Received: 14 November 2024 | Accepted: 3 April 2025

DOI: 10.1111/bmsp.12392

the british psychological society
promoting excellence in psychology

ARTICLE

Score-based tests for parameter instability in ordinal factor models

Franz Classe[1] | **Rudolf Debelak**[2] | **Christoph Kern**[3]

[1]Deutsches Jugendinstitut e.V., Munich, Germany

[2]Department of Psychology, University of Zurich, Zurich, Switzerland

[3]Department of Statistics, LMU Munich, Munich, Germany

Correspondence
Franz Classe, Lerchenauer Straße 5, 80809 Munich, Germany.
Email: classefranz@gmail.com

Abstract

We present a novel approach for computing model scores for ordinal factor models, that is, graded response models (GRMs) fitted with a limited information (LI) estimator. The method makes it possible to compute score-based tests for parameter instability for ordinal factor models. This way, rapid execution of numerous parameter instability tests for multidimensional item response theory (MIRT) models is facilitated. We present a comparative analysis of the performance of the proposed score-based tests for ordinal factor models in comparison to tests for GRMs fitted with a full information (FI) estimator. The new method has a good Type I error rate, high power and is computationally faster than FI estimation. We further illustrate that the proposed method works well with complex models in real data applications. The method is implemented in the *lavaan* package in R.

KEYWORDS

multidimensional item response theory, ordinal factor analysis, parameter instability, score test

1 | INTRODUCTION

Researchers investigating thought processes and cognitive abilities often use *item response theory* (IRT) models to measure multiple unobserved (or latent) variables like personality traits or proficiencies. One of the most widely applied IRT frameworks for observed variables with a small amount of ordered response categories is the *graded response model* (GRM; Samejima, 1969).

However, unidimensional IRT models, that is, models with only one latent variable, are often not able to model the full complexity of conceptually broad personality traits or abilities. Multidimensional item response theory (MIRT) models make it possible to analyse psychological assessment data such that underlying multidimensionality is captured (Reckase, 1997). The potential

of such models for large-scale test and questionnaire evaluation and development has been emphasized numerous times (Bean & Bowen, 2021; Immekus et al., 2019; ten Holt et al., 2010). A major advantage of MIRT models is their flexibility, because latent covariance structures, hierarchical latent variable structures or within-item multidimensionality can be included in the model (Hartig & Höhler, 2009). In this paper, we develop an approach to compute model scores for a special kind of (multidimensional) IRT model, namely the ordinal factor model. This opens up novel avenues in latent variable modelling.

A popular estimation method in IRT is *marginal maximum likelihood* (MML) estimation via the *expectation maximization* (EM) algorithm (Bock & Aitkin, 1981; Jöreskog & Moustaki, 2006). This approach is commonly considered a full information (FI) estimation method because all distinct values on the observed variables are used (Bolt, 2005). However, parameter estimation for MIRT models via this FI method is computationally demanding, especially if there is more than one dimension (i.e. latent variable) (Forero & Maydeu-Olivares, 2009; Muraki & Carlson, 1995) as the complexity of the EM algorithm increases exponentially with the number of latent variables. In contrast, the complexity of the *Metropolis-Hastings Robbins-Monro* (MH-RM) increases linearly with the number of latent variables. It has proven to be accurate and relatively efficient for MIRT model estimation (Cai, 2010; Yavuz & Hambleton, 2017). However, compared to alternative approaches, model estimation with the MH-RM algorithm is still computationally demanding if more than one latent variable is specified in the model.

According to Liu et al. (2018), contemporary MIRT is a convergence of developments from test theory and confirmatory factor analysis (CFA). This means that certain types of CFA models and IRT models are equivalent (Takane & De Leeuw, 1987). Building on this assumption, Muthén (1984) proposed a *limited information*(LI) approach in which the polychoric correlation matrix of the response variables is used for parameter estimation. LI methods are usually computationally more efficient than FI methods and commonly used in practice. *Pairwise maximum likelihood* (PML) is a specific type of LI method which (like MML) uses a likelihood function for parameter estimation and maximizes the log-likelihoods associated with all item pairs (Katsikatsou et al., 2012). In this article, however, we focus on the most widely used LI estimate, which goes back to Muthén (1984). In the following, we therefore use ordinal factor analysis as a broad term for (multidimensional) IRT models estimated via polychorics (Maydeu-Olivares et al., 2011; Shi et al., 2020).

In IRT, it is generally assumed that the item parameters are independent of any covariates of the observed variables in the population of test takers. Such covariates may be demographic characteristics such as age, gender or education level. Violations of this assumption are interpreted as differential item functioning (DIF; Millsap, 2012; Osterlind & Everson, 2009). In practice, DIF may be detected by pre-specifying subgroups for which measurement invariance is not assumed. Alternatively, one can use the score-based test for parameter instability (Zeileis & Hornik, 2007) to detect DIF. This test focuses on identifying parameter instability through an analysis of the relation between model parameters and person covariates. It tests the null hypothesis that model parameters remain invariant across all values of person covariates. The score-based test is computed using the model scores, that is, the partial derivative of the casewise contributions to the objective function with respect to the model parameters (Merkle & Zeileis, 2013). It has been applied to a variety of different psychometric models, including factor analysis (Merkle et al., 2014), Bradley–Terry models (Strobl et al., 2011), binary and polytomous Rasch models (Komboz et al., 2018; Strobl et al., 2015), logistic IRT models (Debelak & Strobl, 2019a), mixed models (Fokkema et al., 2018), as well as the two-parameter normal ogive model via the PML estimation method (Wang, Strobl, et al., 2018). It is, however, currently not applicable to the GRM via ordinal factor analysis (i.e. LI estimation via polychorics).

We propose a method to efficiently approximate individual model scores, that is, the partial derivative of the casewise contributions to the objective function, for ordinal factor models. With this method, it is possible to apply score-based tests to such models. The score-based test for parameter instability can therefore be applied to MIRT models, specifically multidimensional GRMs, with reasonable computational effort.

We simulate data based on two (uni- and multidimensional) GRMs and systematically investigate the performance of the proposed score-based test. We compare our approach to tests based on models fitted with FI estimation under various conditions. Furthermore, we investigate the distribution of the scores estimated with the proposed method by comparing the correlations of model score contributions from different fitting approaches.

In the following, we describe the methodological background of this paper and how score-based tests for parameter instability can be used to detect DIF. We further introduce ordinal factor analysis and subsequently present our approach to compute individual model scores for ordinal factor models. Next, we present simulations with different scenarios to test the performance of score-based tests based on models fitted with both LI and FI estimation. Following this, we apply models fitted via different estimation methods to real data and compare the computation times and the results of the score-based tests for parameter instability. In the last section, the results are discussed.

2 | METHODOLOGICAL BACKGROUND

2.1 | Model definition

In IRT models, the latent variable, denoted as ξ, typically represents the respondent's ability that is assumed to underlie their response patterns. Let the graded responses be represented by the observed variable Y_j, for a given item j. Usually, IRT models are estimated based on ordered observed variables, wherein $i = 1, \ldots, n$, respondents choose from a range of ordered response categories $k_j = 1, \ldots, l_j$, for items $j = 1, \ldots, p$. For simplicity, we assume that all items have the same number of categories, such that $k_j = k \forall j = 1, \ldots, p$. In a multidimensional GRM, $\boldsymbol{\xi}$ is a $m \times 1$ vector containing all latent variables $\xi_q \forall q = 1, \ldots, m$. An observed variable Y_j may be associated with multiple latent variables.

In the GRM, the probability of answering in a category smaller or equal to a certain ordered category k depends on the (multidimensional) distribution of the latent variables as well as on the model's parameters. The threshold parameters τ_{jk} represent the boundaries between the categories. The threshold locations determine the difficulties of the item categories. The discrimination parameters $\boldsymbol{\lambda}_j$ denote the loadings of the items on the latent variables. The relationship between the latent variable and the response variables is defined by the cumulative category response function, that is

$$P(Y_j \leq k | \boldsymbol{\xi}, \theta) = \Phi(\tau_{jk} - \boldsymbol{\lambda}_j' \boldsymbol{\xi}), \tag{1}$$

where Φ is the distribution function of the standard normal distribution. It is used as a link function to convert a linear function into a probability function. The link function is also known as probit function or normal ogive function. Alternatively, a logit function can be used for the GRM (Samejima, 1997).

The model parameter vector θ contains all freely estimated threshold parameters τ_{jk}, all freely estimated discrimination parameters λ_{qj} that make up the $m \times 1$ vector $\boldsymbol{\lambda}_j$, as well as all freely estimated latent variable variances and covariances, such that

$$\begin{aligned} \theta \; = \{ & \tau_{11}, \ldots, \tau_{pl}, \lambda_{11}, \ldots, \lambda_{mp}, \\ & \mathrm{Var}(\xi_1), \ldots, \mathrm{Var}(\xi_m), \\ & \mathrm{Cov}(\xi_1, \xi_2), \ldots, \mathrm{Cov}(\xi_{m-1}, \xi_m) \}. \end{aligned} \tag{2}$$

Note that $\mathrm{Var}(\xi_q)$ is fixed to 1 if λ_{q1} is freely estimated (and vice versa).

2.2 | Differential item functioning

In the context of IRT models, differential item functioning (DIF) arises when an item's characteristics are related to person covariates. For instance, covariates such as ethnicity, education or gender may have an impact on, for example, the difficulty of an item. This means that one or more items of a test have different difficulties for subgroups with different ethnicity, education or gender. Let Z be a covariate that induces such a DIF effect. In this case, the item parameters in $\boldsymbol{\theta}$ deviate across the distribution of Z. If Z is independent of the latent variable, DIF occurs when the probability of responding to an item in a particular category differs between two individuals with the same ability (i.e. the same values on $\boldsymbol{\xi}$) solely due to their different values on Z. Practically, undetected DIF may lead to a misinterpretation of group differences concerning latent variables (Wang, Strobl, et al., 2018). Thus, DIF analyses are important in the practice of test validation (Walker, 2011). Note, however, that DIF is fundamentally different to *impact*, which means that the distribution of the latent variable depends on Z. For example, two subgroups with different ethnicity, education and gender may differ with respect to the values on the latent variable, but the difficulties of the test items may be equal across these groups. If impact of the latent variable is expected, testing for DIF requires a model in which the item parameters can differ between groups while controlling for group differences in the latent variable distribution (Belzak & Bauer, 2020; Sterner et al., 2024).

As mentioned above, DIF is closely related to the concept of measurement invariance, which is a concept primarily used in factor analysis. Measurement invariance in a model is established by the conditional independence of all observed variables and all potentially confounding covariates (Sterner et al., 2024). For a model with p observed response variables, this rule can be expressed as

$$Y_i \perp\!\!\!\perp \mathbf{Z} | \boldsymbol{\xi}_i, \forall i = 1, \ldots, p, \tag{3}$$

where Y_i is the observed response variable for item i, $\mathbf{Z}$ is the vector of all potentially confounding covariates and $\boldsymbol{\xi}_i$ is the vector of latent variables pertaining to item i. For an MIRT model, it follows from Equation (3) that the $\boldsymbol{\xi}_i$-conditional probability of answering to item i is independent from $\mathbf{Z}$, which means that there is no DIF. For simplicity, we refer to DIF as measurement non-invariance in IRT models.

Traditional approaches of empirical testing for DIF require the prespecification of subgroups for which DIF is assumed. For a focal subgroup and a reference subgroup, differences in item parameters can be tested for. This can be done for single items. This way, one can detect items with DIF so that this item can, for instance, be removed from the scale. For example, the subgroups tested for DIF are divided at the median of the metric covariate Z. In this case, two distinct subgroups are defined and the likelihood ratio (LR) test can be applied. With the LR test, an augmented model, permitting variation in all item parameters across the two groups, is tested against a baseline model where all item parameters are constrained to be equal between the reference and focal groups (Bulut & Suh, 2017). If the likelihood ratio of these two models is significantly different from one, researchers must assume the presence of DIF between these two groups. In practice, prior specification of subgroups potentially subjected to DIF can be difficult, especially in situations where there are a multitude of potential splitting points on Z. As researchers might not have strong assumptions which groups might be affected by DIF, certain subgroups exhibiting DIF might remain undiscovered.

2.3 | Score-based test for parameter instability

A solution to this problem was proposed by Zeileis and Hornik (2007) who presented a family of generalized M-fluctuation tests for testing parameter instability with respect to observed metric, ordinal and categorical variables. In the following, we refer to these tests as *score-based tests*. They are applicable to a wide range of IRT models to detect DIF (Debelak & Strobl, 2019a; Schneider et al., 2022). The score-based test is a global test for parameter instability. Usually, all freely estimated model parameters are tested for instability when

the score-based test is applied to a fitted model. The application of the score-based test to MIRT models for DIF detection presupposes that no impact of the latent variable is assumed. If differences in the latent variable are assumed across prespecified groups, one can apply the score-based test to a multiple-group MIRT model, in which the means and variances of the latent variable can differ for predefined subgroups (Bock & Zimowski, 1997; Debelak et al., 2022; Debelak & Strobl, 2019a). Note that such multiple-group models require one or more anchor items to make sure that the latent variable is measured on the same scale across groups. In single-group MIRT models, such group differences in the latent variable distributions are mistaken for DIF if the score-based test is used for DIF detection. In this paper, we only consider single-group MIRT models without differences in the latent variable between subgroups.

Another prerequisite for the score-based test is that an *M-estimator* is used to fit the model. If this is the case, parameter instability of the fitted model with respect to a covariate can be investigated. Following Stefanski and Boos (2002), an M-estimator $\widehat{\boldsymbol{\theta}}$ is defined as the solution to the equation

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{y}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{4}$$

where $\boldsymbol{\psi}$ is a $1 \times \|\boldsymbol{\theta}\|$ matrix. Note that $\|\cdot\|$ denotes vector length.

The function $\boldsymbol{\psi}$ is the first derivative of the objective function that is minimized to estimate the model parameters. In the context of marginal maximum likelihood (MML) estimation, which is a common full information estimation approach for IRT models, the objective function is the negative log-likelihood function. Following Baker and Kim (2004, pp. 160–164), the marginal likelihood L of the observed data is

$$L(\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} P(\mathbf{y}_i), \tag{5}$$

where $\mathbf{y}_i = \{y_{i1}, \ldots, y_{im}\}$ is the response pattern of respondent i. The probability of the individual response pattern of respondent i is

$$P(\mathbf{y}_i) = \int P(\mathbf{y}_i|\boldsymbol{\xi}_i, \boldsymbol{\theta}) g(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \tag{6}$$

where $\boldsymbol{\xi}_i$ are the values of respondent i on the latent variables (in IRT these are also referred to as person parameters). These values are drawn from the specific multivariate distribution $g(\boldsymbol{\xi}_i)$. Under the usual conditional independence assumption of the GRM, $P(\mathbf{y}_i|\boldsymbol{\xi}_i, \boldsymbol{\theta})$ follows from Equation (1). The derivative of the log likelihood with respect to some parameter $\varkappa$ is

$$\frac{\partial \log L(\mathbf{Y}, \varkappa)}{\partial \varkappa} = \sum_{i=1}^{n} \frac{1}{P(\mathbf{y}_i)} \frac{\partial}{\partial \varkappa} P(\mathbf{y}_i) = \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{y}_i, \varkappa) = 0, \tag{7}$$

where $\frac{\partial P(\mathbf{y}_i)}{\partial \varkappa}$ differs for each parameter $\varkappa$ in $\boldsymbol{\theta}$.[1] The individual contributions to the first derivative of the log likelihood with respect to the M-estimator $\widehat{\boldsymbol{\theta}}$ are also referred to as the *score contributions* of the fitted model. This is why the generalized M-fluctuation test is called score-based test.

The null hypothesis of the score-based test, which states that model parameters are invariant, is rejected if the empirical fluctuation during parameter estimation with respect to Z is improbably large. To estimate the empirical fluctuation, the individual model scores $\boldsymbol{\psi}(\mathbf{y}_i, \widehat{\boldsymbol{\theta}})$ are computed for all individuals

1In Debelak and Strobl (2019b), other examples of $\boldsymbol{\psi}$ can be found.

i in the sample. If the model parameters deviate across the distribution of a metric covariate Z, then a transition from positive to negative scores for lower values on Z to higher values on Z (or vice versa) is expected (see left hand side of Figure A3). The scores are then cumulated according to the order of the covariate of interest Z to compute the cumulative score process

$$\mathrm{CSP}(H) = \widehat{\mathbf{B}}^{-1/2} \frac{1}{\sqrt{n}} \sum_{h=1}^{H} \boldsymbol{\psi}(\mathbf{y}_{(h|Z)}, \widehat{\theta}), \tag{8}$$

where $(h|Z)$ denotes the h-th ordered observation with respect to the covariate Z. The transition from positive to negative scores is captured as a clearly noticeable peak in the cumulative sum process (see right hand side of Figure A3). The sum process is scaled by an estimate $\widehat{\mathbf{B}}$ for the covariance matrix $\mathrm{cov}(\boldsymbol{\psi}(\mathbf{Y}, \widehat{\theta}))$ to decorrelate the scores so that the score processes for all parameter estimates in $\widehat{\theta}$ are independent from each other. By analysing the CSP, a possible systematic change from positive to negative scores across the covariate can be detected.

Different kinds of test statistics can be derived from the CSP to capture the fluctuation across all parameter estimates in $\widehat{\theta}$. For metric covariates, the maximum Lagrange multiplier (*maxLM*), the double maximum (*DM*) and the Cramér-von-Mises (*CvM*) test statistics are available. The unordered LM test statistic (*LMuo*), which is based on the sum of the values in each category, is used to assess instability in relation to categorical covariates where it is not possible to order the values (Merkle & Zeileis, 2013). For ordered covariates, the ordered maximum LM (*maxLMo*) and the "weighted" double maximum (*WDMo*) statistic can be used (Merkle et al., 2014). Critical values associated with these test statistics can either be obtained through closed-form solutions of certain functions (*DM*, *WDMo*, *LMuo*), through tables of critical values obtained from simulation (*maxLM*, *CvM*) or through repeated simulation of Brownian Bridges (*maxLMo*). All these test statistics are implemented in the `strucchange` package in R (Zeileis et al., 2015).

As mentioned before, the score-based test for parameter instability is applicable for many different kinds of IRT models. However, MIRT models are commonly fitted via FI estimation, namely with the MML estimator (Schneider et al., 2022), such that individual score contributions can be computed as terms of the derivative of the marginal log-likelihood (Baker & Kim, 2004; Debelak & Strobl, 2019a). For simple IRT models, such as the Rasch model (Rasch, 1960) or the 2PL model by Birnbaum (1968), FI estimation is very efficient and repeated model fittings in a recursive partitioning algorithm are computationally feasible (see Komboz et al., 2018; Strobl et al., 2015). However, this is not the case for complex MIRT models. For these models, LI estimation, as common in ordinal factor analysis, is much quicker (Forero & Maydeu-Olivares, 2009). Therefore, a method for estimating individual score contributions for ordinal factor models is an important prerequisite for the efficient application of the score-based test.

2.4 | Full information estimation

The marginal maximum likelihood (MML) estimation approach via the expectation maximization (EM) algorithm (Bock & Aitkin, 1981; Jöreskog & Moustaki, 2006) iteratively estimates the true probabilities of each observed response pattern. In the first step of the algorithm, the latent variable is estimated (E-step), and in the second step, the model parameters are optimized (M-step). However, for this full information (FI) estimation method, multidimensional integrals are evaluated in the estimation process. Intensive computations are required, especially if latent variables in the MIRT model are correlated (Forero & Maydeu-Olivares, 2009). Efforts to reduce computation time have been made by Meng and Schilling (1996) via the Monte Carlo EM algorithm and later via the Markov Chain Monte Carlo (MCMC) algorithm (Bolt & Lall, 2003; Kim & Bolt, 2007). The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm is building on these advances (Cai, 2010). The algorithm has initially been proposed for exploratory factor analysis. It synthesizes a type of MCMC algorithm, the Metropolis–Hastings algorithm (Hastings, 1970; Metropolis et al., 1953), with the Robbins–Monro

method (Robbins & Monro, 1951) for stochastic approximation. Its complexity increases linearly with the number of latent variables. In the following, we will compare the performance of the MML estimation approach via the MH-RM algorithm with the performance of the limited information estimation approach used for ordinal factor analysis. We will focus on computation time and score-based parameter instability test results.

2.5 | Ordinal factor analysis

Using the classic maximum likelihood approach for CFA (see Jöreskog, 1969) to fit (multidimensional) IRT models introduces model misspecification because the common CFA assumes linear relationships between continuous and normally distributed observed variables and continuous factors (Maydeu-Olivares et al., 2011). Thus, in order to include ordered observed variables in CFA models, continuous latent response variables $\mathbf{Y}^*$ are assumed to underlie the observed ordered variables $\mathbf{Y}$ (Takane & De Leeuw, 1987), that is

$$\mathbf{Y}^* = \boldsymbol{\lambda}'\boldsymbol{\xi} + \boldsymbol{\epsilon}. \tag{9}$$

The mean vector of $\mathbf{Y}^*$ is $E(\mathbf{Y}^*) = \boldsymbol{\lambda}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the mean vector of $\boldsymbol{\xi}$. The (co-)variances matrix of the $\mathbf{Y}^*$ is $\mathrm{Var}(\mathbf{Y}^*) = \boldsymbol{\lambda}'\boldsymbol{\Psi}\boldsymbol{\lambda} + \boldsymbol{\Theta}$, where $\boldsymbol{\Psi}$ is the covariance matrix of $\boldsymbol{\xi}$ and $\boldsymbol{\Theta}$ is the covariance matrix of $\boldsymbol{\epsilon}$. The scale of the latent response variables is a priori indeterminate. Thus, in multi-group models, the latent response variables are usually standardized for at least one group (Muthén, 1984).

The latent response variable of item j is related to the observed ordered variable of item j via a threshold relation, that is

$$Y_j = k \text{if} \tau_{j(k-1)} < y_j^* \leq \tau_{jk}, \tag{10}$$

where $\tau_{j0} = -\infty$ and $\tau_{jl} = \infty$. Thus, for every item j, there is one threshold parameter τ_{jk} less than the total number of ordered categories l within item j. Note that the probability of Y_j being greater than k may be derived from the threshold parameters, that is,

$$P(Y_j > k) = P(Y_j^* > \tau_{jk}) = \Phi(-\tau_{jk}). \tag{11}$$

Building on this assumption, Muthén (1984) proposed a method in which parameters of CFA models including ordered observed variables are estimated by minimizing the discrepancy between the polychoric correlation matrix of the observed variables and the model-implied covariance matrix. Parameter estimation via polychorics is also referred to as a form of limited information (LI) estimation, as it only uses information from bivariate relations of the observed variables. The estimation of the thresholds, as defined in Equation (10), is performed as a first step in the model fitting process. Furthermore, in this phase, bivariate polychoric correlations ρ_{js} are computed for all $j, s = 1, \ldots, p$ when $j \neq s$, following the approach established by Olsson (1979). These polychoric correlations quantify the degree of linear dependence between the variables Y_j^* and Y_s^* for $j \neq s$.

After the estimation of thresholds and polychoric correlations, the model parameters in $\boldsymbol{\theta}$ are estimated through minimization of the objective function

$$F_{\mathrm{OFA}}(\boldsymbol{\theta}) = [\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\theta})]'\mathbf{W}^{-1}[\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\theta})], \tag{12}$$

where $\hat{\boldsymbol{\kappa}}$ and $\boldsymbol{\kappa}(\boldsymbol{\theta})$ are the vectors of the sample and model implied polychoric correlation matrices. Different choices for the positive-definite weight matrix $\mathbf{W}$ lead to different estimators (Shi et al., 2020). In Weighted Least Squares (WLS; Muthén, 1984) estimation, $\mathbf{W}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\kappa}}$. The WLS estimator may produce unstable results for small sample sizes and large models (Flora & Curran, 2004;

Garnier-Villarreal et al., 2021; Wang, Su, et al., 2018). However, it usually performs equally well or better than FI estimation for large sample sizes (Forero & Maydeu-Olivares, 2009). In this paper, we therefore focus on the WLS estimator for ordinal factor analysis.

2.6 | Approximated scores for ordinal factor analysis

As we assume a specific model structure for a multidimensional GRM, we may denote the assumed model (see Equation 1) as the *structured* model, or H_0. It may be tested against the *unstructured* or *saturated* model (H_1) that does not impose any restrictions on the thresholds or the covariance matrix. The vector $\boldsymbol{\kappa}$ contains the saturated model parameters $(\boldsymbol{\tau}, \boldsymbol{\sigma}^*)'$, where $\boldsymbol{\tau}$ is the vector of threshold parameters, and $\boldsymbol{\sigma}^* = \text{vech}[\text{Cov}(\mathbf{Y}^*)]$ contains the vectorized non-redundant elements of the model implied covariance matrix. The size of $\boldsymbol{\kappa}$ is $[p(l-1) + p(p-1)/2] \times 1$ which we refer to as $p^* \times 1$ in the following.

The first derivative of $\boldsymbol{\kappa}$ with respect to θ is

$$\Delta = \frac{\partial \boldsymbol{\kappa}(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \boldsymbol{\tau}(\theta)}{\partial \theta} \\ \frac{\partial \boldsymbol{\sigma}^*(\theta)}{\partial \theta} \end{pmatrix}. \tag{13}$$

We apply the chain rule to get the first derivative of the objective function with respect to θ, that is the $1 \times \|\theta\|$ matrix

$$\frac{\partial F_{\text{OFA}}(\theta)}{\partial \theta} = \frac{\partial F_{\text{OFA}}(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \frac{\partial \boldsymbol{\kappa}(\theta)}{\partial \theta} = -2[\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\theta)]' \mathbf{W}^{-1} \Delta. \tag{14}$$

Note that Equation (14) is not an individual function, meaning that it does not refer to a single observation i and cannot be used for the score-based parameter instability test. To our knowledge, it is not possible, with reasonable effort, to formulate the gradient of Equation (12) as an individual function.

Therefore, to compute scores that can then be used for the score-based parameter instability test, we focus on an alternative approach to MML estimation. Muthén (1997) and Reboussin and Liang (1998) proposed a *generalized estimating equations* (GEE) approach for the estimation of parameters in (multidimensional) latent variable models with ordered response variables. In Appendix S1, we describe how the GEE estimation method is applied to MIRT models based on non-binary response variables. This approach minimizes a set of estimating equations, that is

$$\sum_{i=1}^{n} \Delta' \mathbf{W}_{\text{GEE}}^{-1} \mathbf{e}_i = \mathbf{0}, \tag{15}$$

where $\mathbf{e_i}$ is the vector of empirical deviations of the first- and second-order empirical moments in the data from the true first- and second-order moments (see Equations 10–14 in Appendix S1). The first-order empirical moments in the data are the indicator variables, that is,

$$1_{y_i > k} = \begin{cases} 1, & \text{if } y_i > k \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

for all individuals $i = 1, \ldots, n$, all items $j = 1, \ldots, p$, and all categories minus one $k = 1, \ldots, (l-1)$. The weight matrix used in GEE estimation, that is, $\mathbf{W}_{\text{GEE}}$, is defined as the working covariance matrix of first- and second-order empirical moments of individual i (see Equations 16–18 in Appendix S1). The $\boldsymbol{\Delta}$-matrix is the derivative of the saturated model with respect to the model parameters θ (see Equation 19 in Appendix S1).

In contrast to Equation (14), the estimating equations in Equation (15) are individual functions that each refer to a single observation i and add up to zero. They are the individual contributions to the derivative of the objective function of the GEE approach. The model parameters in $\boldsymbol{\theta}$ are estimated by iteratively updating the estimator, that is, solving the set of quadratic estimating equations for $\boldsymbol{\theta}$. The solution to the set of quadratic estimating equations are the model scores obtained through GEE estimation. Using the GEE estimation approach leads to slightly different parameter estimates than ordinal factor analysis (i.e. WLS). Our goal is to approximate the GEE scores that would have resulted if the parameters estimated using the GEE approach were exactly the same as those estimated using ordinal factor analysis. We claim that these approximated scores can be used for the score-based parameter instability test.

We learn from GEE estimation (e.g. equation 28 in Muthén, 1997) that an empirical deviation vector $\mathbf{e_i}$, defined on the individual level, can be used for the individual estimating function (Equation 15). Let an alternative set of individual estimating equations be

$$\mathbf{s}^*{}_i = \begin{pmatrix} (y^*_{i1} - \tau_1)(y^*_{i2} - \tau_2) \\ (y^*_{i1} - \tau_1)(y^*_{i3} - \tau_3) \\ \vdots \\ (y^*_{ip-1} - \tau_{p-1})(y^*_{ip} - \tau_p) \end{pmatrix}, \quad \sum_{i=1}^{n} \Delta' \mathbf{W}^{-1} \begin{pmatrix} \mathbf{y}^*{}_i - \boldsymbol{\tau} \\ \mathbf{s}^*{}_i - \boldsymbol{\sigma}^* \end{pmatrix} = \mathbf{0}, \tag{17}$$

where $\mathbf{y}^*{}_i$ contains the values of individual i on the latent response variables for all items $j = 1, \ldots, p$. The vector $\mathbf{s}^*{}_i$ can be referred to as the vector of the true second-order moments. Note that both $\mathbf{y}^*{}_i$ and $\mathbf{s}^*{}_i$ cannot be observed. However, for a direct translation of the GEE estimation method to ordinal factor analysis, it would be necessary to observe $\mathbf{y}^*{}_i$ and $\mathbf{s}^*{}_i$. Also, for such a translation, the $p^* \times p^*$ matrix $\mathbf{W}$ in Equation (17) would be an estimator of the working covariance matrix of the vectors $(\mathbf{y}^*{}_i, \mathbf{s}^*{}_i)'$ across all individuals $i = 1, \ldots, n$. In the following, we will show how to approximate the GEE scores without having to observe $\mathbf{y}^*{}_i$ and $\mathbf{s}^*{}_i$.

Let us assume that the latent response variables in the model be normally distributed and that the model's residuals $\epsilon_j = Y^*_j - \boldsymbol{\lambda}'_j \boldsymbol{\xi}$ (see Equation 9) are independent and identically distributed. If this is the case, then $\widehat{\boldsymbol{\kappa}} = \boldsymbol{\kappa}(\boldsymbol{\theta})$, that is, the assumed model fits the data perfectly and the empirical deviation vector in Equation (17) is equal to $\begin{pmatrix} \mathbf{y}^*{}_i - \bar{\mathbf{y}}^* \\ \mathbf{s}^*{}_i - \bar{\mathbf{s}}^* \end{pmatrix}$, where $\bar{\cdot}$ represents the arithmetic mean.

To compute individual score contributions based on Equation (12), we apply the logic of Equation (17) to the non-binary case. The aim is to mimic the scores produced by the estimation function in Equation (17). However, the individual values of the latent response variable distribution $\mathbf{y}^*$ are not identifiable. Thus, the true second-order moments $\mathbf{s}^*$ are not identifiable either. We therefore replace the empirical deviation vector with $\begin{pmatrix} \text{vec}(\mathbf{1}_{\mathbf{y}_i}) - \text{vec}(\overline{\mathbf{1}_\mathbf{Y}}) \\ \mathbf{s}_i - \bar{\mathbf{s}} \end{pmatrix}$. The vector $\text{vec}(\mathbf{1}_{\mathbf{y}_i})$ contains the indicator variables for all items $j = 1, \ldots, p$, and all categories minus one $k = 1, \ldots, (l-1)$ (see Equation 10 in Appendix S1). The vector $\text{vec}(\overline{\mathbf{1}_\mathbf{Y}})$ of size $p(l-1)$ contains the arithmetic means of the indicator variables across all individuals. Furthermore, we replace the weight matrix of Equation (17) with the weight matrix of Equation (12). This way, we account for the multivariate non-normality within the observed variable distribution. Thus, we claim that the individual score contributions of an ordinal factor model fitted using WLS can be estimated as follows

$$\mathbf{s}_i = \begin{pmatrix} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2) \\ (y_{i1} - \bar{y}_1)(y_{i3} - \bar{y}_3) \\ \vdots \\ (y_{ip-1} - \bar{y}_{p-1})(y_{ip} - \bar{y}_p) \end{pmatrix},$$
$$\sum_{i=1}^{n} \tilde{\psi}(\mathbf{y}_i, \theta) = \sum_{i=1}^{n} \Delta' \mathbf{W}^{-1} \begin{pmatrix} \text{vec}(\mathbf{1}_{\mathbf{y}_i}) - \text{vec}(\overline{\mathbf{1}_{\mathbf{Y}}}) \\ \mathbf{s}_i - \bar{\mathbf{s}} \end{pmatrix} = \mathbf{0}. \tag{18}$$

We refer to Equation (18) as the *approximated score function* of the WLS estimation method that can be used for the score-based parameter instability test.

2.7 | Computational details

The R implementation of the proposed method, replication materials for all simulations, all simulation results as well as Appendix S1 are provided in the following OSF repository: https://osf.io/hmwpc/?view{_}only=69ed2919e7a64db2b0354f99243c307c. All simulations and real data applications were executed on a 20 core, 170 GB RAM server. The proposed method to compute individual model scores for ordinal factor models is implemented in the functions `lavScores()` and `estfun.lavaan()` in the latest version (since version 0.6-18) of `lavaan` (Rosseel, 2012).

3 | SIMULATION

We simulated data to fit two different IRT models: a unidimensional model with five observed variables Y_j (Figure A1) and a multidimensional model with nine observed variables Y_j (Figure A2). To simulate model-compliant data, first, true latent variable scores were simulated for all latent variables in the model. Then, values of the conditional probabilities $P(Y_j = k|\boldsymbol{\xi}, \theta)$ were computed for all categories of all items. On the basis of these conditional probabilities, values for five ordinal response variables with k categories each were sampled.

From these conditional probability functions, DIF effect sizes can be calculated. Following Chalmers (2023), the scoring function that is

$$S(\boldsymbol{\xi}, \theta) = \sum_{k=1}^{l} (k-1) \cdot P(Y_j = k|\boldsymbol{\xi}, \theta), \tag{19}$$

is used to compute the DIF effect size of an item j. The *Noncompensoratory DIF* (NCDIF) value quantifies the average deviation of the response function of an item j between a focal group (F) and a reference group (R). It is defined as

$$\text{NCDIF} = \frac{\sum_{i=1}^{n_F} [S(\xi_i, \theta_F) - S(\xi_i, \theta_R)]^2}{n_F}. \tag{20}$$

Using the true values for ξ_i, θ_F and θ_R from the simulation, we are able to compute the true NCDIF values of the items in the simulated data sets. To illustrate how parameter fluctuation affects parameter estimation, we report a DIF effect size, that is, the NCDIF value, for one specific item (Item 2) in 24 different simulation scenarios: two different models, that is, the unidimensional model (Figure A1) and the multidimensional model (Figure A2), four different numbers of threshold parameters $k \in \{1, 2, 4, 6\}$ and three different scenarios for parameter fluctuation in the data (see below). For each scenario, we

simulate 1000 simulation samples of $n = 1000$ in which there is a focal group and a reference group. For each group, the parameters of the (multidimensional) GRM are randomly drawn. For each sample, the NCDIF values are computed. The average NCDIF values (i.e. the arithmetic means) are shown in Table A3.

To test the performance of the score-based test, we created 36 different simulation scenarios for each model: four different numbers of threshold parameters $k \in \{1, 2, 4, 6\}$, which means that the simulated ordinal observed variables Y_j have two, three, five or seven categories, three different sample sizes $n \in \{500, 1000, 2000\}$, and three different scenarios for parameter fluctuation in the data. For each of the simulated samples, we created one numerical covariate (Z_{num}) ranging from 1 to 200, one ordinal (Z_{ord}) and one categorical (Z_{cat}) covariate with scores on a five-point response scale. Each simulation sample consists of a focal and a reference group of size $n/2$ that both fit the corresponding model but have different parameter values.

The three simulated scenarios for parameter fluctuation are: All parameter values differ between the focal and the reference group, only the threshold parameters of the first item τ_{1k} (for the unidimensional model) or the threshold parameters of the first and the second item (for the multidimensional model) differ between the focal and the reference groups, or only the discrimination parameters λ_j (of all items) differ between the two subsets. Thus, each simulation sample for each model for each simulation scenario is of size n and exhibits DIF with respect to the covariates Z_{num}, Z_{ord} and Z_{cat}. This means that all score-based tests for parameter instability which are applied to the covariates in all data sets should result in significant p-values. For each simulation scenario and model, 1000 simulation samples (i.e. repetitions) were generated. We denote the percentage of simulated samples for which the p-value of the score-based test is smaller than .05 as the power of the score-based test.

For each of the simulated samples, ordinal factor models are fitted with the WLS estimator using the `lavaan` package (Rosseel, 2012) and (multidimensional) GRMs are fitted via FI estimation, namely the MML estimator, using the `mirt` package (Chalmers, 2012). With FI estimation, the unidimensional model is fitted via the EM algorithm and the multidimensional model is fitted via the MH-RM algorithm. Each of the fitted models is tested for parameter instability using the *maxLM*, *DM* and *CvM* test statistics on Z_{num}, the *WDMo* and *maxLMo*test statistics on Z_{ord} and the *LMuo* test statistic on Z_{cat}.

We further conduct additional simulations with data that do not exhibit DIF, that is, the values of the covariates were simulated randomly. This means that all score-based tests for parameter instability which are applied to the covariates in all data sets should not result in significant p-values. We denote the percentage of simulated samples for which the p-value of the score-based test is smaller than .05 as the Type I error rate of the score-based test.

To see how the approximated scores of the ordinal factor model are distributed, we additionally simulate two data sets to fit the unidimensional model (Figure A1). One data set has binary response variables and the other data set has response variables with four ordered response categories. We simulate two other data sets to fit the multidimensional model (Figure A2). We then use three different approaches to fit the models to the data: ordinal factor analysis (LI estimation), FI estimation and GEE estimation (see Appendix S1). Subsequently, the models scores are estimated for each model for each data set. The correlations of the model score contributions of each parameter in the respective model are shown in Table A1 (for the unidimensional model) and A2 (for the multidimensional model).

3.1 | Results

The means of the NCDIF values in Table A3 show that DIF effect sizes on one item are considerably lower if only the discrimination parameters differ between the focal and the reference group. This is reflected in the power results of the simulation for both the unidimensional and the multidimensional model. However, if only the thresholds of an item differ between focal and reference group, the DIF effect size of that item is similar to the case in which all parameters differ. From this result, we deduce that the power of the score-based test in the first simulation scenario (all parameters differ) most likely

does not differ significantly from a scenario in which only the threshold values of all items differ. The second simulated scenario for parameter fluctuation thus consists of only the threshold parameters of one item (i.e. 20% DIF in the unidimensional model), respectively, of two items (i.e. 22% DIF in the multidimensional model) differing between the focus and reference groups.

The results of the simulations with data based on the unidimensional model (see Figure A1) show that power generally increases with sample sizes and the number of response categories. For the proposed score-based tests for ordinal factor models as well as for tests based on GRMs fitted via FI estimation, power lies between 98% and 100% when there is parameter fluctuation with respect to all model parameters. Figure A4shows that given fluctuation with respect to the threshold parameters of the first item τ_{1k} only, sample sizes of at least $n = 2000$ are needed for $k = 4$ thresholds and sample sizes of at least $n = 1000$ are needed for $k = 6$ thresholds to achieve power of over 90% for all test statistics. For the simulated data sets with parameter fluctuation with respect to the discrimination parameters λ_j, power results for both ordinal factor models and for GRMs fitted via FI estimation are shown in Figure A5. For $k = 1$, sample sizes of $n = 2000$ are needed to achieve power of over 90% for all tests statistics. In general, with respect to power, the score-based test does not perform better for models fitted via FI estimation as compared to ordinal factor models.

Type I error results for the unidimensional model are generally within the expected range of 3% and 6% for all test statistics for ordinal factor models and for GRMs fitted via FI estimation for all sample sizes and numbers of thresholds. This indicates that the score-based test for ordinal factor analysis performs equally well as for the GRMs fitted via FI estimation when estimating unidimensional IRT models. The computation times for fitting the unidimensional GRM via FI estimation (using the EM algorithm) and the ordinal factor models are very similar (see Table A4).

Computation times for fitting the multidimensional model (see Figure A2) are much higher for FI estimation (using the MH-RM algorithm) compared to ordinal factor analysis with LI estimation (see Table A5), highlighting the benefits of ordinal factor analysis in this setting. The results also show that high power (100%) is achieved for both ordinal factor models and for GRMs fitted via FI estimation when there is parameter fluctuation with respect to all model parameters. Power results for the data sets with parameter fluctuation with respect to only the threshold parameters of item 1 and 2 are shown in Figure A6. Interestingly, the multidimensional model outperforms the unidimensional model in this simulation scenario. Here, sample sizes of $n = 1000$ suffice for models with two response categories to achieve power of over 90% for all test statistics. The power results of the score-based test when the discrimination parameters λ_j of all items differ between the focal and the reference group are shown in Figure A7. When only the discrimination parameters differ in data sets of $n = 500$ and $k = 1$, power lies between 29. 2% and 52. 8%. For data sets with $k = 2$, power is at least 72. 5%. Power results are generally very similar between ordinal factor models and for GRMs fitted via FI estimation. However, there are considerable differences between these two types of models regarding the Type I error (see Figure A8). Type I error rate is higher for the score-based tests applied to GRMs fitted via FI estimation. This is particularly the case for the *CvM*, *maxLM*, *maxLMo* and *WDMo* test statistics.

The correlations of the model scores for the unidimensional model in Table A1 show that the score contributions of the model fitted with the GEE approach correlate negatively with the scores of the models fitted with the LI (i.e. ordinal factor analysis) or the FI approach. This is because of the definition of the $\boldsymbol{\Delta}$-matrix for GEE estimation, where $\frac{\partial \text{vec}(\boldsymbol{\nu})(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ contains negative values (see Equation 19 in Appendix S1). The approximated model score contributions of the ordinal factor model correlate strongly with the score contributions of the model fitted with the GEE approach. The parameter estimates are expected to differ between the two approaches; therefore, perfect correlations of the score contributions are not expected. Interestingly, the correlations of the model score contributions from the GEE approach with the score contributions from the FI approach are lower for discrimination parameters and higher for threshold parameters. The correlations of the model score contributions from the LI approach with those from the FI approach are generally a bit lower than those from the LI approach with those from the GEE approach. The correlations of the model scores for the multidimensional model (Table A2) show a very similar pattern.

4 | REAL DATA APPLICATION

We demonstrate the application of score-based tests with (multidimensional) GRMs using data obtained from the LISS (Longitudinal Internet studies for the Social Sciences) panel administered by Centerdata (Tilburg University, the Netherlands). LISS is a longitudinal survey conducted annually, covering topics such as employment, education, income, housing and personality traits (Scherpenzeel & Das, 2010). We analyse the data from four survey waves that were conducted in 2008, 2009, 2011 and 2013. In the survey waves of 2010 and 2012, certain application-relevant items were not included. A total of 2893 individuals participated across all four waves of the survey. Our analysis focuses on five items from the Satisfaction with Life (SL) scale (Diener et al., 1985), which assesses life satisfaction. We excluded any cases that did not provide responses to all five items, resulting in a final sample size of 2888 individuals. The items were rated on a seven-point response scale. The specific wording of these items is displayed in Table A6.

We apply three different models of different sizes to the data. Model 1 has the same unidimensional GRM model structure used in the simulation (see Figure A1). The five items Y_j represent the SL scale in the first survey wave. Model 2 has a multidimensional GRM model structure with correlated latent variables and is shown in Figure A9. The items Y_{tj} represent the SL scale in survey waves one ($t = 1$) and two ($t = 2$). Model 3 is a *probit multistate IRT model with latent item effect variables for graded responses* (PIEG) in which one reference latent state variable $\boldsymbol{\eta}_t$ is assumed for every time point of measurement and one latent item effect variable $\boldsymbol{\beta}_i$ is defined for every item but the reference item (here: $j = 1$). In this model, the variances and covariances of the latent state variables, as well as the latent item effect variables and the covariances between them, are estimated. The discrimination parameters in the model are all fixed at 1 (Classe & Steyer, 2023). The model is shown in Figure A10.

We fit each model using three different estimation methods: ordinal factor analysis (using the WLS estimator), FI estimation (Model 1 via the EM algorithm, and Model 2 and Model 3 via the MH-RM algorithm) and common factor analysis. For common factor analysis, we use the robust maximum likelihood (MLR) estimator, since here the model fit statistics are corrected for the non-normality of the response variables (Li, 2016).

For every fitted model, we apply the score-based test with respect to three different background variables representing general characteristics of households and household members that participate in the LISS panel: Gender (categorical: "Female", "Male" and "Other"), urban character of place of residence (ordinal: five categories from "extremely urban" to "not urban") and individual age (metric). We do not assume an impact of any of the covariates on satisfaction with life. This is mainly due to methodological considerations. We do not want to specify an anchor item as we assume that the item characteristics of all five items may differ across the subgroups defined by the covariates. For the categorical covariate, we use the *LMuo*; for the ordinal covariate, we use the *WDMo*; and for the metric covariate, we use the *DM* test statistic. All three test statistics can be used with large models as they obtain their critical values through closed-form solutions of certain functions instead of default tables.

We analyse the fitted models with respect to the degree of model fit and the computation time of the model fitting process and apply the score-based test using the outlined covariates. Furthermore, for each model, we analyse the time needed to compute the empirical fluctuation process, which includes the computation of the model scores.

4.1 | Results

The results of the real data application displayed in Table A7 show that computation time increases when fitting larger ordinal factor models compared to smaller ones. However, compared to the considerable increase in computation time for fitting the GRMs via FI estimation, the increase in computation time for larger ordinal factor models is marginal. This agrees with the simulation results shown in Tables A4 and A5 and shows that FI estimation is not computationally efficient

for models with two or more non-orthogonal latent variables. Compared to FI estimation, ordinal factor analysis is computationally efficient, even for large models. When it comes to the results of the score-based tests, models fitted via FI estimation are very similar to ordinal factor models, at least for model 1 and model 2. For model 3, all *p*-values for the score-based test are smaller than 2.20E-16. Also, computing the empirical fluctuation process is particularly expensive for model 3 when fitted via FI estimation.

Comparing the results of the common factor models with the ordinal factor models shows that common factor analysis is computationally faster than ordinal factor analysis, especially for large models. These results are also shown in Table A7. Also, the model fit estimation results of common factor analysis (using the MLR estimator) are similar to those of ordinal factor analysis. However, there are considerable discrepancies with respect to the results of the score-based tests, particularity for the categorical and metric covariates for all model sizes. Note that in using common factor models for categorical data, model misspecification is introduced as, for instance, no threshold parameters are estimated.

5 | DISCUSSION

The results of our simulations show that score-based tests for parameter instability perform equally well for unidimensional GRMs fitted via FI estimation and for ordinal factor models. As there are no considerable differences regarding computation time, we conclude that fitting univariate IRT models and testing them for parameter instability is equally convenient using FI estimation or ordinal factor analysis.

However, the results of the simulation regarding the multidimensional model show that there are considerable differences in computation times when fitting the model via FI estimation (using the MH-RM algorithm) compared to ordinal factor analysis. The limited information method of ordinal factor analysis is 32–91 times faster than the MH-RM algorithm. The power results indicate that the proposed score-based test for unidimensional GRMs as well as for multidimensional GRMs implemented via ordinal factor models performs equally well as tests based on unidimensional GRMs fitted via FI estimation. However, when it comes to multidimensional GRMs, there are considerable specificity problems of the score-based test when applied to models fitted via FI estimation in contrast to ordinal factor analysis. Debelak et al. (2024) point out that increased Type I errors of the score-based test when applied to models fitted via FI estimation could be due to numerical inaccuracies of the MH-RM algorithm. Additional fine-tuning of the implementation of the algorithm in the `mirt` package may help to obtain accurate Type I error rates.

The distribution of the approximated scores of the ordinal factor model are generally very similar to the scores from the GEE estimation method. Note that, unlike LI estimation, the GEE estimation method optimizes the model scores to estimate the model's parameters. This takes a very long time, especially for multidimensional non-binary models. The fact that the scores are distributed similarly to the model scores estimated with the method proposed in this paper indicates that our approach is, in fact, a valid approximation and thus a computationally efficient alternative when it comes to parameter instability tests.

Note that the score-based test may also be applicable for GRMs fitted via PML (which is also a LI method). However, in their application Wang, Strobl, et al. (2018) focused on unidimensional two-parameter normal ogive models for dichotomous response variables.

The real data applications show that the results for the score-based tests are very similar for unidimensional ordinal factor models and models fitted via FI estimation. This matches the results of the simulation. For very large models, however, the discrepancy between score-based tests applied to ordinal factor models and GRMs fitted via FI estimation is considerable. Additionally, it appears that score-based tests for parameter instability produce different results for ordinal factor analysis compared to common factor analysis. We therefore conclude that ordinal factor analysis should be preferred over common factor analysis and GRMs fitted via FI estimation when testing for parameter instability in multidimensional GRMs.

Note that, within our simulated samples, the covariates $\mathbf{Z}$ are always independent from the latent variable distribution $\boldsymbol{\xi}$ (in both the unidimensional and the multidimensional case). This implies that only single-group MIRT models without differences in the latent variable between subgroups are considered. Also for the real data applications in this paper, we assume independence of the covariates from the latent variable distribution. Future research might investigate the performance of the score-based test for multiple-group ordinal factor models.

5.1 | Model-based recursive partitioning

Methods based on the score-based test can be very helpful in scenarios where there are a multitude of metric, ordinal or categorical covariates potentially causing DIF. In such contexts, data-driven methods such as Model-Based Recursive Partitioning (MOB; Zeileis et al., 2008) prove valuable for identifying subgroups in which DIF is present. This algorithm repeatedly splits a sample into subgroups based on covariates Z_r in $Z_1, \ldots, Z_R$ (referred to as partitioning variables) to form a decision tree (see Breiman et al., 1984). The score-based test for parameter instability can be used in such a recursive partitioning algorithm to account for parameter instability. When parameter instability is detected in a tree node during the partitioning process, that is, the score-based test for one of the partitioning variables falls below a predefined significance level, the partitioning variable Z_{r*} associated with the smallest p-value is selected for partitioning. The unique value of a partitioning variable that maximizes the respective score-based test statistic can be used as a split point (see Arnold et al., 2021). The MOB algorithm continues to partition different subgroups until the stopping criteria are met. This is usually the case when there is no more significant instability in a node or when the subsample in a node becomes too small to fit the model. However, the application of MOB in conjunction with ordinal factor models is not yet implemented in the available R packages. The quick computation of MOB trees for MIRT models may, among other things, be relevant for the estimation of unbiased latent variable scores (Classe & Kern, 2024). Thus, future research should further investigate the application of MOB to ordinal factor models, building on the technique proposed in this paper.

5.2 | Outlook

The efficient computation of individual model scores for MIRT models is not only useful for efficient computation of parameter instability tests. The proposed method may also be used to compute robust test statistics based on sandwich covariance matrices (Zeileis, 2006). Such robust corrections are already widely used in structural equation modelling with complete (Savalei, 2014) or incomplete (Savalei & Rosseel, 2022) data. Another possible area of application is model selection of non-nested models via Vuong tests, since the Vuong test statistics are generally calculated on the basis of the individual model scores (Schneider et al., 2020). With the method proposed in this paper, such advances can be extended to ordinal factor models.

AUTHOR CONTRIBUTIONS

Franz Classe: conceptualization; investigation; writing – original draft; methodology; visualization; software; validation; project administration; formal analysis. **Rudolf Debelak:** writing – review and editing; methodology; project administration. **Christoph Kern:** writing – review and editing; methodology; project administration.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at https://osf.io/hmwpc/?view_only=69ed2919e7a64db2b0354f99243c307c.

ORCID

Franz Classe https://orcid.org/0000-0003-1257-1719

REFERENCES

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology*, *11*, 564403.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bean, G. J., & Bowen, N. K. (2021). Item response theory and confirmatory factor analysis: Complementary approaches for scale development. *Journal of Evidence-Based Social Work*, *18*(6), 597–618.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). Springer.

Bolt, D. M. (2005). Limited-and full-information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 73–100). Lawrence Erlbaum Associates Publishers.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395–414.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Brooks/Cole Publishing.

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*, 51.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2023). A unified comparison of IRT-based effect sizes for dif investigations. *Journal of Educational Measurement*, *60*(2), 318–350.

Classe, F., & Kern, C. (2024). Latent variable forests for latent variable score estimation. *Educational and Psychological Measurement*, *84*, 1138–1172.

Classe, F. L., & Steyer, R. (2023). A probit multistate IRT model with latent item effect variables for graded responses. *European Journal of Psychological Assessment*, *40*, 172–183.

Debelak, R., Meiser, T., & Gernand, A. (2024). Investigating heterogeneity in IRTree models for multiple response processes with score-based partitioning. *British Journal of Mathematical and Statistical Psychology*, *78*, 420–439.

Debelak, R., Pawel, S., Strobl, C., & Merkle, E. C. (2022). Score-based measurement invariance checks for Bayesian maximum-a-posteriori estimates in item response theory. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 728–752.

Debelak, R., & Strobl, C. (2019a). Investigating measurement invariance by means of parameter instability tests for 2pl and 3pl models. *Educational and Psychological Measurement*, *79*(2), 385–398.

Debelak, R., & Strobl, C. (2019b). *Investigating measurement invariance by means of parameter instability tests for 2pl and 3pl models*. https://www.zora.uzh.ch/id/eprint/151192/2/AppendixA.pdf

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275.

Garnier-Villarreal, M., Merkle, E. C., & Magnus, B. E. (2021). Between-item multidimensional IRT: How far can the estimation methods go? *Psych*, *3*(3), 404–421.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.
Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4*, 45.
Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.
Jöreskog, K. G., & Moustaki, I. (2006). *Factor analysis of ordinal variables with full information maximum likelihood*. Unpublished report.
Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.
Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.
Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166.
Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949.
Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional item response theory. In P. Irwing & D. J. H. Tom Booth (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 445–493). Wiley Online Library.
Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356.
Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*(435), 1254–1267.
Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*, 569–584.
Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*, 59–82.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.
Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.
Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*(1), 73–90.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.
Muthén, B. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika. https://www.statmodel.com/download/Article_075.pdf
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.
Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
Reboussin, B. A., & Liang, K.-Y. (1998). An estimating equations approach for the liscomp model. *Psychometrika*, *63*, 165–182.
Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36.
Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 1–97.
Samejima, F. (1997). Graded response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.
Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 149–160.
Savalei, V., & Rosseel, Y. (2022). Computational options for standard errors and test statistics with incomplete normal and nonnormal data in sem. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 163–181.
Scherpenzeel, A. C., & Das, M. (2010). "True" longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies*. (pp. 77–104). Taylor & Francis.
Schneider, L., Chalmers, R. P., Debelak, R., & Merkle, E. C. (2020). Model selection of nested and non-nested item response models using Vuong tests. *Multivariate Behavioral Research*, *55*(5), 664–684.
Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*(5), 2101–2113.
Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 1–15.
Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, *56*(1), 29–38.

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*, 747–758.
Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the rasch model. *Psychometrika*, *80*(2), 289–316.
Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153.
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
ten Holt, J. C., van Duijn, M. A., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272–297.
Walker, C. M. (2011). What's the dif? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364–376.
Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behavioral Research*, *53*(3), 403–418.
Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, *83*, 132–155.
Yavuz, G., & Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement*, *77*(2), 263–274.
Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*, 1–16.
Zeileis, A., & Hornik, K. (2007). Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508.
Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.
Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., Hansen, B., Merkle, E. C., & Zeileis, M. A. (2015). *Package 'strucchange'. R package version*, 1–5.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Classe, F., Debelak, R., & Kern, C. (2025). Score-based tests for parameter instability in ordinal factor models. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–29. https://doi.org/10.1111/bmsp.12392

APPENDIX A

TABLES

TABLE A1 Correlation of model scores for a unidimensional GRM (see Figure A1) with binary and non-binary (four ordered categories) response variables fitted on a simulated data set with $n = 2000$ respondents. The model scores of three different fitted models are compared: SC_{OFA} meaning the approximated scores for a ordinal factor model, SC_{FI} meaning the scores for a model fitted with FI estimation and SC_{GEE} meaning the scores of a model fitted with GEE (see Appendix S1).

	$Cor(SC_{OFA}, SC_{GEE})$		$Cor(SC_{FI}, SC_{GEE})$		$Cor(SC_{OFA}, SC_{FI})$	
	Binary	**Non-binary**	**Binary**	**Non-binary**	**Binary**	**Non-binary**
$Var(\eta_1)$	.94	.97	.99	.95	.93	.96
λ_2	.91	.94	.88	.85	.96	.94
λ_3	.94	.93	.93	.80	.98	.91
λ_4	.94	.90	.96	.73	.99	.87
λ_5	.92	.91	.92	.68	.97	.82
τ_{11}	−.98	−.92	−.96	−.94	.99	.76
τ_{12}		−.95		−.97		.90
τ_{13}		−.92		−.99		.90
τ_{21}	−.99	−.94	−.98	−.96	1.00	.84
τ_{22}		−.97		−.99		.95
τ_{23}		−.93		−1.00		.91
τ_{31}	−.99	−.93	−.99	−.96	1.00	.81
τ_{32}		−.97		−.99		.96
τ_{33}		−.92		−.98		.85
τ_{41}	−.97	−.90	−.96	−.99	1.00	.84
τ_{42}		−.97		−.98		.92
τ_{43}		−.92		−.94		.76
τ_{51}	−1.00	−.91	−.99	−.96	1.00	.78
τ_{52}		−.97		−.98		.95
τ_{53}		−.90		−.92		.69

Note: White color means perfect (positive or negative) correlation. Shades of red indicate a deviation from the perfect correlation.

TABLE A2 Correlation of model scores for a multidimensional GRM (see Figure A2) with binary and non-binary (four ordered categories) response variables fitted on a simulated data set with $n = 2000$ respondents. The model scores of three different fitted models are compared: SC_{OFA} meaning the approximated scores for a ordinal factor model, SC_{FI} meaning the scores for a model fitted with FI estimation, and SC_{GEE} meaning the scores of a model fitted with GEE (see Appendix S1).

	$Cor(SC_{OFA}, SC_{GEE})$		$Cor(SC_{FI}, SC_{GEE})$		$Cor(SC_{OFA}, SC_{FI})$	
	Binary	Non-binary	Binary	Non-binary	Binary	Non-binary
$Var(\eta_1)$	.97	.96	.92	.71	.96	.83
$Var(\eta_2)$	.91	.95	.80	.67	.83	.80
$Var(\eta_3)$	.93	.97	.80	.89	.88	.89
$Cov(\eta_1, \eta_2)$	.89	.85	.91	.74	.93	.88
$Cov(\eta_1, \eta_3)$	.91	.84	.90	.85	.95	.93
$Cov(\eta_2, \eta_3)$	.88	.90	.85	.84	.91	.93
λ_{12}	.93	.91	.85	.51	.87	.68
λ_{13}	.92	.87	.89	.77	.87	.89
λ_{22}	.85	.88	.87	.72	.93	.86
λ_{23}	.92	.87	.65	.54	.68	.71
λ_{32}	.93	.88	.90	.74	.93	.81
λ_{33}	.91	.85	.76	.73	.90	.79
τ_{11}	−.98	−.88	−.98	−.97	1.00	.79
τ_{12}		−.92		−.98		.88
τ_{13}		−.89		−.94		.76
τ_{21}	−.98	−.89	−.98	−.94	1.00	.74
τ_{22}		−.93		−.98		.90
τ_{23}		−.90		−.91		.71
τ_{31}	−.99	−.84	−.99	−.88	1.00	.64
τ_{32}		−.83		−.90		.71
τ_{33}		−.83		−.94		.82
τ_{41}	−.95	−.88	−.94	−.94	.99	.71
τ_{42}		−.94		−.99		.93
τ_{43}		−.88		−.98		.80
τ_{51}	−.88	−.88	−.88	−.98	.97	.81
τ_{52}		−.95		−.99		.94
τ_{53}		−.89		−.95		.74
τ_{61}	−.97	−.87	−.97	−.94	.99	.71
τ_{62}		−.93		−.99		.92
τ_{63}		−.88		−.92		.65
τ_{71}	−.81	−.94	−.91	−1.00	.84	.92
τ_{72}		−.93		−1.00		.93
τ_{73}		−.88		−1.00		.87
τ_{81}	−.98	−.90	−.98	−1.00	1.00	.90
τ_{82}		−.85		−1.00		.85
τ_{83}		−.82		−.99		.81
τ_{91}	−.94	−.90	−.92	−1.00	.99	.90
τ_{92}		−.88		−.99		.88
τ_{93}		−.79		−.99		.76

Note: White color means perfect (positive or negative) correlation. Shades of red indicate a deviation from the perfect correlation.

TABLE A3 Means of noncompensatory DIF (NCDIF) effect sizes for Item 2. Results for 1000 simulated samples with sample size of $n = 1000$. Modes: "all" for all parameters differ, "thresholds" for only thresholds differ and "betas" for only discrimination parameters differ between focal group and reference group.

		Model	
Mode		**Unidimensional**	**Multidimensional**
All	$k = 1$	.05	.14
All	$k = 2$	.18	.48
All	$k = 4$	.54	1.75
All	$k = 6$	1.20	3.65
Thresholds	$k = 1$	.04	.13
Thresholds	$k = 2$	.15	.49
Thresholds	$k = 4$	.47	1.65
Thresholds	$k = 6$	1.02	3.46
Lambdas	$k = 1$	.01	.01
Lambdas	$k = 2$	.05	.04
Lambdas	$k = 4$	.15	.14
Lambdas	$k = 6$	.31	.31

TABLE A4 Computation time in seconds for fitting the unidimensional GRM given no parameter fluctuation in the data. FI, meaning full information estimation, corresponds to model estimation with the MML estimator. LI, meaning limited information estimation, corresponds to ordinal factor analysis with the WLS estimator.

	$n = 500$		$n = 1000$		$n = 2000$	
	FI	**LI**	**FI**	**LI**	**FI**	**LI**
$k = 1$	.19	.19	.20	.21	.21	.23
$k = 2$	.23	.20	.25	.22	.27	.26
$k = 4$	.31	.25	.36	.27	.38	.31
$k = 6$	.41	.30	.47	.31	.51	.38

TABLE A5 Computation time in seconds for fitting the multidimensional GRM given no parameter fluctuation in the data. FI, meaning full information estimation, corresponds to model estimation with the MML estimator. LI, meaning limited information estimation, corresponds to ordinal factor analysis with the WLS estimator.

	$n = 500$		$n = 1000$		$n = 2000$	
	FI	**LI**	**FI**	**LI**	**FI**	**LI**
$k = 1$	12.80	.39	17.42	.46	21.02	.39
$k = 2$	14.04	.42	19.53	.40	27.40	.42
$k = 4$	18.38	.52	26.65	.47	40.92	.50
$k = 6$	22.48	.64	50.66	.84	57.21	.63

TABLE A6 Life satisfaction scale items as asked in the LISS panel.

Text: Below are five statements with which you may agree or disagree. Using the 1–7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.	
Item	**Wording**
$i = 1$	In most ways, my life is close to my ideal
1	The conditions of my life are excellent
2	I am satisfied with my life
3	So far I have gotten the important things I want in life
4	If I could live my life over, I would change almost nothing

TABLE A7 Results of the real data application.

			Model 1 (unidim.)	**Model 2 (multidim.)**	**Model 3 (PIEG)**
Ordinal factor analysis	Number of parameters		35	63	156
	RMSEA		.127	.127	.051
	Score-based test *p*-value	Categorical	1.04E-05	2.22E-04	.017
		Ordinal	.918	.694	.689
		Metric	.165	.008	.014
	Computation time	Model	.345	.913	6.189
		Scores	.063	.109	.325
GRM: FI estimation	Number of parameters		35	63	156
	RMSEA		0	0	0
	Score-based test *p*-value	Categorical	3.92E-06	4.02E-05	0
		Ordinal	.181	.445	0
		Metric	.197	.002	0
	Computation time	Model	.541	88.428	301.546
		Scores	.287	15.506	1074.466
Common factor analysis	Number of parameters		10	13	56
	RMSEA		.099	.122	.043
	Score-based test *p*-value	Categorical	.002	.266	.424
		Ordinal	.227	.453	.770
		Metric	.014	.000	.040
	Computation time	Model	.186	.142	.514
		Scores	.089	.194	.247

APPENDIX B

FIGURES

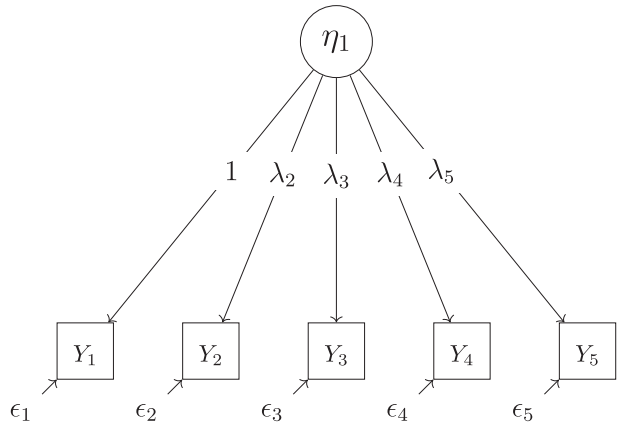

FIGURE A1 Unidimensional graded response model (GRM) with five items.

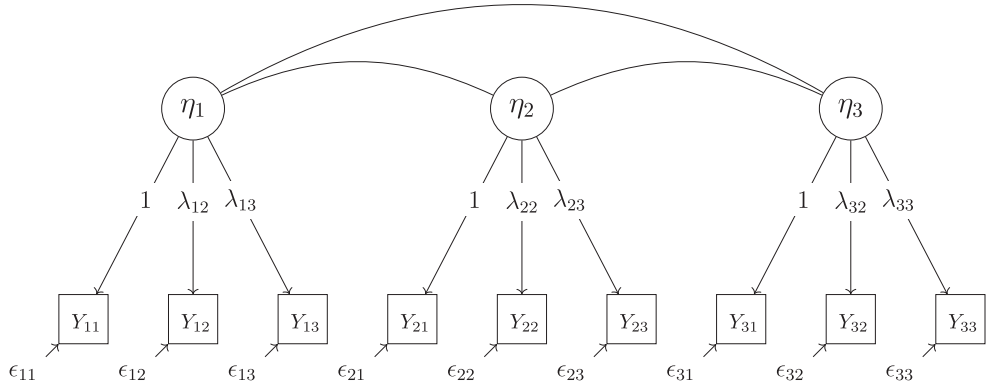

FIGURE A2 Multidimensional graded response model (GRM) with three non-orthogonal latent variables and nine items.

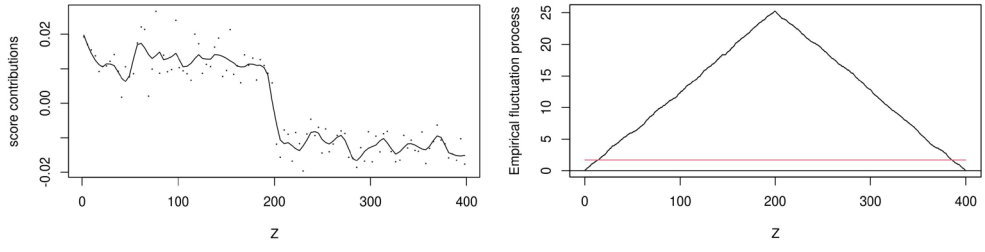

FIGURE A3 Score and CSP distribution (illustration inspired by figure 2 in Strobl et al., 2015).

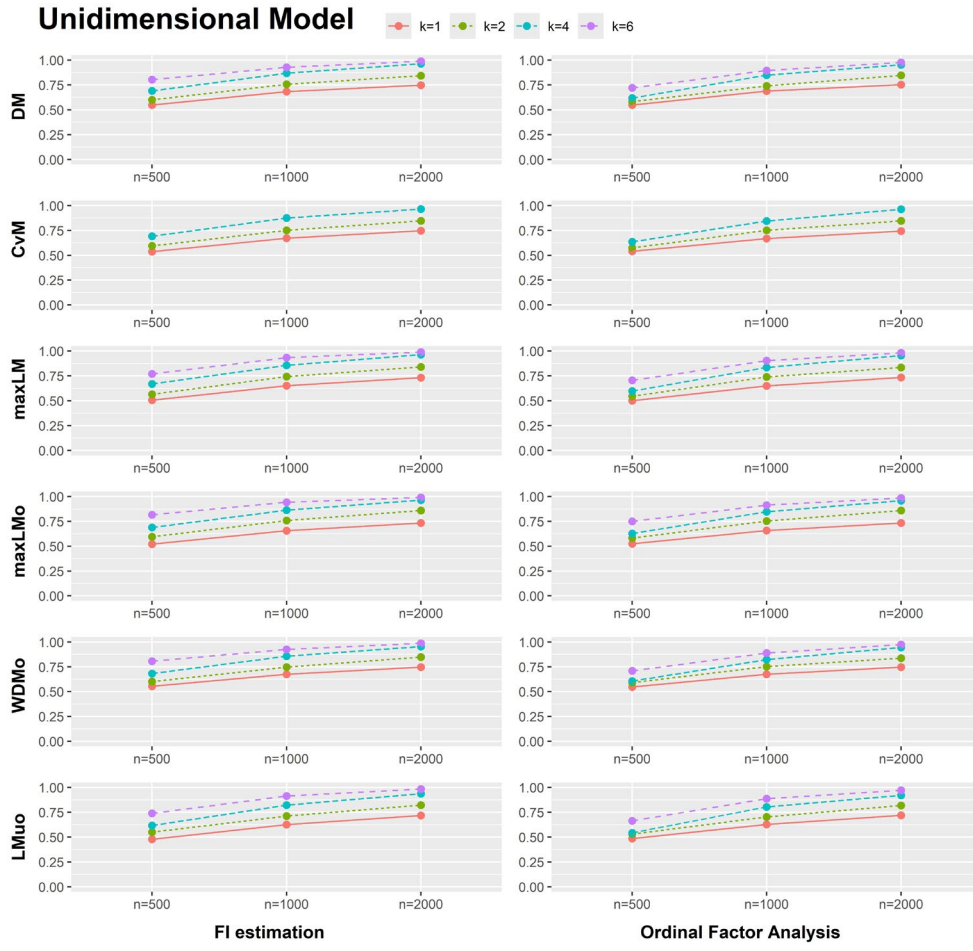

FIGURE A4 Power of score-based test for the unidimensional graded response model (GRM) given fluctuation with respect to the threshold parameters of the first item τ_{1k}.

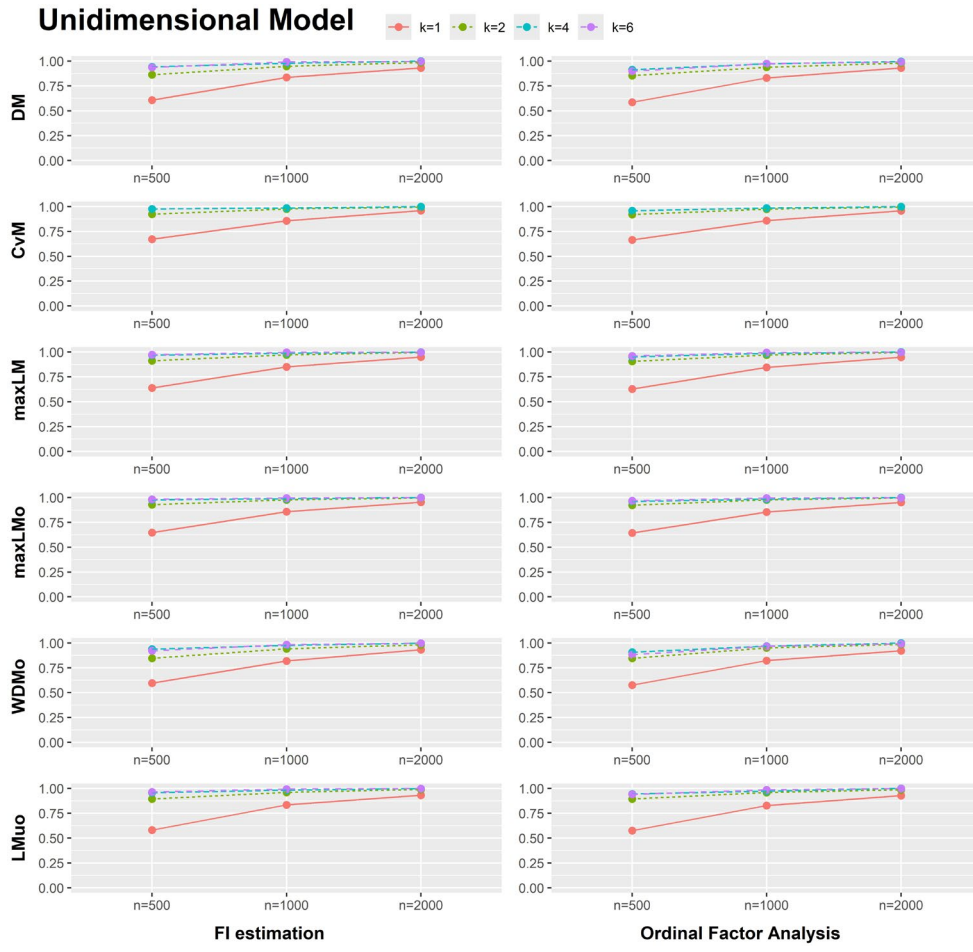

FIGURE A5 Power of score-based test for the unidimensional graded response model (GRM) given fluctuation with respect to the discrimination parameters λ_j.

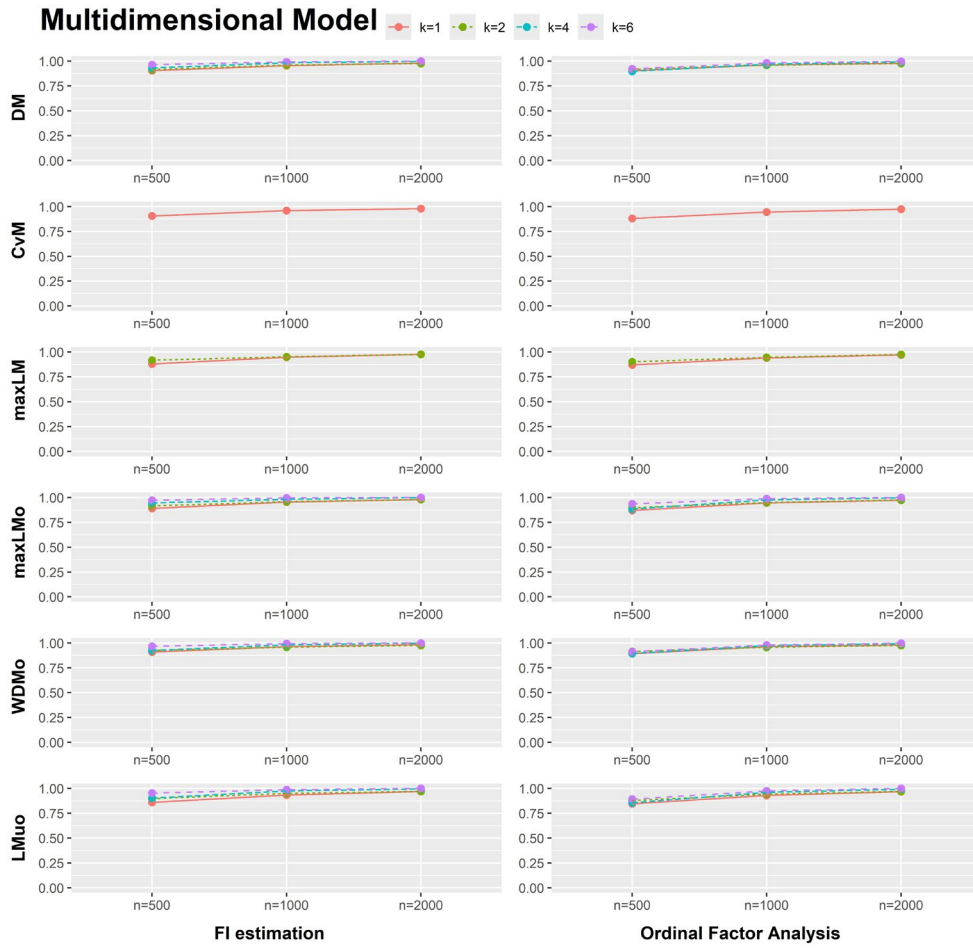

FIGURE A6 Power of score-based test for the multidimensional graded response model (GRM) given fluctuation with respect to the threshold parameters of the first two items, that is, τ_{1k} and τ_{2k}.

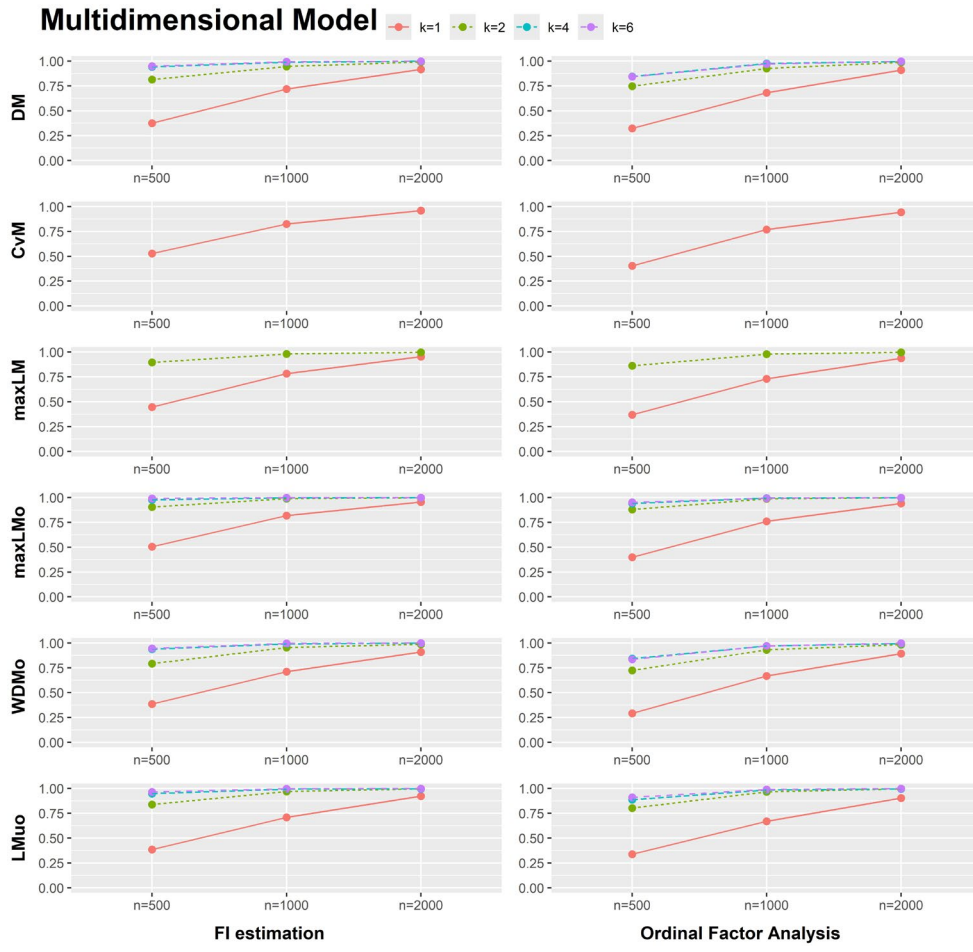

FIGURE A7 Power of score-based test for the multidimensional graded response model (GRM) given fluctuation with respect to the discrimination parameters λ_j.

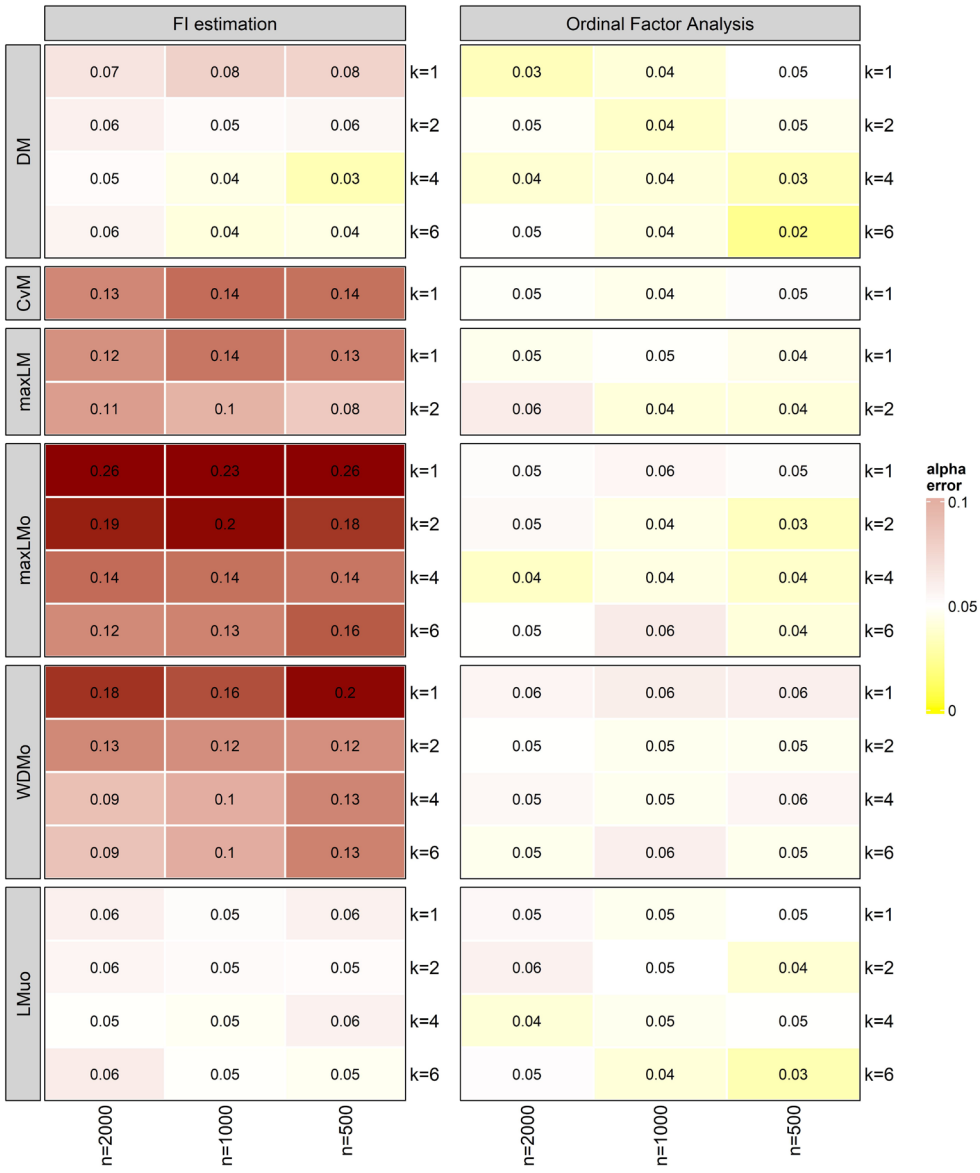

FIGURE A8 Type I errors of score-based test for the multidimensional graded response model (GRM). Note that for the *CvM* test statistic, there are no critical values implemented in the `strucchange` package for models with more than 25 parameters. This also applies for the *maxLM* test statistic for models with more than 40 parameters. Therefore, models with more than 1 (for *CvM*) and 2 (for *maxLM*) threshold parameters are not shown.

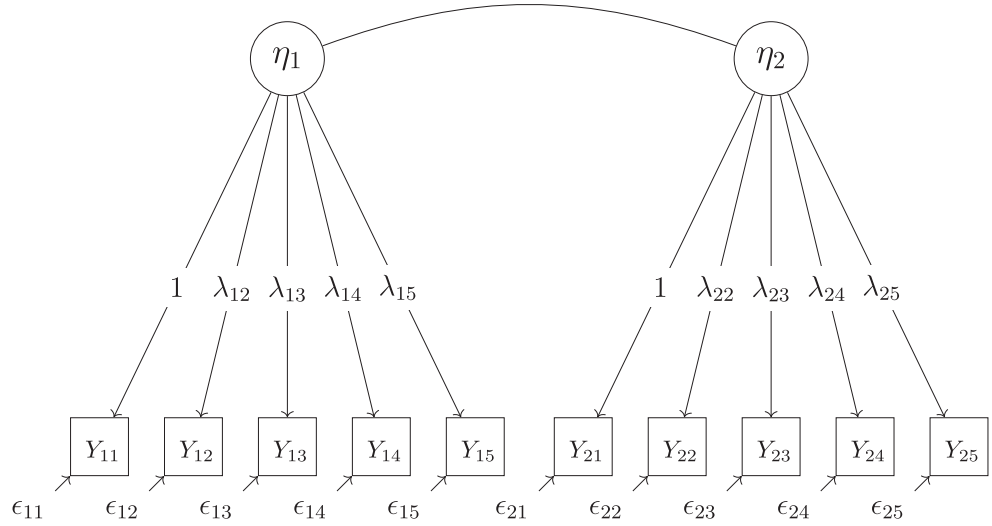

FIGURE A9 Real data application model 2: multidimensional graded response model (GRM) with two latent state variables and five items on two time points.

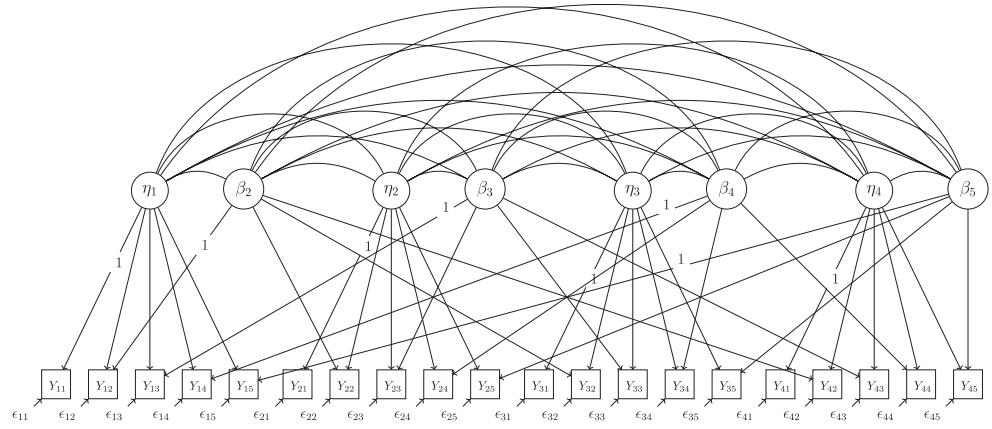

FIGURE A10 Real data application model 3: probit multistate IRT model with latent item effect variables for graded responses (PIEG) with four latent state variables ($\boldsymbol{\eta}_t$), four latent item effect variables ($\boldsymbol{\beta}_i$) and five items on three time points.