.1	0	.	-21638.58	18	M2.0.1 vs. M2.1.1	500.22	24	< 0.001
.	M2.1.1	2	.	-21388.47	42	.	.	.	.
.	M2.0.2	.	0	-26410.49	18	M2.0.2 vs. M2.1.2	415.81	24	< 0.001
.	M2.1.2	.	2	-26202.59	42	.	.	.	.
4	M4.1	1	2	-16080.72	56	M4.1 vs. M3.0	3.69	16	0.999
.	M4.2	2	1	-16084.77	56	M4.2 vs. M3.0	11.77	16	0.760
5	M5.0	1	1	-16084.53	40	M5.0 vs. M3.0	11.29	32	1.000
5R	M5.R1	1^{R1}	1	-16094.05	38	M5.R1 vs. M5.0	19.05	2	< 0.001
.	M5.R2	1^{R2}	1	-16098.34	34	M5.R2 vs. M5.0	27.62	6	< 0.001

(d) Covariate *eduLev*. Note M3.0 = M1.0. Dead end at *step 2*.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-26909.03	22	M1.0 vs. M1.1	54.68	36	0.024
.	M1.1	2	2	-26881.69	58	.	.	.	.
2	M2.0.1	0	.	-38938.99	16	M2.0.1 vs. M2.1.1	24.12	18	0.151
.	M2.1.1	2	.	-38926.93	34	.	.	.	.
.	M2.0.2	.	0	-39685.60	16	M2.0.2 vs. M2.1.2	11.49	18	0.872
.	M2.1.2	.	2	-39679.86	34	.	.	.	.

(e) Covariate *ecoAct*. Note M3.0 = M1.1.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-16156.90	34	M1.0 vs. M1.1	600.17	108	< 0.001
.	M1.1	2	2	-15856.82	142	.	.	.	.
2	M2.0.1	0	.	-21290.43	28	M2.0.1 vs. M2.1.1	697.78	54	< 0.001
.	M2.1.1	2	.	-20941.54	82	.	.	.	.
.	M2.0.2	.	0	-26160.09	28	M2.0.2 vs. M2.1.2	527.06	54	< 0.001
.	M2.1.2	.	2	-25896.56	82	.	.	.	.
4	M4.1	1	2	-15864.46	106	M4.1 vs. M3.0	15.28	36	0.999
.	M4.2	2	1	-15870.69	106	M4.2 vs. M3.0	27.75	36	0.836
5	M5.0	1	1	-15880.40	70	M5.0 vs. M3.0	47.16	72	0.990
5R	M5.R1	1^{R1}	1	-15880.64	68	M5.R1 vs. M5.0	0.49	2	0.782
.	M5.R2	1^{R2}	1	-15928.97	54	M5.R2 vs. M5.0	97.15	16	< 0.001

(f) Covariate: *intMan*.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-27275.89	18	M1.0 vs. M1.1	17.91	12	0.118
.	M1.1	2	2	-27266.93	30	.	.	.	.

(g) Covariate: *jobDur*. Note M3.0 = M1.1. Dead end at *step 5R*.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-15651.26	28	M1.0 vs. M1.1	224.49	72	< 0.001
.	M1.1	2	2	-15539.01	100	.	.	.	.
2	M2.0.1	0	.	-20615.77	22	M2.0.1 vs. M2.1.1	609.86	36	< 0.001
.	M2.1.1	2	.	-20310.84	58	.	.	.	.
.	M2.0.2	.	0	-25496.54	22	M2.0.2 vs. M2.1.2	388.13	36	< 0.001
.	M2.1.2	.	2	-25302.48	58	.	.	.	.
4	M4.1	1	2	-15540.35	76	M4.1 vs. M3.0	2.67	24	1.000
.	M4.2	2	1	-15542.89	76	M4.2 vs. M3.0	7.75	24	0.999
5	M5.0	1	1	-15547.86	52	M5.0 vs. M3.0	17.69	48	1.000
5R	M5.R1	1^{R1}	1	-15549.91	50	M5.R1 vs. M5.0	4.09	2	0.129
.	M5.R2	1^{R2}	1	-15552.07	42	M5.R2 vs. M5.0	8.42	10	0.588
.	M5.R3	0	1	-15554.24	40	M5.R3 vs. M5.0	12.76	12	0.387

(h) Covariate *migBac*.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-27128.3	28	M1.0 vs. M1.1	57.5	72	0.893
.	M1.1	2	2	-27099.5	100	.	.	.	.

(i) Covariate *sofClu*.

step	model	conER	conLFS	LL	Npar	comparison	LRTS	df	p
1	M1.0	0	0	-16252.07	26	M1.0 vs. M1.1	253.32	60	< 0.001
.	M1.1	2	2	-16125.41	86	.	.	.	.
2	M2.0.1	0	.	-21548.69	20	M2.0.1 vs. M2.1.1	514.65	30	< 0.001
.	M2.1.1	2	.	-21291.37	50	.	.	.	.
.	M2.0.2	.	0	-26666.26	20	M2.0.2 vs. M2.1.2	547.21	30	< 0.001
.	M2.1.2	.	2	-26392.66	50	.	.	.	.
4	M4.1	1	2	-16127.06	66	M4.1 vs. M3.0	3.30	20	1.000
.	M4.2	2	1	-16129.87	66	M4.2 vs. M3.0	8.93	20	0.984
5	M5.0	1	1	-16132.96	46	M5.0 vs. M3.0	15.11	40	1.000
5R	M5.R1	1^{R1}	1	-16133.58	44	M5.R1 vs. M5.0	1.24	2	0.538
.	M5.R2	1^{R2}	1	-16141.71	38	M5.R2 vs. M5.0	17.49	8	0.025

Table 17: Results of all tested models for the exhaustive Bayesian information criterion method of all tested covariates. Indicated is the type of differential item functioning (DIF) for the indicators: employment contract type as recorded in the Employment Register (*conER*) and employment contract type as recorded in the Labour Force Survey (*conLFS*); the log-likelihood (LL); the number of estimated parameters (Npar); the Akaike information criterion based on the LL (AIC[LL]); and the Bayesian information criterion based on the LL (BIC[LL]). For the type of DIF, 0, 1, and 2 represent *no DIF*, *uniform DIF*, and *non-uniform DIF*, respectively. Models are sorted by BIC(LL) in descending order. For covariates with derived frequencies, models with parameter restrictions have been tested. The models with parameter restrictions are indicated with grey text.

(a) Covariate *ageGro*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-26707.34	30	53474.68	53705.57
1	2	-26711.53	26	53475.05	53675.16
2	1	-26709.55	26	53471.09	53671.20
0	2	-26712.83	24	53473.65	53658.37
2	0	-26711.24	24	53470.48	53655.19
1	1	-26712.50	22	53469.00	53638.32
1	0	-26719.19	20	53478.38	53632.31
0	1	-26714.65	20	53469.30	53623.23
0	0	-26720.95	18	53477.90	53616.43

(b) Covariate *intMan*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-27266.93	30	54593.86	54824.75
1	2	-27268.80	26	54589.59	54789.70
2	1	-27268.31	26	54588.62	54788.72
0	2	-27270.17	24	54588.35	54773.06
2	0	-27268.73	24	54585.46	54770.18
1	1	-27269.37	22	54582.75	54752.07
1	0	-27271.78	20	54583.55	54737.48
0	1	-27270.87	20	54581.75	54735.68
0	0	-27275.89	18	54587.77	54726.31

(c) Covariate *comSiz*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-16245.84	58	32607.68	33054.07
2^{R1}	2	-16249.57	52	32603.14	33003.35
2^{R2}	2	-16255.48	46	32602.97	32957.00
2	1	-16248.11	46	32588.23	32942.26
1	2	-16247.43	46	32586.85	32940.88
1^{R1}	2	-16252.15	44	32592.30	32930.94
1^{R2}	2	-16255.57	42	32595.14	32918.39
0	2	-16257.40	40	32594.81	32902.66
2^{R2}	0	-16314.05	28	32684.11	32899.61
2	0	-16252.93	40	32585.86	32893.71
2^{R1}	1	-16251.31	40	32582.62	32890.47
0	0	-16337.63	22	32719.25	32888.57
1^{R2}	0	-16314.13	24	32676.26	32860.98
2^{R1}	0	-16259.99	34	32587.98	32849.66
2^{R2}	1	-16256.47	34	32580.94	32842.62
1	1	-16248.80	34	32565.60	32827.28
1	0	-16274.40	28	32604.80	32820.30
1^{R1}	0	-16283.80	26	32619.60	32819.71
1^{R1}	1	-16253.48	32	32570.96	32817.24
1^{R2}	1	-16256.54	30	32573.07	32803.96
0	1	-16261.10	28	32578.20	32793.70

(d) Covariate *eduLev*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-26881.69	58	53879.39	54325.78
1	2	-26890.17	46	53872.34	54226.37
2	1	-26888.24	46	53868.49	54222.52
0	2	-26893.35	40	53866.71	54174.56
2	0	-26891.42	40	53862.84	54170.69
1	1	-26891.65	34	53851.29	54112.97
0	1	-26901.78	28	53859.55	54075.05
1	0	-26901.39	28	53858.77	54074.27
0	0	-26909.03	22	53862.07	54031.39

(e) Covariate *conHou*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-16078.88	72	32301.76	32855.90
2^{R1}	2	-16082.67	66	32297.33	32805.29
2	1	-16084.77	56	32281.53	32712.53
1	2	-16080.72	56	32273.45	32704.44
2^{R2}	2	-16090.26	54	32288.52	32704.12
1^{R1}	2	-16083.52	54	32275.04	32690.64
1^{R2}	2	-16090.30	50	32280.59	32665.41
2^{R1}	1	-16088.29	50	32276.59	32661.41
0	2	-16094.04	48	32284.07	32653.50
2	0	-16092.3	48	32280.60	32650.03
2^{R1}	0	-16101.14	42	32286.28	32609.53
2^{R2}	1	-16098.35	38	32272.69	32565.15
1	1	-16084.53	40	32249.05	32556.91
1^{R1}	1	-16094.05	38	32264.10	32556.56
1^{R2}	1	-16098.34	34	32264.67	32526.35
0	1	-16108.00	32	32280.00	32526.28
2^{R2}	0	-16113.79	30	32287.58	32518.47
1	0	-16104.02	32	32272.04	32518.32
1^{R1}	0	-16108.80	30	32277.60	32508.49
1^{R2}	0	-16113.83	26	32279.66	32479.76
0	0	-16119.61	24	32287.22	32471.94

(f) Covariate *sofClu*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-16125.41	86	32422.82	33084.71
2^{R1}	2	-16126.07	80	32412.14	33027.84
2	1	-16129.87	66	32391.75	32899.71
1	2	-16127.06	66	32386.12	32894.07
1^{R1}	2	-16127.60	64	32383.21	32875.77
2^{R2}	2	-16133.54	62	32391.08	32868.26
2^{R1}	1	-16130.88	60	32381.75	32843.53
1^{R2}	2	-16133.58	58	32383.16	32829.54
0	2	-16141.85	56	32395.70	32826.70
2	0	-16137.48	56	32386.95	32817.95
2^{R2}	0	-16246.42	32	32556.84	32803.12
1^{R2}	0	-16246.51	28	32549.02	32764.52
2^{R1}	0	-16139.05	50	32378.10	32762.92
0	0	-16252.07	26	32556.14	32756.24
1	1	-16132.96	46	32357.93	32711.96
1^{R1}	1	-16133.58	44	32355.17	32693.81
2^{R2}	1	-16141.70	42	32367.40	32690.65
1^{R2}	1	-16141.71	38	32359.42	32651.88
0	1	-16145.89	36	32363.78	32640.85
1	0	-16144.08	36	32360.15	32637.22
1^{R1}	0	-16143.42	34	32354.85	32616.53

(g) Covariate *jobDur*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-15539.01	100	31278.03	32047.66
2^{R1}	2	-15542.02	94	31272.03	31995.49
2	1	-15542.89	76	31237.77	31822.70
1	2	-15540.35	76	31232.70	31817.62
2^{R1}	0	-15623.16	58	31362.31	31808.70
1^{R1}	2	-15543.50	74	31235.00	31804.53
2^{R2}	2	-15547.06	70	31234.13	31772.87
2^{R1}	1	-15545.74	70	31231.47	31770.22
1^{R2}	2	-15547.10	66	31226.19	31734.15
0	2	-15550.13	64	31228.26	31720.82
2	0	-15548.23	64	31224.47	31717.03
1^{R1}	0	-15631.51	38	31339.02	31631.48
1	1	-15547.86	52	31199.72	31599.93
1^{R1}	1	-15549.91	50	31199.81	31584.63
0	0	-15651.26	28	31358.51	31574.01
2^{R2}	1	-15552.04	46	31196.08	31550.11
1^{R2}	1	-15552.07	42	31188.14	31511.39
0	1	-15554.24	40	31188.48	31496.34
1	0	-15552.96	40	31185.91	31493.77
2^{R2}	0	-15557.80	34	31183.60	31445.27
1^{R2}	0	-15557.85	30	31175.70	31406.59

(h) Covariate *migBac*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-27099.52	100	54399.05	55168.68
2	1	-27105.28	76	54362.56	54947.49
1	2	-27104.44	76	54360.88	54945.80
2	0	-27108.77	64	54345.55	54838.11
0	2	-27107.62	64	54343.25	54835.81
1	1	-27110.22	52	54324.44	54724.65
0	1	-27119.68	40	54319.37	54627.22
1	0	-27118.01	40	54316.02	54623.88
0	0	-27128.27	28	54312.55	54528.05

(i) Covariate *ecoAct*.

conER	conLFS	LL	Npar	AIC(LL)	BIC(LL)
2	2	-15856.82	142	31997.63	33090.51
2^{R1}	2	-15861.78	136	31995.56	33042.26
2	1	-15870.69	106	31953.38	32769.19
1	2	-15864.46	106	31940.92	32756.73
1^{R1}	2	-15863.64	104	31935.28	32735.70
2^{R1}	1	-15874.33	100	31948.67	32718.30
2^{R2}	2	-15889.78	94	31967.55	32691.01
2^{R1}	0	-15943.38	82	32050.77	32681.87
0	2	-15901.31	88	31978.62	32655.90
1^{R2}	2	-15889.81	90	31959.62	32652.29
2	0	-15896.09	88	31968.17	32645.45
0	0	-16156.90	34	32381.80	32643.48
2^{R2}	0	-16062.22	40	32204.44	32512.29
1^{R2}	0	-16062.23	36	32196.45	32473.52
1^{R1}	0	-15983.33	50	32066.66	32451.47
1	1	-15880.40	70	31900.79	32439.53
1^{R1}	1	-15880.64	68	31897.28	32420.63
2^{R2}	1	-15928.70	58	31973.39	32419.78
0	1	-15954.14	52	32012.29	32412.50
1^{R2}	1	-15928.97	54	31965.94	32381.54
1	0	-15937.06	52	31978.13	32378.34

9.5 Stepwise regression simulation covariates

Table 18: Stepwise regression for the response variable *found*. With (a) forward selection and (b) backward selection, the effect of the simulation covariates is assessed. To limit the analysis, only main effects are considered. Indicated are the formula representations of the model (formula); the model change assessed (change); the number of free model parameters (Npar); and two model selection criteria: Akaike information criterion (AIC) and Bayesian information criterion (BIC). As both criteria select the same model in each step, the results are combined in a single table.

(a) Forward selection for the response variable *found*.

formula	change	Npar	AIC	BIC
found ~ 1	none	2	1765.54	1775.14
	+ casWei	4	1675.25	1694.46
	+ covEff	10	1456.73	1504.75
	+ effSiz	6	1484.30	1513.11
	+ method	6	1656.65	1685.46
found ~ 1 + covEff	none	10	1456.73	1504.75
	+ casWei	12	1356.02	1413.65
	+ effSiz	14	1136.47	1203.70
	+ method	14	1336.68	1403.91
found ~ 1 + covEff + effSiz	none	14	1136.47	1203.70
	+ casWei	16	970.56	1047.40
	+ method	18	942.24	1028.69
found ~ 1 + covEff + effSiz + method	none	18	942.24	1028.69
	+ casWei	20	697.73	793.78

(b) Backward selection for the response variable *found*.

formula	change	Npar	AIC	BIC
found ~ 1 + covEff + effSiz + method + casWei	none	20	697.73	793.78
	- casWei	18	942.24	1028.69
	- method	16	970.56	1047.40
	- covEff	12	1147.11	1204.74
	- effSiz	16	1218.98	1295.82

9.6 Schematic latent class model

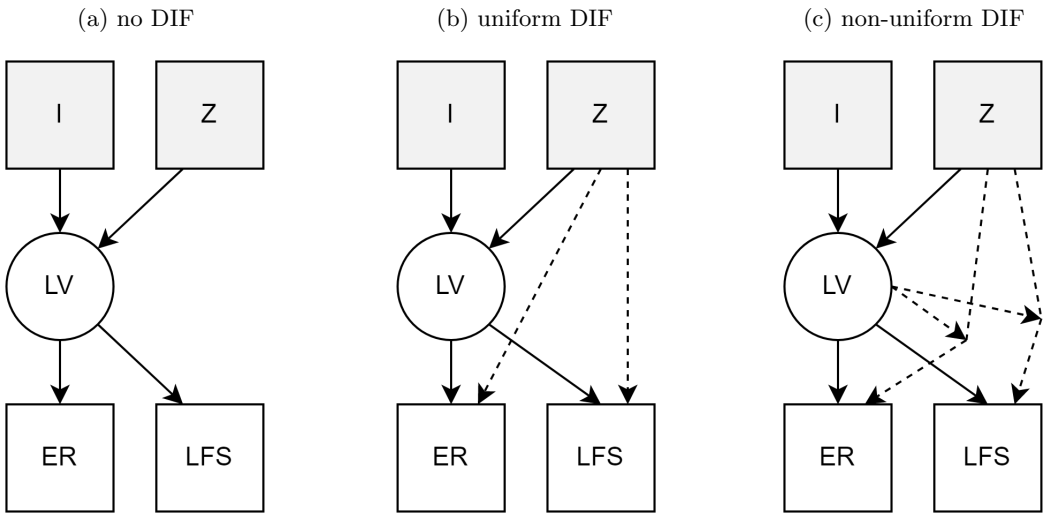

Figure 2: Schematic representation of latent class models for employment contract type with various types of differential item functioning (DIF). Covariates are indicated by a grey square; the latent variable is indicated by a white circle; and indicators are indicated by a white square. Effects are indicated with solid arrows and potential direct effects of covariates on indicators are indicated with dotted arrows. Depicted are (a) no DIF; (b) potential uniform DIF for indicators; and (c) potential non-uniform DIF for indicators. I: identifying covariate, Z: tested covariate, LV: latent variable, LFS: Labour Force Survey, ER: Employment Register.

9.7 Online repository

The code used for this study are collected in a GitHub repository. This repository is available at the URL: https://github.com/golderick/employment-contract-type. The included README.md file gives a concise explanation for all files present in the repository.

Note that the datasets provided by CBS (ER and LFS), all derived datasets, and the model results of Bakker et al. (2021) are not publicly available.