
Machine Learning Approaches to Latent Variable Modeling

Franz Classe



München 2025

Machine Learning Approaches to Latent Variable Modeling

Franz Classe

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Franz Classe
aus Nürnberg

München, den 06.03.2025

Erstgutachter: Prof. Dr. Frauke Kreuter
Zweitgutachter: Prof. Dr. Daniel Oberski
Tag der mündlichen Prüfung: 25.07.2025

Acknowledgements

I would like to express my sincere gratitude to everyone who contributed to this dissertation. Special thanks go to...

- ... Prof. Dr. Christoph Kern for always taking time to advise and to support me on all kinds of research ideas. Thank you for introducing me to the world of algorithmic modeling and machine learning. Without your backing, this dissertation would not have been possible.
- ... Prof. Dr. Frauke Kreuter for giving me the opportunity to write this dissertation and for believing in the potential of this endeavor.
- ... PD Dr. Rudolf Debelak for supporting me in the attempt to use the parameter instability test on ordinal factor models and for advising me in all stages of this project.
- ... Prof. Dr. Daniel Oberski and Prof. Dr. Helmut Küchenhoff for their willingness to be part of the examination committee.
- ... Prof. Yves Rosseel for including my code in his R-package and helping me tackle the mathematical and technical challenges of GEE estimation.
- ... Prof. Dr. Susanne Kuger for giving me the opportunity to pursue this dissertation while being employed at the German Youth Institute.
- ... Prof. Dr. Rolf Steyer for inspiring and patiently supporting me to become a statistician and psychometrician.
- ... my family and friends for believing in me and for building me up so often.

Zusammenfassung

Diese Arbeit enthält vier Beiträge (Manuskripte I bis IV), die jeweils neue methodische Ansätze zum Umgang mit Verzerrungen und Bias in mehrdimensionalen IRT-Modellen einführen. Insbesondere wird das Potenzial nichtparametrischer, maschineller Lernverfahren eingehend untersucht. Die im Rahmen dieser Arbeit verfassten Manuskripte stellen Methoden zur Schätzung von Modellparametern und latenten Variablen-Scores multidimensionaler IRT-Modelle vor. Diese Methoden berücksichtigen die Verzerrung, die ungemessene und/oder gemessene Kovariaten auf die Parameterschätzung haben können.

In Manuskript I wird gezeigt, dass die Einbeziehung von latenten Item-Effekt-Variablen in longitudinale IRT-Modelle für ordinale Antwortvariablen interindividuelle Unterschiede in den Item-Schwierigkeits-Parametern kontrollieren kann. Auf diese Weise wird die Verzerrung, die gemessene oder nicht gemessene Kovariaten auf die Schätzung der Item-Schwierigkeits-Parameter haben können, berücksichtigt.

Außerhalb der Längsschnittforschung ist es nicht möglich, solche Item-Effekt-Variablen zu schätzen. Interindividuelle Unterschiede in den Item-Parametern, die auch als Differential Item Functioning (DIF) bezeichnet werden, können jedoch mit Hilfe von Model Based Recursive Partitioning (MOB) berücksichtigt werden, einem algorithmischen Modellierungsansatz, der aus den Methoden des maschinellen Lernens stammt. Manuskript II zeigt, dass MOB zur Kontrolle von DIF in mehrdimensionalen IRT-Modellen verwendet werden kann. Dies funktioniert, indem automatisch Untergruppen mit stabilen Item-Parameterschätzungen erkannt werden.

Manuskript III stellt eine Methode zur Schätzung latenter Variablen-Scores von Individuen vor, die in Bezug auf bestimmte gemessene Kovariaten unverzerrt sind. Zu diesem Zweck wird ein Ensemble von MOB-Trees gebildet. Innerhalb des MOB-Tree-Ensembles werden Untergruppen mit stabilen Item-Parameter-Schätzungen verwendet, um latente Variablen-Scores zu schätzen, die in Bezug auf relevante Untergruppen in der Population unverzerrt sind. Somit sind diese latenten Variablen-Scores im Hinblick auf systematische Einflüsse dieser gemessenen Kovariablen interpretierbar, ohne durch diese Variablen verzerrt zu werden.

Um einen MOB-Tree zu erstellen, muss ein Parameterinstabilitätstest wiederholt für ein (mehrdimensionales) IRT-Modell berechnet werden. Mehrdimensionale IRT-Modelle werden effizient als ordinale Faktorenmodelle geschätzt. Für das Modell wird die erste Ableitung der Zielfunktion (d.h. die Score-Funktion) verwendet, um die Parameterinstabilität zu schätzen. In Manuskript IV wird daher eine Methode zur Schätzung der individuellen Beiträge zu dieser Funktion für ordinale Faktorenmodelle vorgeschlagen. Dadurch wird es möglich, viele Parameterinstabilitätstests für mehrdimensionale IRT-Modelle in kurzer Zeit zu berechnen.

Die mit diesen vier Beiträgen vorgestellten Methoden ermöglichen die effiziente Berechnung von Parameterinstabilitätstests für mehrdimensionale IRT Modelle, die Schätzung individueller Schwierigkeits-Parameter in Längsschnittkontexten und latenter Variablen-Scores, die außerhalb von Längsschnittkontexten in Bezug auf spezifische gemessene Kovariaten unverzerrt sind.

Summary

This thesis contains four contributions (Papers I to IV) which present approaches to dealing with bias in multidimensional IRT models. In particular, the potential of nonparametric tree-based machine learning methods is examined in detail. The papers written in the scope of this thesis provide methods to estimate model parameters and latent variable scores of multidimensional IRT models while considering the bias that unmeasured and/or measured covariates may have on parameter estimation.

In Paper I, it is shown that the inclusion of latent item effect variables in longitudinal IRT models for ordinal response variables can control for inter-individual differences in item difficulty parameters. This way, the bias that measured or unmeasured covariates may have on the estimation of the item difficulty parameters is taken into account.

Outside of longitudinal research, it is not possible to estimate such item effect variables. However, inter-individual differences in item parameters, also referred to as Differential Item Functioning (DIF), can be accounted for via Model Based Recursive Partitioning (MOB), an algorithmic modeling approach borrowed from the tree-based methods of machine learning. Paper II illustrates that MOB can be used to control for DIF in multidimensional IRT models. For such models, MOB may be used to automatically detect subgroups with stable item parameter estimates.

Paper III introduces a method to estimate latent variable scores of individuals that are unbiased with respect to certain measured covariates. For this, an ensemble of MOB trees is grown. Within the MOB tree ensemble, subgroups with stable item parameter estimates are used to estimate latent variable scores that are unbiased with respect to relevant subgroups in the population. Thus, these latent variable scores are interpretable with respect to systematic influences of specific measured covariates without being biased by these variables.

In order to grow a MOB tree, a parameter instability test must be computed repeatedly for a fitted (multidimensional) IRT model. Multidimensional IRT models are efficiently fitted as ordinal factor models. For the fitted model, the first derivative of the objective function (i.e. the score function) is used to estimate parameter instability. In Paper IV, a method for the estimation of individual contributions to this score function for ordinal factor models is therefore proposed. This makes it computationally feasible to repeatedly compute parameter instability tests for multidimensional IRT models.

The methods introduced with these four contributions make it possible to efficiently compute parameter instability tests for MIRT models, to estimate individual difficulty parameters in longitudinal settings and latent variable scores that are unbiased w.r.t. specific measured covariates outside of longitudinal settings.

Contents

1	General Introduction	1
2	Methodological Background	3
2.1	Item Response Theory and Factor Analysis Models	3
2.1.1	Limited Information Maximum Likelihood	3
2.1.2	Ordinal Factor Analysis	4
2.2	Measurement Invariance and Differential Item Functioning	5
2.3	Tree Based Machine Learning	5
2.4	Model Based Recursive Partitioning	7
3	Paper I: A Probit Multistate IRT Model With Latent Item Effect Variables for Graded Responses	8
3.1	Background	8
3.2	Summary and Contribution	9
4	Paper II: Detecting Differential Item Functioning in Multidimensional Graded Response Models With Recursive Partitioning	10
4.1	Background	10
4.2	Summary and Contribution	11
5	Paper III: Latent Variable Forests for Latent Variable Score Estimation	11
5.1	Background	11
5.2	Summary and Contribution	12
6	Paper IV: Score-Based Tests for Parameter Instability in Ordinal Factor Models	13
6.1	Background	13
6.2	Summary and Contribution	14
7	General Discussion	15
	References	18
A	Formulas for Score Functions	22
B	Attached contributions	23

Contributions of the thesis

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I 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(3), 172–183. <https://doi.org/10.1027/1015-5759/a000751>
- II Classe, F., & Kern, C. (2024). Detecting differential item functioning in multidimensional graded response models with recursive partitioning. *Applied Psychological Measurement*, 48(3), 83-103. <https://doi.org/10.1177/01466216241238743>
- III Classe, F., & Kern, C. (2024). Latent Variable Forests for Latent Variable Score Estimation. *Educational and Psychological Measurement*, 84(6), 1138-1172. <https://doi.org/10.1177/00131644241237502>
- IV Classe, F., Debelak, R., & Kern, C. (2025). Score-based tests for parameter instability in ordinal factor models. *British Journal of Mathematical and Statistical Psychology*. <https://doi.org/10.1111/bmsp.12392>

The full text of Paper I is not included in this manuscript due to licensing restrictions.

Since Paper IV was still being reviewed by the journal at the time of submission of the dissertation, the revised version of the paper originally submitted to the journal is included here. The paper has since been published.

Declaration of the author's specific contributions

Paper I: For the conception of the model, an existing model (i.e. the PIEV model by Thielemann, Sengewald, Kappler, and Steyer (2017)) was revised by the first author Franz Classe for the special case of ordinal response variables. The literature research, the Monte Carlo simulations, the application to real data, the writing of the text and the mathematical notation, as well as the submission were also carried out by the first author. The co-author Rolf Steyer initiated the idea, supervised each of the steps in regular meetings and contributed significantly to the development of the publication by giving advice on how to proceed with the application and simulation, the formulation of text and notation, and the submission. He also proofread the manuscript.

Paper II: The topic of the publication was determined in close collaboration between first author Franz Classe and co-author Christoph Kern. The literature search, the simulations, the application to real data, the writing of the text and mathematical notation, and the submission were also carried out by the first author. The co-author Christoph Kern supervised each of the steps in regular meetings and contributed significantly to the development of the publication by giving advice on how to proceed with the application and simulation, the formulation of the text and the submission. He also proofread the manuscript.

Paper III: The contributions of first author Franz Classe and co-author Christoph Kern are to be described in the same way as in Paper II.

Paper IV: The topic for publication was deemed relevant by first author Franz Classe together with co-author Christoph Kern. During this collaboration, contact was made with co-author Rudolf Debelak. He provided the impetus for further work on the topic. The first author then carried out the literature research and, in an experimental phase, implemented calculation methods in R, which he then evaluated through extensive simulations. The text and the mathematical notation were then written by the first author Franz Classe. The co-author Rudolf Debelak supervised each of the steps in regular meetings and contributed significantly to the development of the publication by advising on the further procedure regarding application and simulation, as well as submission. He also proofread the manuscript. The manuscript was proofread by both co-authors Christoph Kern and Rudolf Debelak.

The declaration of the authors' own contribution was agreed with the co-authors involved in the articles.

1 General Introduction

Attempts to explain or predict events based on phenomena that are not directly observable has always been common practice among humans. In psychometrics, these unobservable phenomena are often characteristics or traits in people. Much of psychometric science is dedicated to formulating theories about such characteristics in individuals or groups. Hypotheses are then tested on the basis of models that formulate the key elements of a theory. In these models, the characteristics are represented by variables. Variables that have no values for individuals, for example because they are not directly observable, are referred to as *latent variables* (Bollen, 2002). According to Bartholomew, Knott, and Moustaki (2011), latent variables are practically relevant as they condense information from many different observations, thus it is impossible to formulate theories about social and psychological phenomena without utilizing such latent variables. To express reasoning about such concepts in the language of mathematics, stochastic models that incorporate latent variables are necessary. Examples for such models are latent class analysis, latent curve, structural equation, factor analysis, and item response theory models.

In general, the variables in latent variable models are not determined in their form. They are, however, determined by the axiom of local independence of the manifest variables. In short, this means that the latent variables in a latent variable model are sufficient to explain the dependencies among the manifest variables. A *factor analysis* model (or factor model) is commonly defined as a latent variable model in which both the manifest and the latent variables are discrete or continuous random variables with real numerical values (i.e. they are not categorical, see Bartholomew et al., 2011, pp. 8–11).

The origins of factor analysis can be found in psychometrics. It was Spearman (1904) who defined a general ability factor to account for the correlation of tests of mental ability. Thurstone (1947) later expanded Spearman’s model to incorporate multiple possibly correlated factors. Several statisticians defined Spearman’s and Thurstone’s assumptions in mathematical terms. Much of the current scientific discussion can be traced back to Jöreskog (1969, 1970), who developed a model of structural equations. This model is a factor model that takes into account linear relationships between the latent variables.

Parallel to this development, in the context of educational testing, a different strand of research emerged. In this, latent variable models were used to relate the probability of a response in a test to a person’s ability as well as to the difficulty of the item. These models were developed in particular in response to classical test theory, which viewed psychometric or educational tests as one entire analytic unit. This led to the development of *item response theory* (IRT) models, which focus on the properties of the test items and regard the test items as a fundamental component of the test design. The original IRT models by Rasch (1960) and Birnbaum (1968) only incorporated one latent variable. However, with increasing understanding of the complexity of psychological and pedagogical processes, more complex multidimensional IRT (MIRT) models became necessary in psychometric science (Reckase & Reckase, 2009).

In contrast to IRT models, factor models consider the characteristics of the items as nuisance effects that should be eliminated for adequate model estimation (Reckase &

Reckase, 2009). Also, IRT models use dichotomous or ordinal manifest variables whereas factor models are most commonly used with metric manifest variables. Additionally, IRT models treat latent variables as vectors of parameters describing the location of a person on a latent scale and not as a random variable (Bartholomew et al., 2011). Despite these differences, the theoretical foundation of factor models and IRT models is very similar as they both define latent variables as determined by the axiom of local independence. Thus, MIRT models can be considered a special case of factor models or structural equation models (see Reckase, 1997; Maydeu-Olivares, 2005).

According to Breiman et al. (2001), the formulation of parametric assumptions about a model such as a factor model is exemplary for *data modeling culture*. Through latent variable models, the researcher tries to derive information about how observed variables are *truly* associated with latent variables. Breiman et al. (2001) claim that the conclusions derived from such stochastic models are about the model's mechanism, not about nature's mechanism. This means that a stochastic model must accurately emulate nature to lead to informative conclusions. Most often though, stochastic models are not complex enough to emulate nature. Machine learning models, on the other hand, are considered part of the *algorithmic modeling culture*. Algorithmic models assume that natural mechanisms that produce data are unknown.

Stochastic models are usually evaluated through goodness-of-fit testing. However, when many different models are considered and fitted to data with complex interactions, the yes-no answers of model fit tests may point to several different models. In this case, choosing one model is a very challenging task. Also, it is very difficult, if not impossible, to formulate a stochastic model that encompasses all rival models. In contrast, an algorithmic model is usually evaluated through assessment of predictive power. It serves the purpose of predicting new or future observations through flexible modeling with minimal assumptions. Algorithmic models need to be flexible enough to approximate the data generating function while also being robust towards changes in the data used to fit the model. This compromise is referred to as *bias-variance trade-off*. Algorithmic models acknowledge the complex and inconceivable ways that nature produces data. They do not need to be fully interpretable, they rather need to provide accurate information (Breiman et al., 2001).

In scientific practice, stochastic models are almost always used as explanatory models while algorithmic models are usually used as predictive models. One should not confuse explanatory models with predictive models although both explanation and prediction are necessary for generating and testing theories. Complicated patterns and relationships in data sets may be hard to hypothesize within the explanatory modeling framework. Therefore, prediction as a scientific endeavor is necessary to grasp the aforementioned complexity of natural mechanisms. Potential new causal mechanisms can be uncovered through the flexibility of predictive models. Also, capturing complex patterns in data can lead to improvements to existing explanatory models (Shmueli et al., 2010).

An ongoing problem in psychometrics is that the increasing knowledge about the natural mechanisms of the mind make it necessary to assume increasingly large and convoluted latent variable models (Reckase & Reckase, 2009). With increasing complexity, it becomes

increasingly difficult to interpret and to estimate the parameters of these latent variable models. The main question of this thesis is how predictive machine learning models can be used to increase the informative value of explanatory latent variable models. For this, it is crucial not to confuse explanatory and predictive modeling but to combine both modeling cultures. The thesis places particular emphasis on the practical application of MIRT models with ordered polytomous observables. In Section 2, I will introduce the most important methodological background on the papers that were published within the scope of this dissertation. In Sections 3 to 6, the contribution of the four papers are introduced individually. In Section 7, limitations, chances and suggestions for further research are discussed.

2 Methodological Background

2.1 Item Response Theory and Factor Analysis Models

A very common type of IRT model that deals with ordered polytomous categories is the *Graded Response Model* (Samejima, 1969). In the multidimensional version of the GRM, several latent variables can be included. A multidimensional GRM models the cumulative category response function. For a test or questionnaire with m items, that is

$$P(Y_i \leq k \mid \boldsymbol{\xi}_i) = \Phi(\alpha_{ik} - \boldsymbol{\lambda}'_i \boldsymbol{\xi}_i), \forall i, \dots, m. \quad (1)$$

This function represents the conditional probability of responding to an item i with a response category smaller or equal to the category k given the latent variables $\boldsymbol{\xi}_i$ that affect item i . It is a two-parameter normal ogive function, with the *item discrimination parameters* $\boldsymbol{\lambda}_i$ constant within an item but variable across items. The *threshold parameters* α_{ik} are variable within and across items.

As mentioned above, IRT models are special cases of factor models. According to Maydeu-Olivares, Cai, and Hernández (2011), there is one type of factor model that is especially suitable of being considered a special case of IRT model: the ordinal factor model. This goes back to Christoffersson (1975) who developed a probabilistic factor model for dichotomous items. The new development was that only the marginal first and second order proportions of the item responses were used to estimate the model parameters, instead of all 2^m possible proportions. Muthén (1978) extended this approach to make parameter estimation even more efficient. The resulting model uses factor loadings and threshold parameters. These parameters are virtually identical to those used in MIRT models.

In Sections 2.1.1 and 2.1.2, we describe the assumptions of the typical metric and ordinal factor model and how their parameters are estimated.

2.1.1 Limited Information Maximum Likelihood

Let Y_i be the observed variable for item i , and $\boldsymbol{\xi}_i$ the $p_i \times 1$ vector of continuous latent variables that affect Y_i , then the basic factor model equation is

$$E(Y_i \mid \boldsymbol{\xi}_i) = \pi_i + \boldsymbol{\gamma}'_i \boldsymbol{\xi}_i, \forall i, \dots, m, \quad (2)$$

where π_i is the intercept, and $\boldsymbol{\gamma}_i$ is the $p_i \times 1$ vector of factor loadings of item i . It is assumed that the distribution of the observed variables is continuous and multivariate normal. The residual variable for item i is defined as $\epsilon_i = E(Y_i | \boldsymbol{\xi}_i) - Y_i$.

The model parameter vector $\boldsymbol{\theta}$ consists of the intercepts, factor loadings, residual variances and latent variable (co-)variances. The estimator $\hat{\boldsymbol{\theta}}$ consists of the parameter estimates that minimize the objective function

$$F_{ML}(\boldsymbol{\theta}) = \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - \ln|\mathbf{S}| - m, \quad (3)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model implied covariance matrix and \mathbf{S} is the sample covariance matrix (Jöreskog, 1969). This means that for parameter estimation, only bi-variate information from the data (in \mathbf{S}) is used. Because of this, this estimation method is referred to as *limited information* method.

2.1.2 Ordinal Factor Analysis

When the observed variables in the analyzed data only have a few response categories, the normality assumption of the classic linear factor model is severely violated (Li, 2016). In this case, a continuous latent response variable Y_i^* is usually assumed to underlie the actual ordered observed variable Y_i for an item i with $k = 1, \dots, l_i$ response categories. The conditional expectation of this latent response variable is defined similarly to the classic linear factor model, that is

$$E(Y_i^* | \boldsymbol{\xi}_i) = \boldsymbol{\lambda}_i' \boldsymbol{\xi}_i, \quad \forall i, \dots, m, \quad (4)$$

where $\boldsymbol{\lambda}_i$ is the $p_i \times 1$ vector of discrimination parameters of item i (which is the equivalent to the factor loadings in the classic linear factor model). Furthermore, instead of intercepts, there are threshold parameters in the ordinal factor model that are defined as

$$Y_i = k \text{ if } \alpha_{i(k-1)} < y_i^* \leq \alpha_{ik}. \quad (5)$$

For each item i , there is always one threshold parameter less than there are response categories l_i . Thus, $\alpha_{i0} = -\infty$ and $\alpha_{il_i} = +\infty$.

The model parameter vector $\boldsymbol{\vartheta}$ consists of the discrimination parameters, threshold parameters, and latent variable (co-)variances. The estimator $\hat{\boldsymbol{\vartheta}}$ consists of the parameter estimates that minimize the weighed least squared type objective function, that is

$$F_{OFA}(\boldsymbol{\vartheta}) = [\hat{\mathbf{k}} - \boldsymbol{\kappa}(\boldsymbol{\vartheta})]' \mathbf{W}^{-1} [\hat{\mathbf{k}} - \boldsymbol{\kappa}(\boldsymbol{\vartheta})]. \quad (6)$$

This function minimizes the discrepancy between $\boldsymbol{\kappa}(\boldsymbol{\vartheta})$, which contains the elements of the model implied covariance matrix as well as the threshold parameters, and $\hat{\mathbf{k}}$, which contains the polychoric correlations and the sample thresholds. The polychoric correlation ρ_{ih} of the ordered observed variables Y_i and Y_h quantify the degree of linear dependence of the latent response variables Y_i^* and Y_h^* for $i \neq h$. Together with the sample thresholds, they can be easily estimated from the data (see Olsson, 1979). Note that the parameters in $\boldsymbol{\vartheta}$ are the same as the parameters of the multidimensional GRM (see Equation 1).

As shown in Paper I to IV, ordinal factor analysis can be used to efficiently estimate the parameters of complex MIRT models. In Paper I, a longitudinal MIRT model with correlated latent variables is introduced. Without ordinal factor analysis, parameter estimation for such models is computationally very demanding.

2.2 Measurement Invariance and Differential Item Functioning

A problem that often occurs in practice is that the parameters estimated for a classic linear or an ordinal factor model deviate between specific subgroups within the population. This is the case if certain covariates of the manifest variables that are not included in the model have an influence on the model's parameters. Sterner, Pargent, Deffner, and Goretzko (2024) describe how establishing *measurement invariance* in a latent variable model is crucial in order to meaningfully compare latent variable scores between groups within the population. Differences between groups should only occur due to true differences in the latent variables and not due to measurement differences. This means that the model's parameters need to be equivalent across groups. They describe measurement invariance as an inherently causal concept established by the ξ_i -conditional independence of Y_i and any observed covariate Z in \mathbf{Z} when the groups within the population are defined as subsets of the covariate space over \mathbf{Z} , that is

$$Y_i \perp\!\!\!\perp \mathbf{Z} \mid \xi_i, \forall i = 1, \dots, m. \quad (7)$$

Equation 7 means that all group differences (denoted by different values on \mathbf{Z}) of the observed variable Y_i are mediated through the latent construct ξ_i . In causal terms, an observed variable Y_i is considered a biased measure of the latent construct ξ_i with respect to a set of manifest variables \mathbf{Z} if there exists an active causal path linking Y_i and \mathbf{Z} that does not pass through ξ_i . However, in an experimental setup by Protzko (2024), it is shown that psychometric tests may exhibit measurement invariance between two groups even if the test does not measure the same thing for both groups. This indicates that testable measurement invariance is at best a necessary condition such causal relations, not a sufficient condition.

In MIRT models, measurement invariance is usually referred to as *Differential Item Functioning* (DIF). It means that if the probability of a particular response to an item is different among equally able test takers because the test takers differ w.r.t. certain covariates, then DIF is present. Apart from making tests unfair for certain groups, DIF can mask true group differences in latent variables (Wang, Su, & Weiss, 2018).

There are algorithmic approaches to detect DIF in latent variable models. However, these methods are not easily applicable for ordinal factor analysis. Papers II to IV deal with the detection of DIF in MIRT models.

2.3 Tree Based Machine Learning

One of the most well-known families of machine learning applications are tree-based machine learning algorithms. They come from predictive customer analytics but are increasingly used in the social science and survey research context, for example for non-response

prediction or missing value imputation (Kern, Klausch, & Kreuter, 2019). Perhaps the most famous representative of this family of machine learning algorithms is *random forest* (Breiman, 2001). During training of a random forest, random subsamples are drawn from the data (i.e. *bagging*) and random selections of covariates \mathbf{Z} are made (i.e. *random split selection*). The randomly sampled data are then used to build a decision tree that reduces variance w.r.t. an outcome variable Y . The terminal nodes (also referred to as *leaves*) of the tree contain a small subsample R_h . The subgroups R_1, \dots, R_H , are defined as subsets of the covariate space over \mathbf{Z} . During prediction, the mean of this subsample $E(Y | R_h)$ is the predicted value for all values of the same subsample R_h . Usually, the selection of partitioning variables is redrawn at every node in a regression tree. This procedure is repeated numerous times. The bagging and the split selection procedures ensure that the trees of the forest are very likely to differ. Also, a high degree of tree complexity is preferred in a random forest. Through the combination of several trees into a robust ensemble, the instability of a single decision tree is leveled out (Kern et al., 2019). Random forest is a purely predictive method in which the true functional form of the relationship between input and observed variables is unknown prior to the procedure and the function approximated by the random forest cannot be interpreted directly. The predictions of a random forests, however, are likely to be more accurate than the predictions of most explanatory models (Fife & D’Onofrio, 2023). If we acknowledge that nature produces data in complex and inconceivable ways, a non-stochastic but accurate function approximated by a random forest might be preferable compared to a stochastic model like a complex MIRT model.

Random forest methodology can be adapted for other purposes. A method that extends the original random forest approach is *causal forest* (see Athey & Imbens, 2016; Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019). It is widely used across many scientific fields (Rehill, 2024). The appeal of causal forests is the possibility to estimate individual causal effects of a treatment. In contrast to random forests, where the trees reduce the variance of Y , the trees in a causal forest reduce the variance of an observed treatment effect $E(Y|X = 1) - E(Y|X = 0)$, where X is a binary treatment variable. The goal of a causal forest is the estimation of

$$CTE_{10}(\mathbf{Z}) = E(\tau_1 - \tau_0 | \mathbf{Z}), \quad (8)$$

where τ_1 is the true-outcome variable of given treatment ($X = 1$) and τ_0 is the true-outcome variable given control ($X = 0$). Equation 8 is the \mathbf{Z} -conditional total treatment effect function. It’s values are the individual causal treatment effects that are always unbiased. For Equation 8 to be true, conditional unconfoundedness, that is

$$\tau_x \perp\!\!\!\perp X | \mathbf{Z} \quad (9)$$

is assumed. This assumption may hold for observational (non-experimental) settings if \mathbf{Z} comprises all potential confounders C_X that could possibly bias the relationship between X and Y . Note that this notion of unconfoundedness is equivalent to the \mathbf{Z} -conditional *strong ignorability* assumption (see Steyer, Mayer, & Lossnitzer, 2023).

Causal forests, as well as random forests, prove to be especially practical if there is a large amount of manifest covariates Z relative to the sample size (Athey & Imbens,

2016). Also, the presence of irrelevant covariates does not severely affect the performance of a causal forest (Wager & Athey, 2018). However, it is not possible to include latent variables in a causal forest.

We want to investigate how tree-based ML methods may be used in conjunction with latent variable models. In Paper III, we investigate how tree-based ML may be used for bias reduction in latent variable models.

2.4 Model Based Recursive Partitioning

A method that combines algorithmic modeling with explanatory modeling is *Model Based Recursive Partitioning* (Zeileis, Hothorn, & Hornik, 2008). A parametric model is the basis to any MOB algorithm. This model is handed to a tree-based algorithm to detect if relevant covariates within \mathbf{Z} should be included in the model definition (Kopf, Augustin, & Strobl, 2013). This is done by recursively partitioning the sample to which the model is fitted to reduce the heterogeneity of the estimated model parameters. The MOB approach combines parametric modeling with the idea of a non-parametric tree structure and is therefore referred to as a *semi-parametric* approach (Strobl, Kopf, & Zeileis, 2015, 3).

In MOB, a subgroup of the sample represents a tree node that is considered as a candidate for potential splitting. A fitted model in a node is tested for parameter instability with respect to any of the covariates in \mathbf{Z} (also called *partitioning variables*). If there is significant parameter instability, the node is eventually split at a point on the covariate with the greatest instability into two locally optimal segments.

To test for parameter instability, parameter changes over a covariate are detected through the ‘generalized M-fluctuation test’ (Zeileis & Hornik, 2007), which is more commonly known as the *score-based test for parameter instability*. For the application of the score-based test, the item parameters are estimated jointly for the entire sample. Then, the individual deviations of the participants from the joint model are detected. If there is overall parameter instability in the current node, the partitioning variable Z^* , that is associated with the greatest instability, is chosen for splitting. In the next step, the objective functions of two rival segmentations are compared until the optimal split point on Z^* is found. This means that the sum of the log-likelihoods for two rival segmentations is maximized at the optimal split point (Zeileis et al., 2008). Note that the score-based test does not require a test for model fit differences (like e.g. the Likelihood-Ratio test) for the decision to split a node into several subgroups.

The partitioning algorithm continues to distinct models for different subgroups via recursive partitioning until the stopping criteria are met. The stopping criteria are usually met when there is no more significant instability in the node or when the subsample becomes too small. The procedure results in a tree structure with one fitted parametric model for each terminal node.

In general, one advantage of MOB is that the researcher does not need to pre-specify the functional form between the covariates and the model. Rather, the tree structure is learned from the data in an exploratory way. Another clear advantage is the ease of interpretation of the resulting subgroups. In this, the method is opposed to the latent

class (or mixture) approach, where parameter differences are tested for all possible subgroups regardless of covariates that provide information about the characteristics of the relevant subgroups (Rost, 1990). The subgroups in the leaves of a MOB-built decision tree, on the other hand, are directly interpretable because they are built on traceable sample splits. Thus, the MOB approach combines the advantages of the mixture approach (no pre-specification of subgroups) with the advantages of the Likelihood-Ratio-approach (interpretable subgroups).

For complex factor models, MOB can be computationally very demanding as the split point on a covariate Z^* is found by comparing all possible rival segmentations w.r.t. the log-likelihood of their fitted models. For covariates with a very large, number of potential split points (e.g. metric variables), this procedure requires the fitting of twice as many factor models as there are potential split points. Therefore, Arnold, Voelkle, and Brandmaier (2021) came up with a method to efficiently find split points within the MOB algorithm by using the test statistic of the score-based test for parameter instability. Using this ‘score-guided’ MOB approach, a factor model has to be fitted only once per tree node.

MOB has been applied to unidimensional IRT models in the past. Strobl et al. (2015) applied the algorithm to Rasch models whereas Komboz, Strobl, and Zeileis (2018) applied it to Partial Credit Models. Both approaches aimed at detecting *Differential Item Functioning* (DIF) in the population.

However, MOB has not been applied to multidimensional GRMs in a computationally efficient manner. In Papers II and IV, we explore different ways to do that. In Paper III we investigate ways to use MOB for bias reduction in latent variable models.

3 Paper I: A Probit Multistate IRT Model With Latent Item Effect Variables for Graded Responses

3.1 Background

A practically relevant application for multidimensional IRT models is research about the change of behavior over time. For this purpose, longitudinal MIRT models can be defined in which there are distinct latent variables for each time point. This makes it possible to differentiate between stable psychological traits and situational fluctuations. Latent State-Trait (LST) theory (Steyer, Ferring, & Schmitt, 1992) does this by defining the latent state as an attribute of a person in a situation that always comprises the latent trait which is the stable attribute of a person. In LST theory, the latent states and traits are defined as conditional expectations of observed variables. The revised version of the theory, i.e. LST-R (Steyer, Mayer, Geiser, & Cole, 2015), includes the notion that past experiences of a person have important implications on the concept and the properties of latent states and traits, state residuals and measurement errors. Based on LST-R theory, longitudinal models with latent variables can be defined in order to generate insights into the dynamics of human attributes.

Another use of MIRT models are multitrait-multimethod (MTMM) models that usu-

ally investigate the validity of specific scales by extracting the effects of certain methods on the measurement of psychometric characteristics. MTMM models can also be used to make causal inferences about the effects of specific methods. In the *method effect model* by Pohl, Steyer, and Kraus (2008), the method effect variable is defined as the difference between the latent true-score variable of the manifest variable measured by a certain method and the latent true-score variable of the same trait measured by a reference method. The scores of this method effect variable may be interpreted as *individual causal effects*. A longitudinal model including method effects is the CT-C(M-1) model by Eid, Lischetzke, Nussbeck, and Trierweiler (2003) that defines item- and trait-specific latent variables as conditional expectations of observed variables.

The model introduced in Paper I has several advantages to the CT-C(M-1) model as it does not define the item-effect as a residual. It rather defines it as a method effect of the item level.

3.2 Summary and Contribution

In Paper I, a longitudinal MIRT model is introduced that includes method effect variables on the item level. We call it the probit multistate model with latent item effect variables for graded responses (PIEG). The latent item-effect variable in the PIEG model for a time point t is defined as

$$\tilde{\beta}_i = \tilde{\tau}_{i1t} - \tilde{\tau}_{11t}, \quad (10)$$

where $\tilde{\tau}_{ikt}$ is the item-, time-, and category-specific probit state variable which is, in essence, a latent true-outcome variable (see Equation 4 to 7 in Paper I).

The PIEG model from Paper I differs from the method effect model by Pohl et al. (2008) in that it is a longitudinal model. In the longitudinal setting of the PIEG model, the units (i.e. the individuals in the sample) are exposed to both treatment and control conditions if treatment is ‘answering to item i ’ and the control is ‘answering to the reference item’. In such settings, one can assume that the causality condition of strong ignorability is met (see Section 2.3). However, Steyer et al. (2023) stresses that the treatment variable X has to be prior to the outcome variable. In the PIEG model in Paper I, the latent item-effect variables are equal across time-points so the treatment variable denoting ‘answering with item i instead of the reference item’ is not bound to a specific time point preceding the outcome. Thus, in contrast to the method effects model, the PIEG model in Paper I focuses on the estimation of *individual item difficulties* instead of causal effects. However, although it does not contain causal effects, the PIEG model shows that, to a certain extent, it is possible to establish causality in latent variable models with the help of repeated measures designs.

In Paper I, it is shown that the integration of latent item effect variables into a latent state model leads to noticeably better model fit in real-data application. Also, extensive Monte Carlo simulations show that only relatively small sample sizes are necessary to estimate stable parameters with the PIEG model.

4 Paper II: Detecting Differential Item Functioning in Multidimensional Graded Response Models With Recursive Partitioning

4.1 Background

As mentioned in Section 2.1.2, it is possible to estimate the parameters of an IRT model, specifically a GRM, via ordinal factor analysis which is a limited information estimation method. However, the method that is most often referred to as the standard estimation method for IRT models (see Forero & Maydeu-Olivares, 2009) is *marginal maximum likelihood* (MML) estimation via the *expectation-maximization* (EM) algorithm (Bock & Aitkin, 1981). This estimation method uses all information contained in the responses $\mathbf{y}_j = \{y_{j1}, \dots, y_{jm}\}$ of all individuals $j = 1, \dots, n$, in a sample. In Section 1, we established that the GRM is determined by the axiom of local independence. From this fundamental assumption, it follows that the $\boldsymbol{\xi}$ -conditional probability responding with the response pattern \mathbf{y}_j is

$$P(\mathbf{Y} = \mathbf{y}_j | \boldsymbol{\xi}) = \prod_{i=1}^m P(Y_i = y_{ji} | \boldsymbol{\xi}), \quad (11)$$

where $P(Y_i = y_{ji} | \boldsymbol{\xi})$ is derived from Equation 1. To estimate the model's parameters $\boldsymbol{\vartheta}$, it is necessary to estimate the marginal response probability for each individual $j = 1, \dots, n$, that is

$$P(\mathbf{Y} = \mathbf{y}_j) = \int_{-\infty}^{\infty} P(\mathbf{Y} = \mathbf{y}_j | \boldsymbol{\xi}) g(\boldsymbol{\xi}) d\boldsymbol{\xi}, \quad (12)$$

where $g(\boldsymbol{\xi})$ is the continuous (multivariate) latent variable distribution in the population. Note that if there are p latent variables in the model, $\boldsymbol{\xi}$ is a p -dimensional distribution and Equation 12 is a p -dimensional integral. The calculation of such integrals is computationally very demanding, especially when the latent variables are assumed to be correlated. Also, the computational burden increases exponentially with a linearly increasing number of latent variables (Forero & Maydeu-Olivares, 2009).

As the observed variables in the GRM are ordered variables with l_i response categories, there are $\prod_{i=1}^m l_i$ possible response patterns. Each individual j in the data can be assigned to one response pattern \mathbf{y}_r and each individual marginal probability can be assigned to the marginal probability of a responses pattern $P(\mathbf{Y} = \mathbf{y}_r)$. Let p_r be the relative frequency of the pattern \mathbf{y}_r in the data. Then, across all response patterns in the data, the objective function

$$F_{MML}(\boldsymbol{\vartheta}) = \sum_r p_r [\ln p_r - \ln P(\mathbf{Y} = \mathbf{y}_r)], \quad (13)$$

is used to successively approximate the best estimation for $\boldsymbol{\vartheta}$ (see Jöreskog & Moustaki, 2006).

As full-information parameter estimation via the EM algorithm is computationally very demanding, the application of MOB to multidimensional GRMs is not easily possible. Note that in MOB a large number of latent variable models are fitted in succession to create a tree.

4.2 Summary and Contribution

To our knowledge, Paper II is the first publication to systematically evaluate the application of MOB to MIRT models. Several different ways to apply MOB to the multidimensional GRM are systematically compared.

First, note that the approach presented by Strobl et al. (2015) and Komboz et al. (2018) is not computationally feasible when applied to multidimensional GRMs. As proposed by Schneider, Strobl, Zeileis, and Debelak (2022), the GRM should be fitted using full information parameter estimation within the MOB algorithm. As we stated in Section 4.1, this is too computationally expensive for complex MIRT models with correlated latent variables.

Thus, to apply MOB to multidimensional GRMs, limited information parameter estimation must be used so that the algorithm remains computationally feasible. When Paper II was published, however, there was no practical way to estimate individual contributions to the score function from a model fitted via ordinal factor analysis (see Equation 6). Thus, the score-based test for parameter instability could not be computed for the multidimensional GRM fitted via ordinal factor analysis. For this reason, the model was fitted with the limited information ML method (see Equation 3) to test for parameter instability within the MOB algorithm. For this approach, the model assumptions of the multidimensional GRM were compromised. Nonetheless, the approach performed very well in recovering subgroups with parameter heterogeneity when applied to simulated data. This shows that MOB can be efficiently applied to detect DIF in MIRT models with non-binary observed variables. Additionally, it was found in Paper II that calculating ensembles of MOB trees can be helpful in recovering subgroups within datasets with complex subgroup structures.

5 Paper III: Latent Variable Forests for Latent Variable Score Estimation

5.1 Background

According to Pearl (2009, p. 25), it is, to a large extent, stability that characterizes causal relationships in models. This means that they describe objective physical constraints, i.e. they are *ontological*. In contrast, probabilistic relationships are *epistemic*, i.e. reflecting what is known about the concepts in the model. Causal relationships in models remain constant as no change has taken place in the environment even when our knowledge about the environment changes. Thus, for the causal interpretation of relationships in latent variables models, latent variables must be understood as real variables instead of hypothetical constructs (see Bollen, 2002) and latent variable models must be interpreted as *functional causal models*. Such models consist of a set of *structural equations* where the latent variables ξ_i directly determine the observed variable Y_i .

The assumptions described above refer to unparametric structural models. In practice, however, latent variable models are usually applied as parametric factor models with

metric and/or ordinal observed variables (see e.g. Jöreskog, 1969; Muthén, 1984). These factor models assume a linear functional form within the latent variable model. In Section 2.2, we established that testing these models for measurement invariance does not ensure that non-ontological relations are ruled out. However, measurement invariance is a necessary condition for a causal, relationship in a parametric latent variable model. Note that a causal relationship in a model is always unbiased.

An important use for latent variable models is to scale individuals on a single construct. For this, it is useful to estimate individual latent variable scores as values of a latent variable. In order to estimate latent variable scores that are unbiased with respect to certain covariates \mathbf{Z} , the relationships within the latent variable model need to be \mathbf{Z} -conditionally stable. In Paper III, we investigate how tree-based machine learning can be used to estimate such unbiased latent variable scores. The estimated latent variable scores may then be used for descriptive and inferential purposes. They may, for example, be used to estimate latent variable effects in factor score regression (FSR) (see Devlieger, Mayer, & Rosseel, 2016; Devlieger, Talloen, & Rosseel, 2019).

5.2 Summary and Contribution

We propose a method to estimate unbiased latent variable scores based on (ordinal) factor models. For this, measurement invariance w.r.t. a predefined vector of covariates \mathbf{Z} is established in relevant subgroups in the sample. For each of these potentially overlapping subgroups, latent variable scores are estimated. This is done by repeatedly growing MOB trees and identifying its terminal nodes as relevant subgroups by testing the models fitted to the subgroups in the terminal nodes for model fit and parameter stability. Model fit is assessed with the RMSEA value and parameter instability is assessed with the score-based test for parameter instability (the same test that is used to grow the MOB tree, see Zeileis & Hornik, 2007). The estimated latent variables are then averaged across all trees for each individual. We call the method *LV Forest*.

To make the algorithm computationally feasible, the trees are grown on the basis of classic linear factor models (see Section 2.1.1). This means that the model fitting processes executed in order to find the splitting variables and split points are linear factor models assuming multivariate normality of the observed variables. Also, score-guided MOB trees are grown (see Section 2.4). After the tree growing process is completed, the models for each terminal nodes are fitted as ordinal factor models (see Section 2.1.2).

The idea for LV Forest arose from the fact that there is no way to include latent variables in causal forests (see Section 2.1.1). Where causal forests are used to estimate individual causal treatment effects (that are always unbiased), LV Forest is used to estimate unbiased latent variable scores. The individual treatment effects can be used to estimate conditional causal effects. The latent variable scores from LV Forest can be used in a similar way. However, as we mentioned before, LV Forest does not fully establish causality within a latent variable model. According to Bollen (1989, pp. 44-67), the assumption of ξ_i being the actual cause of Y_i , encompasses three components: isolation, association, and direction of influence. Where LV Forest may help to establish a certain degree of pseudo-isolation within the relevant subgroups, the other two conditions are not

addressed by the proposed method.

When the contribution of LV Forest is evaluated from the perspective of measurement invariance or DIF research, it becomes clear that the ξ_i -conditional independence of Y_i and \mathbf{Z} is most likely established in the relevant subgroups found by the algorithm. However, in LV Forest, the latent variable means and variances of the fitted models are assumed to differ between the relevant subgroups. This way, *impact* of the covariates on the latent variable is allowed across all subgroups in the sample. Impact means that a covariate Z affects one or more latent variables in the model. For measurement invariance detection, this phenomenon is usually problematic because the researcher has to define a model in which the item parameters can differ between groups while controlling for group differences in the latent variable distribution (Belzak & Bauer, 2020). Often, multiple group factor models are used for this purpose. The use of LV Forest has the advantage that no predefined groups need to be defined for which impact of \mathbf{Z} on ξ is assumed.

We conducted extensive simulations, showing that given parameter instability within the simulated samples, LV Forest increases the accuracy of estimated latent variable scores. However, in some simulation scenarios as well as in the application of LV Forest to real data, there were relatively high level of ‘nonconvergence’. In the context of this paper, nonconvergence of an individual means that this individual is not part of any relevant subgroup found by the algorithm. Therefore, scores are not estimated for this individual.

6 Paper IV: Score-Based Tests for Parameter Instability in Ordinal Factor Models

6.1 Background

The MOB algorithm is a useful tool for the detection of DIF. It may even be used for the estimation of unbiased latent variable scores. However, the application of MOB to MIRT models remains challenging as ordinal factor models cannot easily be used for parameter instability testing.

In Section 2.4, we explain that the the score-based test for parameter instability is used within the MOB algorithm to detect the individual deviations across the individuals in the sample w.r.t. the parameter estimates of the fitted model. The score-based test for parameter instability uses the individual score contributions to test if there is significant parameter instability in the sample (see Zeileis & Hornik, 2007). These individual score contributions are the individual contributions to the score function $\psi(\cdot)$. The score function is the first derivative of the objective function with respect to the parameter vector. For an arbitrary objective function $F(\theta)$ and for an individual j it is true that $E[\psi(\mathbf{y}_j, \theta)] = 0$.

As mentioned in section 2.1.1, a factor model with metric, normally distributed observed variables can be fitted by minimizing the objective function $F_{ML}(\theta)$ (shown in Equation 3), which corresponds to a limited information estimation method. However, $F_{ML}(\theta)$

is not an *individual* function, which means that it does not correspond to an individual j . This makes it impossible to estimate its individual score contributions. Yet, alternatively, the parameters can be estimated by maximizing the log likelihood $\ln L(\mathbf{Y}, \boldsymbol{\theta})$. This full information objective function is shown in Equation 16 in the Appendix. The derivative of this function with respect to the parameter vector $\boldsymbol{\theta}$ are the individual score contributions. From Equation 17 in the appendix it becomes clear that all information necessary to estimate the individual contributions to the score function $\psi(\mathbf{y}_j, \hat{\boldsymbol{\theta}}) \forall j = 1, \dots, n$, are the observations in the data, the model implied means and covariances, and the parameter estimates. Therefore, the score-based test for parameter instability is easily applicable to factor models with metric, normally distributed observed variables fitted with a limited information estimation method. The estimation of the score contributions is computationally efficient.

In Section 4.1, we introduced the full information marginal maximum likelihood estimation method. Equivalent to minimizing the objective function $F_{MML}(\boldsymbol{\vartheta})$ (see Equation 13) is maximizing the log likelihood $\ln L(\mathbf{Y}, \boldsymbol{\vartheta})$ defined in Equation 18 in the Appendix. The derivative of this log likelihood function is defined in Equation 19 in the Appendix. From this Equation, it becomes clear that for the estimation of the individual score contributions for a MIRT model fitted via marginal maximum likelihood, the computationally expensive estimation of the marginal response probability distribution $P(\mathbf{Y} = \mathbf{y}_j) \forall j = 1, \dots, n$, is necessary (see Section 4.1).

This shows that, when it comes to factor models with non-normally distributed, ordinal observed variables, the score-based test is not that easily applicable. The ordinal factor analysis approach introduced in Section 2.1.2 is computationally feasible but the objective function $F_{OFA}(\boldsymbol{\vartheta})$ shown in Equation 6 is not an individual function, just as $F_{ML}(\boldsymbol{\theta})$ (Equation 3). However, ordinal factor models cannot be fitted by maximizing individual log likelihoods. Thus, a model fitted via ordinal factor analysis (see Section 2.1.2) cannot as easily be used to estimate the individual score contributions as a metric factor model. This issue is addressed in Paper IV.

6.2 Summary and Contribution

We introduce a method to estimate individual contributions to the score function of ordinal factor models that is as computationally feasible as the estimation of score contributions for metric factor models.

To compute scores that can then be used for the score-based parameter instability test, we focus on an alternative parameter estimation approach to MML estimation that is also a full information method. In contrast to MML estimation, which is a maximum likelihood estimation method, ordinal factor analysis is a type of generalized least squares estimation method (see Muthén, 1984). It is also a limited information estimation method. A full information generalized least squares estimation method is the *generalized estimating equations* (GEE) approach (Reboussin & Liang, 1998; Muthén, 1997). The method fits latent variable models with ordered response variables. It is supposed to outperform ordinal factor analysis in cases of small sample size and large numbers of observed indicator variables. The estimation approach solves a set of estimating equations defined as follows.

Let \mathbf{e}_i is the vector of empirical deviations of the first and second order empirical moments from the true first and second order moments, that is

$$\mathbf{e}_j = \begin{pmatrix} \mathbf{1}_{\mathbf{y}_j} - \boldsymbol{\nu}(\boldsymbol{\vartheta}) \\ \mathbf{s}_j - \boldsymbol{\sigma}(\boldsymbol{\vartheta}) \end{pmatrix}, \quad (14)$$

where $\mathbf{1}_{\mathbf{y}_j}$ and \mathbf{s}_j are the empirical first and second order moments, and $\boldsymbol{\nu}$ and $\boldsymbol{\sigma}$ are the true first and second order moments for individual j . The estimation equations used in the GEE estimation approach are defined as

$$\sum_{j=1}^n \psi(\mathbf{y}_j, \boldsymbol{\vartheta}) = \sum_{i=1}^n \Delta' \mathbf{W}_{GEE}^{-1} \mathbf{e}_j = \mathbf{0}, \quad (15)$$

where Δ is the matrix of model derivatives, \mathbf{W}_{GEE} is the working covariance matrix of first and second order empirical moments. By construction, the estimating equations in Equation 15 add up to 0 across the sample, we therefore refer to them as the GEE score function. The parameters are estimated by solving this set of quadratic estimating equations for $\boldsymbol{\vartheta}$ by iteratively updating the estimator $\hat{\boldsymbol{\vartheta}}$. We use these estimating equations as basis for our approach to estimate the individual contributions to the score function for ordinal factor analysis.

However, parameter estimation via the GEE approach is not equivalent to parameter estimation via ordinal factor analysis. This means that the estimator $\hat{\boldsymbol{\vartheta}}$ is slightly different when estimated by minimizing Equation 6 compared to the estimator estimated by solving Equation 15. In contrast, the estimator $\hat{\boldsymbol{\theta}}$ of the factor model with metric, normally distributed observed variables is equal when it is approximated by minimizing Equation 3 compared to maximizing Equation 16. Thus, in order to define an individual score function for ordinal factor models, our goal is to approximate the GEE score function that would have resulted if the parameters estimated using the GEE approach were exactly the same as those estimated using ordinal factor analysis. By doing this in Paper IV, we successfully introduce a method to estimate model scores for ordinal factor models.

In the Technical Appendix of Paper IV, we also extend the definitions of the GEE estimation method. Originally, it was only defined by Reboussin and Liang (1998) and Muthén (1997) for binary observed variables. To our knowledge, no specific definitions have been published for non-binary observed variables. We applied the estimation method for the binary and non-binary case and compared the GEE scores to the ordinal factor model scores. The results indicate that the ordinal factor model scores are a valid approximation of the GEE scores.

In simulations, it is shown that the score-based test for parameter instability performs very well with the ordinal factor model scores. In a variety of different simulation scenarios, the test performs equally well or better than the score-based test for MIRT models fitted with the full information MML estimator.

7 General Discussion

The four papers written in the scope of this dissertation contribute to the field of statistics, particularly to *psychometrics*. According to Galton (1879), psychometrics means “the art

of imposing measurement and number upon operations of the mind”. A central goal of psychometric applications is to derive causal inference from behavioral measures. Often enough though, such measurements are inherently biased. All four studies deal with the question of how the bias of psychometric measurements can be reduced. A particular focus of this dissertation is on the potential of algorithmic, predictive machine learning methods for this endeavor.

Bias of psychometric measurements can often be found on the item level, e.g. when it comes to items in a psychometric questionnaire. Therefore, multidimensional IRT modeling is of great importance in the effort to reduce bias. Such explanatory MIRT models are used to test hypotheses about unobserved phenomena of the mind, taking into account the individual characteristics of each item. Paper I shows how in longitudinal settings, the inter-individual differences in item difficulty can be modeled as latent item effect variables. This illustrates that item bias in explanatory models can be reduced through sophisticated MIRT modeling.

When inter-individual differences with respect to item characteristics, like item difficulty, are not accounted for, researchers also speak of differential item functioning. When there is no longitudinal setting and there is little theoretical guidance as to which subgroups exhibit DIF, then algorithmic methods can be used for DIF detection. In Paper II, it is shown that model based recursive partitioning can be used to detect DIF in MIRT models. This algorithmic method, especially when performed as an ensemble, can be considered a type of machine learning model although it is not used for prediction.

Generally, MIRT models are usually not used for prediction. However, as shown in Paper III, together with ML methods, MIRT models can be used to compute unbiased individual scores of latent variables. For the LV Forest approach, we step away from purely explanatory modeling. Just like the individual effects estimated using the causal forest approach, the latent variable scores estimated with LV Forest can be used for descriptive and inferential purposes. Therefore, this dissertation, particularly Paper II and III, show that machine learning methods can be used to reduce bias in MIRT models.

A central aspect of Papers II and III is the computational efficiency of the proposed methods. With the estimation of individual score contributions for ordinal factor models, the computation of accurate MOB tree applications for complex MIRT models becomes computationally efficient. As machine learning in psychometrics needs to be easily applicable, the reduction of computation time is crucial for the future of machine learning in psychometrics. We address this issue in Paper IV.

Generally, there are numerous ways that machine learning techniques can be used in psychometrics. For example, item selection for psychometric questionnaires (Gonzalez, 2021b, 2025), individual classification of respondents to psychological diagnoses (Gonzalez, 2021a; Yan, Ruan, & Jiang, 2022), or the estimation of construct and criterion validity (Trognon, Cherifi, Habibi, Demange, & Prudent, 2022) may be facilitated using ML methods. However, machine learning isn’t regularly used in conjunction with latent variable models. This dissertation is a contribution to this area. Future research should focus more on this promising field.

Future research may also focus on the development of the thoughts that are introduced in this thesis. More specifically, researchers could find a way to reduce the non-convergence rate of LV Forest applications. This may be done by using a flexible RMSEA cutoff value so that all individuals in the sample are part of at least one subgroup in which a fitted MIRT model has stable parameter estimates. Also, to increase the accuracy of LV Forest, the estimation of score contributions as proposed in Paper IV needs to be implemented for score-guided *SEMTrees*. Furthermore, the method proposed in Paper IV should be extended for the use with models fitted with weighted least squares estimation methods like diagonally weighted least squares (DWLS), mean- and variance-adjusted weighted least squares (WLSMV), or unweighted least squares (ULS). Furthermore, the performance of the method proposed in Paper IV in combination with Vuong tests and robust test statistics based on sandwich estimators should be investigated. Finally, the GEE estimation method, that we implemented in the technical appendix of Paper IV, should be systematically evaluated, especially for MIRT models with non-binary observed variables.

This dissertation shows that objectives from explanatory modeling can be combined with objectives from predictive modeling. The potential of ML approaches for psychometrics has most likely not yet been fully exploited.

References

- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in psychology*, 11, 564403.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178.
- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.
- 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*. 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.
- Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, 53(1), 605–634.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Breiman, L., et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199–231.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5–32.
- Debelak, R., & Strobl, C. (2019). *Investigating measurement invariance by means of parameter instability tests for 2pl and 3pl models - online appendix*. Retrieved from <https://www.zora.uzh.ch/id/eprint/151192/2/AppendixA.pdf>
- Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76(5), 741–770.
- Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, 79(6), 1017–1037.
- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C (M-1) model. *Psychological Methods*, 8(1), 38–60.
- Fife, D. A., & D'Onofrio, J. (2023). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, 55(5), 2447–2466.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of irt graded response models: limited versus full information methods. *Psychological methods*, 14(3), 275.
- Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology*, 2(2),

- 149–162.
- Gonzalez, O. (2021a). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods*, 26(2), 236.
- Gonzalez, O. (2021b). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research*, 56(6), 903–919.
- Gonzalez, O. (2025). Combining psychometric and machine learning approaches to select items and score responses. *Behaviormetrika*, 1–34.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.
- Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2), 239–251.
- Jöreskog, K. G., & Moustaki, I. (2006). Factor analysis of ordinal variables with full information maximum likelihood. *unpublished report*.
- Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. In *Survey research methods* (Vol. 13, p. 73).
- Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous rasch models. *Educational and Psychological Measurement*, 78(1), 128–166.
- Kopf, J., Augustin, T., & Strobl, C. (2013). The potential of model-based recursive partitioning in the social sciences: Revisiting ockham’s razor. In *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 75–95). Routledge.
- 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.
- Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item response theory and factor analysis: a unified framework. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for roderick p. mcdonald* (pp. 73–102). Lawrence Erlbaum Associates Publishers.
- 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.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4), 551–560.
- 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*.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1), 41–63.
- Protzko, J. (2024). Invariance: What does measurement invariance allow us to

- claim? *Educational and Psychological Measurement*. Retrieved from <https://doi-org.emedien.ub.uni-muenchen.de/10.1177/00131644241282982>
- 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.
- Reckase, M. D., & Reckase, M. D. (2009). Historical background for multidimensional item response theory (MIRT). In *Multidimensional item response theory* (pp. 57–77). Springer.
- Rehill, P. (2024). How do applied researchers use the causal forest? a methodological review of a method. *arXiv preprint arXiv:2404.13356*.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
- 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.
- 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.
- Shmueli, G., et al. (2010). To explain or to predict? *Statistical science*, 25(3), 289–310.
- Spearman, C. (1904). “General Intelligence”, Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201–292.
- Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12.
- Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*.
- Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits – Revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- Steyer, R., Mayer, A., & Lossnitzer, C. (2023). Causal inference on total, direct, and indirect effects. In *Encyclopedia of quality of life and well-being research* (pp. 1–26). Springer.
- 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.
- Thielemann, D., Sengewald, M.-A., Kappler, G., & Steyer, R. (2017). A probit latent state IRT model with latent item-effect variables. *European Journal of Psychological Assessment*.
- Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of the vectors of mind*. University of Chicago Press.
- Trogonon, A., Cherifi, Y. I., Habibi, I., Demange, L., & Prudent, C. (2022). Using machine-learning strategies to solve psychometric problems. *Scientific Reports*, 12(1), 18922.

- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
- 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.
- Yan, W.-J., Ruan, Q.-N., & Jiang, K. (2022). Challenges for artificial intelligence in recognizing mental disorders. *Diagnostics*, 13(1), 2.
- 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.

A Formulas for Score Functions

The full information objective function for factor models with metric, normally distributed variables is the log likelihood function, that is

$$\begin{aligned}\ln L(\mathbf{Y}, \boldsymbol{\theta}) &= \sum_{j=1}^n \ln L(\mathbf{y}_j, \boldsymbol{\theta}) = \\ &= \sum_{j=1}^n -\frac{1}{2} \left(\ln |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + (\mathbf{y}_j - \boldsymbol{\mu}(\boldsymbol{\theta}))' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\mathbf{y}_j - \boldsymbol{\mu}(\boldsymbol{\theta})) \right),\end{aligned}\tag{16}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model implied covariance matrix and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the vector of model implied means. Maximizing Equation 16 is equivalent to minimizing $F_{ML}(\boldsymbol{\theta})$. By the chain rule, the first derivative of Equation with respect to $\boldsymbol{\theta}$ 16 is

$$\begin{aligned}\frac{\partial \ln L(\mathbf{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{j=1}^n \frac{\partial \ln L(\mathbf{y}_j, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^n \frac{\partial \ln L(\mathbf{y}_j, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\beta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \\ &= \sum_{j=1}^n \left(\frac{\partial \ln L(\mathbf{y}_j, \boldsymbol{\beta})}{\partial \text{vec}(\boldsymbol{\Sigma})}, \frac{\partial \ln L(\mathbf{y}_j, \boldsymbol{\beta})}{\partial \boldsymbol{\mu}} \right) \frac{\partial \boldsymbol{\beta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \\ &= \sum_{j=1}^n \left(.5 \left(\boldsymbol{\Sigma}^{-1} (\mathbf{y}_j - \boldsymbol{\mu}) (\mathbf{y}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \right), (\mathbf{y}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} \right) \frac{\partial \boldsymbol{\beta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{j=1}^n \psi(\mathbf{y}_j, \boldsymbol{\theta}) = \mathbf{0},\end{aligned}\tag{17}$$

where $\boldsymbol{\beta} = (\text{vec}(\boldsymbol{\Sigma}), \boldsymbol{\mu})$ is the vector of model parameters of the *saturated model*. In contrast to the *structured model* (with parameter vector $\boldsymbol{\theta}$), the saturated model does not impose any restrictions on the means or the covariance matrix. The matrix of model derivatives $\frac{\partial \boldsymbol{\beta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is often referred to as $\Delta(\boldsymbol{\theta})$ (see Appendix B in Savalei & Rosseel, 2022).

The the marginal log likelihood function is defined as

$$\ln L(\mathbf{Y}, \boldsymbol{\vartheta}) = \ln \left(\prod_{j=1}^n P(\mathbf{Y} = \mathbf{y}_j) \right).\tag{18}$$

The derivative of the marginal log likelihood with respect to the parameter vector $\boldsymbol{\vartheta}$ is defined as

$$\begin{aligned}\frac{\partial \ln L(\mathbf{Y}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} &= \sum_{j=1}^n \frac{\partial \ln L(\mathbf{y}_j, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \sum_{j=1}^n \frac{1}{P(\mathbf{Y} = \mathbf{y}_j)} \frac{\partial P(\mathbf{Y} = \mathbf{y}_j)}{\partial \boldsymbol{\vartheta}} = \\ &= \sum_{j=1}^n \psi(\mathbf{y}_j, \boldsymbol{\vartheta}) = \mathbf{0},\end{aligned}\tag{19}$$

see Debelak and Strobl (2019).

B Attached contributions

Paper I:	p. 24 (full text not included)
Paper I - Supplementary Material:	p. 25 (full text not included)
Paper II:	p. 26–47
Paper II - Supplementary Material:	p. 48–63
Paper III:	p. 64–99
Paper IV:	p. 100–147
Paper IV - Supplementary Material:	p. 148–158

Paper I:

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(3), 172–183. <https://doi.org/10.1027/1015-5759/a000751>

Full text not included due to licensing restrictions.

Paper I - Supplementary Material:

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(3, Suppl.), 172–183. <https://doi.org/10.1027/1015-5759/a000751>

Full text not included due to licensing restrictions.

Paper II:

Classe, F., & Kern, C. (2024). Detecting differential item functioning in multidimensional graded response models with recursive partitioning. *Applied Psychological Measurement*, 48(3), 83-103. <https://doi.org/10.1177/01466216241238743>

*Article*

Detecting Differential Item Functioning in Multidimensional Graded Response Models With Recursive Partitioning

Applied Psychological Measurement
2024, Vol. 48(3) 83–103
© The Author(s) 2024



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216241238743
journals.sagepub.com/home/apm



Franz Classe¹  and Christoph Kern²

Abstract

Differential item functioning (DIF) is a common challenge when examining latent traits in large scale surveys. In recent work, methods from the field of machine learning such as model-based recursive partitioning have been proposed to identify subgroups with DIF when little theoretical guidance and many potential subgroups are available. On this basis, we propose and compare recursive partitioning techniques for detecting DIF with a focus on measurement models with multiple latent variables and ordinal response data. We implement tree-based approaches for identifying subgroups that contribute to DIF in multidimensional latent variable modeling and propose a robust, yet scalable extension, inspired by random forests. The proposed techniques are applied and compared with simulations. We show that the proposed methods are able to efficiently detect DIF and allow to extract decision rules that lead to subgroups with well fitting models.

Keywords

differential item functioning, multidimensional item response theory, graded response model, categorical analysis, surveys, algorithmic modeling, machine learning

Introduction

Multi-item batteries are frequently used in social scientific surveys to examine latent traits. Examples include the measurement of creativity (Jauk et al., 2014), social anxiety (Prenoveau et al., 2011), and personality disorders (Drislane & Patrick, 2017). Some traits, such as self-leadership (Furtner et al., 2015), may include multiple dimensions and can involve complex (i.e., multidimensional) measurement structures. If these latent traits are to be meaningfully used for substantive analyses, one must assume measurement invariance. This requires that the

¹Deutsches Jugendinstitut, Munchen, Germany

²Ludwig-Maximilians-University of Munich, Munchen, Germany

Corresponding Author:

Franz Classe, Deutsches Jugendinstitut, Nockherstraße 2, Munchen 81541, Germany.

Email: classefranz@gmail.com

association between items of the questionnaire and latent traits of individuals do not depend on group membership. However, especially in the context of large scale surveys, the measurement invariance assumption rarely holds because of the heterogeneous nature of survey samples (Van De Schoot et al., 2015). Furthermore, a researcher can rarely identify and control all factors that jeopardize this assumption.

Measurement non-invariance is also referred to as *differential item functioning* (DIF). If group differences are found in latent factors measured by a survey questionnaire, it cannot be ruled out that this effect is only an artifact due to unnoticed DIF. That is, if DIF remains undetected, group differences can be misinterpreted. The common methods used to test for DIF usually require pre-specification of the subgroups in which DIF is assumed (Hambleton et al., 1991, p. 110). The decision which subgroups to consider for assumed DIF is often driven by theoretical priors, strong convention and biases (see Brand et al., 2019). This lets many potential relevant subgroups undetected if they do not reflect the researcher's assumptions. Therefore, more flexible, data-driven approaches can complement traditional approaches for detecting DIF.

By using data-driven, algorithmic approaches, it is possible to detect subgroups with DIF when little theoretical guidance on the relevant subgroups is available. This strand of research includes the work of Vaughn and Wang (2010) and Schauburger and Tutz (2016), who propose data-driven methods for detecting DIF for single dichotomous items in tests or questionnaires. A particularly promising method to algorithmically account for heterogeneity is *model-based recursive partitioning* (MOB), which embeds model estimation and subgroup detection in one methodological framework (Zeileis et al., 2008). In this case, the researcher only needs to specify a set of partitioning variables along with the statistical model, which are then used to iteratively search for relevant subgroups. Tutz and Berger (2016) as well as Strobl et al. (2015) present the usage of MOB for detecting DIF in the Rasch model. Komboz et al. (2018) propose a MOB-based approach for the Partial Credit Model, called *PCM Tree*, in which a single latent variable that may be susceptible to DIF is assumed. Similar in spirit, *structural equation model tree* (SEMTree) approaches have been proposed to detect homogeneous subgroups in latent variable modeling via recursive partitioning (Arnold et al., 2021; Brandmaier et al., 2013). However, there is little guidance on how recursive partitioning may be best utilized for multidimensional measurement models with ordinal response variables.

In this study, we propose and compare recursive partitioning techniques for detecting DIF with a focus on measurement models with multiple latent variables. In terms of the response variables, we consider ordinal scales, for example, Likert or ratings scales, that are often used in social scientific applications. Such models may be referred to as *multidimensional graded response* (MGR) models. Table 1 gives an overview of the methods considered in this paper. Starting from PCM Tree, we will demonstrate that a direct analogue for graded response models using full information estimation (MML) is hardly feasible to use in practice due to its high computational costs. We therefore propose a MOB for MGR approach that eases computational burden in the multidimensional setting by focusing on limited information estimation (ML, WLS). Furthermore, we compare different algorithmic approaches provided by the partykit and the semtree packages.

In addition, we address the instability issues of single tree approaches when modeling DIF. Due to MOB's hierarchical nature, small changes in the data can severely affect which subgroups are eventually identified in the splitting process (Brandmaier et al., 2016). While PCM Tree as well as the partykit and semtree approaches are susceptible to such changes, a random forest-like extension to MOB for MGR models, is analyzed that allows to robustly identify subgroups with DIF in multidimensional latent variable models.

We test and compare the outlined methods in simulations. Multiple simulation scenarios are considered that vary in the complexity of the partitioning task. The simulation results show that the proposed methods are able to correctly retrieve subgroups with distinct sets of model parameters.

Table I. Comparison of tree-based methods for detecting DIF in MGR models.

Method	Estimation	Multiple latent variables?	Complex models*?	Computationally practical?	Robustified approach?	Uncom-promised model assumptions?
PCM Tree ^a	CML (FIML)	✗	✗	✓	✗	✓
Partykit ^b	MML (FIML) ^c	✓	✗	✗	✗	✓
	ML, WLS (LIML)	✓	✓	✓	✗	✗
Naive semtree ^d	ML, WLS (LIML)	✓	✓	✗	✗	✗
Score-guided semtree ^e	ML, WLS (LIML)	✓	✓	✓	✗	✗
Partykit forest	ML, WLS (LIML)	✓	✓	✗	✓	✗
Score-guided semtree forest	ML, WLS (LIML)	✓	✓	✓	✓	✗

*Multivariate models with correlated latent variables or hierarchical structure.

^aProposed by [Komboz et al. \(2018\)](#).

^bproposed by [Zeileis et al. \(2008\)](#).

^cfollowing [Schneider et al. \(2021\)](#).

^dproposed by [Brandmaier et al. \(2013\)](#).

^eproposed by [Arnold et al. \(2021\)](#).

While partykit and semtree correctly identify subgroups in settings with clean partitioning structures, their multi-tree extensions are able to retrieve complex groups that could not have been recovered by a single decision tree. Nonetheless, computation time varies considerably across all considered methods.

Methodology

Methodological Background

Stochastic models which specify the relationship between single items with a limited amount of response categories and a continuous latent variable are consolidated under the term *item response theory* (IRT). Usually, in IRT models, a latent variable represents the ability of the respondent. This ability is assumed to underlie their response behavior ([Steyer & Eid, 2013](#)). In the following, we refer to this latent variable as ξ . Let the graded response to item i be denoted by the response variable Y_i . In IRT models for ordered response variables, as opposed to dichotomous response variables, ξ is measured by a number of items $i = 1, \dots, m$, to which the respondent answers by choosing one of the ordered response categories $k_i = 0, \dots, l_i$. The most widely applied IRT framework for items with a small amount of ordered response categories is the *graded response model* (GRM) ([Samejima, 1969](#)). Furthermore, in a multidimensional IRT framework (also referred to as MIRT, see [Forero and Maydeu-Olivares \(2009\)](#)) a response variable Y_i may be linked to more than one latent variable. In the following, we refer to the *multidimensional GRM* as MGR model. For an MGR model, ξ is a $p \times 1$ vector containing all latent variables $\xi_g \forall g = 1, \dots, p$.

The fact that the latent variables are measured by graded responses on items means that the probability of answering in a category smaller or equal to a certain ordered category k_i depends on the (multidimensional) distribution of the latent variables. In the MGR model, this relationship is defined by the cumulative category response function, that is the ξ -conditional probability function

$$P(Y_i \geq k_i | \xi) = \Phi(\beta'_i \xi - \alpha_{ik}). \quad (1)$$

The link function Φ is the distribution function of the standard normal distribution. The threshold parameter α_{ik} is the location on the underlying latent variable space where $P(Y_i \geq k_i | \xi) = 0.5$. The threshold parameters are, per definition, ordered in size, so that $\alpha_{i1} < \alpha_{i2} < \dots < \alpha_{il}$. Note that for every item i there is one threshold parameter α_{ik} less than the total number of ordered categories l_i within item i . The discrimination parameters β_{ig} , that make up the $p \times 1$ vector β_i , can be interpreted as the slope parameters of the multidimensional probability function $P(Y_i \geq k_i | \xi)$ for all categories $k_i = 0, \dots, l_i$ of item i . Because IRT parameters specify the relation between items and latent variables, we will refer to the MGR model parameters as item parameters, which form the item parameter vector, that is

$$\vartheta = \{\alpha_{11}, \dots, \alpha_{ml}, \beta_{11}, \dots, \beta_{mp}, \text{Var}(\xi_1), \dots, \text{Var}(\xi_p), \text{Cov}(\xi_1, \xi_2), \dots, \text{Cov}(\xi_{p-1}, \xi_p)\}. \quad (2)$$

Note that $\text{Var}(\xi_p)$ is fixed to 1 if β_{1p} is freely estimated (and vice versa). Also, estimating covariances between latent variables has an impact on the estimation of item threshold and discrimination parameters. We therefore consider latent variable variances and covariances as item parameters.

In IRT models, *differential item functioning* (DIF) occurs if an item parameter depends on covariates of \mathbf{Y} , that is a $m \times 1$ vector of observed response variables. Such covariates can take the form of characteristics of the individuals who respond to the items. Different scores on these covariates classify different subgroups in the population. The item parameters for each of these subgroups may differ. The difficulty of an item may, for example, depend on ethnicity, education, or gender. Differential item functioning means that the item parameter vector ϑ depends on the covariate vector \mathbf{Z} . It does not necessarily mean that the latent variable vector ξ also depends on \mathbf{Z} . This implies that DIF is present when the probability of responding to an item is different for two individuals with the same ability, only because of their group membership.

In practice, DIF can be very problematic because the number of relevant covariates may be large. Also, there is an even greater amount of possible values or value ranges of these covariates for which the item parameters might differ. In addition, complex interactions within the covariate vector \mathbf{Z} are possible so that subgroups may only be detected by considering several covariates jointly. If DIF remains undetected, group differences with respect to the latent variables can be misinterpreted (Komboz et al., 2018).

Usually, the hypothesis $\vartheta_1 \neq \vartheta_2$, where $h = 2$ stands for a focal subgroup and $h = 1$ stands for a reference group, can be tested empirically. Let's assume that, in this exemplary case, the subgroups that are tested for DIF are split at the median on the metric covariate Z_1 . In this situation, the *Likelihood Ratio* (LR) test can be applied to test if an *augmented* model, where all item parameters are allowed to vary across the two groups, outperforms a *template* model, in which all item parameters are constrained to be equal across the reference and the focal group (Bulut & Suh, 2017). If this is the case, the researcher must assume DIF for these two groups.

Turning to model parameter estimation, social scientists often use *confirmatory factor analysis* (CFA) to operationalize and estimate latent variable models with Likert-scale items (Li, 2016). In a classic CFA model, the observed items are assumed to be measured on a continuous (metric) scale. The basic factor analytic model with intercepts is

$$Y = \pi + \beta' \zeta + \epsilon, \quad (3)$$

where ϵ is the $m \times 1$ vector of residual variables and π is the $m \times 1$ vector of intercepts representing the expected values of $Y_i \forall i = 1, \dots, m$, when the values of ζ are zero (Jöreskog, 1969). Note that model fit is not affected by the estimation of intercepts. In the factor analytic framework, the model parameter vector is

$$\theta = \{\pi_1, \dots, \pi_m, \beta_{11}, \dots, \beta_{mp}, \text{Var}(\zeta_1), \dots, \text{Var}(\zeta_p), \text{Cov}(\zeta_1, \zeta_2), \dots, \text{Cov}(\zeta_{p-1}, \zeta_p), \text{Var}(\epsilon_1), \dots, \text{Var}(\epsilon_m)\}. \quad (4)$$

The CFA approach can also be used to estimate MGR model parameters. For this, a continuous, normally distributed latent response variable Y_i^* is assumed to underlie each observed response variable Y_i for item i (Muthén, 1984). In the factor analytic approach for ordinal items, the latent response variable Y_i^* of item i is related to the observed categorical response variable Y_i via a threshold relation, that is

$$Y_i = k_i \text{ if } \alpha_{ik} < y_i^* < \alpha_{i(k+1)}. \quad (5)$$

It is assumed that a respondent chooses a response category k_i when the respondent's latent response value y_i^* lies between thresholds α_{ik} and $\alpha_{i(k+1)}$.

Parameter estimation in the factor analytic framework for metric items is usually done with the maximum likelihood (ML) estimator (Jöreskog, 1969). The use of ML estimation in SEM requires the assumption that the observed variables follow a multivariate normal distribution (Li, 2016). Note that this assumption rarely holds for ordinal items. In the factor analytic framework for metric items only univariate and bivariate information is used for parameter estimation. For this, the objective function F_{ML} is minimized, that is

$$F_{ML}(\theta) = \ln|\Sigma(\theta)| + \text{tr}(\mathcal{S}\Sigma^{-1}(\theta)) - \ln|\mathcal{S}| - m, \quad (6)$$

where $\Sigma(\theta)$ is the model implied covariance matrix and \mathcal{S} is the sample covariance matrix (Jöreskog, 1969). This approach for parameter estimation is thus called *limited information approach* (LIML) and is computationally more efficient than the *full information approach* (FIML, see SupplementalMaterial S3).

Calculating the log-likelihood function for every single individual j in the sample, that is

$$\ln L(y_j, \theta) = -\frac{1}{2} \left\{ \ln|\Sigma(\theta)_j| + (y_j - \pi_j)^T \Sigma(\theta)_j^{-1} (y_j - \pi_j) \right\}, \quad (7)$$

$$\forall j = 1, \dots, n,$$

where π_j denotes the subvector of the model-implied mean vector and $\Sigma(\theta)_j$ denotes the submatrix of the model-implied covariance matrix with respect to y_j . Summing the results of equation (7) across the whole sample and maximizing the results yields asymptotically equivalent parameter estimates to limited information maximum likelihood estimation (Lee & Shi, 2021). The derivative of equation (7) can easily be derived from a model that has been fitted with F_{ML} . This derivative is also referred to as the score function and is particularly important for parameter instability testing.

It is also possible to use the limited information approach to parameter estimation for factor analysis with ordinal items. As mentioned above, normal distribution of the observed response variables cannot be assumed in this case. However, through the use of an asymptotically distribution free *weighted least squares* (WLS) estimator, normal distribution of the observed

response variables need not be assumed. Prior to parameter estimation, the thresholds that define the relation of Y^* to Y (see equation (5)) are estimated through bivariate contingency tables. Additionally, bivariate polychoric correlations are estimated in this step (Muthén, 1984). A polychoric correlation captures the strength of the considered linear dependence between Y_i^* and Y_s^* for $i \neq s$. The model parameters are then estimated through minimization of the WLS fit function, that is

$$F_{WLS}(\theta) = [\hat{\kappa} - \kappa(\theta)]' \hat{W} [\hat{\kappa} - \kappa(\theta)], \quad (8)$$

where $\kappa(\theta)$ contains the vectorized elements of the lower half of the model implied covariance matrix $\Sigma(\theta)$ and $\hat{\kappa}$ is a vector of corresponding polychoric correlation estimates below the diagonal of the polychoric correlation matrix K . The weight matrix \hat{W} is the asymptotic covariance matrix of the polychoric correlation estimates $\hat{\kappa}$. The weight matrix is supposed to account for distributional variability among the observed variables (Li, 2016).

Both CFA and MGR models can be consolidated under the *structural equation model* (SEM) framework, as both models hypothesize about multivariate constructs by specifying relationships between observable and latent variables.

Model Based Recursive Partitioning to Detect Differential Item Functioning

The application of tests such as the LR test to detect DIF requires a priori specification of the analyzed groups. Often though there are several numerical or categorical covariates and a large number of possible splitting points and the researcher may not have specified theoretical priors for all of the possible subgroups. Consequently, some subgroups with DIF might remain uncovered. In cases like this, recursive partitioning can be used as a data-driven method to uncover relevant groups for DIF. Recursive partitioning methods follow tree-based, algorithmic approaches (Breiman et al., 1984). In recursive partitioning, the full sample sits at the root of a decision tree. This root is considered a candidate for potential splitting into subgroups with respect to any of the covariates Z_r in $\{Z_1, \dots, Z_R\}$ (also called partitioning variables). A subgroup represents a tree node, which in turn is a candidate for further splitting. The algorithm may continue splitting until certain predefined stopping criteria are met. This is usually the case when there is no more significant instability in a tree node or when the subsample becomes too small. The terminal nodes of a decision tree are also called leaves. There are several methods that can be grouped under the umbrella term *Model Based Recursive Partitioning* (MOB), which we present below.

Originally, *Structural equation model trees* (SEM Trees), as presented by Brandmaier et al. (2013), combine recursive partitioning with the LR test. The algorithm searches through all partitioning variables to find subgroups that differ with respect to the model parameters. It is implemented in the *semtree* package (Brandmaier et al., 2015).

With the original (or “naive”) *semtree* approach, the parameters in θ are first estimated jointly for the entire sample using an M-estimator (like the ML estimator, see section methodological background). Then, the augmented models for all possible split points of all partitioning variables Z_r in $\{Z_1, \dots, Z_R\}$ are fitted. Note that especially if there are several (unordered) categorical and numerical partitioning variables, this means that there is a large number of augmented models to fit. However, this step is necessary to compute the log likelihood ratio for every augmented model against the template model. For every partitioning variable, the maximum log likelihood ratio is used to set the optimal split point. Then, the LR test is performed for every partitioning variable. The partitioning variable Z_{r*} with the smallest p -value in the LR test is then chosen for splitting. If none of the partitioning variables show a significant p -value, the partitioning process is stopped.

Bonferroni adjustments may be used to account for multiple comparisons. The procedure results in a tree structure with one fitted SEM for each terminal node.

One clear advantage of the naive semtree approach, compared to the LR test, is that the researcher does not need to pre-specify the functional form between the covariates and DIF. Rather, the tree structure is learned from the data in an exploratory way (Brandmaier et al., 2013). Another advantage is the ease of interpretation of the resulting subgroups. They are directly interpretable because they are built on traceable sample splits. Thus, the advantage that no pre-specification of subgroups is necessary, as in mixture models (Rost, 1990), are combined with the advantage of the LR approach, that the resulting subgroups are interpretable with respect to covariates. However, the high computational cost of this method can make its application on large data sets and complex models unfeasible.

A similar recursive partitioning approach is provided in the partykit package by Hothorn and Zeileis (2015). In contrast to the naive semtree approach, partykit tests a fitted model in a node for parameter instability with respect to any of the partitioning variables. If there is significant parameter instability, the node is eventually split at a point on the covariate with the greatest instability into two locally optimal segments. If an M-estimator is used to fit the model, parameter instability of the fitted model with respect to a covariate can be detected through the generalized M-fluctuation test (Zeileis & Hornik, 2007). The null hypothesis of the generalized M-fluctuation test is rejected if the empirical fluctuation during parameter estimation with respect to a covariate is improbably large.

Following Stefanski and Boos (2002), an M-estimator $\hat{\theta}$ is defined as the solution to the equation

$$\sum_{j=1}^n \psi(y_j, \theta) = 0. \quad (9)$$

In the context of SEM, ψ is a $(k \times 1)$ -function where k denotes the number of parameters estimated in a SEM model. The estimator $\hat{\theta}$ is the solution that minimizes the model's objective function (e.g., F_{ML} or F_{WLS} , see equation (6) and (8)). For ML estimation, $\psi(y_j, \hat{\theta})$ is the derivative function of the individual contributions to the model's log likelihood with respect to the parameter vector (see equation (7)). For $\hat{\theta}$, the derivatives add up to zero across all individuals in the sample. For k parameters in the latent variable model, the derivative function is

$$\psi(y_j, \hat{\theta}) = \left(\frac{\partial \ln L(y_j, \hat{\theta})}{\partial \hat{\theta}_1}, \dots, \frac{\partial \ln L(y_j, \hat{\theta})}{\partial \hat{\theta}_k} \right), \forall j = 1, \dots, n. \quad (10)$$

The generalized M-fluctuation test uses the function $\psi(y_j, \hat{\theta})$ to derive tests statistics that capture the empirical fluctuation process across all parameter estimates in $\hat{\theta}$. For this, different kinds of test statistics can be used. For example, for numerical covariates, partykit uses a test statistic that is equivalent with the *maxLM* statistic from Merkle and Zeileis (2013). To assess instability with respect to categorical or ordinal covariates, different kinds of test statistics based on the sum of the scores in every category are used.

The generalized M-fluctuation test rejects the null hypothesis of “no structural change” when the empirical fluctuation process becomes exceptionally large in comparison to the fluctuation of the limiting process. This limiting process is represented by the limiting distribution which can be approximated as closed form solutions to certain functions. If closed form solutions are not possible, critical values for hypothesis testing can be simulated “on the fly” (Zeileis, 2006a). Although solutions in closed form are faster, the p -values can be calculated very quickly in this

way. The generalized M-fluctuation test is provided in the *strucchange* package (Zeileis et al., 2002).

Note that the function $\psi(y_j, \hat{\theta})$ is easily obtained for ML estimation. As mentioned in section 2.1, from SEM models fitted with the limited information ML method, individual log-likelihood values (equation (7)) can be easily derived (see Zeileis, 2006b). However, this is not (yet) the case for SEM models fitted with the limited information WLS method. Parameter instability tests for MGR models fitted with WLS are not yet available. In this paper, we therefore do not directly apply the M-fluctuation test to models fitted with WLS.

In every node of a decision tree partykit tests for parameter instability. If there is overall parameter instability in the current node, that is, if the instability test for any of the partitioning variables falls below a prespecified significance level, the partitioning variable Z_* that is associated with the smallest p -value is chosen for splitting. To find the optimal split point in a binary partykit decision tree, the segmented objective functions of two rival segmentations are compared until the optimal split point on Z_* is found (Zeileis et al., 2008, p. 498f.). Note that this requires fitting as many models as there are possible segmentations of the partitioning variable Z_* .

Compared to the naive *semtree* approach, one advantage of partykit is reduced computation time. To apply the generalized M-fluctuation test to all partitioning variables, the model needs only be fitted once. Split point selection, however, is more time consuming because the model has to be fit for all possible segmentations of the selected partitioning variable.

The idea of testing a fitted model in a node for parameter instability with respect to the partitioning variables is also used in the “score-guided” *semtree* approach (Arnold et al., 2021), which supersedes naive *semtree*. As with the partykit method, the first step of the algorithm is to select the partition variable. This is done in the same way as in partykit, through the generalized M-fluctuation test.

The key difference between partykit and score-guided *semtree* is that the latter performs a different procedure than partykit for selecting the split point given a selected partitioning variable. Instead of calculating the log likelihoods for all possible rival segmentations, score-guided *semtree* identifies which of the unique values of a partitioning variable maximizes the respective score-based test statistic (Arnold et al., 2021, p. 8). As a result, the model only needs to be fitted once at each node of the decision tree. Compared to the partykit method, score-guided *semtree* can further reduce computation time in the construction of the decision tree. For the generalized M-fluctuation test, both partykit and score-guided *semtree* use the *supLM* (or equivalently *maxLM*) test statistic for metric covariates and the *LMuo* statistic for categorical variables (see Merkle & Zeileis, 2013). Score-guided *semtree* uses the *maxLM* statistic for ordered variables (*maxLMo*) (Merkle et al., 2014). All these test statistics are implemented in the *strucchange* package.

A drawback of naive and score-guided *semtree* as well as partykit is their instability towards small changes in the data because of the hierarchical nature of the tree growing process. The position of a split point in the partition determines how the sample is split up in new nodes. The position of the split point as well as the selection of the splitting variable, however, strongly depend on the particular distribution of the data. The entire structure of the tree could be altered if one splitting variable or split point was chosen differently (Strobl et al., 2009).

Recursive Partitioning for Multidimensional Graded Response Models

As mentioned in Section 2.2, recursive partitioning can be applied to any kind of parametric model that is fitted using an M-estimator (e.g., maximum-likelihood). Komboz et al. (2018) propose a recursive partitioning algorithm to detect DIF in the *Partial Credit Model* (PCM), called *PCM Tree*. The PCM is another model from the IRT framework. The PCM Tree algorithm includes a

global test for measurement invariance. If there is significant item parameter instability with respect to any of the covariates Z_r in \mathbf{Z} , then the assumption of measurement invariance (no DIF) should be rejected.

In PCM Tree, only one latent variable ζ can be considered in the models that are associated with the tree's nodes and thus multidimensional graded response (MGR) models cannot be handled. A direct analogue to PCM Tree for MGR models would draw on full information parameter estimation in the tree growing process (see [Schneider et al., 2021](#)). In [Supplemental Material S3](#), however, we establish that model based recursive partitioning for MGR models using the full information approach is rarely feasible due to enormous computational costs. Thus, in order to conduct MOB for MGR models, computationally efficient approaches are needed.

We present and compare practicable methods to test and control for differential item functioning for complex survey scales and large scale survey data. Particularly, we suggest to combine the limited information approach for parameter estimation (Section 2.1) and recursive partitioning algorithms (Section 2.2) in order to efficiently compute MGR model based decision trees and to evaluate the resulting models with regard to model fit.

Recursive Partitioning for Multidimensional Graded Response Models: Single Tree. In this section, we introduce different ways to efficiently compute a single recursive partitioning tree for MGR models. We distinguish between the tree growing process (first step) and the terminal node model estimation process (second step). On this basis, we draw on different estimators to detect subgroups with DIF and to estimate fit indices and parameter estimates in an MGR modeling context. We present three algorithms, utilizing the *semtree* and the *partykit* packages (Section 2.2). The proposed methods are summarized schematically in [Supplemental Material S1](#) in Algorithm 1, 2, and 3. Note that the algorithms differ with respect to the tree growing process as implied by the different packages used.

To start tree growing with the naive *semtree* approach, numerous models have to be fitted for which the log likelihoods are then compared with the template model. In the first step of the *partykit* method and the score-guided *semtree* method, the score function (see equation (10)) is used to build the tree structure. Usually, the MML estimation method is too computationally expensive for these approaches (see [Supplemental Material S3](#)). To efficiently calculate log-likelihoods for naive *semtree* and the score function for *partykit* and score-guided *semtree*, we propose to use (limited information) ML estimation in the tree growing process, that is, parameter estimates are computed by minimizing the objective function of the ML estimator (equation (6)). Thus for all three algorithms, we compromise on our assumptions about the distribution of the response variables. In the first step of the proposed recursive partitioning approaches for MGR models, information is used that is based on the assumption that the observed variables follow a continuous multivariate distribution. This may lead to problems in the tree growing process. In this study, we therefore analyze tree stability using data with simulated numeric response variables (based on a traditional CFA model) and compare the resulting trees to those grown using data with ordinal response variables (based on a MGR model).

Note that for *partykit* and *semtree* for MGR models, the M-fluctuation test uses the partial derivative of the objective function with respect to the model parameter vector θ (as opposed to the item parameter vector ϑ). This means that individual contributions to the score function include individual deviations with respect to residual variances and nodes are split to minimize the interindividual variance with respect to these parameters. However, these parameters don't exist in the original GRM. In the MGR model, DIF occurs if the item parameter vector ϑ depends on covariates of the response variables (see Section Methodological Background). Thus, strictly speaking, the partial score function with respect to the item parameter vector, $\psi(\mathbf{Y}, \hat{\vartheta})$, needs to be considered for DIF detection through *partykit* or *semtree*. In [Supplemental Material S3](#), we apply

the MOB method to detect DIF with respect to the item parameter vector $\boldsymbol{\theta}$. However, this method turned out to be nearly infeasible due to high computational costs as outlined above. The estimation is computationally expensive because multidimensional integrals have to be solved in order to minimize the objective function. Using this full information approach, however, individual contributions to the minimization of the objective function are considered and the function $\psi(\mathbf{Y}, \hat{\boldsymbol{\theta}})$ is derived.

In the second step of our proposed algorithms, the parameter and model fit estimates of the models that are stored in the terminal nodes of the decision tree are calculated using the distribution free weighted least squares (WLS) estimator. Thus, for evaluation of the resulting decision tree, the model fit indices in the terminal nodes are estimated under consideration of non-normally distributed response variables and the existence of the threshold relation between the response variable vector \mathbf{Y} and latent response variable vector \mathbf{Y}^* . Thus, parameters and standard errors are only estimated for models that fit the data within the subgroup. Along with sufficient sample size, this is very important for correctly estimating parameters and standard errors. Parameters in models in which the parameters are stable but which don't fit the data are unlikely to be interpretable.

Recursive Partitioning for Multidimensional Graded Response Models: Forests. While the outlined methods allow to efficiently grow a single decision tree, this method may be slightly inaccurate because MGR model assumptions are compromised. At some splitting points in the decision tree, variable and split point selection may be different if the objective function considered all parameters and distributional assumptions of the MGR model (see also [Supplemental Material S2](#)). Also, a single decision tree can be vulnerable to small changes to the data and to the set of partitioning variables. This is a consequence of the hierarchical nature of the splitting process ([Brandmaier et al., 2016](#); [Kern et al., 2019](#))—the selection of one particular partitioning variable Z_{r*} at the root node determines the entire tree structure.

Using the computation time saving method described above, we are able to tackle the problem of unstable and potentially inaccurate trees by computing several structurally different trees and evaluating the compiled results of the tree ensemble. As the computation of a decision tree using partykit and score-guided semtree is considerably less time consuming compared to the naive semtree approach ([Arnold et al., 2021](#)) we only consider these methods (i.e., [Supplemental Material S1](#), Algorithm 2 and 3) as base learner in the ensemble.

We are guided by the concept of random forests, a method that uses an ensemble of decision trees rather than a single one to enhance prediction performance ([Breiman, 2001a](#)). We use random split selection to grow decorrelated trees for the ensemble that are structurally different from each other. In this procedure, random selections of partitioning variables are made. The selection of partitioning variables is redrawn at every node in a decision tree. This way, we encourage that all partitioning variables are considered at least once, even if a small number of trees are computed. Another technique used in the random forest framework is bagging. If bagging is used, the tree growing algorithm is applied to a bootstrap sample drawn from the full sample at every iteration. However, we refrain from using bagging together with recursive partitioning for MGR models. We want to ensure that the parameter estimates in the subgroups that are found by the algorithm are directly replicable. This is necessary to ensure that the fit indices of the fitted models are comparable between the trees.

The steps performed to grow a forest of partykit trees or score-guided semtrees for MGR models are summarized in [Supplemental Material S1](#) in Algorithm 4. Multiple decision trees are grown using either partykit or semtree for MGR models (see section Recursive Partitioning for MGR models: Single Tree) with random sampling of partitioning variables at each node. After

multiple decision trees are grown, the fit indices of the fitted models in the terminal nodes of each decision tree are evaluated. In this step, fitted models in terminal nodes that don't exceed a predefined cutoff criterion (χ^2 -test p -value or RMSEA cutoff) are selected. The forest outputs a list of subgroups for which the proposed MGR model holds and DIF is present.

Simulations

Measurement Model

We test and compare the presented recursive partitioning techniques for MGR models with simulations. For this, a multidimensional graded response model needs to be defined. In the following, the simulated data is created based on the assumptions of the *probit multistate IRT model with latent item effect variables for graded responses* (PIEG, Classe & Steyer, 2023a).

The PIEG model is a multistate model with latent item effect variables for ordinal observables. For every category of a response variable, one category-specific latent state variable τ_{ikt} for category k of item i at time point t is defined in the PIEG model. One reference latent state variable η_t , which is equal to the latent state variable of the reference item τ_{11t} , is assumed for every time point of measurement. The latent item effect variable β_i is defined as the difference between the latent state variable of the reference item and the latent state variable of another item. Thus, there are as many latent item effect variables as there are items, minus the reference item. In this model, variances and covariances of latent state variables, and latent item effect variables as well as the covariances between latent item effect variables and latent state variables are estimated. The model's discrimination parameters are all fixed at 1. For our application, all threshold parameters are freely estimated.

To simulate data on the basis of the PIEG model, we define three reference latent state variables η_t and two latent item effect variables β_i . We are thus mimicking a longitudinal setting with data collected for three time points. The proposed latent variables are derived from three items, respectively, resulting in nine five-category ordinal response variables Y_{it} . The model structure is shown in Figure 1 in [Supplemental Material S1](#). The cumulative category response function of the PIEG model is

$$P(Y_{it} \geq k_i | \eta_t, \beta_i) = \Phi(\eta_t + \beta_i - \kappa_{ikt}), \quad (11)$$

$$\forall k = 1, \dots, 4, \forall i = 2, \dots, 3, \forall t = 1, \dots, 3.$$

In this model, there are 36 free threshold parameters (4 for every five-category item), 10 free covariances between the latent variables, and 5 free variances of the latent variables, resulting in 51 free parameters in total.

To additionally simulate data with which ML estimation can be performed without compromising model assumptions, we define a traditional CFA model for which the response variables are numerical and follow the normal distribution. The model function is

$$Y_{it} = \pi_{it} + \eta_t + \beta_i + \epsilon_i, \quad (12)$$

$$\forall i = 2, \dots, 3, \forall t = 1, \dots, 3.$$

where π_{it} is an item- and time-specific intercept and ϵ_i is an item-specific residual variable. In this model, there are 10 free covariances between the latent variables, 5 free variances of the latent variables, 9 free intercepts and 9 free residual variances resulting in 33 free parameters in total.

For all data sets, several partitioning variables $Z_r \forall r = 1, \dots, R$ are simulated. Different subgroups $R_h \forall h = 1, \dots, H$ for which DIF is present may be defined as different areas on the (multidimensional) distribution of these partitioning variables.

Simulation Setup

We create simulated data to test and compare the performance of partykit, naive and score-guided semtree for MGR models. Single decision tree approaches are applied to the first set of simulations (simulation 1) while ensemble techniques are applied to the second set of simulations (simulation 2). We conduct additional simulations to test the performance of the generalized M-fluctuation test under misspecification in Supplemental Material S2. R implementations of the proposed methods and replication materials for all simulations are provided in the following OSF repository: https://osf.io/sv35m/?view_only=6cdde2777b914322b32ca00ad567ff2b.

Simulation 1. The samples of simulation 1 each consist of 2000 observations with values on 17 variables. There are no missing data points in the samples. We simulate 9 response variables in two ways: One set of samples with ordinal response variables that are based on the model function in equation (11) (see Figure 1). We also created a set of numeric samples that are based on the model function in equation (12). For each (ordinal and numeric) sample, we created two ordinal variables (cat1 and cat2) with scores on a five-point Likert Scale and one numerical variable (num1) ranging from 1 to 200. Those three variables are relevant partitioning variables. This means that they allow to distinguish between four subgroups with 500 observations in each group. Additionally, for each sample five random partitioning variables (rand1 to rand5) were simulated that do not systematically differentiate among the four subgroups. There are two numerical and three ordinal random partitioning variables.

We simulated 100 ordinal samples and 100 numeric samples. For each sample, the data for each subgroup is simulated with a different set of parameters so that the model function in equation (11) (for the ordinal data) or equation (12) (for the numerical data) is true for each subgroup, but there is DIF in the overall sample. The true group-specific parameters differ between the samples. Intercepts and threshold parameters were sampled from a normal distribution, and latent variable variances and covariances were sampled from a uniform distribution. Further details and code for replication purposes is provided in the OSF repository. For each subgroup within one single sample, the values on the relevant partitioning variables are simulated such that each subgroup is exclusive with respect to the values of the relevant partitioning variables. Additionally, the structure of the simulated sample can be broken down by a single decision tree. The subgroups are defined as

$$\begin{aligned} R_1 &:= \{\{\text{num1} < 100\} \cap \{\text{cat1} \in \{1, 5\}\}\}, \\ R_2 &:= \{\{\text{num1} < 100\} \cap \{\text{cat1} \in \{2, 3, 4\}\}\}, \\ R_3 &:= \{\{\text{num1} \geq 100\} \cap \{\text{cat2} \leq 2\}\}, \\ R_4 &:= \{\{\text{num1} \geq 100\} \cap \{\text{cat2} \geq 3\}\}. \end{aligned}$$

All subgroups within one single sample fit the assumed model very well (RMSEA of 0.05 or lower for the models shown in equations (11) and (12), respectively).

We conduct the simulation analysis in two steps. In the first step, we apply partykit, naive and score-guided semtree for MGR models to one single ordinal sample of simulation 1 to test if the methods are able to detect DIF and to compare runtime results for a sample that has a clear subgroup structure. In the respective model setup, we do not impose constraints on the minimum sample size in the terminal nodes. Bonferroni adjustments are applied at every node to correct for the multiple comparisons arising from the repetition of the generalized M-fluctuation test (for partykit and score-guided semtree) or of the LR-test (for naive semtree). The number of hypothesis repeated at every node is equal to the number of partitioning variables used.

The PIEG model fit the four subsets of this sample very well (R_1 : RMSEA < .001, 95% C.I. = .000 – .034, R_2 : RMSEA < .001, 95% C.I. = .000 – .033, R_3 : RMSEA = .025, 95% C.I. = .000 – .048, R_4 : RMSEA = .013, 95% C.I. = .000 – .041). Through Monte Carlo

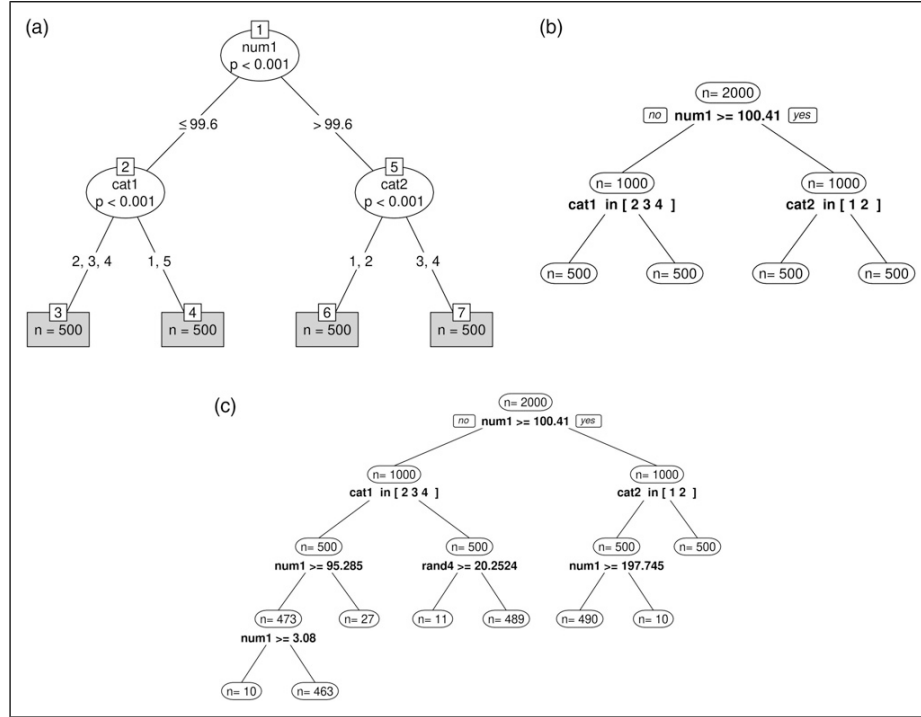


Figure 1. Results of single sample application of simulation I. (a) Partykit for MGR models. (b) Score-guided semtree for MGR models. (c) Naive semtree for MGR models.

simulation, [Classe and Steyer \(2023b\)](#) found that the quality of the parameter estimates and standard errors associated with the PIEG model are very good for sample sizes of 500, given the model fits the data. We therefore assume that recovery of the simulated subgroups, in which the models fit very well, results in accurate parameter estimation within these subgroups. The input parameters for all subgroups (R_1 to R_4) in this sample that are used for data generation are shown in Tables 1 and Table 2 in [Supplemental Material S1](#).

In the second step, we apply partykit and score-guided semtree to all 100 ordinal and 100 numerical samples and analyze tree stability across simulations.

Simulation 2. The samples of simulation 2 each consist of 2000 observations on 18 variables. Again, there are five random partitioning variables in these samples. In addition, there are four relevant partitioning variables: cat1 (categorical), cat2 (ordinal), num1 (numerical) and dichol1 (dichotomous). The relevant partitioning variables differentiate among two (exclusive) subgroups defined as

$$R_1 = \{\{\text{cat } 2 \geq 3\} \cap \{\text{cat } 2 \leq 4\} \cap \{\text{num1} \leq 50\}\},$$

$$R_2 = \{\{\text{dichol1} = 0\} \cap \{\text{cat1} \in \{1, 4, 5\}\}\}.$$

The subgroups R_1 and R_2 consist of 500 observations each (within one single sample). The data for the subgroups are simulated to fit the PIEG model well but with different sets of parameters such that DIF is present. Again, 100 ordinal samples as well as 100 numeric samples are created. The values of the simulated response variables for the remaining half of each sample of simulation 2 are random. For the ordinal samples, this means that values between 1 and 5 were randomly sampled for all response variables for all individuals that did not belong to R_1 or R_2 . For the numerical samples, the values were randomly sampled from a uniform distribution with a minimum of -3 and a maximum of 3 . Consequently, the PIEG model only holds true for subgroups R_1 and R_2 . Additionally, the simulated subgroup structure of the sample of simulation 2 cannot be recovered by one single decision tree.

We again proceed in two steps. In the first step, we apply partykit and score-guided semtree forests to a single ordinal sample of simulation 2 to test whether the methods are able to detect DIF in a sample in which the subgroup structure is complex and the assumed MGR model does not hold for every individual in the sample. The data of half of that sample includes the same response variables as the initial sample of simulation 1 (i.e., except for the randomly generated data points). The partitioning variables are re-simulated. The input parameters are shown in [Supplemental Material S1](#) in Table 1 and 2 in column R_1 and R_2 . For every computed decision tree, we refit the models in each terminal node using the WLS estimator, and gather the model fit information. We compute an ensemble of 50 trees and set an RMSEA cutoff criterion of 0.05. The minimal size of the subgroups in the terminal nodes is set to 100 such that model parameters and fit indices can be estimated properly. Additionally, we set the number of variables randomly sampled as candidates at each split point to 3. For this data set, we defined the *cat2* variable as categorical so that only two splits are necessary to retrieve the simulated subgroup R_1 in a terminal node of a decision tree

$$R_1 : = \{ \{ \text{cat2} \in \{3, 4\} \} \cap \{ \text{num1} \leq 50 \} \}.$$

In the second step, we compute score-guided semtree forests for all 100 ordinal data sets and for 100 numeric data sets and analyze the method's ability to retrieve the two simulated subgroups from a complex sample structure across multiple samples. We computed ensembles of 20 trees using the same hyperparameters as in the single sample application.

Simulation Results

The results of the single sample application of simulation 1 are shown in [Figure 1\(a\)](#) (partykit), 1b (score-guided semtree), and 1c (naive semtree). When using partykit and score-guided semtree for MGR models ([Figures 1\(a\) and \(b\)](#)), all subgroups (R_1 to R_4) were retrieved correctly. For the naive semtree ([Figure 1\(c\)](#)), however, the algorithm did not stop splitting although the parameters in a terminal node are stable. These results indicate that partykit as well as score-guided semtree may be used for DIF detection in a sample that has a clear subgroup structure and for which the assumed MGR model is generally true. For the naive semtree method, on the other hand, it seems like the LR-test does not perform well with respect to numerical covariates.

When it comes to computation time, there are considerable differences between the three methods. The computation of the partykit tree took 361.5 seconds (6 minutes), the computation of score-guided semtree took 7.8 seconds, and the computation of the naive semtree algorithm took 4357 seconds (1.2 hours). These applications were conducted on a processor with a single core and 8 GB RAM. The runtime results show that naive semtree algorithm is computationally demanding and not a reasonable candidate for growing a decision tree ensemble. The modern, score-guided semtree, on the other hand, appears to be a considerably more practical method for the detection of DIF in MGR models, also in comparison to partykit. As it allows to choose from different types of

score-based test statistics, semtree appears to be a good candidate to efficiently calculate robust tree ensembles.

We analyze and compare tree stability results of 100-fold simulations between partykit and score-guided semtree as well as between ordinal and numerical response data. We define three levels of tree stability. A stable tree is defined as a tree in which all splits have been performed at the correct split points using the correct partitioning variables and all individuals in the sample are correctly distributed among the terminal nodes. An example for such a perfect split result is shown in Figures 1(a) and (b). The second level of tree stability is defined as a tree in which the split point on the numerical variable num1 has not been perfectly detected so that not all individuals in the sample are correctly distributed among the terminal nodes. An example for such an imperfect split result is shown in Figure 2(a). The third level of tree stability is defined as a tree in which one or more faulty splits have been performed. An example for such an incorrect split result is shown in Figure 2(b).

The results of applying partykit and score-guided semtree to 100 numeric and 100 ordinal samples are shown in Table 2. The tree stability patterns show no strong differences between partykit and semtree. However, there are apparent differences when comparing the applications on ordinal and numerical response data. With numerical response data, more trees were perfectly stable. However, this is only due to a higher rate of inaccuracies in split point selection and not due to more (fully) incorrect splits with ordinal response data.

The samples for simulation 2 included two subgroups with DIF (R_1 and R_2) and random data such that the PIEG model only holds for a portion of the sample. In addition, the simulated subgroups are not retrievable through one single decision tree. In a single sample application, we first investigate if a forest of decision trees is able to correctly detect the simulated subgroups in the data set. As shown in Tables 3 and 4, both methods are successful in the retrieval of the two subgroups as those subgroups are repeatedly identified with best model fit. However, there were other subgroups that also fit the data well (i.e., model fit estimate fell under RMSEA cutoff) although these subgroups were not explicitly simulated to fit the data. It becomes apparent that the other subgroups identified by the forests are (random) subsets of either R_1 or R_2 . This result indicates that not all of the subgroups with acceptable model fit indices in the tree ensembles should be strictly interpreted as subgroups in which the assumed model is inherently true. Those groups with the best model fit that do not share any subset with another subgroup in the list, however, may be interpreted as subgroups in which the assumed model holds.

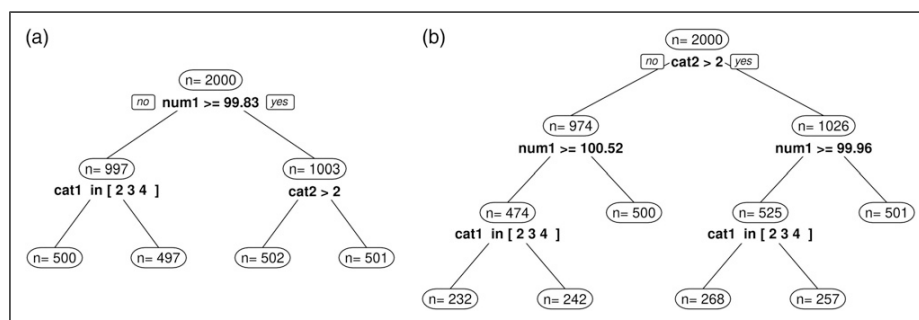


Figure 2. Examples for tree instability in simulation 1. (a) Inaccurate split point selection. (b) Incorrect splits performed.

The runtime of partykit and semtree forest depend on the number of trees of the ensemble. Thus, holding the number of trees constant, semtree forest take considerably less time to compute than partykit forest. In simulation 2, the computation time of the single trees in the ensembles were on average comparable to the computation times in simulation 1, as some trees grew deeper and others stopped splitting at the root node. Note that growing a forest can be parallelized and therefore the computation time of recursive partitioning forests also depends on the number of available processing cores.

Repeating the application of semtree forests with 20 trees in each ensemble on 100 ordinal data sets resulted in 95% of the forests recovering at least one simulated subgroup (R_1 or R_2). Furthermore, 41% of the forests recovered both R_1 and R_2 . The same application on 100 numeric data sets resulted in 78% of the forests recovering at least one simulated subgroup and only 18% recovering both subgroups. Thus, the problem of inaccurate selection of split points in the decision tree for ordinal data seems to be solved by using partitioning tree ensembles.

Discussion

Heterogeneity in survey samples is a common challenge when latent variable models are used to measure latent traits in substantive research. Survey data may include multiple, complex subgroups which can be subject to differential item functioning, and/or for which the implied measurement model does not hold altogether. Following the work of Strobl et al. (2015) and Komboz et al. (2018), we investigate several approaches for accounting for DIF in the most prominent type of multidimensional polytomous IRT model: the multidimensional graded response (MGR) model. By focusing on ordinal response scales and allowing for multiple latent

Table 2. Tree stability across repetitions in simulation 1.

	Ordinal data			Numerical data		
	Perfect splits	Inaccurate split point	Incorrect splits	Perfect splits	Inaccurate split point	Incorrect splits
Partykit	42%	48%	10%	76%	14%	10%
Semtree	40%	47%	13%	69%	13%	18%

Table 3. Results of the application of semtree forest for MGR models to the sample of simulation 2. Subgroups with best model fit are shown. The column label "Freq." refers to the number of decision trees in the forest that identified the respective subgroup.

Sim.	Decision rule	Freq	n	RMSEA	p-value χ^2 -test
✓	$R_{\text{semtree}_1} := \{\{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.88\}\}$	3	499	0	0.592
✓	$R_{\text{semtree}_2} := \{\{\text{dicho1} = 0\} \cap \{\text{cat1} \in \{1, 4, 5\}\}\}$	1	500	0	0.568
✗	$R_{\text{semtree}_3} := \{\{\text{dicho1} = 1\} \cap \{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.99\}\}$	1	254	0	0.792
✗	$R_{\text{semtree}_4} := \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} \leq 49.51\} \cap \{\text{cat2} \in \{3, 4\}\}\}$	4	146	0	0.462
✗	$R_{\text{semtree}_5} := \{\{\text{num1} \geq 53.82\} \cap \{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{dicho1} = 0\}\}$	7	373	0.011	0.402
✗	$R_{\text{semtree}_6} := \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} \geq 49.51\} \cap \{\text{dicho1} = 0\}\}$	2	382	0.011	0.398
✗	$R_{\text{semtree}_7} := \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} \leq 49.98\} \cap \{\text{cat2} \in \{3, 4\}\}\}$	2	147	0.016	0.411
✗	$R_{\text{semtree}_8} := \{\{\text{cat2} \in \{1, 2, 5\}\} \cap \{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{dicho1} = 0\}\}$	3	349	0.027	0.198
✗	$R_{\text{semtree}_9} := \{\{\text{cat1} \in \{2, 3\}\} \cap \{\text{num1} \leq 49.91\} \cap \{\text{cat2} \in \{3, 4\}\}\}$	4	353	0.028	0.179

Table 4. Results of the application of partykit forest for MGR models to the sample of simulation 2. Subgroups with best model fit are shown. The column label “Freq.” refers to the number of decision trees in the forest That identified the respective subgroup.

Sim. Subgrp	Decision rule	Freq	n	RMSEA	p-value χ^2 -test
✓	$R_{mab_1} : = \{\{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.87\}\}$	2	499	0	0.592
✓	$R_{mab_2} : = \{\{\text{dichol} = 0\} \cap \{\text{cat1} \in \{1, 4, 5\}\}\}$	7	500	0	0.568
✗	$R_{mab_3} : = \{\{\text{dichol} = 1\} \cap \{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.39\}\}$	1	252	0	0.823
✗	$R_{mab_4} : = \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} \leq 49.51\} \cap \{\text{cat2} \in \{3, 4\}\}\}$	2	146	0	0.462
✗	$R_{mab_5} : = \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} \leq 49.39\} \cap \{\text{dichol} = 1\} \cap \{\text{cat2} \in \{2, 4\}\}\}$	1	102	0	0.523
✗	$R_{mab_6} : = \{\{\text{num1} > 53.77\} \cap \{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{dichol} = 0\}\}$	2	373	0.011	0.402
✗	$R_{mab_7} : = \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{num1} > 49.39\} \cap \{\text{dichol} = 0\}\}$	4	382	0.011	0.398
✗	$R_{mab_8} : = \{\{\text{dichol} = 0\} \cap \{\text{cat1} \in \{2, 3\}\} \cap \{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.87\}\}$	1	246	0.022	0.321
✗	$R_{mab_9} : = \{\{\text{cat2} \in \{1, 2, 5\}\} \cap \{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{dichol} = 0\}\}$	2	349	0.027	0.198
✗	$R_{mab_{10}} : = \{\{\text{cat1} \in \{2, 3\}\} \cap \{\text{num1} \leq 49.89\} \cap \{\text{cat2} \in \{3, 4\}\}\}$	3	353	0.028	0.179
✗	$R_{mab_{11}} : = \{\{\text{cat1} \in \{2, 3\}\} \cap \{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 49.87\}\}$	1	352	0.029	0.164
✗	$R_{mab_{12}} : = \{\{\text{cat1} \in \{1, 4, 5\}\} \cap \{\text{dichol} = 1\} \cap \{\text{cat2} \in \{3, 4\}\} \cap \{\text{num1} \leq 45.9\}\}$	1	133	0.045	0.184

variables, recursive partitioning for MGR models aims to tackle DIF in modeling contexts that are common in social scientific survey settings. We draw on three different recursive partitioning algorithms: naive and score-guided semtree (Arnold et al., 2021; Brandmaier et al., 2013) as well as partykit (Zeileis et al., 2008). As we utilize limited information estimation in building decision trees, we also propose practicable multi-tree extensions of partykit and semtree for MGR models. These approaches allow to account for instabilities in the tree growing process while maintaining computational feasibility.

In simulation 1, we demonstrated that partykit and score-guided semtree can be used to correctly find subgroups with DIF in MGR models. Comparing the algorithms using data in which the assumptions underlying ML estimation are compromised (i.e., ordinal response data) versus data in which these assumptions are not compromised (i.e., numeric response data) showed that there are not more incorrect splits performed with ordinal data. The results of the simulation study performed in [Supplemental Material S2](#) support this finding as they indicate that different struchange tests used on ordinal data do not perform worse than the same tests used on numeric data. However, compromising the MGR model assumptions during the tree growing process can lead to more inaccurate split points, at least for numerical partitioning variables.

The results of simulation 2 showed that a forest of semtrees is computationally more practical than a forest of partykit trees. The repeated application of semtree ensembles indicated that it is possible to retrieve subgroups with DIF from data with complex subgroup structures using tailored tree ensemble approaches. Our simulation also showed that applying such a tree ensemble method to numeric response data does not lead to better subgroup recovery. This result suggests that an ensemble method may be able to account for the instabilities of the tree caused by the compromised MGR model assumptions during tree growth. Note that in real applications, samples consist of complex subgroup structures anyway, and tree instability may be present even if the assumptions of the underlying model are not compromised. We may thus conclude that partykit and (ensembles of) semtree for MGR models represent useful tools for researchers working with multidimensional latent variable models and ordinal items in survey data.

Note that in extending recursive partitioning for MGR models to a tree ensemble method, we do no longer focus exclusively on detecting DIF. We rather consider the possibility that the assumed model structure underlying the ordinal items does not hold for all subgroups of the sample. Additionally, we acknowledge that the subgroup structure may be too complex to be disentangled by a single decision tree. In other words, an ensemble of recursive partitioning trees for MGR models recognizes that traditional data models, such as MGR models, are often not complex enough to accurately represent the internal processes of all respondents in deciding which categories to check off on survey scale items. It is rather likely that the assumption of a fixed model structure with stable parameters does not hold for every individual in every context. In these cases, parameter heterogeneity and model fit heterogeneity can be expected.

For this reason, we use a hybrid approach that includes an algorithmic model (random forest) and a data model (multidimensional GRM). Methods from the algorithmic modeling culture assume that natural mechanisms that produce data are unknown. Algorithmic models are usually used as “black boxes” to predict outcomes of such natural mechanisms (Breiman, 2001b, p. 205). Models from the data modeling culture, on the other hand, are typically restrictive explanatory models used to estimate parameters that are then used to test causal explanations. Algorithmic models need to be flexible enough to approximate the data generating mechanism well while also being robust to changes in the data. This compromise is referred to as the bias-variance trade-off in the algorithmic modeling literature (Hastie et al., 2009, p. 37). A recursive partitioning ensemble reduces bias by identifying various decision rules and associated parameter values for which the assumed model fits. It is these decision rules that lead to conditions under which controlling for DIF in MGR models actually reduces bias. Variance in tree ensembles for MGR models, on the

other hand, can be controlled via the minimum size of the subgroups in the terminal nodes. Further extensions to this end could include the use of bootstrap resampling in tree ensembles for MGR models.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

Supplemental Material

Supplemental material for this article is available online.

References

- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology, 11*, 564403. <https://doi.org/10.3389/fpsyg.2020.564403>.
- Brand, J. E., Xu, J., Koch, B., & Geraldo, P. (2019). *Uncovering sociological effect heterogeneity using machine learning*. arXiv.
- Brandmaier, A. M., Prindle, J. J., & Arnold, M. (2015). semtree: Recursive partitioning of structural equation models in R [Computer software manual]. CRAN. <https://cran.r-project.org/web/packages/semtree/semtree.pdf>
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*(4), 566–582. <https://doi.org/10.1037/met0000090>
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86. <https://doi.org/10.1037/a0030001>
- Breiman, L. (2001a). Random forests. *Machine Learning, 45*(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- 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. <https://doi.org/10.3389/educ.2017.00051>.
- Classe, F. L., & Steyer, R. (2023a). A probit multistate irt model with latent item effect variables for graded responses. *European Journal of Psychological Assessment*. <https://econtent.hogrefe.com/doi/10.1027/1015-5759/a000751>.
- Classe, F. L., & Steyer, R. (2023b). *A probit multistate irt model with latent item effect variables for graded responses*. [supplementary material]. Hogrefe Publishing. https://www.oee.comm/oee_factors.html
- Drislane, L. E., & Patrick, C. J. (2017). Integrating alternative conceptions of psychopathic personality: A latent variable model of triarchic psychopathy constructs. *Journal of Personality Disorders, 31*(1), 110–132. https://doi.org/10.1521/pedi_2016_30_240

- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275–299. <https://doi.org/10.1037/a0015825>
- Furtner, M. R., Rauthmann, J. F., & Sachse, P. (2015). Unique self-leadership: A bifactor model approach. *Leadership, 11*(1), 105–125. <https://doi.org/10.1177/1742715013511484>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research, 16*(1), 3905–3909.
- Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The road to creative achievement: A latent variable model of ability and personality predictors. *European Journal of Personality, 28*(1), 95–105. <https://doi.org/10.1002/per.1941>
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series, 1967*(2), 183–202. <https://doi.org/10.1002/j.2333-8504.1967.tb00991.x>
- Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods, 13*(1), 73–93.
- Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous rasch models. *Educational and Psychological Measurement, 78*(1), 128–166. <https://doi.org/10.1177/0013164416664394>
- Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods, 26*(4), 466–485. <https://doi.org/10.1037/met0000381>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79*(4), 569–584. <https://doi.org/10.1007/s11336-013-9376-7>
- Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika, 78*(1), 59–82. <https://doi.org/10.1007/s11336-012-9302-4>
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115–132. <https://doi.org/10.1007/bf02294210>
- Prenoveau, J. M., Craske, M. G., Zinbarg, R. E., Mineka, S., Rose, R. D., & Griffith, J. W. (2011). Are anxiety and depression just as stable as personality during late adolescence? Results from a three-year longitudinal latent variable study. *Journal of Abnormal Psychology, 120*(4), 832–843. <https://doi.org/10.1037/a0023939>
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282. <https://doi.org/10.1177/014662169001400305>
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika monograph supplement.
- Schauberger, G., & Tutz, G. (2016). Detection of differential item functioning in rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology, 69*(1), 80–103. <https://doi.org/10.1111/bmsp.12060>
- Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2021). An R toolbox for score-based measurement invariance tests in irt models. *Behavior Research Methods, 54*(5), 2101–2113. <https://doi.org/10.3758/s13428-021-01689-0>
- Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician, 56*(1), 29–38. <https://doi.org/10.1198/000313002753631330>
- Steyer, R., & Eid, M. (2013). *Messen und testen*. Springer-Verlag.

- 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. <https://doi.org/10.1007/s11336-013-9388-3>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, 81(3), 727–750. <https://doi.org/10.1007/s11336-015-9488-3>
- Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, 6(883), 1064. <https://doi.org/10.3389/fpsyg.2015.01064>
- Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement*, 70(6), 941–952. <https://doi.org/10.1177/0013164410379326>
- Zeileis, A. (2006a). Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics and Data Analysis*, 50(11), 2987–3008. <https://doi.org/10.1016/j.csda.2005.07.001>
- Zeileis, A. (2006b). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9), 1–16. <https://doi.org/10.18637/jss.v016.i09>
- Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4), 488–508. <https://doi.org/10.1111/j.1467-9574.2007.00371.x>
- Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational & Graphical Statistics*, 17(2), 492–514. <https://doi.org/10.1198/106186008x319331>
- Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2), 1–38. <https://doi.org/10.18637/jss.v007.i02>

Paper II - Supplementary Material:

Classe, F., & Kern, C. (2024). Detecting differential item functioning in multidimensional graded response models with recursive partitioning. *Applied Psychological Measurement*, 48(3, Suppl.), 83-103. <https://doi.org/10.1177/01466216241238743>

Supplementary Material

November 13, 2023

S1 Additional Tables and Figures

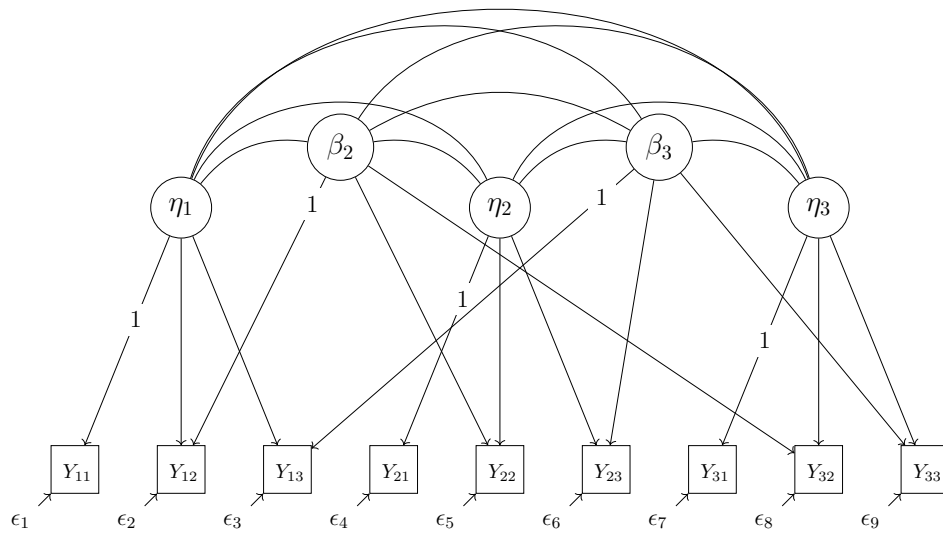


Figure 1: PIEG model for three time points t and three items i . One latent state variable η_t is assumed for each time point. Item 1 serves as reference item so that β_2 and β_3 are the only latent item effect variables in the model.

Table 1: Input variances and covariances for simulation 1 (R_1 to R_4) and simulation 2 (R_1 and R_2).

Parameter	Subgroup			
	R_1	R_2	R_3	R_4
$Var(\eta_1)$	0.27	0.37	0.51	0.47
$Var(\eta_2)$	0.27	0.21	0.30	0.28
$Var(\eta_3)$	0.27	0.34	0.42	0.24
$Var(\beta_2)$	0.34	0.49	0.44	0.39
$Var(\beta_3)$	0.22	0.31	0.39	0.25
$Cov(\eta_1, \eta_2)$	0.10	-0.05	0.15	0.26
$Cov(\eta_1, \eta_3)$	0.06	-0.03	0.11	-0.14
$Cov(\eta_2, \eta_3)$	-0.08	0.12	0.11	0.04
$Cov(\eta_1, \beta_2)$	-0.06	-0.09	0.03	-0.33
$Cov(\eta_1, \beta_3)$	0.07	0.09	0.10	-0.17
$Cov(\eta_2, \beta_2)$	-0.21	-0.05	0.03	-0.10
$Cov(\eta_2, \beta_3)$	-0.12	-0.03	0.14	-0.10
$Cov(\eta_3, \beta_2)$	0.06	0.21	-0.29	0.12
$Cov(\eta_3, \beta_3)$	0.05	0.12	-0.17	0.04
$Cov(\beta_2, \beta_3)$	0.09	0.19	0.06	-0.01

Table 2: Input threshold parameters for simulation 1 (R_1 to R_4) and simulation 2 (R_1 and R_2).

Parameter		Subgroup			
		R_1	R_2	R_2	R_2
Y_{11}	κ_{111}	-0.57	-1.16	-1.21	0.19
	κ_{112}	-0.14	-0.53	-0.58	0.81
	κ_{113}	0.27	-0.06	0.02	1.30
	κ_{114}	0.73	0.42	0.69	1.89
Y_{12}	κ_{121}	-1.18	-0.22	-0.96	-0.01
	κ_{122}	-0.47	0.49	0.02	0.39
	κ_{123}	0.07	1.16	0.86	0.72
	κ_{124}	0.67	1.88	1.72	1.14
Y_{13}	κ_{131}	-0.65	-1.43	-1.83	-1.23
	κ_{132}	0.14	-0.49	-0.81	-0.68
	κ_{133}	0.74	0.24	0.01	-0.19
	κ_{134}	1.51	1.00	0.96	0.33
Y_{21}	κ_{211}	-2.02	0.20	-0.19	0.16
	κ_{212}	-1.56	0.58	0.33	0.61
	κ_{213}	-1.14	0.94	0.74	0.97
	κ_{214}	-0.69	1.36	1.24	1.46
Y_{22}	κ_{221}	-0.61	-0.47	-2.25	0.19
	κ_{222}	-0.23	0.14	-1.49	0.75
	κ_{223}	0.12	0.80	-0.81	1.33
	κ_{224}	0.58	1.42	-0.02	1.88
Y_{23}	κ_{231}	-2.29	-1.83	0.32	-1.34
	κ_{232}	-1.80	-1.26	1.15	-0.76
	κ_{233}	-1.39	-0.72	1.93	-0.35
	κ_{234}	-0.96	-0.12	2.83	0.19
Y_{31}	κ_{311}	-0.46	-0.51	-0.93	0.42
	κ_{312}	0.01	-0.01	-0.39	0.81
	κ_{313}	0.42	0.43	0.13	1.20
	κ_{314}	0.84	0.96	0.68	1.61
Y_{32}	κ_{321}	-1.45	-1.73	-0.73	-1.34
	κ_{322}	-0.67	-0.75	-0.26	-0.59
	κ_{323}	-0.05	0.07	0.12	0.15
	κ_{324}	0.72	1.14	0.58	1.02
Y_{33}	κ_{331}	-0.67	-1.09	-1.58	-1.15
	κ_{332}	0.04	-0.21	-1.06	-0.44
	κ_{333}	0.67	0.51	-0.49	0.16
	κ_{334}	1.33	1.32	0.14	0.81

Algorithm 1: Naive **semtree** for MGR models

Initialization: Assign data to root node
Parameters: minimum sample size in terminal node, p-value threshold

- 1 Estimate model parameters in θ for the sample in the current node using the ML estimator (template model);
- 2 Compute augmented models for all possible split points for all partitioning variables;
- 3 Compute log-likelihood ratio of all augmented models against template model;
- 4 Set optimal split point for every partitioning variable;
- 5 Perform LR test for every partitioning variable;
- 6 **if** *minimum p-value exceeds threshold OR min node size reached* **then**
- 7 end partitioning;
- 8 **else**
- 9 select partitioning variable with lowest p-value in LR test;
- 10 split node into two subnodes at optimal split point;
- 11 **for** *each node of current tree* **do**
- 12 continue partitioning process;
- 13 **end**
- 14 **end**
- 15 **for** *each terminal node* **do**
- 16 re-fit models using WLS estimator;
- 17 **end**

Algorithm 2: partykit for MGR models

Initialization: Assign data to root node

Parameters: minimum sample size in terminal node, p-value threshold

```

1 Estimate model parameters in  $\theta$  for the current node using ML estimation;
2 Assess item parameter instability though generalized M-fluctuation test with
  respect to each covariate  $Z_1, \dots, Z_R$ ;
3 if minimum p-value exceeds threshold OR min node size reached then
4   | end partitioning;
5 else
6   | detect covariate  $Z_{r^*}$  with the strongest instability;
7   | select the unique value as split point that maximizes the sum of the
   | objective functions of the two segmentations;
8   | split node into two subnodes at split point;
9   | for each node of current tree do
10  |   | continue partitioning process;
11  | end
12 end
13 for each terminal node do
14  | re-fit models using WLS estimator;
15 end

```

Algorithm 3: Score-guided **semtree** for MGR models

Initialization: Assign data to root node**Parameters:** minimum sample size in terminal node, p-value threshold

```

1 Estimate model parameters in  $\theta$  for the current node using ML estimation;
2 Assess item parameter instability through generalized M-fluctuation test with
  respect to each covariate  $Z_1, \dots, Z_R$ ;
3 if minimum p-value exceeds threshold OR min node size reached then
4   | end partitioning;
5 else
6   | detect covariate  $Z_{r^*}$  with the strongest instability;
7   | select the unique value as split point that maximizes the score-based test
   | statistic;
8   | split node into two subnodes at split point;
9   | for each node of current tree do
10  |   | continue partitioning process;
11  | end
12 end
13 for each terminal node do
14  | re-fit models using WLS estimator;
15 end
```

Algorithm 4: Recursive partitioning forest for MGR models

Parameters: minimum sample size in terminal node, M-fluctuation test p-value cutoff, number of trees B , partitioning variable subset size, χ^2 -test p-value or RMSEA cutoff

```

1 for  $b = 1$  to  $B$  do
2   | Grow recursive partitioning tree using partykit or semtree for MGR with
   | random draws from partitioning variables;
3   | save decision rules and model fit indices for terminal nodes;
4 end
5 Select exclusive subgroups with model fit indices that don't exceed cutoff;
```

S2 Performance of the Generalized M-fluctuation Test with Ordinal Data

We generate multiple samples to test the performance of the generalized M-fluctuation test with numerical and ordinal data. `partykit` and `semtree` use the results of the generalized M-fluctuation test to decide if the sample should be split into groups. The results of the test also guide the selection of the partitioning variable Z_{r^*} . `semtree` even uses the test statistic associated with the M-fluctuation test to determine the split point. The M-fluctuation test, in turn, draws on the scores of the fitted model. It is thus crucial for `partykit` and for `semtree` that the M-fluctuation test detects parameter stability correctly, even if parameter estimates derived from ordinal data are based on model assumptions of a common CFA model for metric items.

We simulate samples with 250, 500, 750 and 1000 observations with numerical response variables for which a model holds that has the same structure as the outlined PIEG model. Furthermore, corresponding samples with ordinal response variables are simulated for which the PIEG model is true. The parameters are stable for all simulated observations, i.e., there is no DIF. All samples are based on the same input parameters for latent variable variances, latent variable covariances, and mean structure. For the ordinal data set, 36 input threshold parameters are created instead of input intercepts. For both types of response variables (numerical and ordinal) and all four sample sizes (250, 500, 750, 1000), we repeat the sampling process 1000 times to compile the final set of simulated data sets.

Next, the common CFA model for numerical data (with 33 parameters) is fitted using the ML estimator to all data sets and the generalized M-fluctuation test is applied, using one random numerical and one random categorical partitioning variable. We use all six test statistics that are offered in the `semtree` R-package (Arnold et al., 2021) to compute the result of the generalized M-fluctuation test. This includes three test statistics for the numerical covariate and three test statistics for the categorical covariate (see Merkle & Zeileis, 2013; Merkle et al., 2014). With this setup, we can determine how the generalized M-fluctuation test performs when the assumptions of a common CFA model are tested

with data that follows a MGR model, and which test statistics are least susceptible to this type of misspecification.

Results. The simulation results are shown in Table 3. We calculated the percentage across all simulated data sets for which the generalized M-fluctuation test is significant (p-value below 0.05). Because the parameters in the simulated samples are stable, we denote this number as the dropout rate. Notably, none of the generalized M-fluctuation tests for models fitted to simulated numerical response variables performed considerably better than the tests for models fitted to simulated ordinal response variables. Even for small sample sizes, there is no test statistic that yields larger dropout rates for simulated ordinal responses. With numeric covariates, the *CvM* test statistic yields very high dropout rates of around 50% for both ordinal and numeric response variables. However, this is due to the fact that critical values of the *CvM* statistic are not provided for models with more than 25 parameters. The best tests statistics for multivariate latent variable models with a considerable amount of parameters (33 in our simulation) seem to be the *maxLM* and *DM* statistic for numerical covariates (see Merkle & Zeileis, 2013) and the *LM* statistic for categorical covariates (see Merkle et al., 2014).

Table 3: Results of simulation. The proportion of p-values of the generalized M-fluctuation test across simulated data sets that are smaller than 0.05 are shown. The column label ‘numerical’ indicates that numerical response variables were simulated, the label ‘ordinal’ indicates that ordinal response variables were simulated.

n	Numerical covariate					
	<i>DM</i>		<i>CvM</i>		<i>maxLM</i>	
	numerical	ordinal	numerical	ordinal	numerical	ordinal
250	2.3%	2.2%	50.3%	48.2%	1.9%	2.0%
500	3.8%	4.5%	45.3%	46.0%	4.2%	3.1%
750	4.5%	3.6%	46.5%	45.9%	4.6%	3.1%
1000	5.2%	3.5%	49.4%	46.4%	4.6%	4.1%

n	Categorical covariate					
	<i>LM</i>		<i>WDM</i>		<i>maxLM_O</i>	
	numerical	ordinal	numerical	ordinal	numerical	ordinal
250	3.4%	4.0%	6.4%	4.8%	4.8%	4.0%
500	4.6%	4.4%	7.5%	5.0%	5.7%	5.1%
750	5.8%	4.9%	5.6%	4.2%	6.2%	4.2%
1000	5.0%	3.9%	5.9%	6.3%	5.9%	5.3%

S3 Model Based Recursive Partitioning for MGR Models with Full Information Estimation

S3.1 Methodology

For a small number of items with a small number of response categories, there are a multitude of unique possible response patterns for the individual respondent. A response pattern y_r indicates a sequence of k_i , that is

$$y_r = \{k_1, k_2, \dots, k_m\}. \quad (1)$$

For m items with l_i response categories, there are $\prod_{i=1}^m l_i$ different response patterns. A full information approach to estimating the parameters of the MGR model uses all the information contained in these response patterns (Forero & Maydeu-Olivares, 2009). The standard full information estimation method for MGR models is the *marginal maximum likelihood* (MML) method that is usually computed via the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981).

In the MGR model, it is assumed that there is local independence, so that within a group of respondents with the same values for $\boldsymbol{\xi}$, the distributions of item responses are independent of each other (Samejima, 1997). Therefore, the $\boldsymbol{\xi}$ -conditional probability of answering in response pattern y_r is

$$P(\mathbf{Y} = y_r \mid \boldsymbol{\xi}) = \prod_{i=1}^m P(Y_i = k_i \mid \boldsymbol{\xi}), \quad (2)$$

For a random subject sampled from a population with a continuous multivariate ability distribution $g(\boldsymbol{\xi})$, the unconditional probability of answering in response pattern y_r is

$$P(\mathbf{Y} = y_r) = \int_{-\infty}^{\infty} P(Y = y_r \mid \boldsymbol{\xi}) g(\boldsymbol{\xi}) \, d\boldsymbol{\xi}, \quad (3)$$

where \int is a p -dimensional multiple integral. The EM algorithm estimates the probability $P(Y = y_r)$ at every iteration through numerical approximation of the p -dimensional

integral. A disadvantage of this approach is the considerable amount of computing power required. The computational burden increases exponentially with an increasing number of latent variable dimensions (Forero & Maydeu-Olivares, 2009).

The MML method is used to find the best estimates for the item parameters in $\boldsymbol{\vartheta}$ (see Equation 2) that maximize the probability for all respondents to answer in their respective response patterns. Maximizing the log likelihood is equivalent to minimizing the objective function $F_{MML}(\boldsymbol{\vartheta})$ through the EM algorithm. Let n be the sample size and $p_r = n_r/n$ be the relative frequency of occurrence of response pattern y_r . In a sense, the objective function represents the difference between the relative frequency p_r of a certain response pattern y_r and the unconditional probability of answering in that response pattern, that is

$$F_{MML}(\boldsymbol{\vartheta}) = \sum_r p_r [\ln p_r - \ln P(Y = y_r)]. \quad (4)$$

The minimalization algorithm generates successive parameter estimations $\boldsymbol{\vartheta}^{(1)}, \boldsymbol{\vartheta}^{(2)}, \dots$, such that

$$F_{MML}[\boldsymbol{\vartheta}^{(s+1)}] < F_{MML}[\boldsymbol{\vartheta}^{(s)}]. \quad (5)$$

At every other iteration of the minimization algorithm, the gradient of the objective function is used as the search direction (gradient descent approach). This way, the next set of parameter estimates can be chosen so that the objective function $F_{MML}(\boldsymbol{\vartheta})$ decreases (Jöreskog & Moustaki, 2006).

When the objective function F_{MML} is used, the overall model fit can be tested for by using the test statistic $T_{MML} = 2NF_{MML}(\boldsymbol{\vartheta})$. Thus, T_{MML} is $2N$ times the minimum value of the fit function $F_{MML}(\boldsymbol{\vartheta})$. The test statistic T_{MML} is asymptotically χ^2 distributed with degrees of freedom equal to the number of different response patterns minus one minus the number of independent elements of $\boldsymbol{\vartheta}$ (Jöreskog & Moustaki, 2006). It can then be used to test the model against the associated saturated model in which all possible parameters are freely estimated. This way, a test statistic for global model fit is obtained.

Schneider et al. (2021) show that it is possible to perform the generalized M-fluctuation

test for several multidimensional polytomous IRT models that are fitted using MML estimation. We can thus perform **partykit** for MGR models while using MML estimation for growing the decision tree. We call this approach *GRM Tree*. The steps performed by GRM Tree are shown in Algorithm 5.

Algorithm 5: **partykit** for MGR models using MML estimation (GRM Tree)

Initialization: Assign data to root node
Parameters: minimum sample size in terminal node, p-value threshold

- 1 Estimate model parameters in ϑ for the current node using MML estimation;
- 2 Assess item parameter instability through generalized M-fluctuation test with respect to each covariate Z_1, \dots, Z_R ;
- 3 **if** *minimum p-value exceeds threshold OR min node size reached* **then**
- 4 | end partitioning;
- 5 **else**
- 6 | detect covariate Z_{r^*} with the strongest instability;
- 7 | select the unique value as split point that maximizes the sum of the objective functions of the two segmentations;
- 8 | split node into two subnodes at split point;
- 9 **for** *each node of current tree* **do**
- 10 | continue partitioning process;
- 11 **end**
- 12 **end**

S3.2 Simulations

Measurement model. The computational requirements of MML estimation for MGR models are particularly high when the model's latent variables are correlated (Forero & Maydeu-Olivares, 2009). This is the case for the original measurement model defined in Section 3.1. In order to apply GRM Tree to the PIEG model, it must be redefined as a MGR model with orthogonal latent variables. We thus define an orthogonal PIEG model and fix the covariances of the latent variables at 0 and the variances at 1. Discrimination parameters are freely estimated.

As in Section 3.1, we consider three reference latent state variables η_t and two latent item effect variables β_i . The latent variables are derived from three items at three time points resulting in nine five-category ordinal response variables Y_{it} . The cumulative

category response function of the orthogonal PIEG model is

$$P(Y_i \geq k_i | \eta_t, \beta_i) = \Phi(\lambda_{it}\eta_t + \delta_{it}\beta_i - \kappa_{ikt}), \quad (6)$$

$$\forall k = 1, \dots, 4, \forall i = 2, \dots, 3, \forall t = 1, \dots, 3.$$

In this model, there are 36 free threshold parameters (4 for every five-category item) and 15 free discrimination parameters, resulting in 51 free parameters in total. The model structure is shown in Figure 2.

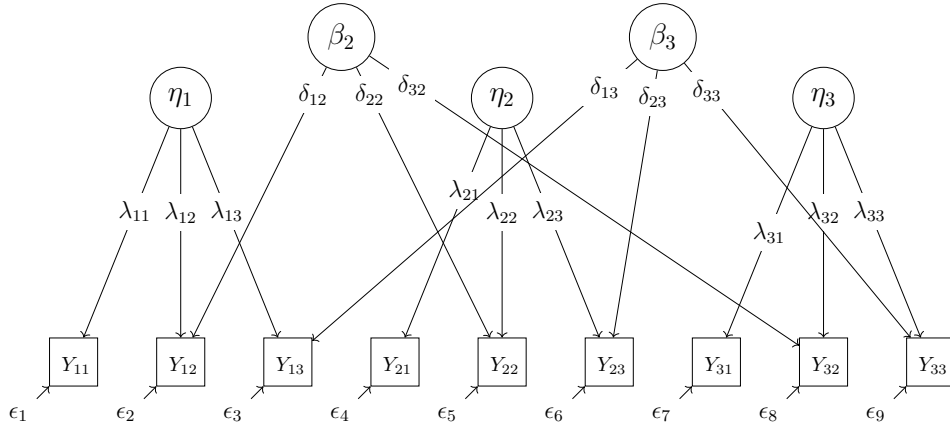


Figure 2: PIEG model with orthogonal latent variables for three time points t and three items i . One latent state variable η_t is assumed for each time point. Item 1 serves as reference item so that β_2 and β_3 are the only latent item effect variables in the model. All latent variable variances are fixed at 1.

Simulation Setup. To test GRM Tree, we simulate a sample with a similar subgroup structure as the sample of simulation 1 in Section 3.2. The only difference with respect to the subgroup structure is that the numeric partitioning variable `num1` is replaced by the categorical partitioning variable `cat3`. The structure of the entire simulated sample

can be broken down by a single decision tree. The simulated subgroups are defined as

$$\begin{aligned} R_1 &:= \{\{\text{cat3} \in \{1, 3\}\} \cap \{\text{cat1} \in \{1, 5\}\}\}, \\ R_2 &:= \{\{\text{cat3} \in \{1, 3\}\} \cap \{\text{cat1} \in \{2, 3, 4\}\}\}, \\ R_3 &:= \{\{\text{cat3} \in \{2, 4, 5\}\} \cap \{\text{cat2} \in \{1, 2\}\}\}, \\ R_4 &:= \{\{\text{cat3} \in \{2, 4, 5\}\} \cap \{\text{cat2} \in \{3, 4\}\}\}. \end{aligned}$$

Each subgroup consists of 250 observations, and thus the full sample size is 1000. The minimum size of the terminal nodes in GRM Tree is set to 100. The outlined changes in comparison to the setup in Section 3.2 (regarding subgroup structure, sample size, and minimum terminal node size) were conducted in order to reduce the computational burden in the application of GRM Tree.

Simulation Results. The results of the GRM Tree application are shown in Figure 3. It becomes apparent that the algorithm does not retrieve the simulated subgroups correctly. The partitioning variable `cat3` is (wrongly) chosen as partitioning variable at the tree's inner nodes 2 and 7. This indicates that the results from the generalized M-fluctuation test may not be as accurate when all threshold parameters are part of the score function (see Equation 10).

An additional disadvantage of GRM Tree, compared to `partykit` or score-based `SEMTree` for MGR models, is the immense computation cost. The computation GRM Tree for the simulated sample described above took 450 minutes (7.5 hours) on a processor with a single core and 170GB RAM. Considering that many limitations were imposed on this particular simulation to keep computation time low, we may conclude that GRM Tree proved to be computationally impractical.

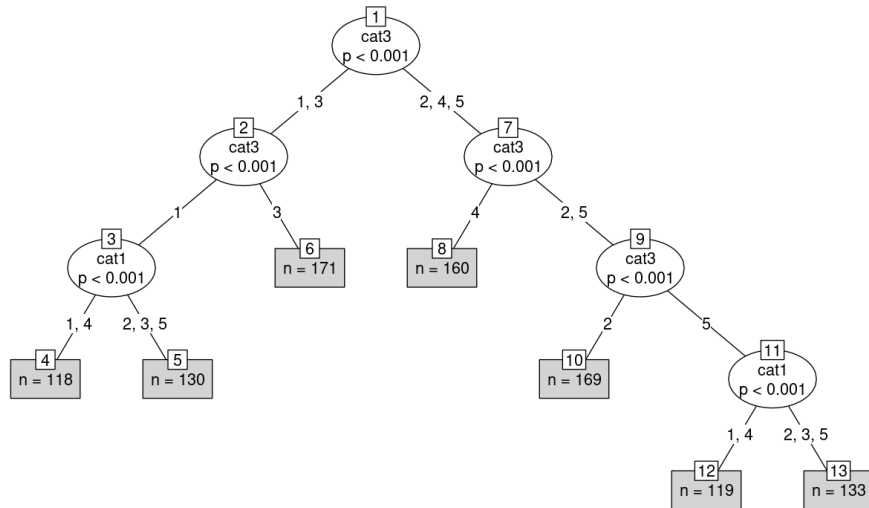


Figure 3: Results of the application of GRM Tree to simulated data.

References

- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology*, *11*, 564403.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological methods*, *14*(3), 275.
- Jöreskog, K. G., & Moustaki, I. (2006). Factor analysis of ordinal variables with full information maximum likelihood. *unpublished report*.
- Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584.
- Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82.
- Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.
- Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2021). An R toolbox for score-based measurement invariance tests in irt models. *Behavior Research Methods*, 1–13.

Paper III:

Classe, F., & Kern, C. (2024). Latent Variable Forests for Latent Variable Score Estimation. *Educational and Psychological Measurement*, 84(6), 1138-1172. <https://doi.org/10.1177/00131644241237502>

*Original Research Article*

Latent Variable Forests for Latent Variable Score Estimation

Educational and Psychological
Measurement

1–35

© The Author(s) 2024



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00131644241237502

journals.sagepub.com/home/epmFranz Classe¹  and Christoph Kern²

Abstract

We develop a *latent variable forest* (LV Forest) algorithm for the estimation of latent variable scores with one or more latent variables. LV Forest estimates unbiased latent variable scores based on *confirmatory factor analysis* (CFA) models with ordinal and/or numerical response variables. Through parametric model restrictions paired with a nonparametric tree-based machine learning approach, LV Forest estimates latent variable scores using models that are unbiased with respect to relevant subgroups in the population. This way, estimated latent variable scores are interpretable with respect to systematic influences of covariates without being biased by these variables. By building a tree ensemble, LV Forest takes parameter heterogeneity in latent variable modeling into account to capture subgroups with both good model fit and stable parameter estimates. We apply LV Forest to simulated data with heterogeneous model parameters as well as to real large-scale survey data. We show that LV Forest improves the accuracy of score estimation if parameter heterogeneity is present.

Keywords

differential item functioning, item response theory, machine learning, confirmatory factor analysis, factor scores

Introduction

The use of psychological questionnaires or tests in research usually involves the assumption of a latent variable measured by the questionnaire items. Latent variable modeling provides a versatile toolkit for measuring such latent traits. There are two

¹Deutsches Jugendinstitut e.V., Munchen, Germany

²Department of Statistics at Ludwig-Maximilians-University of Munich, Germany

Corresponding Author:

Franz Classe, Deutsches Jugendinstitut e.V., Nockherstrasse 2, Munchen 81541, Germany.

Email: classefranz@gmail.com

main areas where latent variables, and particularly latent variable scores, are used: Scaling individuals on a single construct, and estimating latent variable effects in *factor score regression* (FSR) (see Devlieger et al., 2016, 2019) applications.

The first purpose of psychometric latent variable modeling, individual assessment of psychological traits, is a critical component of the cognitive and behavioral sciences (American Psychological Association [APA], 2014). Individual latent variable scores based on observed responses to items of psychological tests are used for psychopathological diagnoses as well as assessment of abilities and personality in occupations and education. However, a major problem is the validity of psychological tests, especially with respect to social minorities (Reynolds et al., 2021). Generally, validity means that a variable measures what it is supposed to measure. Evidence against test validity usually relies on the hypothesis of construct underrepresentation or construct-irrelevant variance, meaning that a variable measures more or less than it should (APA, 2014, p. 12).

Providing evidence for validity usually includes taking into account deviating response behavior in subgroups. Systematic deviations may indicate that the functioning of the scale item differs with regard to certain construct-irrelevant variables. This phenomenon is referred to as measurement noninvariance (Van De Schoot et al., 2015) or differential item functioning (DIF, Bulut & Suh, 2017), and it is present if item parameters differ between subgroups. An item identified as exhibiting DIF is considered biased if the source of variability is irrelevant to the trait being assessed by the test (i.e., construct-irrelevant). However, because any individual characteristic could be defined as construct irrelevant, controlling for item bias may cause real group differences on these variables to be interpreted as bias (see Davies, 2010).

Latent variable scores can be estimated based on item response theory (IRT) (Hartig & Höhler, 2009; Immekus et al., 2019) or confirmatory factor analysis (CFA) (Li, 2016) models (Bhaktha & Lechner, 2021). Practically, construct underrepresentation can be tested for through model fit tests of CFA or IRT models (APA, 2014). Because parameter heterogeneity leads to parameter instability, the assumption of measurement invariance may be investigated via parameter instability tests (Zeileis & Hornik, 2007). However, such a parameter test usually requires a hypothesis about the covariates that negatively affect the parameter stability of a model. In other words, it requires a priori specification of the subgroups for which DIF is suspected.

In recent years, tree-based machine learning methods have been proposed to algorithmically control for DIF in unidimensional IRT models (Komboz et al., 2018; Strobl et al., 2015) through recursive partitioning (Zeileis et al., 2008). Machine learning methods have also been developed to deal with effect heterogeneity in experimental and observational studies (Athey et al., 2019; Athey & Imbens, 2016; Wager & Athey, 2018). As these methods touch on (distinct) aspects of construct validity, they form the ingredients of our approach that focuses on the estimation of unbiased latent variable scores.

We propose *latent variable forest* (LV Forest) for estimating latent variable scores. LV Forest tackles parameter heterogeneity in latent variable models with

ordinal and/or numerical response variables by splitting the original data set to reduce parameter heterogeneity. This way, parameter stability with respect to relevant subgroups is established. LV Forest automatically detects relevant subgroups *within* which parameters do not differ w.r.t. construct-irrelevant variables. LV Forest outputs latent variable score estimates from latent variable models with good model fit estimated separately for each relevant subgroup. However, the estimated latent variable scores may differ *between* these relevant subgroups. This way, latent variable scores may be estimated without true-value group differences being misinterpreted as bias. In psychometric testing, the opportunities and the treatment for examinees as well as the assessment and interpretation of test scores need to be comparable across all individuals and groups in a population. For the stages between assessment and interpretation of test scores this means that construct-irrelevant variables as well as construct underrepresentation have no systematic effect on latent variable scores (Xi, 2010). However, relevant subgroups in which this is the case usually have to be defined a priori. LV Forest overcomes this limitation by automatically creating suggestions for structures of relevant subgroups. Thus, the proposal of this method fills a gap in test methodology. LV Forest is based on the *SEMTree* algorithm to ensure computational efficiency (Arnold et al., 2021; A. M. Brandmaier et al., 2013).

LV Forest comes with a number of favorable properties that allow to take complex heterogeneities in the context of latent variable modeling into account. First, LV Forest uses a data-driven approach for detecting groups that are subject to parameter heterogeneity. The researcher only needs to specify a set of construct-irrelevant partitioning variables for which she suspects differences in model parameters. The partitioning variables are then used to algorithmically search for subgroups with conditionally stable parameters in a decision tree-like fashion. This approach is particularly valuable in situations in which a priori specification of all relevant subgroups based on theoretical assumptions may not be feasible and/or is likely to be insufficient. Second, LV Forest computes multiple decision trees to account for the instability of single trees to small changes in the data to detect relevant subgroups robustly. This approach is inspired by random forests and includes random split selection and bagging to increase tree diversity (Breiman, 2001a). Third, decision trees in LV Forest are heavily pruned. This means that subgroups that are subject to parameter heterogeneity are only selected if the model fits the data and the model parameters are stable with respect to a prespecified vector of covariates.

When applying LV Forest in practice, the algorithm iteratively learns which subgroups in the sample are relevant for estimation and uses these subgroups to repeatedly estimate latent variable scores. Thus, LV Forest can be used for latent variable score estimation especially if the assumed latent variable model does not fit the (full) data and/or includes parameter estimates that are unstable with respect to construct-irrelevant covariates. We show that LV Forest estimates accurate scores in complex settings and outperforms naive and single tree approaches in simulations.

In section “Combining Factor Analytic Modeling and Item Response Theory,” we describe the methodological background of this paper and how the ideas of IRT

and Confirmatory Factor Analysis (CFA) can be merged. In section “Parameter Heterogeneity,” the issues of parameter heterogeneity are described and the M-fluctuation test is introduced. In section “Tree-based Machine Learning,” we briefly introduce tree-based machine learning methods and how the algorithmic modeling perspective can be used to account for heterogeneity. Subsequently, our LV Forest approach is described (section “LV Forest”). In sections “Simulation” and “Real Data Application,” simulations as well as an empirical application of LV Forest with survey data are presented. The advantages and limitations of the proposed method are discussed in section “Discussion.”

Latent Variable Modeling and Score Estimation

Stochastic models which specify the relationship between individual responses to items with a limited amount of response categories and an underlying continuous latent variable are consolidated under the term IRT. Note that IRT was originally developed to examine the response process of individuals. *Confirmatory factor analysis* (CFA), however, is commonly used to formulate assumptions about items within a model that is supposed to reflect a common unobservable phenomenon. The adequacy of these assumptions is usually tested for by testing model quality (Bean & Bowen, 2021). However, modern estimation methods merge the two traditions of latent variable modeling so that certain variants of CFAs are equivalent to an IRT model (Kamata & Bauer, 2008; ten Holt et al., 2010). This means that IRT models may be used for scale evaluation, that is, to determine whether a set of items measures a latent variable. The advantage of an IRT approach is that it better maps the response process to ordinal or dichotomous response variables.

Combining Factor Analytic Modeling and Item Response Theory

Usually, in IRT models, a latent variable represents the ability of the respondent. This ability is assumed to underlie the response behavior (Steyer & Eid, 2013). In the following, we refer to this latent variable as η . In the multidimensional GRM (see Immekus et al., 2019; Samejima, 1969), a multidimensional IRT model (MIRT) for graded responses which can cover various model structures, several latent variables are measured by response variables $Y_i \forall i = 1, \dots, m$, with ordered response categories. The latent variables are comprised in the vector $\boldsymbol{\eta}$. This means that the probability of answering in a category *smaller or equal to* a certain ordered category k_i depends on the (multidimensional) distribution of the latent variables. This relationship is described by the *cumulative category response function*, that is the $\boldsymbol{\eta}$ -conditional probability function:

$$P(Y_i \geq k_i | \boldsymbol{\eta}) = \Phi(\boldsymbol{\beta}'_i \boldsymbol{\eta} - \alpha_{ik}). \quad (1)$$

The link function Φ is the distribution function of the standard normal distribution. The *threshold parameter* α_{ik} may be interpreted as the item-category-specific

intercept whereas the *discrimination parameters* β_{ij} , that make up the $p \times 1$ vector $\boldsymbol{\beta}_i$, can be interpreted as the slope parameters of the multidimensional probability function in Equation 1.

It is possible to efficiently estimate MIRT parameters via CFA modeling. This means that assumptions of an MIRT model can be translated into a special CFA model and parameters can then be estimated in a computationally efficient manner that is common in the CFA framework (limited information approach, see Li, 2016). For this, a continuous, normally distributed latent response variable Y_i^* is assumed to underlie each nonnumerical observed response variable/endogenous variable Y_i . The relation between the latent response variable Y_i^* and the (multidimensional) distribution of the latent variables is described by the conditional expectation function:

$$E(Y_i^* | \boldsymbol{\eta}) = \boldsymbol{\beta}_i' \boldsymbol{\eta}. \quad (2)$$

Note that in this model, the *discrimination parameters* β_{ij} are equivalent to the factor loadings in a CFA model. In the factor analytic approach to MIRT modeling, the latent response variable Y_i^* of item i is related to the observed categorical response variable Y_i via a threshold relation, that is

$$Y_i = k_i \text{ if } \alpha_{ik} < Y_i^* < \alpha_{i(k+1)}. \quad (3)$$

Using the factor analytic approach makes it possible to estimate MIRT parameters through *weighed least squares* (WLS) estimation (Muthén, 1984). Note that WLS estimation makes it possible to include numerical and ordinal endogenous variables within one model. For a numerical response variable Y_i , the basic factor analytic model is

$$Y_i = \pi_i + \boldsymbol{\beta}_i' \boldsymbol{\eta} + \epsilon_i, \quad (4)$$

where π_i is the intercept and ϵ_i is the residual variable for item i . The conditional expectation function $E(Y_i | \boldsymbol{\eta})$ is estimated such that the threshold relationship shown in equation 3 is omitted.

For simplicity, we refer to CFA models with continuous and/or categorical variables as well as multidimensional GRMs as *latent variable models* in this paper. In IRT, the location of an individual on a construct and specific item characteristics are the only factors that account for a person's response (Immekus et al., 2019; Reeve & Fayers, 2005). From this point of view, it is usually desirable to determine the level of a person in relation to the construct. When using the limited information approach to parameter estimation of the CFA framework, one has to create scores to represent each individual's placement on the latent variable. These *latent variable scores* are estimated from fitted models and can be used as dependent or independent variables in regression analyses (DiStefano et al., 2009).

The latent variable score estimates in $\hat{\eta}$, however, do not represent a unique solution to the latent variable η . For any single factor η in a model, there is an infinite number of sets of scores that are equally consistent with the model's parameters. A

latent variable score estimate may not even have identical rankings on different sets of factor scores for the same latent variable. Due to this problem, that is referred to as *indeterminacy*, one can regard $\hat{\eta}$ only as an indicator of η that contains measurement error (Bollen, 1989, p. 305). Thus, the degree to which latent scores are interpretable highly depends on the degree of indeterminacy.

The indeterminacy of latent variable scores varies widely across different models, applications and methods for latent variable score estimation. It may depend, for example, on the degree of commonality between latent variables and response variables (Grice, 2001). It is suggested by Grice (2001), to examine the correlational relationship between η and $\hat{\eta}$ (referred to as *validity*) as well as the correlational accuracy among the scores of all latent variables within the model to evaluate the degree of indeterminacy of latent variable scores. This could, for example, be done through simulation studies.

Parameter Heterogeneity

In MIRT models, DIF occurs when an item- or category-specific parameter depends on covariates of the manifest variables (i.e., response variables). Such covariates may take the form of characteristics of the individuals responding to the items. For example, the difficulty of an item may depend on ethnicity, education, or gender. Conditioning on such covariates is equivalent to analyzing separately certain subgroups defined by different values on these covariates. Similarly, in CFA models the structural parameters determining the relation between latent variables and endogenous variables may differ between subgroups. We refer to between-subgroup differences of parameters in both MIRT and CFA models as *parameter heterogeneity*.

Let \mathbf{Z} be the vector of covariates (Z_1, \dots, Z_R) that contribute to parameter heterogeneity. Let R_1, \dots, R_H , be the subgroups for which there is parameter heterogeneity and let the subgroups be defined as subsets of the covariate space over \mathbf{Z} and let the model parameters be different across all subgroups. In this case, the association with a subgroup R_h corresponds to the event $\{\mathbf{Z} = R_h\}$. The model parameters in a subgroup R_h are homogeneous.

Controlling for parameter heterogeneity for ordinal dependent variables in latent variable models can be formalized by assuming η -conditional probability functions of the category k_i on the response variable Y_i given membership to the subgroup R_h , that is

$$P^{\mathbf{Z} = R_h}(Y_i \geq k_i | \eta) = \Phi(\boldsymbol{\beta}'_{ih} \eta - \alpha_{ikh}). \quad (5)$$

Accordingly, for a numeric response variable Y_i , the η -conditional expectation is assumed to depend on membership to the subgroup R_h , that is

$$E^{\mathbf{Z} = R_h}(Y_i | \eta) = \pi_{ih} + \boldsymbol{\beta}'_{ih} \eta. \quad (6)$$

If the latent variables are properly defined, the latent variable vector $\boldsymbol{\eta}$ does not depend on the covariate vector \mathbf{Z} within the subgroups $R_h \forall h = 1, \dots, H$ in which the parameters are homogeneous, only the model parameters do. This shows that parameter heterogeneity is present when the conditional probability of responding to an item (or the conditional expectation of an item) is different for two individuals *with the same ability*, only because of their group membership.

In practice, parameter heterogeneity can be very problematic because the number of relevant covariates may be very large. Also, there is an even greater amount of possible values or value ranges of these covariates for which model parameters may differ. In addition, complex interactions within the covariate vector \mathbf{Z} are possible so that subgroups may only be detected by considering several covariates jointly. If parameter heterogeneity remains undetected, group differences with respect to the latent variables could be misinterpreted (Komboz et al., 2018), meaning they may be due to bias not due to real latent variable score differences.

Systematic parameter instability with regard to a covariate Z_r can be tested with the generalized M-fluctuation test (Zeileis & Hornik, 2007). The test is applicable for latent variable models that were fitted to a data set via maximum likelihood (ML). The null hypothesis of the M-fluctuation test is rejected if the empirical fluctuation during parameter estimation is improbably large. To represent the empirical fluctuation process, the partial derivatives of the individual log-likelihood function $\ln L(\mathbf{y}_j, \hat{\boldsymbol{\theta}})$ are used. For k parameters in the latent variable model, this is given by the score function:

$$\psi(\mathbf{y}_j, \hat{\boldsymbol{\theta}}) = \left(\frac{\partial \ln L(\mathbf{y}_j, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_1}, \dots, \frac{\partial \ln L(\mathbf{y}_j, \hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_k} \right), \forall j = 1, \dots, n. \quad (7)$$

Summing this function across the sample and maximizing the results yields asymptotically equivalent parameter estimates to limited information maximum likelihood estimation in CFA models for metric variables (maximum likelihood estimation, see Lee & Shi, 2021). Thus, in the estimation process, the score function ψ leads to the parameter estimates $\hat{\boldsymbol{\theta}}$ via the condition $\sum_{i=1}^n \psi(\mathbf{y}_i, \hat{\boldsymbol{\theta}}) = 0$. The M-fluctuation test checks for systematic fluctuations of the scores, ordered with regard to a covariate Z_r . If parameter heterogeneity is present, the scores will differ for different subgroups that are defined as subsets of Z_r . Thus, a test statistic is derived from the scaled cumulative sum of the ordered scores and critical values are obtained from simulation (Wang et al., 2014). Given multiple covariates $Z_r \in \mathbf{Z}$, the generalized M-fluctuation test should be applied for all covariates using a Bonferroni-corrected α -level.

Tree-based Machine Learning

In section “Parameter Heterogeneity,” we introduced the problem of parameter heterogeneity in latent variable models. We assume that reducing parameter instability

by conditioning on a set of covariates \mathbf{Z} will lead to several latent variable models with stable parameters. However, we must assume that the relation between the model parameters in a latent variable model and the covariates \mathbf{Z} could be nonlinear and that associations may be complex. Thus, we need a method for which no hypotheses or assumptions about the functional form of parameter heterogeneity need to be prespecified. In other words, we need an exploratory method that is able to resemble the complex nature of parameter heterogeneity in a latent variable model. For this, we draw on tree-based machine learning methodology.

Machine learning models are considered parts of the *algorithmic modeling culture*. As a counterpart to models from the *data modeling culture*, algorithmic models assume that natural mechanisms, which produce data, are unknown. Data models like latent variable models, however, are stochastic models that are supposed to represent how response variables are *truly* associated with latent variables. Most often though, stochastic models are not complex enough to emulate the true nature of the association between latent variables and response variables (Shmueli, 2010).

In contrast, algorithmic models serve the purpose of predicting new or future observations through flexible modeling with minimal assumptions. Algorithmic models need to be flexible enough to approximate the data generating function while also being robust toward changes in the data used to fit the model. This compromise is referred to as the *bias-variance trade-off* (Hastie et al., 2009). Algorithmic models acknowledge the complex and inconceivable ways that nature produces data. They do not need to be fully interpretable, they rather need to provide accurate information (Breiman, 2001b).

Decision trees represent a popular set of nonparametric machine learning methods that are usually used for prediction of an outcome variable. A predictive model (referred to as a *tree*) is built by recursively partitioning the covariate space over \mathbf{Z} into a set of nodes (referred to as *leaves*) in which the outcome is considered homogeneous (Kern et al., 2019).

Score-based structural equation model trees, as presented by Arnold et al. (2021), combine tree-based machine learning with latent variable modeling. The algorithm searches through all partitioning variables to find subgroups that differ with respect to the model parameters. The aim is to find nodes in which the model parameters are considered homogeneous. For this, the generalized M-fluctuation test with respect to any of the partitioning variables is performed at every node of the tree. If there is significant parameter instability, the node is eventually split at a point on the covariate with the greatest instability into two locally optimal segments. The split point is identified as the location on a partitioning variable at which splitting maximizes the respective score-based test statistic (Arnold et al., 2021, p. 8). As a result, the model only needs to be fitted once at each node of the decision tree. Thus, score-based structural equation model trees are computationally efficient methods for parameter heterogeneity reduction. For simplicity we refer to them as *SEMTrees*.

For the purpose of iteratively reducing parameter heterogeneity, it is important not to overfit a decision tree. At first, a minimum sample size within the terminal nodes

(leaves) of the tree must be established so that parameters for latent variable models can be properly estimated for the subsamples in the terminal nodes. Then, only splits that significantly reduce parameter heterogeneity (according to the generalized M-fluctuation test) should be performed, otherwise spurious parameter heterogeneities may be induced for the models in the terminal nodes.

A popular extension to single decision trees is random forests. They are purely predictive methods where the true functional form of the relationship between input and response variables is assumed to be unknown before the procedure is applied and the function approximated by random forest is not directly interpretable. The predictions of a random forest, however, are likely to be more accurate than the predictions of most data models (Fife & D’Onofrio, 2021; Shmueli, 2010). If we acknowledge that nature produces data in complex and inconceivable ways, the approximation through a nonstochastic but accurate function by random forest might be preferable compared with data models.

Random forest methodology can be tailored to serve other purposes. For example, *SEM forests* by A. Brandmaier et al. (2016) can be used for selection of variables that predict differences across individuals w.r.t. parameters in Structural Equation Models (SEMs). The method can also be used for outlier detection and clustering. Another method that extends Breiman’s random forest algorithm is the causal forest approach (see also Athey et al., 2019; Athey & Imbens, 2016; Wager & Athey, 2018) that is used for the estimation of individual treatment effects. Given such tailored extensions, tree-based machine learning methods are being applied more commonly in the social science and survey research context (Buskirk, 2018; Kern et al., 2019).

LV Forest

We develop a tree-based algorithm for latent variable score estimation: LV Forest. The proposed algorithm is outlined in Figure 1. We begin our considerations with the assumption that the parameters of the proposed latent variable model are not equal for all participants in the population. Parameter heterogeneity in the latent variable model may imply unintended influence of construct-irrelevant variables on the relations within the model. Furthermore, we presume that the proposed latent variable model does not fit the data equally well for all subgroups of the population. With the proposed algorithm, we aim to detect subgroups relevant to bias in estimated latent variable scores, and only latent variable models that fulfill conditional independence from construct-irrelevant variables as well as achieve adequate model fit are chosen for latent variable score estimation. This way, we establish both unbiasedness with respect to construct-irrelevant variables in latent variable score estimation and latent variable scores are not estimated with an underrepresented model. We combine the limited-information approach for parameter estimation (section “Combining Factor Analytic Modeling and Item Response Theory”) and the SEMTree algorithm (section “Tree-based Machine Learning”) to efficiently compute an ensemble of decision trees, in which each tree reduces parameter instability. We then prune the resulting

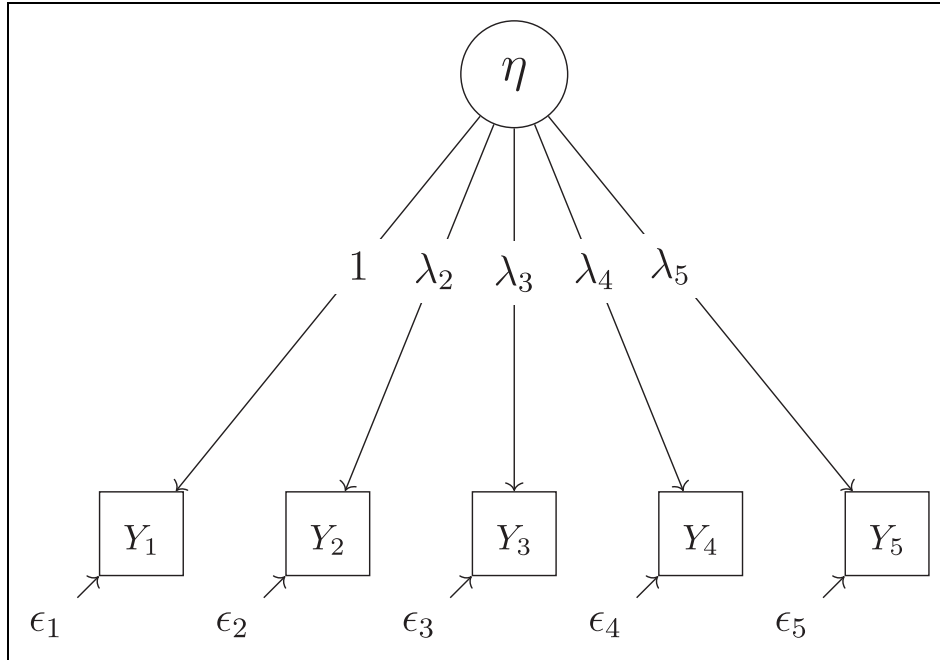


Figure 1. Univariate GRM Model.

Item 1 serves as reference item so that λ_1 is fixed to 1.

trees to detect subgroups in which the model fits the data and the parameter estimations are stable. Note that we do assume that the proposed latent variable model fulfills the criteria for latent variable score determinacy (Grice, 2001, section “Latent Variable Modeling and Score Estimation”).

First, an SEMTree (section “Tree-based Machine Learning”) is grown. Note that for the computation of ψ , which is necessary for the generalized M-fluctuation test (see section “Parameter Heterogeneity”), we need to fit the model with the maximum likelihood (ML) estimator. We then re-assess parameter instability for all construct-irrelevant variables Z_1, \dots, Z_R using the M-fluctuation test. Second, the latent variable model is re-fit in each terminal node of the tree. The parameter and model fit estimates in the terminal nodes of the decision tree are calculated using the distribution free weighted least squares (WLS) estimator. Using only the models in the terminal nodes of the tree that fulfill the criteria for model fit and stability of parameter estimates, latent variable scores are then computed via empirical Bayes estimation. For the computation of these Empirical Bayes Modal (EBM) scores, information about response patterns and model parameters are combined with a prior distribution to obtain a posterior distribution. This method is still appropriate if there are categorical or ordinal response variables and it has performed well in simulations (Bhaktha & Lechner, 2021).

Algorithm 1: LV Forest

Parameters: minimum sample size in terminal node, RMSEA-cutoff, number of trees in ensemble, number of partitioning variables to sample at each node

```

1 do
2   fit model for current sample with ML estimator;
3   randomly sample set of partitioning variables;
4   assess item parameter instability through generalized M-fluctuation test with
   respect to each selected partitioning variable;
5   if parameters are instable AND stopping criteria are not met then
6     detect covariate  $Z_{r^*}$  with the strongest instability;
7     select unique value as split point that maximizes the score-based test statistic;
8     split node into two subnodes at split point;
9     for each node of current tree do
10      continue partitioning process;
11    end
12  end
13 else
14   stop splitting;
15 end
16 for each terminal node do
17   re-assess parameter instability w.r.t. each covariate  $Z_1, \dots, Z_R$  (all partitioning
   variables considered);
18   re-fit model for subgroup in terminal node with the WLS estimator;
19   if minimum RMSEA-cutoff exceeds AND parameters are stable then
20     estimate latent variable scores for subgroup in terminal node;
21   end
22 end
23 while number of iteration < number of trees in ensemble;

```

We might say that the decision trees in the ensemble are heavily “pruned,” leaving only those leaves that are most likely to contain models that are adequate for latent variable score estimation. Specifically, this means that, we exclude terminal nodes for which (a) the proposed model does not fit the data, and (2) the model’s parameters are instable w.r.t. the covariates. For (a), an RMSEA-cutoff value is defined (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003) and all models that exceed this cutoff are excluded. For (b), the generalized M-fluctuation test for parameter instability (Zeileis & Hornik, 2007) is performed. Classe and Kern (2024) show that the performance of the generalized M-fluctuation test for ordinal data is as good as for metric data and thus can be used for ML-based models.

We learn from the machine learning literature that a single decision tree may be vulnerable to small changes in the training data and the set of partitioning variables (Breiman, 2001a). For the most part, this is a consequence of the hierarchical nature of the decision tree (A. Brandmaier et al., 2016; Kern et al., 2019). In addition, if an SEMTree is grown with ordinal data this can lead to inaccuracies in the partitioning

process because the ML estimator is used for the computation of the fitted model scores (section “Parameter Heterogeneity”) at the beginning of the tree growing process. For parameter estimation via maximum likelihood, the dependent variables are assumed to be normally distributed. This assumption rarely holds for ordinal data (Li, 2016). We account for the problem of unstable and potentially inaccurate trees by computing several structurally different decision trees and evaluating the compiled results of this ensemble of trees. We use *random split selection* together with *bootstrap aggregating (bagging)* to ensure that the decision trees in the ensemble are structurally different from each other. For random split selection, random selections of partitioning variables are made. The selection of partitioning variables is redrawn at every node in a decision tree. The researcher can specify the number of partitioning variables that are drawn at each node and thus determine the variability between trees. Bagging means that subsamples are randomly drawn from the full data to grow an individual decision tree. This process is repeated to build an ensemble of multiple trees.

After computing all trees in the ensemble, the estimated latent variable scores are accumulated for each individual over all relevant subgroups in the tree ensemble. This means that across all relevant subgroups found by the algorithm that contain individual i , the scores are averaged.

For the application of LV Forest, the R function `lvforest` was written. In summary, it computes an ensemble of SEMTrees, automatically estimates latent variable scores and tests them for independence of potential construct-irrelevant variables. The R implementation of the proposed method and replication materials for all simulations are provided in the following OSF repository: https://osf.io/gs562/?view_only=c5c715e8e1594445884bb5a1dec27406.

Simulation

Setup

We test the performance of LV Forest with simulated data. We carried out three simulations. For Simulation 1, the data are simulated based on a simple univariate latent GRM model, that is

$$\begin{aligned} P(Y_i \geq k_i | \eta) &= \Phi(\lambda_i \eta - \kappa_{ik}), \\ \forall k &= 1, \dots, 6, \quad \forall i = 1, \dots, 5. \end{aligned} \quad (8)$$

In this model, the variance and mean of the latent variable are estimated. The discrimination parameter pertaining to item 1 (i.e., λ_1) is fixed at 1. Also, the first threshold parameter pertaining to the first item κ_{11} is fixed at 1. We simulated five items with seven response categories each, thus there are six threshold parameters κ_{ik} for each response variable. The model is shown in Figure 1.

The data set used in Simulation 1 consist of 10 model-compliant subsamples ($R_h \forall h = 1, \dots, 10$), each with 500 data points. To simulate model-compliant data,

first, true latent variable scores $\dot{\eta}$ were simulated. Furthermore, values of the conditional probabilities $P(Y_i \geq k_i | \dot{\eta})$ were computed for all categories of all items. On the basis of these conditional probabilities, values for five ordinal response variables with seven categories each were sampled.

In addition, for each of the simulated subsamples, we created one numerical covariate (num_h) ranging from 1 to 200, one ordinal (ord_h) and one categorical (cat_h) covariate with scores on a 5-point scale. These covariates serve as partitioning variables. For each subsample, the range of values on all partitioning variables were fixed, such that

$$\begin{aligned} R_h := & \{num_h \leq 50\} \cap \{cat_h \in \{1, 3, 5\}\} \cap \{ord_h \geq 4\} \cap \\ & \{ord_{is} \leq 3 \mid i \in R_s = \{num_h \leq 50\} \cap \{cat_h \in \{1, 3, 5\}\}\}, \\ & \forall h, s = 1, \dots, 10, s \neq h. \end{aligned}$$

This means that the values on num_h , cat_h , and ord_h are only fixed for those individuals that belong to subgroup R_h except of those individuals i belonging to any other subgroup R_s and happen to fall within the range of values of $num_h \cap cat_h$ to which R_h is fixed. Those individuals are fixed w.r.t. cat_h . This way, given a complete simulated data set, the model-compliant subsamples are recoverable in the terminal nodes of the decision trees of a tree ensemble. It is, however, not possible to recover all model-compliant subsamples in a single decision tree. In simulating the data this way, we want to mimic the complex data structure produced by natural mechanisms.

All input model parameters that were used to simulate the data differ between all subgroups R_h (see Tables 1 and 2). This way, overall parameter instability between the model-compliant partitions of the simulated data set is simulated. The simulation is set up such that the model (see Equation 8) fits the model-compliant subgroups very well (see Table 2).

We apply LV Forest to the simulated data set and compute a forest of 10000 decision trees. All covariates ($num_h, cat_h, ord_h \mid \forall 1, \dots, 10$) are included as partitioning variables. The minimum sample size of the terminal nodes of the trees is set to 200, random split selection is set to 2. We set the model fit cutoff to a RMSEA value of 0.05 to make sure that only the decision rules for well-fitted models are considered when estimating latent variable scores.

Latent variable score estimation accuracy is evaluated by comparing the true simulated latent variable scores $\dot{\eta}$ to latent variable score estimates based on different methods: one fitted model for the entire data set, that is, the *naive* model ($\hat{\eta}_{naive}$), a single SEMTree, that is, one fitted model for each terminal node of the single tree ($\hat{\eta}_{SEMTree}$), LV Forest ($\hat{\eta}_{LVForest}$), and distinct models for the simulated subgroups, that is, 10 separately fitted latent variable models ($\hat{\eta}_{dist.models}$). Note that latent variable score estimation using models fitted on the simulated subgroups individually is not possible in practice as usually the subgroups that are subject to parameter heterogeneity are unknown.

Table I. Input Threshold Parameters for Simulation.

R_1	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	R_6	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\dot{\kappa}_{11}$	0.00	1.54	0.27	-0.45	-1.24	$\dot{\kappa}_{11}$	0.00	1.50	-0.17	-1.56	-0.17
$\dot{\kappa}_{12}$	0.31	1.73	0.79	0.13	-0.74	$\dot{\kappa}_{12}$	0.75	1.94	0.43	-1.15	0.10
$\dot{\kappa}_{13}$	0.90	2.08	0.86	0.42	-0.09	$\dot{\kappa}_{13}$	1.34	2.06	0.92	-0.58	0.70
$\dot{\kappa}_{14}$	1.20	2.40	1.21	0.84	0.35	$\dot{\kappa}_{14}$	2.06	2.34	1.47	0.09	1.66
$\dot{\kappa}_{15}$	1.47	2.51	1.31	1.20	1.13	$\dot{\kappa}_{15}$	2.33	2.47	1.68	0.73	2.07
$\dot{\kappa}_{16}$	1.96	2.76	1.69	1.37	1.55	$\dot{\kappa}_{16}$	2.95	2.70	2.17	1.49	2.78
R_2	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	R_7	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\dot{\kappa}_{11}$	0.00	1.06	-0.43	-1.03	-0.14	$\dot{\kappa}_{11}$	0.00	0.99	1.98	-1.86	2.66
$\dot{\kappa}_{12}$	0.25	1.69	-0.03	-0.40	0.20	$\dot{\kappa}_{12}$	0.83	1.96	2.43	-0.68	3.89
$\dot{\kappa}_{13}$	0.85	1.98	0.11	0.18	0.40	$\dot{\kappa}_{13}$	1.49	2.65	2.52	-0.19	4.22
$\dot{\kappa}_{14}$	1.20	2.24	0.37	0.94	0.98	$\dot{\kappa}_{14}$	1.95	3.73	2.68	0.23	5.14
$\dot{\kappa}_{15}$	1.49	2.51	0.69	1.31	1.29	$\dot{\kappa}_{15}$	2.46	4.33	2.94	0.84	5.48
$\dot{\kappa}_{16}$	2.00	3.25	0.80	1.80	1.73	$\dot{\kappa}_{16}$	2.80	5.34	3.08	1.43	6.04
R_3	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	R_8	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\dot{\kappa}_{11}$	0.00	0.60	0.37	1.94	0.78	$\dot{\kappa}_{11}$	0.00	0.20	-1.69	1.54	0.00
$\dot{\kappa}_{12}$	0.56	1.22	0.78	2.16	1.27	$\dot{\kappa}_{12}$	0.64	1.44	-1.23	1.98	0.71
$\dot{\kappa}_{13}$	1.11	1.80	0.96	2.30	1.69	$\dot{\kappa}_{13}$	1.03	1.63	-0.54	2.12	1.32
$\dot{\kappa}_{14}$	1.74	2.15	1.30	2.54	2.34	$\dot{\kappa}_{14}$	1.45	2.33	0.23	2.37	1.84
$\dot{\kappa}_{15}$	1.91	2.85	1.53	2.70	3.03	$\dot{\kappa}_{15}$	1.87	3.16	0.62	2.47	2.25
$\dot{\kappa}_{16}$	2.59	3.43	1.92	2.95	3.63	$\dot{\kappa}_{16}$	2.45	4.04	1.48	2.64	2.76
R_4	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	R_9	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\dot{\kappa}_{11}$	0.00	-1.02	-0.43	0.33	0.30	$\dot{\kappa}_{11}$	0.00	-0.94	0.99	-0.93	-1.39
$\dot{\kappa}_{12}$	0.33	-0.49	-0.22	0.56	0.71	$\dot{\kappa}_{12}$	0.52	-0.89	1.76	-0.63	-0.74

(continued)

Table I (continued)

R_4	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	R_9	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$\dot{\kappa}_{13}$	0.54	-0.19	0.13	0.84	1.23	$\dot{\kappa}_{13}$	0.78	-0.71	1.93	-0.07	-0.30
$\dot{\kappa}_{14}$	1.07	0.38	0.76	1.17	1.56	$\dot{\kappa}_{14}$	1.16	-0.58	2.60	0.34	0.50
$\dot{\kappa}_{15}$	1.41	0.59	1.08	1.34	1.85	$\dot{\kappa}_{15}$	1.54	-0.41	3.30	0.94	0.77
$\dot{\kappa}_{16}$	1.77	1.19	1.52	1.54	2.17	$\dot{\kappa}_{16}$	2.08	-0.17	3.60	1.69	1.63
R_5	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	R_{10}	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$\dot{\kappa}_{11}$	0.00	-1.85	1.45	-0.13	0.44	$\dot{\kappa}_{11}$	0.00	1.28	1.52	0.51	-0.91
$\dot{\kappa}_{12}$	0.32	-1.24	1.76	0.30	1.00	$\dot{\kappa}_{12}$	0.64	1.51	2.01	1.45	0.22
$\dot{\kappa}_{13}$	0.71	-0.99	1.96	0.83	1.65	$\dot{\kappa}_{13}$	0.96	1.81	2.37	1.77	0.60
$\dot{\kappa}_{14}$	1.14	-0.33	2.12	1.27	2.14	$\dot{\kappa}_{14}$	1.46	2.00	3.01	2.22	1.47
$\dot{\kappa}_{15}$	1.38	0.30	2.22	1.68	2.50	$\dot{\kappa}_{15}$	1.84	2.30	3.40	2.54	2.01
$\dot{\kappa}_{16}$	1.91	1.00	2.49	2.17	3.21	$\dot{\kappa}_{16}$	2.09	2.62	4.17	3.27	2.43

Table 2. Model Fit Indicators and Input Discrimination Parameters of Simulated Data Sets.

χ^2	p-value	RMSEA	$Var(\hat{\eta})$	$E(\hat{\eta})$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
R_1	0.116	0.039	0.66	0.99	fixed to 1	0.67	0.57	0.80	1.47
R_2	0.508	0.000	0.67	0.95		0.80	0.58	1.36	0.89
R_3	0.641	0.000	1.13	1.21		1.20	0.48	0.37	1.17
R_4	0.178	0.032	0.68	0.77		1.12	1.12	0.66	0.97
R_5	0.390	0.009	0.54	0.80		1.51	0.50	1.54	1.33
R_6	0.715	0.000	1.04	1.76		0.39	0.90	1.00	1.19
R_7	0.285	0.022	1.15	1.67		1.46	0.39	1.23	1.21
R_8	0.900	0.000	0.68	1.21		1.57	1.46	0.52	1.36
R_9	0.122	0.038	0.54	1.03		0.40	1.44	1.32	1.51
R_{10}	0.513	0.000	0.70	1.25		0.74	1.30	1.21	1.50

For Simulation 2, we simulated 100 data sets in a simplified form of the procedure described above. We simulated data based on an univariate IRT model with eight items with five categories each (instead of five items with seven categories like in Simulation 1). Furthermore, each of the simulated data sets consist of three model-compliant subsamples for each of which one ordinal (ord_h) and one numerical (num_h) partitioning variable were created. Each of the simulated subgroups consists of 500 data points so the full data set size is $n=1500$ (instead of $n=5000$ in Simulation 1). The range of values on these partitioning variables is fixed, such that

$$R_h := \{num_h \leq 50\} \cap \{ord_h \geq 4\} \cap \{ord_{is} \leq 3 \mid i \in R_s = \{num_h \leq 50\}\}, \\ \forall h, s = 1, \dots, 3, s \neq h.$$

Thus, the model-compliant subsamples are recoverable in the terminal nodes of several decision trees, but not in the terminal nodes of a single decision tree. In Simulation 2, we reduce the number of partitioning variables per simulated data set to six (instead of 30 in Simulation 1).

We apply LV Forest to each of the simulated data sets using the same hyperparameters as in Simulation 1, except that we compute 20 trees per ensemble (instead of 10,000 in Simulation 1). Furthermore, we apply LV Forest to each of the simulated data sets and randomly select 5 out of the 6 relevant partitioning variables to be generally available for the computation of the ensemble. This way, we want to find out how the absence of relevant partitioning variables affects latent variable score estimation with LV Forest. Note that this is not random split selection, but it is a simulation scenario in which not all relevant partitioning variables can be used by the algorithm. We also apply a single SEMTree to each of the simulated data sets, fit a separate model for each of the terminal nodes and estimate latent variable scores using these fitted models.

The accuracy of the latent variable score estimations are evaluated by comparing the true simulated latent variable scores η to five kinds of latent variable score estimates based on: a naive model ($\hat{\eta}_{naive}$), a single decision tree ($\hat{\eta}_{SEMTree}$), an LV Forest with absence of one relevant partitioning variable ($\hat{\eta}_{part.LVForest}$), an LV Forest including all relevant partitioning variables ($\hat{\eta}_{LVForest}$), and three distinct models fitted on each of the subgroups ($\hat{\eta}_{dist.models}$). In addition, we evaluate the *nonconvergence* rate of each of the five estimation methods on each of the simulated data sets. The nonconvergence rate describes the relative frequency of individuals in a sample for which latent variable score estimation was not possible, for example, because the model fitting process did not converge. Note that in an LV Forest, “nonconvergence” of latent variable score estimation for individual i means that i is not part of any relevant subgroup found by the forest and thus scores are not estimated.

For Simulation 3, we simulated one data set in a similar way as in Simulation 1, but now the full data set is simulated using a single set of parameters. We simulated the data to fit a univariate model with five response variables with seven response categories each. We simulate three covariates (num_1, cat_1, ord_1) with random values. We apply LV Forest to this data set and compute a forest of 10 decision trees. All covariates are included as partitioning variables. The hyperparameters are set to the same values as in Simulation 1.

Results

The application of LV Forest with the simulated data resulted in a tree ensemble in which 425 out of 10,000 decision trees included at least one terminal node in which the assumed model fits well and the model parameters are stable w.r.t. the partitioning variables. Overall, there are 439 terminal nodes in which these two conditions apply. These terminal nodes remained for the estimation of latent variable scores for the whole sample. On a 20 core, 170GB RAM server, LV Forest took 5.89 hours (353.5 minutes) of computation time.

The estimation of the single SEMTree with the simulated data of Simulation 1 took 5.01 minutes on a 20-core, 170GB RAM server. The tree structure is shown in Figure 2. It is obvious that the single SEMTree did not reproduce the simulated subgroup structure. The RMSEA values of the models in the 16 terminal node range from 0.02 to 0.18 but only two of the models have a RMSEA lower than 0.05.

To estimate the naive latent variable scores $\hat{\eta}_{naive}$, we fit the model in Equation 8 to the whole data set. As suspected, the naive model does not fit the entire data set well (RMSEA = .090, 95% C.I. = .080 – .101). The RMSEA values of the distinct subgroups in the data set range from 0.00 to 0.04.

The correlation matrix of the four sets of latent variable score estimates and the true simulated latent variable scores are shown in Table 3. We used Spearman’s rank correlation coefficient because the latent variable score estimations may not follow a normal distribution. The accuracy of latent variable score estimations, that is, the correlations with the simulated latent variable scores η , are highlighted. The

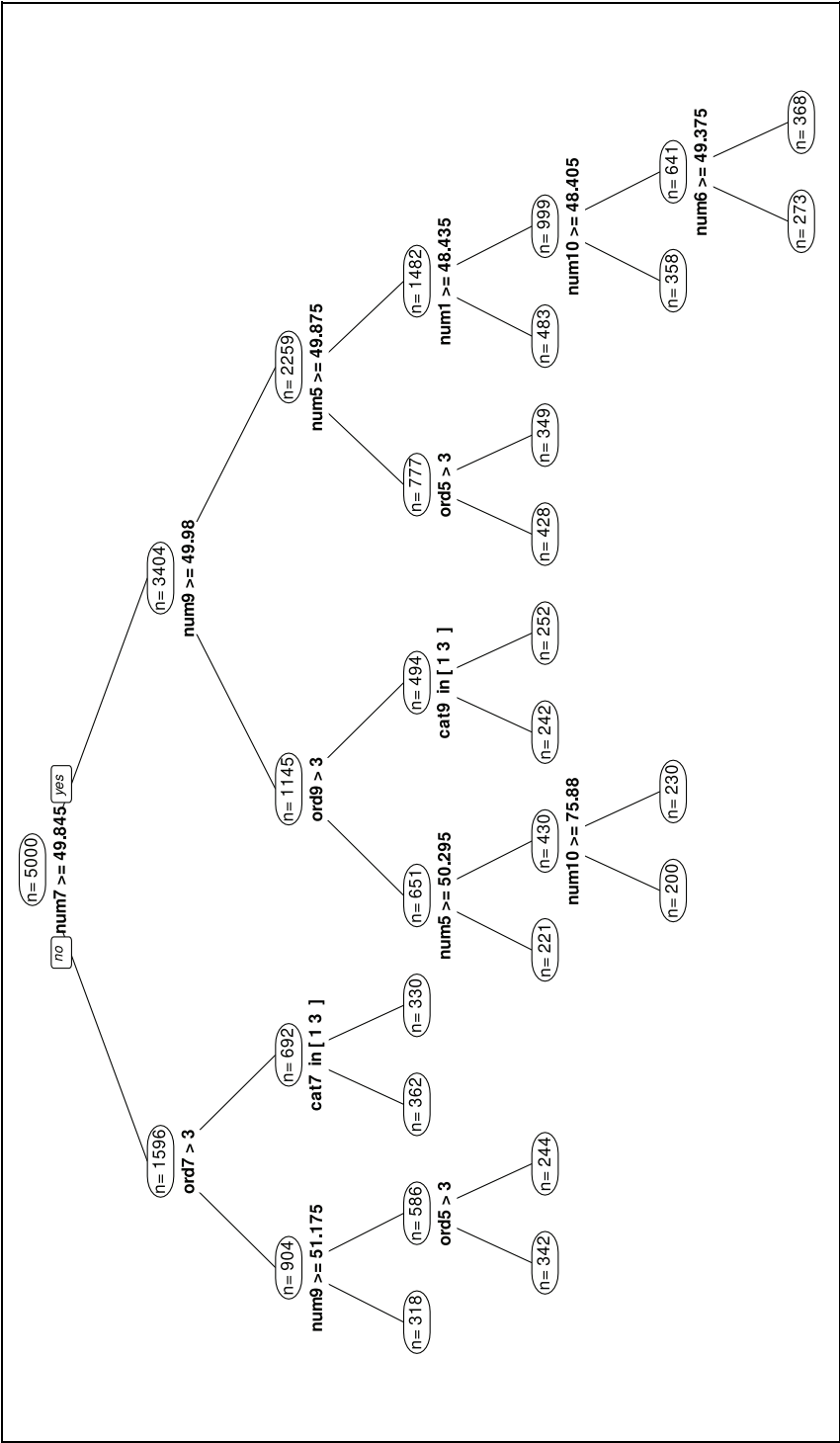


Figure 2. Single SEMTree on Simulated Data.
(a) Evaluation w.r.t. the correlation of simulated latent variable score estimates..
(b) Evaluation w.r.t. nonconvergence of latent variable score estimation for individuals in the sample.

Table 3. Correlations of Estimated Latent Variable Scores From Simulation 1.

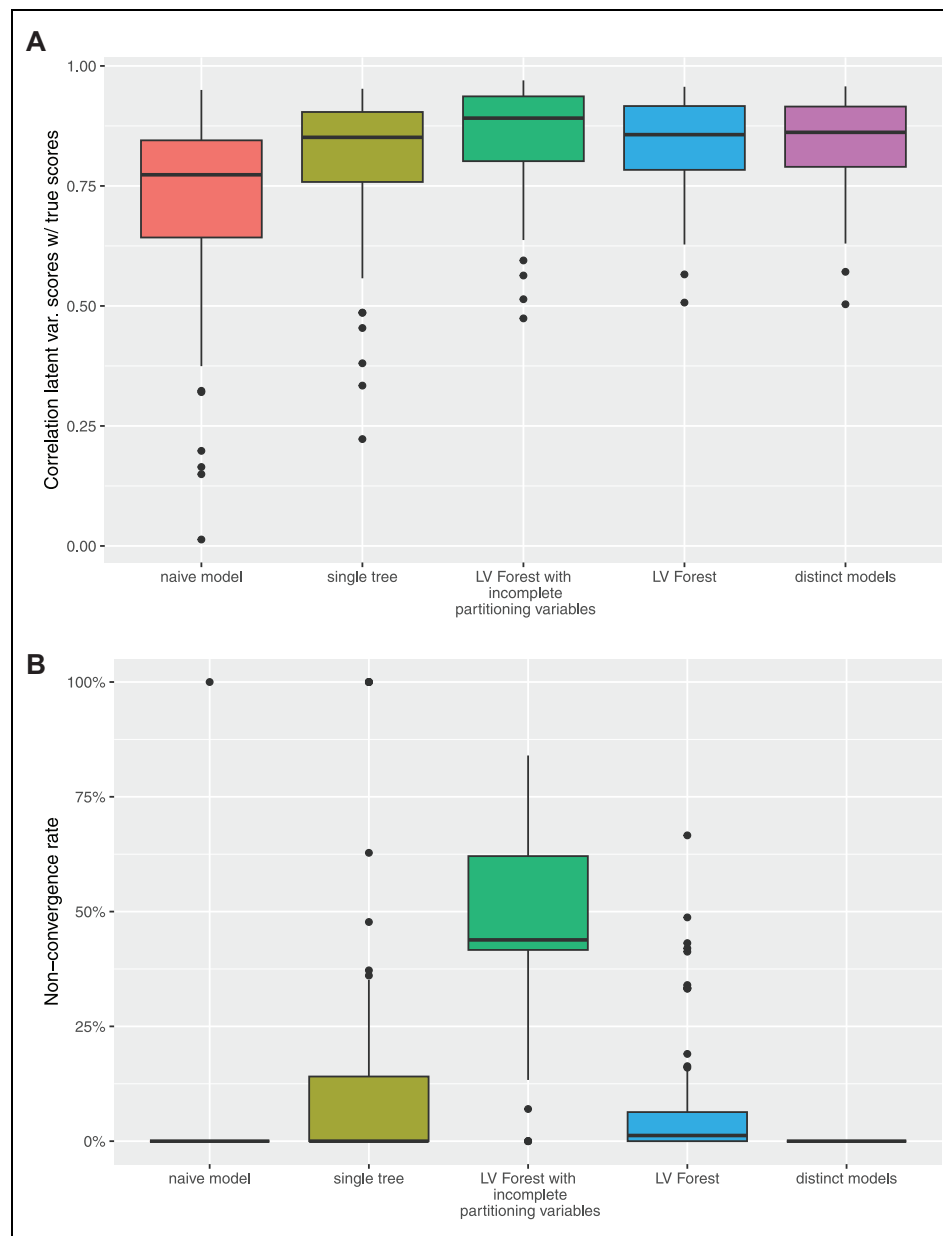
	$\hat{\eta}$	$\hat{\eta}_{dist.models}$	$\hat{\eta}_{LVForest}$	$\hat{\eta}_{SEMTree}$	$\hat{\eta}_{naive}$
$\hat{\eta}$	1.000				
$\hat{\eta}_{dist.models}$	0.830	1.000			
$\hat{\eta}_{LVForest}$	0.816	0.986	1.000		
$\hat{\eta}_{SEMTree}$	0.761	0.922	0.950	1.000	
$\hat{\eta}_{naive}$	0.728	0.899	0.936	0.940	1.000

correlation of $\hat{\eta}$ with $\hat{\eta}_{naive}$ (Row 2) is lower than the correlations of $\hat{\eta}_{SEMTree}$ (row 4) and $\hat{\eta}_{LVForest}$ (Row 3) with $\hat{\eta}$. The correlations of $\hat{\eta}_{dist.models}$ (row 2) and $\hat{\eta}_{LVForest}$ (Row 3) with $\hat{\eta}$ are very similar and noticeably different from the correlations with $\hat{\eta}_{SEMTree}$ and $\hat{\eta}_{naive}$ with $\hat{\eta}$. This suggests that latent variable scores estimated by LV Forest may be more accurate than latent variable scores estimated by a single model fitted to the entire data set when there is substantial parameter heterogeneity in the sample. Also, if there is a complex subgroup structure underlying the data, latent variable scores estimated by LV Forest may be more accurate than those estimated by a single SEMTree. Note, however, that the accuracy of latent variable scores depends on the degree of score indeterminacy (see section “Latent Variable Modeling and Score Estimation”). It is still possible that latent variable scores estimated on the basis of a model that fits the data and has stable parameters are inaccurate.

The results of Simulation 2 in terms of accuracy are shown in Figure 3a and the results in terms of nonconvergence are shown in Figure 3b. The application of the different latent variable score estimation methods on 100 simulated data sets shows that the accuracy of latent variable score estimation based on a naive model ($\hat{\eta}_{naive}$) is, on average, lower than the accuracy of the other methods. In terms of nonconvergence, the naive model did not estimate latent variable scores on one data set. The SEMTree algorithm did not converge on 4 data sets such that no latent variable score estimations were made. For 27 data sets, the nonconvergence rate is larger than 10% as individual models in the terminal nodes did not converge. For 6 data sets, the accuracy of the latent variable scores $\hat{\eta}_{SEMTree}$ is lower than 0.5. The accuracy of $\hat{\eta}_{LVForest}$, and $\hat{\eta}_{dist.models}$ is, on average, higher than the accuracy of $\hat{\eta}_{SEMTree}$. Also, there are no outliers with accuracy lower than 0.5 for $\hat{\eta}_{LVForest}$. Overall, $\hat{\eta}_{LVForest}$ seems to be very similar to $\hat{\eta}_{dist.models}$ in terms of accuracy. However, the analysis of the nonconvergence rates show that there are 17 data sets for which the nonconvergence rate of LV Forest is larger than 10%. Note that if all relevant partitioning variables are included, nonconvergence can be reduced to 0% if more than 20 trees are computed in an ensemble. The analysis of $\hat{\eta}_{part.LVForest}$ indicates that the high accuracy of LV Forest is not affected if not all partitioning variables are available. However, 95 of 100 data sets exhibit a nonconvergence rate of more than 10% and 30 data sets exhibit a nonconvergence rate of over 50%. This indicates that when using LV Forest, the lack of

20

Educational and Psychological Measurement 00(0)

**Figure 3.** Results of Simulation 2.

Latent variable score estimation for 100 data sets based on five different methods. The five methods estimate $\hat{\eta}_{naive}$ (naive model), $\hat{\eta}_{SEMTree}$ (SEMTree), $\hat{\eta}_{LVForest}$ (LV Forest), $\hat{\eta}_{part.LVForest}$ (LV Forest with incomplete partitioning variables), and $\hat{\eta}_{dist.models}$ (distinct models).

Table 4. 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
$i=2$	The conditions of my life are excellent
$i=3$	I am satisfied with my life
$i=4$	So far I have gotten the important things I want in life
$i=5$	If I could live my life over, I would change almost nothing

relevant partitioning variables does not affect the accuracy of the estimated scores, but it does affect the convergence rate and thus the coverage of the scores that are estimated.

Over all 100 samples, the mean computation time of a single SEMTree was 15.17 seconds. The mean computation time of LV Forest was 37.86 seconds. Note that the computations were executed on a 20 core, 170GB RAM server and the trees were computed in parallel.

The results of Simulation 3 show that no splits were performed in any of the 10 LV Forest trees. Thus, in the absence of parameter heterogeneity, the scores estimated by LV Forest are equal to the scores of the naive model.

Real Data Application

We demonstrate the application of LV Forest using data obtained from the LISS (Longitudinal Internet studies for the Social Sciences) panel administered by Centerdata (Tilburg University, The Netherlands). LISS is a comprehensive longitudinal survey conducted annually, encompassing a wide range of topics such as employment, education, income, housing and personality traits (Scherpenzeel, 2018). For this application, we analyze the data from the first survey wave in 2008. In this wave, 8,722 household members were contacted and 6808 individuals responded. We focus on five items from the satisfaction with life scale (Diener et al., 1985) measuring life satisfaction. We excluded all cases that did not respond on all of the five items which leads to a final sample of $n=6626$. The items were rated on a 7-point Likert-type scale. The wording of the items is shown in Table 4.

We analyze the data using the same univariate GRM model structure that Simulation 1 is based on (see Equation 8 and Figure 1). First we fit such a model to the whole data set and refer to it as the naive model. We then we apply LV Forest.

For the application of LV Forest, we choose 11 background variables representing the construct-irrelevant variables for our latent variable model. These variables describe the general characteristics of households and household members that

participate in the LISS panel. They encode characteristics on the individual level (such as gender, age or civil status) as well as on the household level (such as household income, domestic situation or type of dwelling). The variables are shown in Table 5. We apply LV Forest to the data set using these background variables as partitioning variables and compute an ensemble of 1,000 trees. To reduce computation time and to ensure that LV Forest outputs a manageable number of relevant subgroups w.r.t. post hoc analysis, we set the cutoff RMSEA value to .03. Minimum terminal node size is set to 200 and random split selection to 2.

As a sensitivity check, we additionally apply LV Forest with the same data but different partitioning variables. We apply an ensemble with the same hyperparameters as described above while using only the first six variables in Table 5 (*geslacht* to *woonvorm*) as partitioning variables.

To illustrate the conditional independence of the estimated latent variable scores, we perform post hoc tests for independence between the estimated latent variable scores and the construct-irrelevant variables within the subgroups found by LV Forest. For this, we apply a test based on the d-variable Hilbert Schmidt independence criterion (Pfister et al., 2018). With this kernel-based nonparametric test, we test for stochastic independence (instead of e.g., linear independence).

As the estimated latent variable scores are accumulated for each individual over all relevant subgroups, the resulting latent variable scores are not expected to be independent of construct-irrelevant partitioning variables for the full sample. Within the relevant subgroups, however, the latent variable scores are expected to be independent of construct-irrelevant variables. Thus, any overall effects of background variables on latent variable scores imply real differences between the relevant subgroups. To analyze such effects on the latent variable scores, we apply regression models using the 11 background variables as individual predictors. We do this for three different outcome variables: the LV Forest scores using all partitioning variables, the LV Forest scores using only a subset of partitioning variables and the latent variable scores estimated with the naive model.

We fit the naive model using the WLS estimator (see Section “Combining Factor Analytic Modeling and Item Response Theory”). The model does not fit the data well ($RMSEA = .122$, $C.I.(95\%) = .113 - .131$).

In the LV Forest ensemble, 15 trees (1.5% of the ensemble) each generated one terminal node that contained a subsample for which the univariate GRM model fits the data and all parameter estimates are stable w.r.t. all 11 background variables. The model fit indices for all subgroups are shown in Table 6. For these relevant subgroups, latent variable scores were estimated, such that score estimates were available for $n = 2631$ individuals (39.7% of the entire sample). On a 20-core, 170GB RAM server, LV Forest took 32 minutes of computation time. For the LV Forest application with only 6 partitioning variables, score estimates were available for $n = 1310$ individuals. The results of independence tests within the subgroups found by LV Forest using the full set of partitioning variables are shown in Table 7. There is only one construct-irrelevant variable (*partner*) in subgroup R_8 that is likely to

Table 5. Partitioning Variables Used in LV Forest Application With LISS Panel Data.

Variable	Variable names in LISS data	Level	Type	Values	Value labels
Gender	geslacht	individual	categorical	1 2 3 0	Male Female Other No
Household head lives together with partner	partner	household	categorical		
Civil status	burgstat	individual	categorical	1 1 2 3 4 5 1	Yes Married Separated Divorced Widow or widower Never been married Self-owned dwelling
Type of dwelling	woning	household	categorical		
Urban character of place of residence	sted	household	ordered	2 3 4 9 1	Rental dwelling Sub-rented dwelling Cost-free dwelling Unknown (missing) Extremely urban
Domestic situation	woonvorm	household	categorical	2 3 4 5 1 2	Very urban Moderately urban Slightly urban Not urban Single (Un)married co-habitation, without

(continued)

Table 5 (continued)

Variable	Variable names in LISS data	Level	Type	Values	Value labels
Number of household members	aantalhh	household	ordered	3	child(ren) (Un)married co-habitation, with child(ren)
				4	Single, with child(ren)
				5	Other
				1	One person
				2	Two persons
				3	Three persons
				4	Four persons
				5	Five persons
				6	Six persons
				7	Seven persons
Number of living-at-home children	aantalki	household	ordered	8	Eight persons
				9	Nine persons or more
				0	None
				1	One child
				2	Two children
				3	Three children
				4	Four children
				5	Five children
				6	Six children
				7	Seven children
Age category	lftdcat	individual	ordered	8	Eight children
				9	Nine children or more
				1	14 years and younger
				2	15–24 years
				3	25–34 years
				4	35–44 years

(continued)

Table 5 (continued)

Variable	Variable names in LISS data	Level	Type	Values	Value labels
Highest level of education	op1zon	individual	ordered	5	45–54 years
				6	55–64 years
				7	65 years and older
				1	primary school
				2	intermediate secondary education
				3	higher secondary education/ preparatory university education
				4	intermediate vocational education
Net monthly income category	nettocat	individual	ordered	5	higher vocational education
				6	university
				7	other
				0	No income
				1	EUR 500 or less
				2	EUR 501 to EUR 1000
				3	EUR 1001 to EUR 1500
				4	EUR 1501 to EUR 2000
				5	EUR 2001 to EUR 2500
				6	EUR 2501 to EUR 3000
				7	EUR 3001 to EUR 3500
				8	EUR 3501 to EUR 4000
				9	EUR 4001 to EUR 4500
				10	EUR 4501 to EUR 5000
				11	EUR 5001 to EUR 7500
				12	More than EUR 7500
				13	I don't know
				14	I prefer not to say

Table 6. Relevant Subgroups Found in LV Forest LISS Data Application.

Subgroup	Tree	Node	n_r	RMSEA	Decision rule
R_1	108	24	215	0.000	$\{aantalki > 1\} \cap \{ffdcac > 2\} \cap \{sted < 2\} \cap \{geslacht = 1\}$
R_2	119	8	206	0.000	$\{woonvorm \notin \{1, 4\}\} \cap \{ffdcac > 2\} \cap \{aantalhh < 3\} \cap$ $\{nettoac < 3\} \cap \{woning \notin \{2\}\} \cap \{burgstat \notin \{1\}\}$
R_3	161	11	456	0.022	$\{woonvorm \in \{2, 3\}\} \cap \{aantalki < 1\} \cap \{woning \notin \{2\}\} \cap$ $\{ffdcac < 5\} \cap \{ffdcac < 3\}$
R_4	209	10	216	0.000	$\{ffdcac > 2\} \cap \{aantalhh > 1\} \cap \{woning \notin \{-99, 2\}\} \cap$ $\{aantalhh < 3\} \cap \{geslacht = 1\} \cap \{burgstat \notin \{1\}\}$
R_5	241	16	264	0.000	$\{geslacht = 1\} \cap \{oplzon < 4\} \cap \{partner \notin \{0\}\} \cap$ $\{burgstat \notin \{1, 2, 3, 4\}\}$
R_6	255	11	205	0.000	$\{nettoac < 3\} \cap \{woonvorm \notin \{1, 4\}\} \cap \{oplzon > 2\} \cap$ $\{ffdcac > 2\} \cap \{aantalki < 1\} \cap \{burgstat \notin \{1\}\}$
R_7	355	8	288	0.000	$\{nettoac < 3\} \cap \{aantalki < 1\} \cap \{aantalhh > 1\} \cap$ $\{woonvorm \notin \{2\}\} \cap \{burgstat \in \{1\}\}$
R_8	660	5	336	0.000	$\{aantalki < 1\} \cap \{woning \notin \{1\}\} \cap \{burgstat \notin \{3, 4\}\} \cap$ $\{woonvorm \notin \{2, 3\}\}$
R_9	677	21	203	0.011	$\{geslacht = 1\} \cap \{woning \in \{-99, 1\}\} \cap \{aantalhh > 1\} \cap$ $\{aantalki < 1\} \cap \{ffdcac < 3\}$
R_{10}	713	12	320	0.026	$\{burgstat \notin \{1\}\} \cap \{ffdcac > 3\} \cap \{burgstat \notin \{2, 4\}\} \cap$ $\{partner \in \{0\}\} \cap \{woning \notin \{1\}\}$

(continued)

Table 6 (continued)

Subgroup	Tree	Node	n_r	RMSEA	Decision rule
R_{11}	745	10	265	0.028	$\{aantalhh > 1\} \cap \{woning \notin \{-99, 2\}\} \cap \{nettocat < 3\} \cap$ $\{geslacht \neq 1\} \cap \{opizon < 2\} \cap \{woonvorm \notin \{2\}\}$
R_{12}	790	13	214	0.015	$\{woonvorm \in \{2, 3\}\} \cap \{burgstat \in \{1, 2, 3, 4\}\} \cap \{aantalhh < 3\} \cap$ $\{opizon < 4\} \cap \{opizon < 2\} \cap \{sted < 3\} \cap \{geslacht = 1\}$
R_{13}	818	19	285	0.000	$\{woning \in \{1\}\} \cap \{nettocat < 3\} \cap \{geslacht = 1\} \cap$ $\{woonvorm \in \{1, 3, 4\}\} \cap \{burgstat \notin \{1, 3, 4\}\}$
R_{14}	877	31	208	0.000	$\{aantalhh > 1\} \cap \{woning \in \{2\}\} \cap$ $\{burgstat \in \{1, 2, 3, 4\}\} \cap \{sted > 3\}$
R_{15}	934	8	200	0.000	$\{woning \notin \{-99, 2\}\} \cap \{aantalhh > 1\} \cap \{woonvorm \notin \{3\}\} \cap$ $\{lfdcat < 4\} \cap \{geslacht = 1\}$

be stochastically dependent on the estimated latent variable scores of R_8 . This is the case although the parameter estimates of the fitted model used for latent variable score estimations are stable w.r.t. all construct-irrelevant variables. This result may be due to latent variable score indeterminacy. The results of the other tests indicate that parameter stability of well-fitting models w.r.t. construct-irrelevant partitioning variables generally leads to latent variable scores that are independent of construct-irrelevant partitioning variables given the affiliation to a relevant subgroup. We conclude that these relevant subgroups are found by LV Forest.

We analyzed the effect of the background variables on the different latent variable score estimations (naive model vs. LV Forest vs. LV Forest with subset of partitioning variables). The results are shown in Table 8. The regression coefficients for the scores of the LV Forest with all partitioning variables indicate a linear effect of two variables (partnership status and domestic situation). For these same variables, the regression coefficients for the scores of the reduced LV Forest show a significant effect. Also, the Spearman's correlation of the scores of the LV Forest with all partitioning variables and the scores of the reduced LV Forest is 0.99. In contrast, the coefficients for the naive scores additionally show significant effects of four other variables (civil status, age, gender, or urban character of dwelling). This indicates that the effect of partnership status and domestic situation on life satisfaction may not be due to bias. The effect of civil status, age, gender or urban character of dwelling, however, may be due to bias w.r.t. the background variables.

Discussion

In this study, we proposed LV Forest, an algorithmic approach to latent variable score estimation. We focused on a setting in which a naive latent variable model is subject to parameter heterogeneity. In this case, fitting a latent variable model and estimating latent variable scores on the basis of this model can lead to false conclusions. The proposed latent variable model may, however, not violate measurement invariance within subgroups that can be defined by covariates. Since tree-based methods have successfully been applied to account for DIF (Komboz et al., 2018; Strobl et al., 2015), we utilized the algorithmic machine learning perspective for handling complex subgroup structures in the context of latent variable score estimation. Assuming that the latent variable scores of a proposed model are determinate (section "Latent Variable Modeling and Score Estimation"), we argue that scores should only be estimated if the latent variable in the proposed model is not underrepresented and independent from construct-irrelevant variables. Construct-irrelevant variables may have an effect on latent variable scores estimated using LV Forest. However, this effect may not be due to bias but due to real differences w.r.t. the latent variable scores between relevant subgroups. We build on the growing body of research that utilizes techniques from the field of machine learning to flexibilize stochastic models when they are confronted with complex covariate structures.

Table 7. Results of Kernel-Based Independence Test: Dependence of Latent Variable (Life Satisfaction) on Construct-Irrelevant Variables Within Relevant Subgroups.

	Gender	Partner	Marital status	Housing	Urban area	Dom. situation	No. of HH members	No. of children	Age	Education	Net income	n_r
	geslacht	partner	burgstat	woning	sted	woonvorm	aantalhh	aantalki	lftdcat	oplzon	nettoct	
R_1	–	0.43	0.71	0.42	0.15	0.43	0.23	0.36	0.72	0.80	0.62	215
R_2	0.41	0.68	0.30	0.81	0.23	0.35	0.31	0.30	0.77	0.18	0.73	206
R_3	0.77	1.00	0.76	0.47	0.41	0.53	0.59	0.53	0.24	0.84	0.47	456
R_4	–	0.47	0.70	0.79	0.18	0.50	0.33	0.52	0.83	0.68	0.41	216
R_5	–	–	–	0.25	0.77	0.74	0.86	0.88	0.71	0.75	0.77	264
R_6	0.82	0.67	0.14	0.10	0.12	0.35	0.33	0.32	0.79	0.32	0.35	205
R_7	0.44	0.73	1.00	0.52	0.87	0.66	0.86	0.78	0.02	0.53	0.23	288
R_8	0.29	0.00*	0.82	0.46	0.36	0.32	0.34	0.39	0.54	0.30	0.86	336
R_9	–	0.28	0.77	–	0.39	0.26	0.38	0.18	0.25	0.65	0.35	203
R_{10}	0.23	1.00	0.79	0.61	0.44	0.86	0.84	0.67	0.21	0.17	0.55	320
R_{11}	–	0.57	0.20	–	0.72	0.57	0.36	0.21	0.16	0.57	0.33	265
R_{12}	–	–	0.53	0.02	0.40	0.75	0.73	0.75	0.88	0.12	0.49	214
R_{13}	–	0.30	–	–	0.55	0.18	0.50	0.27	0.56	0.44	0.15	285
R_{14}	0.73	0.64	0.78	–	0.43	0.66	0.62	0.63	0.70	0.32	0.63	208
R_{15}	–	0.25	0.50	0.78	0.73	0.24	0.38	0.16	0.52	0.80	0.54	200

Table 8. Regression Coefficients of Covariates on Latent Variable Scores in the Real Data Application.

	Naive model	LV Forest w. all part. vars	LV Forest w. subset of part. vars
Geslacht	0.07*	0.05	0.00
Partner	0.42*	0.14*	0.20*
Burgstat	−0.63*	−0.47	−0.15
Woning	0.54	0.33	−0.12
Sted	0.07*	0.03	−0.03
woonvorm	0.41*	0.11*	0.21*
Aantalhh	0.09	−0.40	−0.41
Aantalki	−0.04	−0.46	−0.48
Lftdcat	0.07*	−0.02	0.06
Opizon	−0.08	−0.13	−0.07
Nettocat	0.19	0.16	0.67
N	6626	2631	1310

The LV Forest with a subset of partitioning variables uses the partitioning variables `geslacht` to `woonvorm`.

In psychological assessment, bias refers to systematically under- or overestimating personality traits or abilities. Especially cultural bias has been a polarizing issue for many years. The controversy lies in the explanations given for the measured systematic differences in traits and abilities between specific subgroups. Are they based on an interaction of genes and environment (i.e., genuinely different ability levels in different groups) or on different cognitive structures requiring different test characteristics, that is, test bias (see Reynolds et al., 2021). According to Bollen (1989), causality, and therefore validity, is only possible if there are no systematic differences in a latent ability or trait with respect to variables outside of the latent variable model. Thus, if systematic differences between groups are not part of the assumed model, they are attributable to test bias. This way, no real differences of the latent variable scores w.r.t. construct-irrelevant variables are interpretable. As virtually all individual characteristics can be such construct-irrelevant variables, this notion is problematic (see, e.g., Davies, 2010). We propose a solution to this problem by proposing a method to estimate latent variables scores whose subgroup differences w.r.t. construct-irrelevant variables are estimable and interpretable.

Latent variable scores estimated using LV Forest are also very useful when it comes to complex SEMs that include measurement paths between latent variables. In these models, *spurious relations* or *suppressor relations* from response variables to latent variables are likely to occur (Bollen, 1989, pp. 51–53). These unmodelled relations distort the other parameters in the model. Therefore, the estimation of effects between two latent variables should rather be performed via FSR (Devlieger et al., 2019) with LV Forest being used for latent variable score estimation.

We applied LV Forest to simulated data to test whether the method is suitable for finding simulated subgroups based on fitting IRT models with stable parameters. The

results show that the method works well for an univariate GRM model. We also show that latent variable score accuracy depends, to some degree, on model fit and parameter stability of a latent variable model. Furthermore, we show that latent variable score estimation via a single SEMTree does not perform as good as LV Forest if the subgroup structure behind the sample cannot be recovered by a single tree. Another advantage of LV Forest is that a 0% convergence rate is very unlikely. However, non-convergence rates are likely to be larger for LV Forest compared with a naive model. However, if there are not many partitioning variables in the data and/or if the data set is not very large, one might prefer using a single SEMTree over LV Forest to estimate latent variable scores.

Furthermore, we applied LV Forest to real data from a large-scale survey. We analyzed five items measuring satisfaction with life and used background variables to recursively partition the sample. As a result, latent variable scores were estimated for 40% of the sample. When the number of partitioning variables was reduced, scores were only estimated for 20% of the sample. This shows that LV Forest may be limited when it comes to exhaustively estimating latent variable scores for the entire sample. In reality, there may always be individuals for which the proposed latent variable model does not apply and relevant partitioning variables are not measured. Our simulations, however, suggest that the accuracy of LV Forest scores is still high, even given considerable nonconvergence. When this is the case, the researcher may increase the RMSEA-cutoff to reduce the nonconvergence rate, but potentially compromise on latent variable score accuracy.

The fact that the estimated latent variable scores were predominantly R_h -conditionally independent from all construct-irrelevant variables in the real data application shows that controlling for DIF w.r.t. construct-irrelevant variables leads to latent variable scores with no systematic effects regarding construct-irrelevant variables *within* relevant subgroups. That is, within these subgroups, all covariance from construct-irrelevant variables is interpreted as bias. *Between* those subgroups, there may be systematic differences regarding construct-irrelevant variables. These differences can be smaller when fewer partitioning variables and a stricter RMSEA-cutoff are used, that is, when fewer relevant subgroups are found. LV Forest estimates latent variable scores that can be interpreted w.r.t. systematic effects of construct-irrelevant variables without inducing bias.

Comparison to Related Methods

Another tree-based machine learning approach to identify and account for parameter heterogeneity, which is also applicable to different types of latent variable models, is called *Model Based Recursive Partitioning* (MOB) (Zeileis et al., 2008). MOB is designed to grow single trees that avoid overfitting and bias. The MOB algorithm applies the M-fluctuation test (see section “Combining Factor Analytic Modeling and Item Response Theory”) at every node of the tree. Splitting is only performed if parameter heterogeneity is significant with regard to at least one covariate. The

covariate with the lowest p-value is selected for splitting. However, splitting is performed in such a way that the sum of the log-likelihood of the two resulting models is maximized. Thus, as many models have to be fit as there are possible split points on a variable chosen for splitting. This is computationally more expensive than score-based SEMTree.

Limitations

In the LV Forest framework, we focus on latent variable models that may be subject to parameter heterogeneity. Simultaneously, we claim that we only use models with R_h -conditionally unbiased measurement paths for latent variable score estimation. For this, we test for parameter homogeneity using the M-based fluctuation test. However, it is controversial to rely on this test too much if ordinal response variables are used because ML estimation is necessary for the computation of the test. If categorical response variables are used, the assumption of normality of the response variables may be violated. However, ordinal response variables are relevant for many applications and (Classe & Kern, 2024) showed that the results of the M-fluctuation test can be reliable for ordinal response variables.

Practical limitations stem from the fact that it is impossible in many cases to measure all construct-irrelevant variables that may confound the measurement paths of a presumed model. The scores estimated by LV Forest should be interpreted with regard to the fact that there may still be potential construct-irrelevant variables that were not collected in the study. The simulation showed that the absence of relevant partitioning variables may lead to nonconvergence score estimation for individuals in the sample. Thus, if not all relevant partitioning variables are measured, it may not be possible to estimate unbiased scores for every individual in the sample. We additionally note that large sample sizes are needed for LV Forest to be efficient. The sample needs to be large enough that sample sizes in terminal nodes in complex trees are sufficient to estimate model parameters, as well as to accurately perform M-fluctuation tests. The simulation also showed that if the subgroup structure of a sample is complex, many trees and therefore long computation times are needed. In practice, if an assumed model does not fit the data and/or has unstable parameters it may be viable for the researcher to adjust the model assumptions before turning to LV Forest. We also acknowledge that LV Forest does not return an inherently interpretable model function. Like random forests, LV Forest allows to model highly complex structures of subgroups. However, a direct interpretation of the composition of these subgroups would lead to results that are unlikely to be generally applicable. Our proposed method therefore explicitly focuses on the estimation of latent variable scores.


Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

References

- American Psychological Association. (2014). *Standards for psychological and educational testing*. American Educational Research Association.
- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology, 11*, Article 564403.
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353–7360.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148–1178.
- 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, 6*, 597–618.
- Bhaktha, N., & Lechner, C. M. (2021). To score or not to score? a simulation study on the performance of test scores, plausible values, and SEM, in regression with socio-emotional skill or personality scales as predictors. *Frontiers in Psychology, 12*, Article 679481.
- Bollen, K. A. (1989). *Structural equations with latent variables (Vol. 210)*. John Wiley.
- Brandmaier, A., Prindle, J., Mcardle, J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*, 566–582.
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86.
- Breiman, L. (2001a). Random forests. *Machine Learning, 45*(1), 5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.
- 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*, Article 51.
- Buskirk, T. D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Survey Practice, 11*(1), 1–13.
- Classe, F., & Kern, C. (2024). Detecting differential item functioning in multidimensional graded response models with recursive partitioning. *Applied Psychological Measurement*. <https://doi-org.emedien.ub.uni-muenchen.de/10.1177/01466216241238743>
- Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171–176.
- Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement, 76*(5), 741–770.

- Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, 79(6), 1017–1037.
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, 14(1), 20.
- Fife, D., & D'Onofrio, J. (2021). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, 55, 2447–2466.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430–450.
- Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, 35(2–3), 57–63.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- 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, Article 45. <https://doi.org/10.3389/educ.2019.00045>
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 136–153.
- Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73–93.
- Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, 78(1), 128–166.
- Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods*, 26, 466–485.
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.
- Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 5–31.
- Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trials: Methods of Practice*, 2, 55–73.
- Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Mastering modern psychological testing* (pp. 573–613). Springer.

- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34, 1–97.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Scherpenzeel, A. C. (2018). “True” longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester & L. Kaczmarek (Eds.), *Social and behavioral research and the internet* (pp. 77–104). Routledge.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- Steyer, R., & Eid, M. (2013). *Messen und testen* [Measuring and Testing]. Springer-Verlag.
- 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.
- 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, 272–297.
- Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, 6, Article 1064.
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
- Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: Use in practice. *Frontiers in Psychology*, 5, Article 438.
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147–170.
- 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.

Paper IV:

Since Paper IV was still being reviewed by the journal at the time of submission of the dissertation, the revised version of the paper originally submitted to the journal is included here. The paper has since been published.

Classe, F., Debelak, R., & Kern, C. (2025). Score-based tests for parameter instability in ordinal factor models. *British Journal of Mathematical and Statistical Psychology*. <https://doi.org/10.1111/bmsp.12392>

Running head: SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

1

Score-Based Tests for Parameter Instability in Ordinal Factor Models

Franz Classe^{1,*}, Christoph Kern², and Rudolf Debelak³

¹Deutsches Jugendinstitut e.V., Germany

²Department of Statistics, LMU Munich, Germany

³Department of Psychology, University of Zurich, Switzerland

*Email: classefranz@gmail.com

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

2

Abstract

We present a novel approach for computing model scores for ordinal factor models, i.e. 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: Ordinal Factor Analysis, Multidimensional Item Response Theory, Parameter Instability, Score Test

Score-Based Tests for Parameter Instability in Ordinal Factor Models

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, i.e. 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 analyze 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; ten Holt et al., 2010; Immekus et al., 2019). 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 modeling.

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) (Muraki & Carlson, 1995; Forero & Maydeu-Olivares, 2009) 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 (Yavuz & Hambleton, 2017; Cai, 2010). 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 (Shi et al., 2020; Maydeu-Olivares et al., 2011).

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, i.e., the partial derivative of the casewise contributions to the objective function w.r.t. 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 (Strobl et al., 2015; Komboz et al., 2018), 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 (T. Wang et al., 2018). It is, however, currently not applicable to the Graded Response Model via ordinal factor analysis (i.e., LI estimation via polychorics).

We propose a method to efficiently approximate individual model scores, i.e. 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.

Methodological Background

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 $i=1, \dots, n$, respondents choose from a range of ordered response categories $k_j=1, \dots, l_j$, for items $j=1, \dots, p$. For simplicity, we assume that all items have the same number of categories, such that $k_j=k \forall j=1, \dots, p$. In a multidimensional GRM, $\boldsymbol{\xi}$ is a $m \times 1$ vector containing all latent variables $\xi_q \forall q=1, \dots, m$. 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

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

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

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

7

well as all freely estimated latent variable variances and covariances, such that

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

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

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, e.g., 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 θ 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 ξ) solely due to their different values on Z . Practically, undetected DIF may lead to a misinterpretation of group differences concerning latent variables (T. Wang 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, Pargent, Deffner, & Goretzko, 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

$$Y_i \perp\!\!\!\perp \mathbf{Z} \mid \boldsymbol{\xi}_i, \forall i = 1, \dots, p, \quad (3)$$

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 a 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.

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 w.r.t. 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 (Schneider et al., 2022; Debelak & Strobl, 2019a). 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 (Debelak & Strobl, 2019a; Debelak et al., 2022; Bock & Zimowski, 1997). 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 $\hat{\theta}$ is defined as the solution to the equation

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

where ψ is a $1 \times \|\theta\|$ matrix. Note that $\|\cdot\|$ denotes vector length.

The function ψ 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

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

10

(2004, p.160–164), the marginal likelihood L of the observed data is

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

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

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

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, $P(\mathbf{y}_i | \boldsymbol{\xi}_i, \theta)$ follows from Equation 1. The derivative of the log likelihood with respect to some parameter x is

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

where $\frac{\partial P(\mathbf{y}_i)}{\partial x}$ differs for each parameter x in θ .¹ The individual contributions to the first derivative of the log likelihood with respect to the M-estimator $\hat{\theta}$ 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 $\psi(\mathbf{y}_i, \hat{\theta})$ are computed for all individuals 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 B3). The scores are then cumulated according to the order of the covariate of interest Z to compute the cumulative score

¹ In Debelak and Strobl (2019b), other examples of ψ can be found.

process

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

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 B3). The sum process is scaled by an estimate $\hat{\mathbf{B}}$ for the covariance matrix $cov(\psi(\mathbf{Y}, \hat{\theta}))$ to decorrelate the scores so that the score processes for all parameter estimates in $\hat{\theta}$ are independent from each other. By analyzing 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 $\hat{\theta}$. 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 (Debelak & Strobl, 2019a; Baker & Kim, 2004). 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 Strobl et al., 2015; Komboz et al., 2018). 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.

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 (Metropolis et al., 1953; Hastings, 1970), 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.

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, a continuous latent response variable Y_j^* is assumed to underlie each observed ordered variable Y_j for item j (Takane & De Leeuw, 1987). This latent response variable is related to the observed ordered variable via a threshold relation, that is

$$Y_j = k \text{ if } \tau_{j(k-1)} < y_j^* \leq \tau_{jk}, \quad (9)$$

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, i.e.

$$P(Y_j > k) = P(Y_j^* > \tau_{jk}) = \Phi(-\tau_{jk}). \quad (10)$$

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 9, is performed as a first step in the model fitting process.

Furthermore, in this phase, bivariate polychoric correlations ρ_{js} are computed for all $j, s = 1, \dots, p$ 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

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

14

θ are estimated through minimization of the objective function

$$F_{OFA}(\theta) = [\hat{\mathbf{\kappa}} - \mathbf{\kappa}(\theta)]' \mathbf{W}^{-1} [\hat{\mathbf{\kappa}} - \mathbf{\kappa}(\theta)], \quad (11)$$

where $\hat{\mathbf{\kappa}}$ and $\mathbf{\kappa}(\theta)$ 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 $\hat{\mathbf{\kappa}}$. The WLS estimator may produce unstable results for small sample sizes and large models (Flora & Curran, 2004; C. Wang et al., 2018; Garnier-Villarreal et al., 2021). 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.

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 $\mathbf{\kappa}$ contains the saturated model parameters $(\boldsymbol{\tau}, \boldsymbol{\sigma}^*)'$, where $\boldsymbol{\tau}$ is the vector of threshold parameters, and $\boldsymbol{\sigma}^* = \text{vech}[Cov(\mathbf{Y}^*)]$ contains the vectorized non-redundant elements of the model implied covariance matrix. The size of $\mathbf{\kappa}$ 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 $\mathbf{\kappa}$ with respect to θ is

$$\Delta = \frac{\partial \mathbf{\kappa}(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \boldsymbol{\tau}(\theta)}{\partial \theta} \\ \frac{\partial \boldsymbol{\sigma}^*(\theta)}{\partial \theta} \end{pmatrix}. \quad (12)$$

We apply the chain rule to get the first derivative of the objective function with respect

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

15

to θ , that is the $1 \times \|\theta\|$ matrix

$$\frac{\partial F_{OFA}(\theta)}{\partial \theta} = \frac{\partial F_{OFA}(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \frac{\partial \boldsymbol{\kappa}(\theta)}{\partial \theta} = -2[\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\theta)]' \mathbf{W}^{-1} \Delta. \quad (13)$$

Note that Equation 13 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 11 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 the Technical Appendix, 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

$$\sum_{i=1}^n \Delta' \mathbf{W}_{GEE}^{-1} \mathbf{e}_i = \mathbf{0}, \quad (14)$$

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 to 14 in the Technical Appendix). The first order empirical moments in the data are the indicator variables, i.e.,

$$1_{y_i > k} = \begin{cases} 1, & \text{if } y_i > k \\ 0, & \text{otherwise,} \end{cases} \quad (15)$$

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

In contrast to Equation 13, the estimating equations in Equation 14 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 θ are estimated by iteratively updating the estimator, i.e. solving the set of quadratic estimating equations for θ . 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 14). Let an alternative set of individual estimating equations be

$$\begin{aligned} \mathbf{s}_i^* &= \begin{pmatrix} (y_{i1}^* - \tau_1)(y_{i2}^* - \tau_2) \\ (y_{i1}^* - \tau_1)(y_{i3}^* - \tau_3) \\ \vdots \\ (y_{ip-1}^* - \tau_{p-1})(y_{ip}^* - \tau_p) \end{pmatrix}, \\ \sum_{i=1}^n \Delta' \mathbf{W}^{-1} \begin{pmatrix} \mathbf{y}_i^* - \boldsymbol{\tau} \\ \mathbf{s}_i^* - \boldsymbol{\sigma}^* \end{pmatrix} &= \mathbf{0}, \end{aligned} \tag{16}$$

where \mathbf{y}_i^* contains the values of individual i on the latent response variables for all items $j = 1, \dots, p$. Note that in this case, the $p^* \times p^*$ matrix \mathbf{W} is an estimator of the working covariance matrix of the vectors $(\mathbf{y}_i^*, \mathbf{s}_i^*)'$ across all individuals $i = 1, \dots, n$. The vector \mathbf{y}_i^* contains the values of individual i on the latent response variables for all items $j = 1, \dots, p$. The vector \mathbf{s}_i^* can be referred to as the vector of the true second-order moments.

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 1 in the Technical

Appendix) are independent and identically distributed. If this is the case, then $\hat{\boldsymbol{\kappa}} = \boldsymbol{\kappa}(\theta)$, i.e. the assumed model fits the data perfectly and the empirical deviation vector in Equation 16 is equal to $\begin{pmatrix} y_{i1}^* - \bar{y}_1^* \\ s_i^* - \bar{s}^* \end{pmatrix}$, where $\bar{\cdot}$ represents the arithmetic mean. To compute individual score contributions based on Equation 11, we apply the logic of Equation 16 to the non-binary case. The aim is to mimic the scores produced by the estimation function in Equation 16. However, the individual values of the latent response variable distribution y^* are not identifiable. Thus, the true second-order moments s^* are not identifiable either. We therefore replace the empirical deviation vector with $\begin{pmatrix} \text{vec}(\mathbf{1}_{\mathbf{y}_i}) - \text{vec}(\bar{\mathbf{1}}_{\mathbf{Y}}) \\ s_i - \bar{s} \end{pmatrix}$. The vector $\text{vec}(\mathbf{1}_{\mathbf{y}_i})$ contains the indicator variables for all items $j = 1, \dots, p$, and all categories minus one $k = 1, \dots, (l - 1)$ (see Equation 10 in the Technical Appendix). The vector $\text{vec}(\bar{\mathbf{1}}_{\mathbf{Y}})$ of size $p(l - 1)$ contains the arithmetic means of the indicator variables across all individuals. Furthermore, we replace the weight matrix of Equation 16 with the weight matrix of Equation 11. 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

$$s_i = \begin{pmatrix} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2) \\ (y_{i1} - \bar{y}_1)(y_{i3} - \bar{y}_3) \\ \vdots \\ (y_{ip-1} - \bar{y}_{p-1})(y_{ip} - \bar{y}_p) \end{pmatrix}, \quad (17)$$

$$\sum_{i=1}^n \tilde{\psi}(\mathbf{y}_i, \theta) = \sum_{i=1}^n \Delta' \mathbf{W}^{-1} \begin{pmatrix} \text{vec}(\mathbf{1}_{\mathbf{y}_i}) - \text{vec}(\bar{\mathbf{1}}_{\mathbf{Y}}) \\ s_i - \bar{s} \end{pmatrix} = \mathbf{0}.$$

We refer to Equation 17 as the *approximated score function* of the WLS estimation method that can be used for the score-based parameter instability test.

Computational Details

The R implementation of the proposed method, replication materials for all simulations, all simulation results as well as the Technical Appendix 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, 170GB 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).

Simulation

We simulated data to fit two different IRT models: a unidimensional model with five observed variables Y_j (Figure B1) and a multidimensional model with nine observed variables Y_j (Figure B2). 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

$$S(\boldsymbol{\xi}, \theta) = \sum_{k=1}^l (k-1) \cdot P(Y_j = k | \boldsymbol{\xi}, \theta), \quad (18)$$

is used to compute the DIF effect size of an item j . The *Noncompensatory 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

$$\text{NCDIF} = \frac{\sum_{i=1}^{n_F} [S(\xi_i, \theta_F) - S(\xi_i, \theta_R)]^2}{n_F}. \quad (19)$$

Using the true values for ξ_i , θ_F , and θ_R from the simulation, we are able to compute the

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

19

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, i.e., the NCDIF value, for one specific item (Item 2) in 24 different simulation scenarios: two different models, i.e., the unidimensional model (Figure B1) and the multidimensional model (Figure B2), 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 w.r.t. 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 0.05 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, i.e., 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 0.05 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 B1). 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 B2). We then use three different approaches to fit the models to the data: ordinal factor analysis (LI estimation), FI estimation, and GEE estimation (see Technical Appendix). 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).

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 B1) 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 w.r.t. all model parameters. Figure B4 shows that given fluctuation w.r.t. 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 w.r.t. the discrimination parameters λ_j , power results for both ordinal factor models and for GRMs fitted via FI estimation are shown in Figure B5. For $k = 1$, sample sizes of $n = 2000$ are needed to achieve power of over 90% for all tests statistics. In general, w.r.t. 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 TESTS FOR ORDINAL FACTOR MODELS

22

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 B2) 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 w.r.t. all model parameters. Power results for the data sets with parameter fluctuation w.r.t. only the threshold parameters of item 1 and 2 are shown in Figure B6. 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 B7. 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 B8). 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. 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.

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, 2018). We analyze 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 B1). 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 B9. 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 η_t is assumed for every time point of

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

24

measurement and one latent item effect variable β_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 (F. L. Classe & Steyer, 2023).

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 w.r.t. 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 analyze the fitted models w.r.t. 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 analyze the time needed to compute the empirical fluctuation process, which includes the computation of the model scores.

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. 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 w.r.t. the results of the score-based tests, particularly 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.

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 to 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, Meiser, and Gernand (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 T. Wang 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 ξ (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.

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, \dots, 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, i.e. 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

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

28

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 (F. 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.

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 modeling 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.

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*. 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 *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* (p. 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*. Monterey, CA: 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. In *Frontiers in education* (Vol. 2, p. 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. doi: 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*.
- Classe, F. L., & Steyer, R. (2023). A probit multistate irt model with latent item effect variables for graded responses. *European Journal of Psychological Assessment*.
- 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*.
- 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*. Retrieved from <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-Villarre, 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).
- 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.

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

31

- 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*.
- 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*, 400–407.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. doi: 10.18637/jss.v048.i02
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.
- Samejima, F. (1997). Graded response models. In *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. (2018). “true” longitudinal and probability-based internet panels: Evidence from the netherlands. In *Social and behavioral research and the internet* (pp. 77–104). Routledge.

- 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*, 1–12.
- 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.

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

33

Appendix A

Tables

	$Cor(SC_{OFA}, SC_{GEE})$		$Cor(SC_{FI}, SC_{GEE})$		$Cor(SC_{OFA}, SC_{FI})$	
	binary	non-binary	binary	non-binary	binary	non-binary
$Var(\eta_1)$	0.94	0.97	0.99	0.95	0.93	0.96
λ_2	0.91	0.94	0.88	0.85	0.96	0.94
λ_3	0.94	0.93	0.93	0.80	0.98	0.91
λ_4	0.94	0.90	0.96	0.73	0.99	0.87
λ_5	0.92	0.91	0.92	0.68	0.97	0.82
τ_{11}	-0.98	-0.92	-0.96	-0.94	0.99	0.76
τ_{12}		-0.95		-0.97		0.90
τ_{13}		-0.92		-0.99		0.90
τ_{21}	-0.99	-0.94	-0.98	-0.96	1.00	0.84
τ_{22}		-0.97		-0.99		0.95
τ_{23}		-0.93		-1.00		0.91
τ_{31}	-0.99	-0.93	-0.99	-0.96	1.00	0.81
τ_{32}		-0.97		-0.99		0.96
τ_{33}		-0.92		-0.98		0.85
τ_{41}	-0.97	-0.90	-0.96	-0.99	1.00	0.84
τ_{42}		-0.97		-0.98		0.92
τ_{43}		-0.92		-0.94		0.76
τ_{51}	-1.00	-0.91	-0.99	-0.96	1.00	0.78
τ_{52}		-0.97		-0.98		0.95
τ_{53}		-0.90		-0.92		0.69

Table A1

Correlation of model scores for a unidimensional GRM (see Figure B1) 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 Technical Appendix).

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

34

	$Cor(SC_{OFA}, SC_{GEE})$		$Cor(SC_{FI}, SC_{GEE})$		$Cor(SC_{OFA}, SC_{FI})$	
	binary	non-binary	binary	non-binary	binary	non-binary
$Var(\eta_1)$	0.97	0.96	0.92	0.71	0.96	0.83
$Var(\eta_2)$	0.91	0.95	0.80	0.67	0.83	0.80
$Var(\eta_3)$	0.93	0.97	0.80	0.89	0.88	0.89
$Cov(\eta_1, \eta_2)$	0.89	0.85	0.91	0.74	0.93	0.88
$Cov(\eta_1, \eta_3)$	0.91	0.84	0.90	0.85	0.95	0.93
$Cov(\eta_2, \eta_3)$	0.88	0.90	0.85	0.84	0.91	0.93
λ_{12}	0.93	0.91	0.85	0.51	0.87	0.68
λ_{13}	0.92	0.87	0.89	0.77	0.87	0.89
λ_{22}	0.85	0.88	0.87	0.72	0.93	0.86
λ_{23}	0.92	0.87	0.65	0.54	0.68	0.71
λ_{32}	0.93	0.88	0.90	0.74	0.93	0.81
λ_{33}	0.91	0.85	0.76	0.73	0.90	0.79
τ_{11}	-0.98	-0.88	-0.98	-0.97	1.00	0.79
τ_{12}		-0.92		-0.98		0.88
τ_{13}		-0.89		-0.94		0.76
τ_{21}	-0.98	-0.89	-0.98	-0.94	1.00	0.74
τ_{22}		-0.93		-0.98		0.90
τ_{23}		-0.90		-0.91		0.71
τ_{31}	-0.99	-0.84	-0.99	-0.88	1.00	0.64
τ_{32}		-0.83		-0.90		0.71
τ_{33}		-0.83		-0.94		0.82
τ_{41}	-0.95	-0.88	-0.94	-0.94	0.99	0.71
τ_{42}		-0.94		-0.99		0.93
τ_{43}		-0.88		-0.98		0.80
τ_{51}	-0.88	-0.88	-0.88	-0.98	0.97	0.81
τ_{52}		-0.95		-0.99		0.94
τ_{53}		-0.89		-0.95		0.74
τ_{61}	-0.97	-0.87	-0.97	-0.94	0.99	0.71
τ_{62}		-0.93		-0.99		0.92
τ_{63}		-0.88		-0.92		0.65
τ_{71}	-0.81	-0.94	-0.91	-1.00	0.84	0.92
τ_{72}		-0.93		-1.00		0.93
τ_{73}		-0.88		-1.00		0.87
τ_{81}	-0.98	-0.90	-0.98	-1.00	1.00	0.90
τ_{82}		-0.85		-1.00		0.85
τ_{83}		-0.82		-0.99		0.81
τ_{91}	-0.94	-0.90	-0.92	-1.00	0.99	0.90
τ_{92}		-0.88		-0.99		0.88
τ_{93}		-0.79		-0.99		0.76

Table A2

Correlation of model scores for a multidimensional GRM (see Figure B2) 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 Technical Appendix).

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

35

Mode		Model	
		unidimensional	multidimensional
all	k=1	0.05	0.14
all	k=2	0.18	0.48
all	k=4	0.54	1.75
all	k=6	1.20	3.65
thresholds	k=1	0.04	0.13
thresholds	k=2	0.15	0.49
thresholds	k=4	0.47	1.65
thresholds	k=6	1.02	3.46
lambdas	k=1	0.01	0.01
lambdas	k=2	0.05	0.04
lambdas	k=4	0.15	0.14
lambdas	k=6	0.31	0.31

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.

	n=500		n=1000		n=2000	
	FI	LI	FI	LI	FI	LI
k=1	0.19	0.19	0.20	0.21	0.21	0.23
k=2	0.23	0.20	0.25	0.22	0.27	0.26
k=4	0.31	0.25	0.36	0.27	0.38	0.31
k=6	0.41	0.30	0.47	0.31	0.51	0.38

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	12.80	0.39	17.42	0.46	21.02	0.39
k=2	14.04	0.42	19.53	0.40	27.40	0.42
k=4	18.38	0.52	26.65	0.47	40.92	0.50
k=6	22.48	0.64	50.66	0.84	57.21	0.63

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.

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

36

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 A6

Life satisfaction scale items as asked in the LISS panel.

			Model 1 (unidim.)	Model 2 (multidim.)	Model 3 (PIEG)
Ordinal Factor Analysis	Number of Paramers		35	63	156
	RMSEA		0.127	0.127	0.051
	Score-based test p-value	categorical	1.04E-05	2.22E-04	0.017
		ordinal	0.918	0.694	0.689
		metric	0.165	0.008	0.014
	Computation time	model	0.345	0.913	6.189
		scores	0.063	0.109	0.325
GRM: FI estimation	Number of Paramers		35	63	156
	RMSEA		0	0	0
	Score-based test p-value	categorical	3.92E-06	4.02E-05	0
		ordinal	0.181	0.445	0
		metric	0.197	0.002	0
	Computation time	model	0.541	88.428	301.546
		scores	0.287	15.506	1074.466
Common Factor Analysis	Number of Paramers		10	13	56
	RMSEA		0.099	0.122	0.043
	Score-based test p-value	categorical	0.002	0.266	0.424
		ordinal	0.227	0.453	0.770
		metric	0.014	0.000	0.040
	Computation time	model	0.186	0.142	0.514
		scores	0.089	0.194	0.247

Table A7

Results of the real data application.

Appendix B

Figures

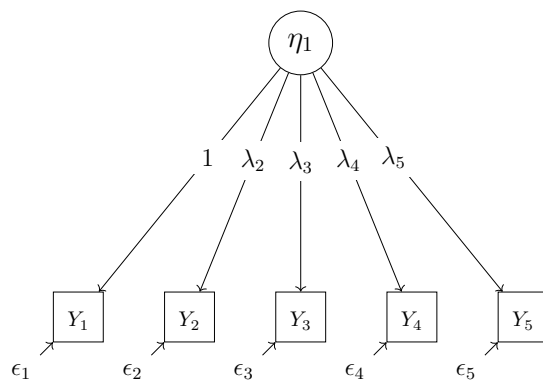


Figure B1. Unidimensional graded response model (GRM) with five items.

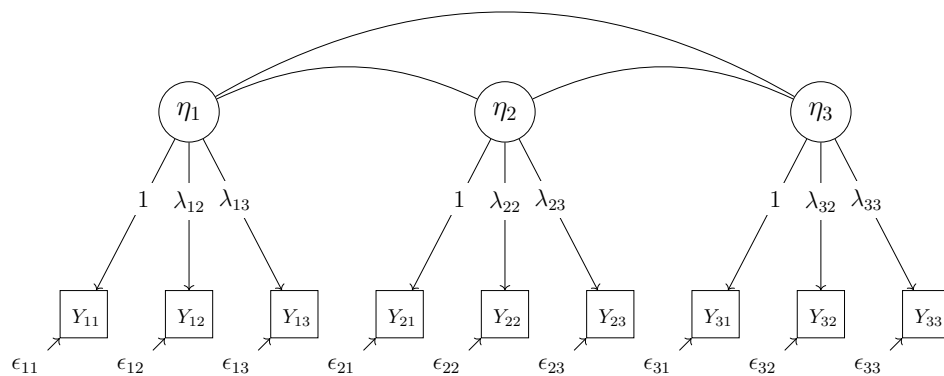


Figure B2. Multidimensional graded response model (GRM) with three non-orthogonal latent variables and nine items.

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

38

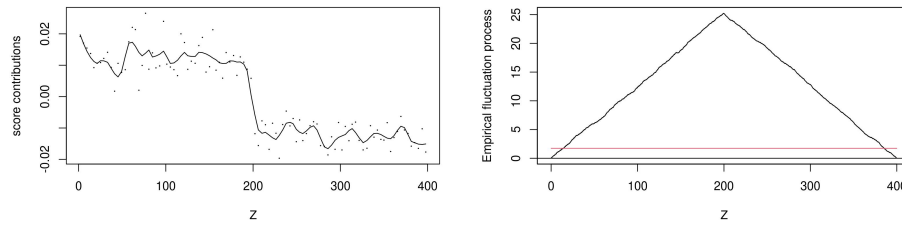


Figure B3. Score and CSP distribution (illustration inspired by Figure 2 in Strobl et al., 2015).

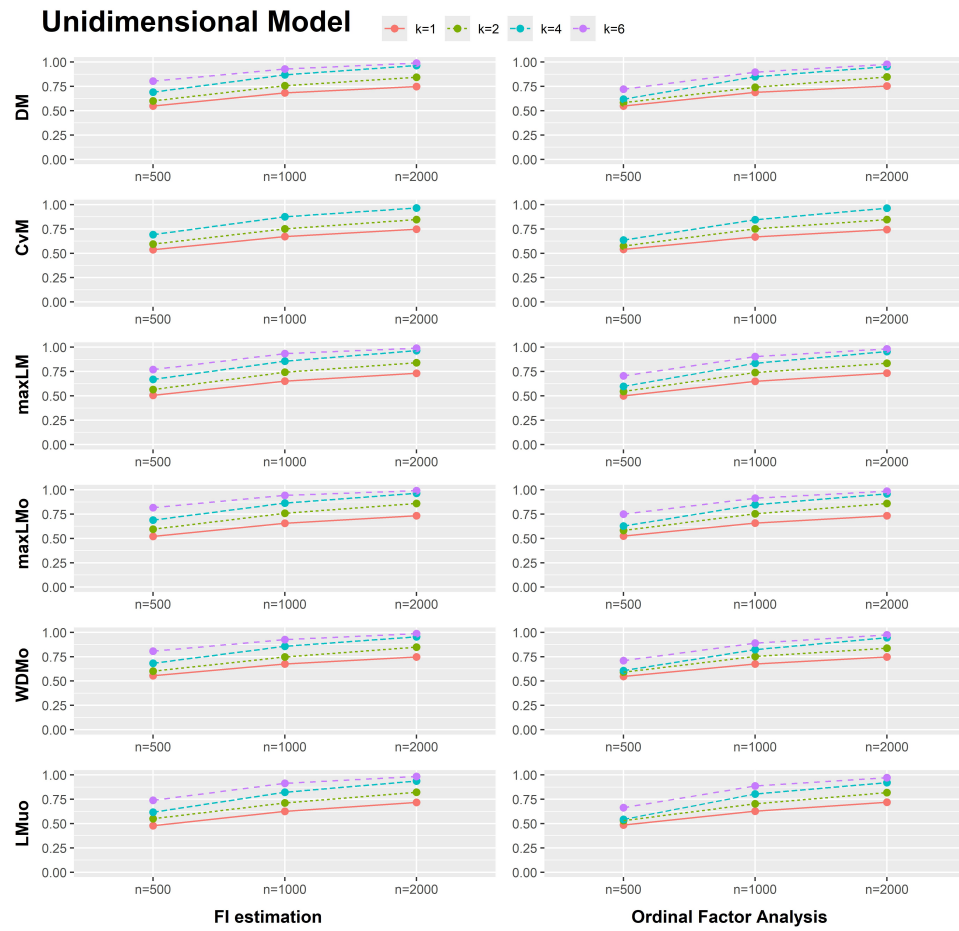


Figure B4. Power of score-based test for the unidimensional GRM model given fluctuation w.r.t. the threshold parameters of the first item τ_{1k} .

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

39

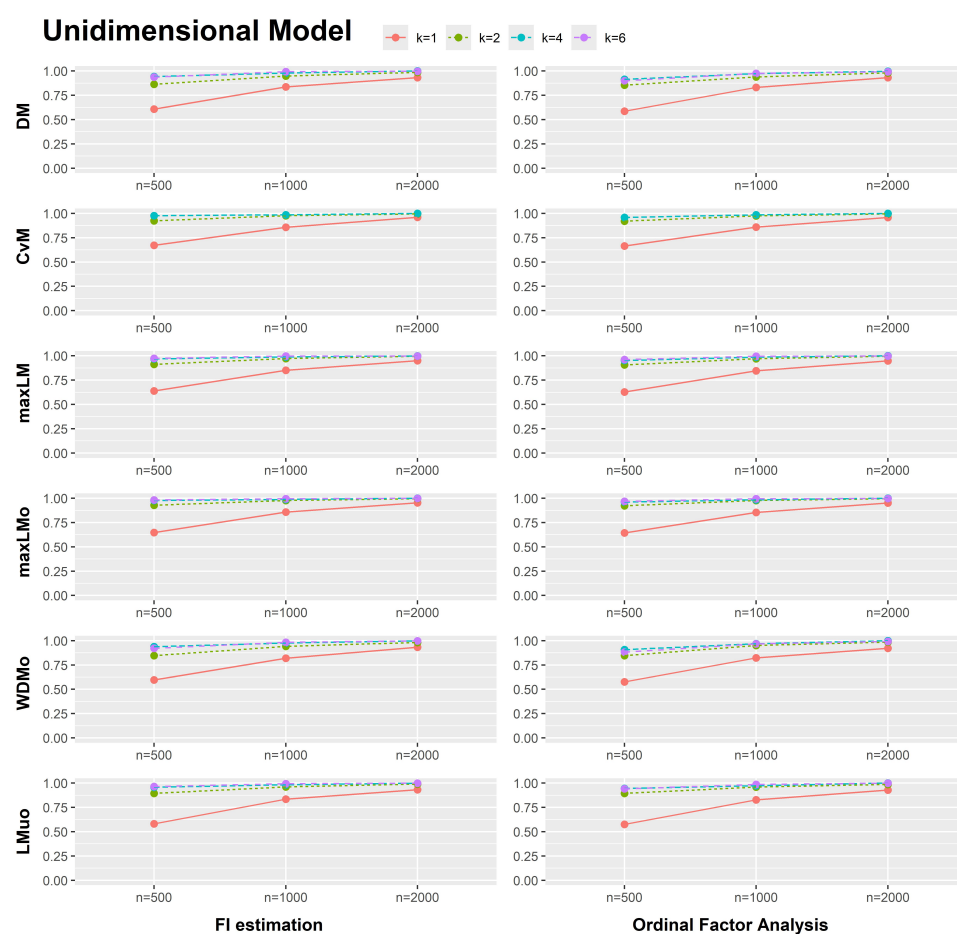


Figure B5. Power of score-based test for the unidimensional GRM model given fluctuation w.r.t. the discrimination parameters λ_i .

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

40

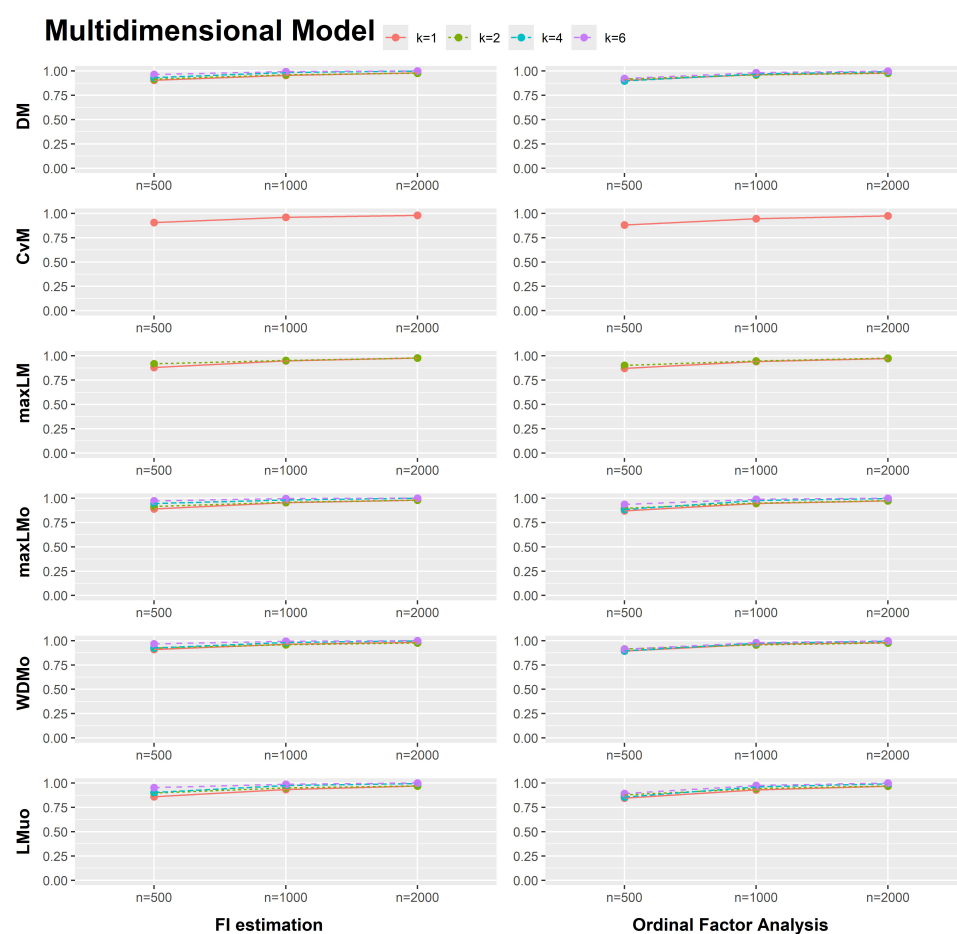


Figure B6. Power of score-based test for the multidimensional GRM model given fluctuation w.r.t. the threshold parameters of the first two items, i.e. τ_{1k} and τ_{2k} .

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

41

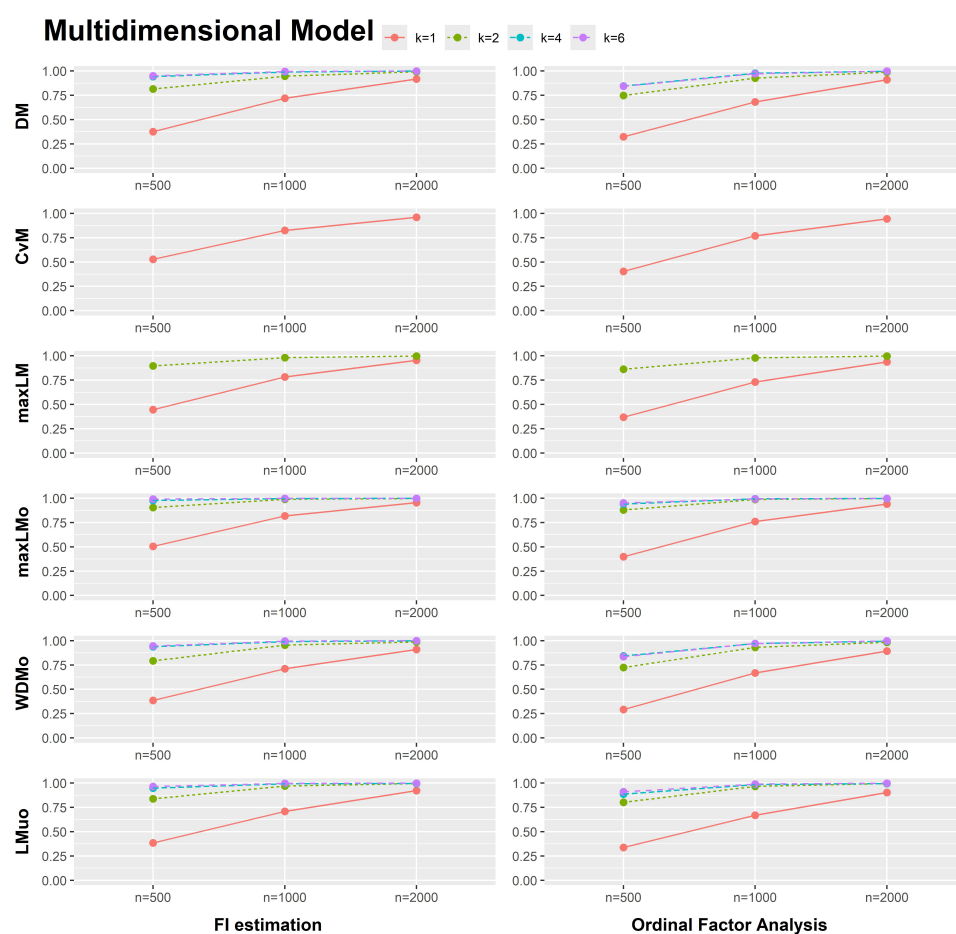


Figure B7. Power of score-based test for the multidimensional GRM model given fluctuation w.r.t. the discrimination parameters λ_i .

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

42

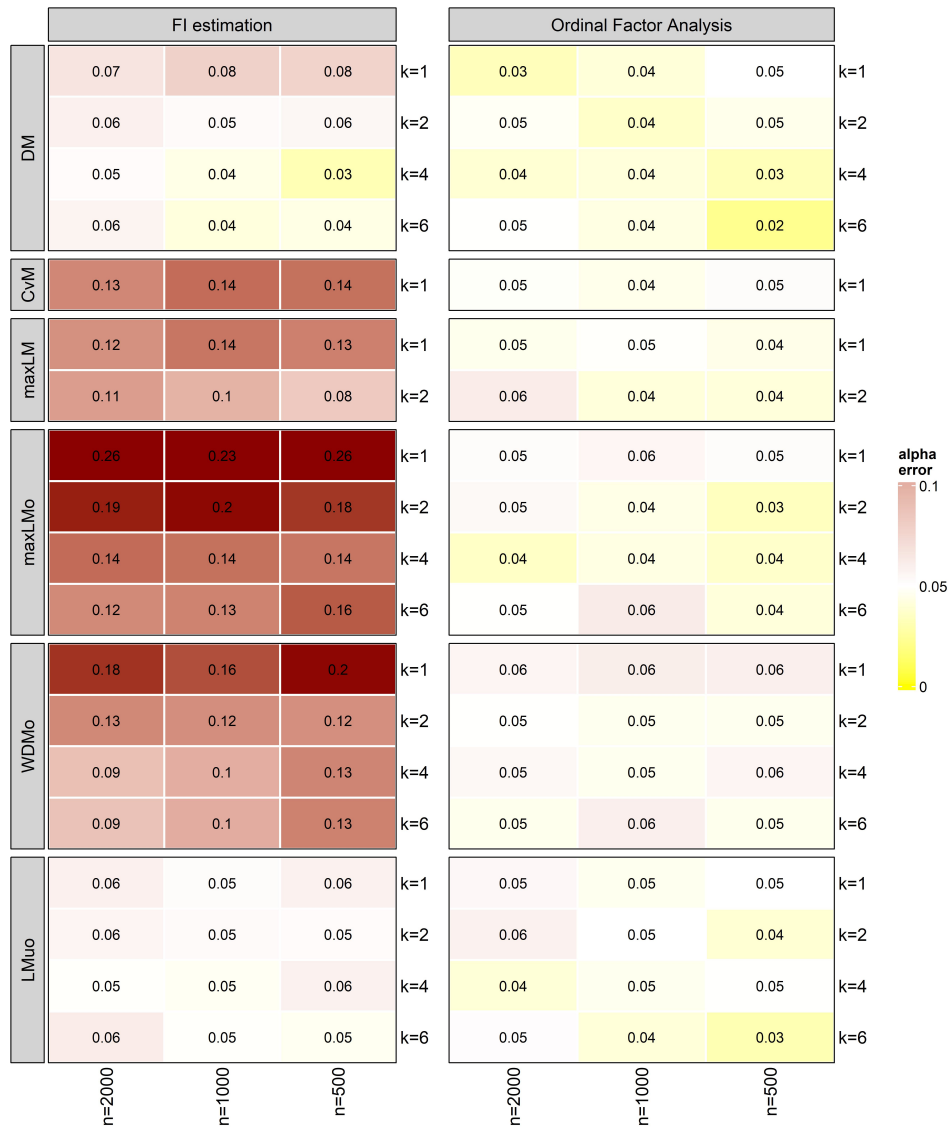


Figure B8. Type I errors of score-based test for the multidimensional GRM model. Note that for the *CvM* test statistic, there are no critical values implemented in the **strucchange** package for models with more than 25 parameters. This also applies for the *maxLM* test statistic for models with more than 40 parameters. Therefore, models with more than 1 (for *CvM*) and 2 (for *maxLM*) threshold parameters are not shown.

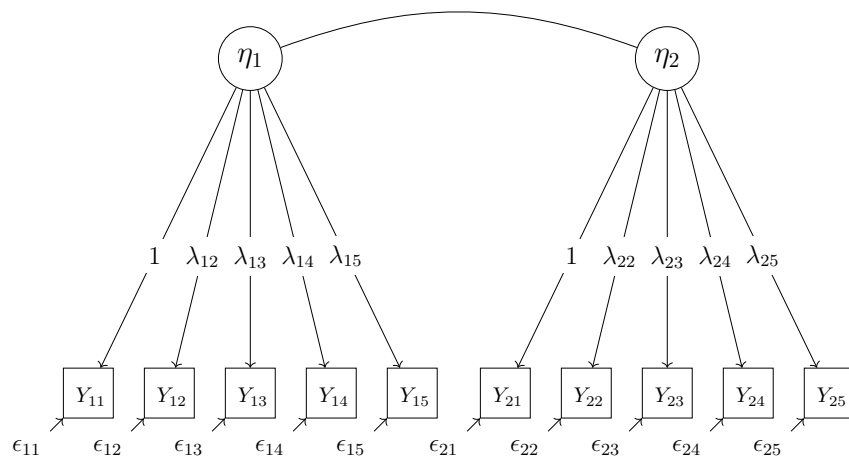


Figure B9. Real Data Application Model 2: Multidimensional graded response model (GRM) with two latent state variables and five items on two time points.

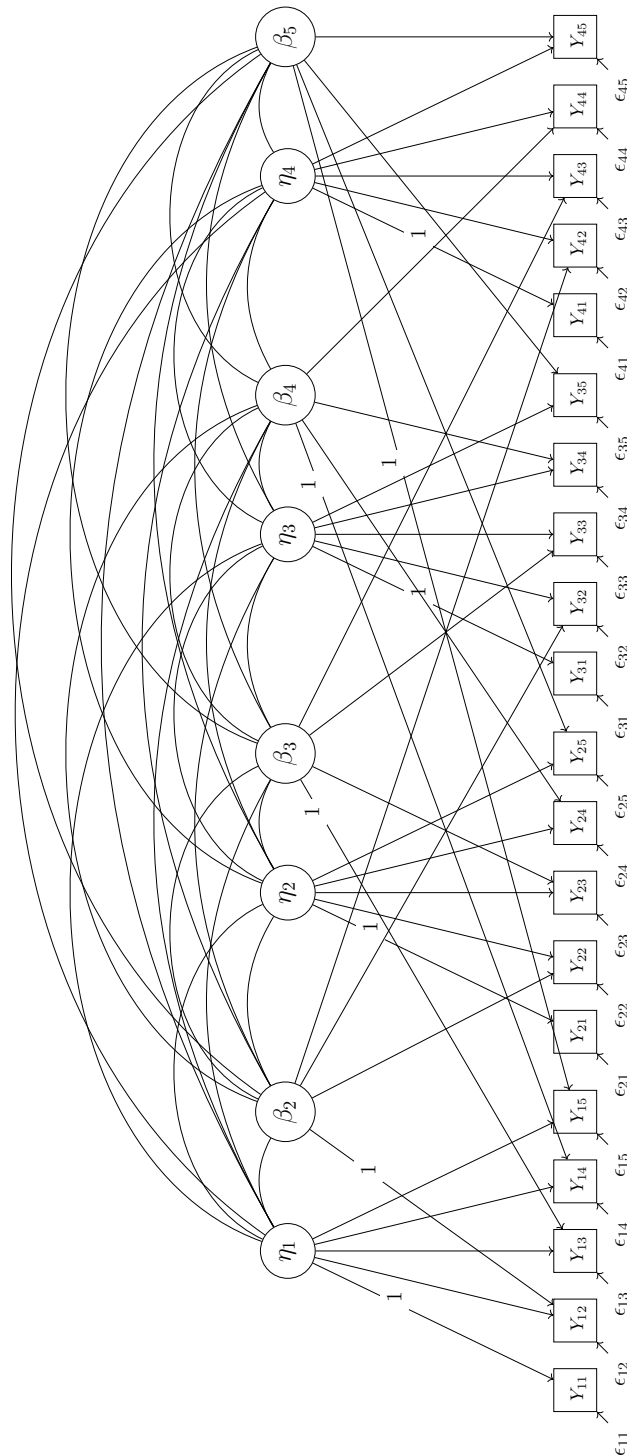


Figure B10. Real Data Application Model 3: Probit Multistate IRT Model With Latent Item Effect Variables for Graded Responses (PIEG) with four latent state variables (η_h), four latent item effect variables (β_i), and 5 items on three time points.

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

45

Appendix C

*

List of Tables

A1	Correlation of model scores for a unidimensional GRM (see Figure B1) 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 Technical Appendix).	33
A2	Correlation of model scores for a multidimensional GRM (see Figure B2) 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 Technical Appendix).	34
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.	35
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.	35

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

46

- 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. 35
- A6 Life satisfaction scale items as asked in the LISS panel. 36
- A7 Results of the real data application. 36

SCORE-BASED TESTS FOR ORDINAL FACTOR MODELS

47

Appendix D

*

List of Figures

B1	Unidimensional graded response model (GRM) with five items.	37
B2	Multidimensional graded response model (GRM) with three non-orthogonal latent variables and nine items.	37
B3	Score and CSP distribution (illustration inspired by Figure 2 in Strobl et al., 2015).	38
B4	Power of score-based test for the unidimensional GRM model given fluctuation w.r.t. the threshold parameters of the first item τ_{1k}	38
B5	Power of score-based test for the unidimensional GRM model given fluctuation w.r.t. the discrimination parameters λ_i	39
B6	Power of score-based test for the multidimensional GRM model given fluctuation w.r.t. the threshold parameters of the first two items, i.e. τ_{1k} and τ_{2k}	40
B7	Power of score-based test for the multidimensional GRM model given fluctuation w.r.t. the discrimination parameters λ_i	41
B8	Type I errors of score-based test for the multidimensional GRM model. Note that for the <i>CvM</i> test statistic, there are no critical values implemented in the strucchange package for models with more than 25 parameters. This also applies for the <i>maxLM</i> test statistic for models with more than 40 parameters. Therefore, models with more than 1 (for <i>CvM</i>) and 2 (for <i>maxLM</i>) threshold parameters are not shown.	42
B9	Real Data Application Model 2: Multidimensional graded response model (GRM) with two latent state variables and five items on two time points.	43
B10	Real Data Application Model 3: Probit Multistate IRT Model With Latent Item Effect Variables for Graded Responses (PIEG) with four latent state variables (η_t), four latent item effect variables (β_i), and 5 items on three time points.	44

Paper IV - Supplementary Material:

Since the supplementary material of Paper IV was still being reviewed by the journal at the time of submission of the dissertation, the revised version of the paper's supplementary material originally submitted to the journal is included here. The paper's supplementary material has since been published.

Classe, F., Debelak, R., & Kern, C. (2025). Score-based tests for parameter instability in ordinal factor models. *British Journal of Mathematical and Statistical Psychology*. <https://doi.org/10.1111/bmsp.12392>

Running head: TECHNICAL APPENDIX

1

Score-Based Tests for Parameter Instability in Ordinal Factor Models: Technical
Appendix

TECHNICAL APPENDIX

2

Score-Based Tests for Parameter Instability in Ordinal Factor Models: Technical
Appendix

Generalized Estimating Equations (GEE) for Ordinal Factor Analysis

In the introduction of the main text, we highlight that parameter estimation for multidimensional IRT (MIRT) models via polychorics is also referred to as limited-information (LI) estimation, as it only uses information from bivariate relations of the observed variables. We then present the WLS estimator, as originally introduced by Muthén (1983, 1984). An alternative LI estimation method is presented in this chapter. To our knowledge, this estimation method had not yet been described or implemented for non-binary data by prior research and is thus discussed in this section. Let \mathbf{Y} be a $p \times 1$ vector of ordered observed variables. For simplicity, we assume that all ordered observed variables have l response categories denoted by the index k . In an ordinal factor model (Maydeu-Olivares, 2005), a continuous, normally distributed latent observed variable Y_j^* is assumed to underlie each observed ordered variable $Y_j \forall j = 1, \dots, p$. Just as in a conventional factor model, a linear measurement structure is assumed. That is

$$\mathbf{Y}^* = \boldsymbol{\lambda}'\boldsymbol{\xi} + \boldsymbol{\epsilon}, \quad (1)$$

where $\boldsymbol{\xi}$ is the the $m \times 1$ vector of continuous latent variables, and $\boldsymbol{\lambda}$ is the $m \times p$ matrix of discrimination parameters (also referred to as factor loadings). Also, $\boldsymbol{\epsilon}$ is a $p \times 1$ vector of residuals. In the following, we refer to $Cov(\mathbf{Y}^*)$ as the *model implied* covariance matrix.

The latent observed variable Y_j is related to the observed ordered variable via a threshold relation, that is

$$Y_j = k_j \text{ if } \tau_{j(k-1)} < y_j^* \leq \tau_{jk}. \quad (2)$$

This means that a respondent chooses a response category k_j when the respondent's latent response value y_j^* on item j lies between the thresholds $\tau_{j(k-1)}$ and τ_{jk} , where

TECHNICAL APPENDIX

3

$\tau_{j0} = -\infty$ and $\tau_{jl} = +\infty$.

The model parameter vector θ contains all freely estimated model parameters. That includes the threshold parameters τ_{jk} for all items $j = 1, \dots, p$, belonging to the item categories $k = 1, \dots, l - 1$. Note that, for each item, there is one threshold parameter less than there are item categories. Furthermore, θ contains all freely estimated discrimination parameters λ_{qj} for all latent variables $q = 1, \dots, m$, and all items $j = 1, \dots, p$, as well as all freely estimated latent variable variances and covariances, such that

$$\begin{aligned} \theta = \{ & \tau_{11}, \dots, \tau_{pl}, \lambda_{11}, \dots, \lambda_{mp}, \\ & Var(\xi_1), \dots, Var(\xi_m), \\ & Cov(\xi_1, \xi_2), \dots, Cov(\xi_{m-1}, \xi_m) \}. \end{aligned} \quad (3)$$

In the original approach, a three-stage generalized least squares procedure for parameter estimation was proposed. However, Reboussin and Liang (1998) claim that this estimation method may not perform well for small sample sizes and large numbers of indicator variables. They therefore propose an alternative estimation procedure based on an quadratic estimating equations. Muthén (1997) referred to this method as the *Generalized Estimating Equation* (GEE) approach.

For the purpose of finding a computationally feasible way to compute model scores for ordinal factor models, the GEE approach seems very promising. Parameter estimation via GEEs is closely related to the idea of the *M-estimator* (Stefanski & Boos, 2002), which is an estimator $\hat{\theta}$ that satisfies

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

where \mathbf{y}_i is the $p \times 1$ vector of observed responses of individual i .

From Muthén (1997) and Reboussin and Liang (1998), we derive that the (model

TECHNICAL APPENDIX

4

implied) true mean μ_j of the observed variable Y_j is

$$\mu_j(\theta) = \mu_j = E(Y_j) = \sum_{k=1}^l k \cdot P(Y_j = k). \quad (5)$$

The true category probabilities of Y_j can be computed on the basis of the threshold parameters τ_{jk} , such that

$$\begin{aligned} P(Y_j = k) &= \Phi(\tau_{j1}) \quad \text{if } k = 1, \\ P(Y_j = k) &= \Phi(\tau_{jk}) - \Phi(\tau_{j;k-1}) \quad \text{if } 1 < k < l, \\ P(Y_j = k) &= 1 - \Phi(\tau_{j(l-1)}) \quad \text{if } k = l. \end{aligned} \quad (6)$$

The true joint probability $P(Y_j = k, Y_s = h)$ and thus the true joint expectation $E(Y_j Y_s)$, can be calculated as follows

$$\begin{aligned} P(Y_j = k, Y_s = h) &= \Phi_2(\tau_{jk}, \tau_{sh}, \sigma_{js}^*) - \\ &\quad \Phi_2(\tau_{j;k-1}, \tau_{sh}, \sigma_{js}^*) - \\ &\quad \Phi_2(\tau_{jk}, \tau_{s;h-1}, \sigma_{js}^*) - \\ &\quad \Phi_2(\tau_{j;k-1}, \tau_{s;h-1}, \sigma_{js}^*), \\ E(Y_j Y_s) &= \sum_{k=1}^l \sum_{h=1}^g k \cdot h \cdot P(Y_j = k, Y_s = h), \end{aligned} \quad (7)$$

where σ_{js}^* are elements in the model implied covariance matrix, i.e. $\sigma_{js}^* = \text{Cov}(Y_{js}^*)$.

This definition is derived from Equation 4 in Olsson (1979).

Furthermore, let the second order moment of Y_j and Y_s be

$$\sigma_{js} = E(Y_j Y_s) - \mu_j \mu_s. \quad (8)$$

For the case of non-binary observed variables, the first order moments of Y_j are made up from the indicator variables $1_{Y_j > k} \forall k = 1, \dots, l-1$. The true mean of $1_{Y_j > k}$ is

$$\nu_{jk} = E(1_{Y_j > k}) = P(Y_j > k) = P(Y_j^* > \tau_{jk}) = \Phi(-\tau_{jk}). \quad (9)$$

TECHNICAL APPENDIX

5

Note that $1_{Y_j > k} = Y_j$ in the special case of dichotomous observed variables (as reported in Muthén, 1997; Reboussin & Liang, 1998).

To fit an ordinal factor models via GEEs, first and second order individual empirical moments are defined. Let the $(l-1) \times p$ matrix $\mathbf{1}_{\mathbf{y}_i}$ contain the first order empirical moments of individual i , that is

$$\mathbf{1}_{\mathbf{y}_i} = \begin{pmatrix} 1_{y_{i1} > 1} & \dots & 1_{y_{ip} > 1} \\ 1_{y_{i1} > 2} & \dots & 1_{y_{ip} > 2} \\ \vdots & \ddots & \vdots \\ 1_{y_{i1} > l-1} & \dots & 1_{y_{ip} > l-1} \end{pmatrix}. \quad (10)$$

The first order moments of \mathbf{Y} are in the $(l-1) \times p$ matrix

$$\boldsymbol{\nu} = \begin{pmatrix} \nu_{11} & \nu_{21} & \dots & \nu_{p1} \\ \nu_{12} & \nu_{22} & \dots & \nu_{p2} \\ \vdots & \ddots & \ddots & \vdots \\ \nu_{1;l-1} & \nu_{2;l-1} & \dots & \nu_{p;l-1} \end{pmatrix}. \quad (11)$$

Moreover, let $\boldsymbol{\sigma}$ be a vector of second order moments of \mathbf{Y} . This vector includes all non-redundant, off-diagonal elements of the true covariance matrix of \mathbf{Y} , i.e.

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{12} \\ \sigma_{13} \\ \vdots \\ \sigma_{p-1;p} \end{pmatrix}. \quad (12)$$

The second order empirical moments of individual i make up the $p(p-1)/2 \times 1$ vector \mathbf{s}_i , that is

$$\mathbf{s}_i = \begin{pmatrix} (y_{i1} - \mu_1)(y_{i2} - \mu_2) \\ (y_{i1} - \mu_1)(y_{i3} - \mu_3) \\ \vdots \\ (y_{ip-1} - \mu_{p-1})(y_{ip} - \mu_p) \end{pmatrix}. \quad (13)$$

TECHNICAL APPENDIX

6

The vector of first and second order empirical deviations for individual i is

$$\mathbf{e}_i = \begin{pmatrix} \text{vec}(\mathbf{1}_{\mathbf{y}_i}) - \text{vec}(\boldsymbol{\nu}) \\ s_i - \boldsymbol{\sigma} \end{pmatrix}. \quad (14)$$

The size of \mathbf{e}_i is $[p(l-1) + p(p-1)/2] \times 1$ which we refer to as $p^* \times 1$ in the following.

An ordinal factor model that does not assume a specific model structure is referred to as a saturated model. The $p^* \times 1$ parameter vector of the saturated model is

$\beta = (\text{vec}(\boldsymbol{\nu}), \boldsymbol{\sigma})'$. Equations 5 to 9 show that β is a function of θ . The parameters of the structured model θ can therefore be estimated through minimization of an objective function, that is

$$F_{GEE}(\theta) = \sum_{i=1}^n \mathbf{e}_i' \mathbf{W}^{-1} \mathbf{e}_i. \quad (15)$$

This GEE fitting function minimizes the deviations of the individual empirical first and second order moments from the saturated model parameters.

The weight matrix \mathbf{W} is defined as the working covariance matrix of $\mathbf{1}_{\mathbf{y}_i}$ and s_i . From Reboussin and Liang (1998), we derive that a choice for this matrix, that is adequate for the case of non-binary observed variables, is

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{pmatrix}. \quad (16)$$

\mathbf{W}_1 is the working covariance matrix of $\mathbf{1}_{\mathbf{y}_i} \forall i = 1, \dots, n$, that is

$$\begin{aligned} [\mathbf{W}_1]_{jks h} &= \mu_j(1 - \mu_j) && \text{if } j = s, k = h, \\ P(Y_j > k) - \nu_{jk}\nu_{jh} &= \Phi(-\tau_{jk}) - \nu_{jk}\nu_{jh} && \text{if } j = s, k > h, \\ P(Y_j > k, Y_s > h) - \nu_{jk}\nu_{sh} &= \Phi_2(-\tau_{jk}, -\tau_{sh}, \sigma_{js}^*) - \nu_{jk}\nu_{sh} && \text{if } j \neq s, k \neq h. \end{aligned} \quad (17)$$

\mathbf{W}_2 is the diagonal working covariance matrix of $s_i \forall i = 1, \dots, n$, with all non-diagonal elements equal to zero and all diagonal elements equal to

$$[\mathbf{W}_2]_{js, js} = \frac{\sum_{i=1}^n (w_{ijs}^2)}{n} - \sigma_{js}^2. \quad (18)$$

TECHNICAL APPENDIX

7

Let Δ be the first derivative of β with respect to θ , that is

$$\Delta = \frac{\partial \beta(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \text{vec}(\boldsymbol{\nu})(\theta)}{\partial \theta} \\ \frac{\partial \boldsymbol{\sigma}(\theta)}{\partial \theta} \end{pmatrix}. \quad (19)$$

Then, the first derivative of Equation 15 with respect to θ is

$$\frac{\partial F_{GEE}(\theta)}{\partial \theta} = \frac{\partial F_{GEE}(\beta)}{\partial \beta} \frac{\partial \beta(\theta)}{\partial \theta} = \sum_{i=1}^n -2\mathbf{e}_i' \mathbf{W}^{-1} \Delta. \quad (20)$$

Transposing the set of $1 \times p$ row vectors resulting from Equation 20 leads to the following set of estimating equations:

$$\sum_{i=1}^n \psi(\mathbf{y}_i, \theta) = \sum_{i=1}^n \Delta' \mathbf{W}^{-1} \mathbf{e}_i = \mathbf{0}. \quad (21)$$

The model parameters in θ are estimated by solving this set of quadratic estimating equations for θ by iteratively updating the estimator via a modified Fisher's scoring algorithm

$$\hat{\theta}^{r+1} = \hat{\theta}^r + (n \cdot \Delta' \mathbf{W}^{-1} \Delta)^{-1} \sum_{i=1}^n \psi(\mathbf{y}_i, \hat{\theta}), \quad (22)$$

where $\hat{\theta}^r$ denotes the parameter estimates at the r^{th} iteration.

Equation 21 is the *score function* (see Stefanski & Boos, 2002) of the GEE estimation method that can be used for the score-based parameter instability test.

Relation of GEE and WLS Model Scores

In order to estimate model parameters via GEEs, all three components of Equation 21 need to be updated step by step. This means that Equation 5 to 22 need to be computed at every iteration. In contrast to this, the WLS estimation method introduced by Muthén (1983, 1984) works without iteratively updating \mathbf{W} , Δ , or \mathbf{e}_i .

In the first estimation step in the model fitting process, the first and second order sample statistics are estimated following the approach established by Olsson (1979).

These sample statistics consist of the sample thresholds t_{jk} , and the bivariate polychoric

	Type I error rate		Power (only λ -fluctuation)	
	GEE	WLS	GEE	WLS
DM	0.03	0.05	0.32	0.68
CvM	0.03	0.05	0.43	0.60
maxLM	0.03	0.04	0.40	0.63
maxLMo	0.04	0.05	0.42	0.60
WDMo	0.06	0.06	0.32	0.71
LMuo	0.02	0.05	0.35	0.66

Table 1

Power and Type I error rate of score-based test for a multidimensional GRM model with 9 dichotomous observed variables and 3 latent variables. Comparison of GEE scores vs. WLS scores.

correlations ρ_{js} for all $k=1, \dots, l$, $j, s=1, \dots, p$ when $j \neq s$. They make up the $p^* \times 1$ vector $\hat{\mathbf{\kappa}}$.

In the third estimation step, the model parameters in θ are estimated through minimization of the objective function

$$F_{OFA}(\theta) = [\hat{\mathbf{\kappa}} - \boldsymbol{\kappa}(\theta)]' \mathbf{W}^{-1} [\hat{\mathbf{\kappa}} - \boldsymbol{\kappa}(\theta)]. \quad (23)$$

Note the distinction between the thresholds that are estimated as sample statistics in the first two steps of the model fitting process and the threshold parameters τ_{jk} in θ . Furthermore, \mathbf{W} is a consistent estimator asymptotic covariance matrix of $\hat{\mathbf{\kappa}}$ (see Muthén & Satorra, 1995). The weight matrix accounts for multivariate non-normality in the observed variables. This idea goes back to Browne (1984) who focussed primarily on continuous, non-normal observed variables. Muthén (1984) then extended this approach to ordered categorical observed variables. Thus, the WLS estimator is often referred to as an *asymptotically distribution free* estimator (Flora & Curran, 2004; Kyriazos et al., 2018).

In the main text, we show that the application of $\tilde{\psi}(\mathbf{y}_i, \hat{\theta})$ to score-based parameter instability tests is computationally efficient and has a low Type I error rate as well as high power. In fact, it outperforms score-based parameter instability tests for models fitted with a full information approach on all metrics.

TECHNICAL APPENDIX

9

In Table 1, the performance of the score functions of Equation 21 and the pseudo score function the main text applied to the score-based parameter instability test are compared. To measure the performance of the score-based test for the GEE estimation method, we used the same simulation setup as in the main text but only for a simple multidimensional GRM model with 3 latent variables and dichotomous observed variables. Also, just one scenario is applied for parameter fluctuation in the data: only the discrimination parameters λ_j differ within a single data set. The results shown in Table 1 indicate that the WLS estimation method outperforms the GEE estimation method with respect to test power.

TECHNICAL APPENDIX

10

References

- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British journal of mathematical and statistical psychology*, 37(1), 62–83.
- 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.
- Kyriazos, T. A., et al. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (efa, cfa) and sem in general. *Psychology*, 9(08), 2207.
- Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item response theory and factor analysis: a unified framework. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for roderick p. mcdonald* (pp. 73–102). Lawrence Erlbaum Associates Publishers.
- Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22(1-2), 43–65.
- 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*.
- Muthén, B., & Satorra, A. (1995). Technical aspects of muthén's liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, 60(4), 489–503.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460.
- Reboussin, B. A., & Liang, K.-Y. (1998). An estimating equations approach for the liscomp model. *Psychometrika*, 63, 165–182.
- Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, 56(1), 29–38.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 27.08.2025

Franz Classe