A R T I C L E

the british
psychological society
promoting excellence in psychology

Investigating heterogeneity in IRTree models for multiple response processes with score-based partitioning

Rudolf Debelak[1] | Thorsten Meiser[2] | Alicia Gernand[3]

[1]University of Zurich, Zurich, Switzerland

[2]University of Mannheim, Mannheim, Germany

[3]RPTU Kaiserslautern-Landau, Landau, Germany

Correspondence
Rudolf Debelak, Department of Education, University of Zurich, Zurich, Switzerland.
Email: rudolf.debelak@ife.uzh.ch

Abstract

Item response tree (IRTree) models form a family of psychometric models that allow researchers to control for multiple response processes, such as different sorts of response styles, in the measurement of latent traits. While IRTree models can capture quantitative individual differences in both the latent traits of interest and the use of response categories, they maintain the basic assumption that the nature and weighting of latent response processes are homogeneous across the entire population of respondents. In the present research, we therefore propose a novel approach for detecting heterogeneity in the parameters of IRTree models across subgroups that engage in different response behavior. The approach uses score-based tests to reveal violations of parameter heterogeneity along extraneous person covariates, and it can be employed as a model-based partitioning algorithm to identify sources of differences in the strength of trait-based responding or other response processes. Simulation studies demonstrate generally accurate Type I error rates and sufficient power for metric, ordinal, and categorical person covariates and for different types of test statistics, with the potential to differentiate between different types of parameter heterogeneity. An empirical application illustrates the use of score-based partitioning in the analysis of latent response processes with real data.

K E Y W O R D S

IRTree models, item response theory, model-based recursive partitioning, parameter heterogeneity, response styles, score-based tests
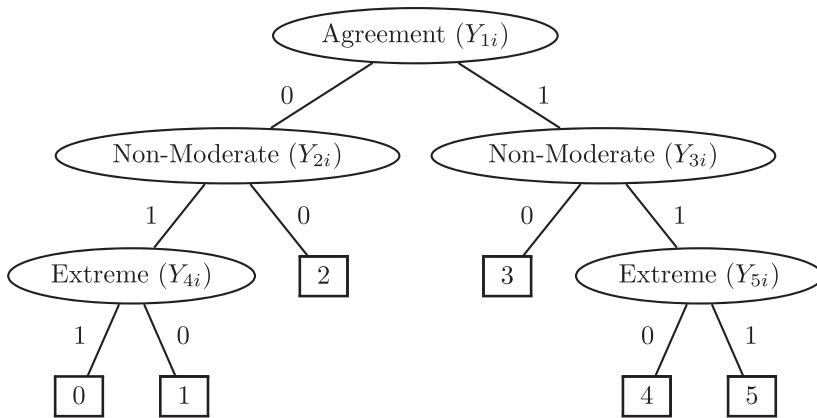
FIGURE 1 IRTree model for Likert items with a six-point rating scale. Theoretical nodes in ellipses denote assumed decision processes, terminal nodes in squares denote observed item responses.

1 | INTRODUCTION

Likert response formats are commonly used to assess latent traits, but it is widely known that the response categories can elicit individual response processes that affect the measurement independently of item content. Examples of such additional response processes, or response styles, include midscale and extreme responding (Baumgartner & Steenkamp, 2001; Mellenbergh, 2011), which lead to the methodological challenge of modeling effects in such a way as to disentangle the latent traits of interest from potential biases.

To account for response styles, traditional IRT models like the partial credit model (Masters, 1982) and nominal response model (Takane & De Leeuw, 1987) have been extended with additional trait dimensions (Bolt et al., 2014; Falk & Cai, 2016) or random threshold parameters (Jin & Wang, 2014; Wang & Wu, 2011) that capture individual differences in the use of response categories (for an overview, see Henninger & Meiser, 2020). An alternative approach to modeling response styles comes from item response tree (IRTree) models, that is, decision trees with IRT parameterizations of decision nodes (Böckenholt, 2012; De Boeck & Partchev, 2012). IRTree models provide a general framework for specifying multiple response processes that lead to an observed response category (Jeon & De Boeck, 2016), and they can be used to model the joint influences of substantive traits and response styles on ordinal responses (Böckenholt & Meiser, 2017; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014). Unlike partial credit and nominal response models, IRTree models do not fall into the class of divide-by-total models and thus allow independent specifications of the category probabilities on a Likert scale (Schoenmakers et al., 2024).

1.1 | IRTree models for ordinal responses with response styles

Conceptually, IRTree models for ordinal responses assume a sequence of logically contingent decisions, where latent traits and response styles determine the outcomes of theoretical decision nodes that terminate in an observed Likert response. Figure 1 illustrates the structure of an IRTree model for Likert items with six response categories ranging from 0 to 5 (Meiser et al., 2019; see Böckenholt, 2017, for a similar model structure). The first decision node concerns the general agreement (Categories 3, 4, 5) versus disagreement (Categories 0, 1, 2) with the item content. Conditional on the first decision, the second decision refers to the choice of a nonmoderate (i.e., .1 or 4.5) versus moderate (i.e., 2 or 3) category of (dis)agreement. Given a nonmoderate response, the third decision differentiates between extreme (i.e., 0 or 5) and nonextreme (i.e., 1 or 4) (dis)agreement.

In IRTree modeling, the decision nodes are specified as pseudo-items with an appropriate IRT parameterization. The IRTree in Figure 1 contains five binary pseudo-items for each Likert item i. The pseudo-item at the first node Y_{1i} distinguishes between disagreement with $Y_{1i} = 0$ and agreement with $Y_{1i} = 1$, and it can be parameterized, for instance, by a one-parameter logistic function of the substantive trait θ and an intercept (or easiness) parameter δ_{1i} (Rasch, 1960). The probability that person v will agree with item i is thus

$$p(y_{1vi} = 1) = \frac{\exp(\theta_v + \delta_{1i})}{1 + \exp(\theta_v + \delta_{1i})}. \tag{1}$$

Given disagreement or agreement, respectively, the pseudo-items Y_{2i} and Y_{3i} reflect moderate responses close to the midscale for $Y_{2i} = 0$ and $Y_{3i} = 0$, as opposed to nonmoderate responses for $Y_{2i} = 1$ and $Y_{3i} = 1$. Because the choice of a nonmoderate versus moderate response can be driven by an individual response style to prefer or avoid clear-cut responses in combination with the actual strength of (dis)agreement, the pseudo-items Y_{2i} and Y_{3i} are parameterized through a multidimensional IRT model (Reckase, 2009; see Meiser et al., 2019, for details). The response style to prefer nonmoderate over moderate response categories is denoted by η^{nm}, while the strength of agreement is reflected by the substantive trait θ. As stronger agreement with the item content decreases the probability of a nonmoderate disagreement response but increases the probability of a nonmoderate agreement response, the trait θ enters the model equations for Y_{2i} and Y_{3i} with opposite signs. The probabilities of observing a nonmoderate disagreement or agreement response for person v in item i follow as

$$p(y_{2vi} = 1) = \frac{\exp(\eta_v^{\mathrm{nm}} - \alpha^{\mathrm{nm}} \cdot \theta_v + \delta_{2i})}{1 + \exp(\eta_v^{\mathrm{nm}} - \alpha^{\mathrm{nm}} \cdot \theta_v + \delta_{2i})}, \tag{2}$$

and

$$p(y_{3vi} = 1) = \frac{\exp(\eta_v^{\mathrm{nm}} + \alpha^{\mathrm{nm}} \cdot \theta_v + \delta_{3i})}{1 + \exp(\eta_v^{\mathrm{nm}} + \alpha^{\mathrm{nm}} \cdot \theta_v + \delta_{3i})}, \tag{3}$$

where the loading parameter $\alpha^{\mathrm{nm}} > 0$ allows for a different weight of the trait θ in the nuanced selection of responses within the subset of disagreement or agreement categories compared to the overall (dis)agreement decision on the first node (Meiser et al., 2019).

Contingent on a nonmoderate disagreement or agreement response, the subsequent pseudo-items Y_{4i} and Y_{5i} model the selection of a nonextreme category with $Y_{4i} = 0$ and $Y_{5i} = 0$, or an extreme category with $Y_{4i} = 1$ and $Y_{5i} = 1$. Like the previous decision of choosing a (non-)moderate category, the decision to select a (non-)extreme category can be driven by a combination of an individual response style and the substantive trait. The response style to prefer extreme response categories is denoted by η^{e}, and the probabilities of observing an extreme disagreement or agreement response for person v in item i are given by

$$p(y_{4vi} = 1) = \frac{\exp(\eta_v^{\mathrm{e}} - \alpha^{\mathrm{e}} \cdot \theta_v + \delta_{4i})}{1 + \exp(\eta_v^{\mathrm{e}} - \alpha^{\mathrm{e}} \cdot \theta_v + \delta_{4i})}, \tag{4}$$

and

$$p(y_{5vi} = 1) = \frac{\exp(\eta_v^{\mathrm{e}} + \alpha^{\mathrm{e}} \cdot \theta_v + \delta_{5i})}{1 + \exp(\eta_v^{\mathrm{e}} + \alpha^{\mathrm{e}} \cdot \theta_v + \delta_{5i})}. \tag{5}$$

Following the same rationale as in the decision for (non-)moderate responses, the trait θ affects extreme disagree and agree responses with opposite signs, and the loading parameter $\alpha^e > 0$ reflects the weight of θ in the selection of (non-)extreme responses relative to the overall agreement node.

In the model Equations (1)–(5), we used constant loadings of the latent dimensions across rating items i, thereby following the Rasch model because of its parsimony and unique measurement properties of specific objectivity (see also Meiser et al., 2019). The parameterization of the pseudo-items can be naturally extended, however, to include item-specific loading parameters when the strict assumptions of the Rasch model are violated. For the substantive trait θ, variability of loading parameters across rating items can be captured by introducing item-specific discrimination parameters α_i in (1) and item-specific loading parameters α_i^{nm} and α_i^e in Equations (2)–(5), respectively. Because the effects of response styles can be defined as being independent of item content, however, we decided to maintain constant loadings of the response style dimensions η^{nm} and η^e for theoretical reasons (see Böckenholt & Meiser, 2017, for a similar rationale).

1.2 | Modeling heterogeneity in response processes

The probability of a given response category for item i results as the product over the pseudo-items that are involved in the response process along the branches of the IRTree (Böckenholt & Meiser, 2017; Jeon & De Boeck, 2016). Because the model Equations (1)–(5) contain different person parameters across the pseudo-items and even within individual pseudo-items, the IRTree model accommodates quantitative differences between respondents on several dimensions, that is, the trait to be measured and response styles. However, IRTree models preserve the assumption that the relative weighting of the latent dimensions and the structural item parameters are homogeneous, that is, invariant across the population of respondents. Accounting for differences in the response processes and model parameters between subgroups of respondents can be essential, however, for investigating the nature of response behavior and adjusting psychometric measurement to heterogeneous response strategies.

So far, heterogeneity in the parameters of IRTree models has been analyzed with discrete mixture distribution models for latent subpopulations showing different response processes, like subpopulations with a purely trait-based response process as opposed to subpopulations with a combination of trait-based processes and response styles (Alagöz & Meiser, 2023; Khorramdel et al., 2019; Kim & Bolt, 2021; Tijmstra et al., 2018). In the present research, we propose a novel approach to testing for heterogeneity in IRTree models via model-based partitioning along person covariates. The partitioning method makes it possible to detect different relative weights α^{nm} and α^e of the substantive trait θ in the selection of (dis)agreement categories between subgroups of respondents. This makes the new method more flexible than the previous mixture distribution approaches, in that all subgroups of respondents can use combinations of latent trait and response style dimensions θ and η, though with different strengths. In addition, partitioning along observed covariates facilitates the analysis of sources of heterogeneity and enhances the interpretability of identified subgroups.

Although in the present research the main focus is on heterogeneity in the loading parameters α^{nm} and α^e, the partitioning method also allows for detecting the variability of other model parameters like the intercept parameters δ_{1i} to δ_{5i} in Equations (1)–(5). While heterogeneity in α^{nm} and α^e reflects differential strengths in the effects of individual trait values on the choices of disagree and agree responses, heterogeneity in the intercepts of pseudo-items indicates overall trends to select specific kinds of response categories that differ among subgroups. In either case, the analysis of heterogeneity in the structural loading and intercept parameters can reveal distinct characteristics of the underlying response behavior in terms of the differentiatedness and overall tendencies of judgments in subgroups of respondents.

In the following sections, we first outline the score-based partitioning method for testing the invariance of IRT parameters with person covariates. Then we provide evidence for the method's validity for

multidimensional IRTree models in simulation studies and illustrate the analysis of heterogeneity with an empirical application of IRTrees for response styles.

2 | SCORE-BASED TESTS FOR PARAMETER HETEROGENEITY

Score-based tests are a widely used family of statistical tests that aim at checking the invariance of maximum likelihood estimates for statistical models in psychometrics that were originally developed in the fields of statistics and econometrics (Andrews, 1993; Hjort & Koning, 2002; Zeileis & Hornik, 2007). Technically, these tests are closely related to the score test (for a technical and historical overview: Rao, 2005). In the field of IRT, in particular Glas suggested numerous applications of the score test (Glas, 1998, 1999, 2001, 2010; Glas & Suárez-Falcón, 2003; Glas & van der Linden, 2010), although earlier applications of this test in psychometrics can be found (Satorra, 1989). Later work applied score-based tests for checking parameter invariance in a variety of models, including Bradley–Terry models (Strobl et al., 2011), Rasch and IRT models (Debelak & Strobl, 2019; Komboz et al., 2018; Strobl et al., 2015), models of factor analysis (Merkle et al., 2014; Merkle & Zeileis, 2013; Sterner & Goretzko, 2023; Wang et al., 2014), structural equation models (Arnold et al., 2021; Brandmaier et al., 2013, 2016), and multilevel models (Wang & Merkle, 2023). Variations of score-based tests were also proposed for Bayesian estimation methods in item response theory (IRT) (Debelak et al., 2022). In another branch of research, these tests were embedded in algorithms for model-based recursive partitioning (Zeileis et al., 2008), which uses an approach similar to decision trees to detect subpopulations for which model parameters are invariant. For IRT models, such approaches were exemplified by Strobl et al. (2015) and others.

2.1 | Score-based tests in IRT(ree) models

The underlying reasoning of score-based tests in the context of IRT can be described as follows. We consider the responses observed in a sample, together with a metric, ordinal or categorical person covariate, and we are interested in testing the null hypothesis that the item parameter estimates are stable with regard to this person covariate. In a first step, we order the responses with regard to the person covariate. In a second step, we want to investigate fluctuations of the maximum likelihood estimates along the sequence of test takers. This step is based on the premise that for item parameters that are invariant with regard to the person covariate, maximum likelihood estimates do not depend on the value of the person covariate. For any sample taken from the population of respondents, estimates for the item parameters would thus fluctuate randomly around their true values. If, on the other hand, the item parameters are not invariant with regard to a person covariate, there is no single true value of the item parameters, but their true values depend on the value of the person covariate. In this scenario, estimates for the item parameters would fluctuate not around a single true value but around a value that depends on the person covariate. If we would calculate common maximum likelihood estimates for the item parameters, our maximum likelihood estimates would therefore vary systematically from this common estimate.

In the framework of score-based tests, the score, that is, the gradient of the log-likelihood, is used to obtain a measure for the fluctuation of maximum likelihood estimates with regard to a person covariate. By definition, maximum likelihood estimates maximize the log-likelihood; therefore, the gradient of the log-likelihood at the point of the maximum likelihood estimate has to be 0. For the case of IRTree models, we summarize all item parameters of the pseudo-items as a vector $\mathbf{\Psi}$, for which we get estimates $\hat{\mathbf{\Psi}}$. We further use $\mathbf{y}_1, \ldots, \mathbf{y}_N$ to denote the response vectors of the N respondents, which summarize the personwise responses to the individual items.

In the context of IRT and other psychometric models, it can further be shown that the score s of the log-likelihood can be represented as a sum of N individual score contributions, with N being the overall sample size:

$$s(\widehat{\Psi}; \boldsymbol{y}_1, ..., \boldsymbol{y}_N) = \sum_{v=1}^{N} s(\widehat{\Psi}; \boldsymbol{y}_v) = \boldsymbol{0}. \tag{6}$$

Following the outlined reasoning for score-based tests, we now order these individual score contributions with regard to a person covariate of interest and calculate the cumulative sums of the individual score contributions. In what follows, we write the index in round brackets (v) to indicate this ordering of the individual score contributions. To obtain the cumulative sums, we first use $\lfloor \rfloor$, denoting the floor function, and a real value t from the interval $[0, 1]$ to obtain the value $\lfloor Nt \rfloor$. This value represents the count of persons in the lowest t fraction of the sample, ordered by the person covariate. The cumulative sums now correspond to terms of the following form:

$$\sum_{v=1}^{\lfloor Nt \rfloor} s(\widehat{\Psi}; \boldsymbol{y}_{(v)}). \tag{7}$$

We further decorrelate the individual score contributions, so that their covariance matrix over the whole sample is equal to the unit matrix. If $\hat{I}$ is a consistent estimate of the covariance matrix of the individual score contributions, we obtain the following term:

$$B(t, \widehat{\Psi}) = \hat{I}^{-1/2} N^{-1/2} \sum_{v=1}^{\lfloor Nt \rfloor} s(\widehat{\Psi}; \boldsymbol{y}_{(v)}). \tag{8}$$

In this study, we derived $\hat{I}$ by initially computing the individual score contributions of all item parameters for all respondents after the convergence of the item parameter estimation via numerical quadrature, using the `estfun.AllModelClass()` function of `mirt` (Chalmers, 2012). Subsequently, we obtained $\hat{I}$ by estimating the covariance matrix of this sample of individual score contributions, that is, we estimated the covariance matrix using the observed cross-product of the gradients (Falk & Monroe, 2018).

2.2 | Testing for heterogeneity across different kinds of covariates

It can be shown that the distribution of $B(t, \widehat{\Psi})$ can be described by a standard stochastic process, namely, a Brownian bridge (Hjort & Koning, 2002; Zeileis et al., 2008). This property makes it possible to compare the observed distribution of the individual score contributions with the distribution expected under the null hypothesis. At this point, we need to define a test statistic that summarizes any deviations of the observed distribution of the cumulative score contributions from their expected distribution. Depending on the type of person covariate, a canon of test statistics was proposed for this purpose (Merkle et al., 2016; Wang et al., 2014). Important examples include the double-maximum statistic, the Cramer–von Mises statistic, and the maximum Lagrange multiplier statistic for settings with a metric person covariate. If we denote the observed score contributions by a matrix $B(\widehat{\Psi})_{vj}$, with v corresponding to the individual respondents and j to the individual item parameters, these statistics are given by

$$DM = \max_{v} \max_{j} |B(\widehat{\Psi})_{vj}|, \tag{9}$$

$$CvM = \frac{1}{N} \max_{v} \max_{j} (B(\widehat{\Psi})_{vj})^2, \tag{10}$$

$$maxLM = \max_{v} \left\{ \frac{v}{N} \left(1 - \frac{v}{N} \right) \right\}^{-1} \cdot (B(\widehat{\Psi})_{vj})^2, \tag{11}$$

CvM and *maxLM* are only available for models with a limited number of item parameters. Since IRTree models as multidimensional IRT models usually use a large number of item parameters, these statistics would only be applicable for tests with very few items. For this reason, we did not include these statistics in the study at hand. For settings with an ordinal categorical covariate with m categories, there are two test statistics available, namely, the ordered weighted double-maximum test statistic and the maxLM test for ordinal covariates. To define these test statistics, we use $t_l, l = 1, \ldots, m-1$ to denote the proportion of respondents in the first l categories, and $v_l = \lfloor N \cdot t_l \rfloor$. Using this notation, we define these statistics as

$$WDM_o = \max_{v \in \{v_1, \ldots, v_{m-1}\}} \left\{ \frac{v}{N} \left(1 - \frac{v}{N} \right) \right\}^{-1/2} \max_{j=1,\ldots,k} |B(\widehat{\Psi})_{vj}|, \tag{12}$$

$$maxLM_o = \max_{v \in \{v_1, \ldots, v_{m-1}\}} \left\{ \frac{v}{N} \left(1 - \frac{v}{N} \right) \right\}^{-1} \sum_{j=1,\ldots,k} (B(\widehat{\Psi})_{vj})^2. \tag{13}$$

For settings with an unordered categorical covariate with m categories, an unordered Lagrange Multiplier test is available as a test statistic. This statistic is given by Merkle et al. (2014) and Wang et al. (2014)

$$LM_{uo} = \sum_l \sum_j \left(B(\widehat{\Psi})_{v_l j} - B(\widehat{\Psi})_{v_{l-1} j} \right)^2. \tag{14}$$

For any of these test statistics, *p*-values for testing the null hypothesis of parameter invariance can be obtained by analytical considerations ($DM, WDM_o,$ and LM_{uo}) or simulations ($maxLM, CvM, maxLM_o$) via the `strucchange` package in R (Zeileis et al., 2002). Technical details on the calculation of the *p*-values for the various test statistics are provided, for instance, by Wang et al. (2014). By calculating these test statistics only for a subset of the item parameters, it is further possible to check the invariance of individual item parameters or selections of parameters (Schneider et al., 2022). This property can be used for specific tests of invariance of item parameters corresponding to factor loadings and intercepts of pseudo-items in an IRTree model by calculating these test statistics selectively for loading and intercept parameters, respectively.

In the remainder of this article, we will evaluate the approach with simulation studies and an empirical application. In particular, the *DM* statistic is used when testing for parameter heterogeneity in a metric covariate, the LM_{uo} statistic when testing for parameter heterogeneity in an unordered categorical person covariate, and the WDM_o and $maxLM_o$ statistics when testing for parameter heterogeneity in an ordinal person covariate.

After detecting heterogeneity in the model parameters, we further used the following partitioning algorithm, which is based on an algorithm of Strobl et al. (2015), to detect at which value of the person covariate the parameters change. First, we defined all possible split points for the person covariate that allowed us to split the sample into two subsamples of 300 or more respondents. Second, we estimated for each of these split points the item parameters of the IRTree models in both subsamples separately. Here, all item parameters are allowed to differ between both estimated models. This step leads to two IRTree models for each split point. In a final step, we calculate the sum of the log-likelihoods of the two IRTree models for each cutpoint and select the cutpoint for which this sum is maximized. This cutpoint is interpreted as the value where the model parameters change.

3 | GOALS OF CURRENT RESEARCH

The proposed new approach to testing for parameter heterogeneity in IRTree models will allow researchers to identify subgroups of respondents that differ in response behavior and, thus, to tailor psychometric models to optimally measure the traits of interest. In particular, heterogeneity can be observed with respect to the relative weighting of trait-based response processes in the selection of

disagree and agree categories via the parameters α^{nm} and α^{e} (Equations 2–5). Whereas previous mixture distribution approaches for IRTree models analyzed differences in the use of latent trait and response style dimensions in a binary yes-or-no fashion by constraining specific dimension weights to zero (Alagöz & Meiser, 2023; Khorramdel et al., 2019; Kim & Bolt, 2021; Tijmstra et al., 2018), the new score-based partitioning algorithm can identify subgroups with more subtle differences in the influence of response dimensions. Moreover, while the focus of the present research is on differences in the loading parameters α^{nm} and α^{e}, the partitioning algorithm can also reveal heterogeneity in the intercept parameters of pseudo-items δ, reflecting differences between subgroups concerning overall tendencies to select or avoid certain kinds of response categories.

The proposed application of score-based tests differs significantly from previous applications of score-based tests in IRT (e.g., Debelak & Strobl, 2019) in at least three respects. The underlying IRT models are (a) significantly more complex and (b) estimated with the Metropolis–Hastings Robbins–Monro estimation algorithm. Finally, our tests (c) aim at detecting changes in specific item parameters instead of being omnibus tests for parameter heterogeneity. To demonstrate the validity and suitability of the new approach, in what follows, we evaluate score-based tests for IRTree models of response styles in simulation studies and illustrate their application with an empirical example. The R Code for the new method, simulations, and empirical example is available in an OSF repository.

4 | SIMULATION DESIGN

To evaluate the new method, we carried out three simulation studies that investigated the power and Type I error of score-based tests for parameter heterogeneity in IRTree models for response styles between two groups. The first simulation study investigated parameter heterogeneity with regard to an unordered categorical covariate.

We investigated all combinations of the following conditions:

- *Test length*: The test consisted of 30 or 50 pseudo-items. As each rating item is represented by five pseudo-items as defined in Equations (1)–(5), this corresponds to scenarios with 6 or 10 rating items, respectively.
- *Sample size*: The test was worked on by 1000 or 2000 respondents.
- *Factor loading of trait in nonmoderate responses*: The factor loading α^{nm} for the tendency to give nonmoderate responses (Equations 2 and 3) was either .2 or .8.
- *Factor loading of trait in extreme responses*: The factor loading α^{e} for the tendency to give extreme responses (Equations 4 and 5) was either .2 or .8.
- *Parameter heterogeneity*: In our simulations, either all model parameters were invariant or the parameters α^{nm} and α^{e} increased or decreased by .2 or by .5 for all respondents of the second group, which represents the focal group.

Each respondent was randomly assigned to one of the two groups, with each member having a probability of .5 of being assigned to each group. There were two latent personality traits of interest, of which one had nonzero factor loadings on the first and the other on the second half of the rating items. Together with the two latent factors that modeled response styles, we therefore obtained four latent dimensions, which were assumed to be drawn from a multivariate normal distribution with mean vector **0**. The variance covariance matrix was chosen to be

$$\mathrm{Cov} = \begin{pmatrix} 2.89 & 1.02 & .00 & .00 \\ 1.02 & 1.44 & .00 & .00 \\ .00 & .00 & 1.00 & .75 \\ .00 & .00 & .75 & 2.25 \end{pmatrix},$$

with latent correlations of $\rho = .5$ between the two traits (Dimensions 1 and 2) and the two response styles (Dimensions 3 and 4). The simulated scenario thereby resembled earlier empirical analyses with two correlated traits of different variance and with latent dimensions of nonmoderate and extreme response styles, where the variance of extreme responding exceeds the variance of nonmoderate responding (Böckenholt & Meiser, 2017; Meiser et al., 2019).

The intercept parameters were drawn from a uniform distribution $U(-1, 1)$ for δ_{i1} in the agreement node, from a uniform distribution $U(-1, .5)$ for δ_{i2} and δ_{i3} in the nonmoderate response nodes, and from a uniform distribution $U(-2.5, -1.5)$ for δ_{i4} and δ_{i5} in the extremity nodes (Equations 1–5). The slope and intercept parameters remained constant for all conditions with tests of the same length.

The second simulation study investigated the detection of parameter heterogeneity with regard to a metric covariate, which was drawn from a uniform distribution $U(20, 80)$. If parameter heterogeneity was present, it affected the persons up to a cutoff value of 40. The other settings were identical to those of the first simulation study.

In the third simulation study, we focused on detecting parameter heterogeneity with regard to an ordinal covariate, which was represented by integer values ranging from 1 to 5. We simulated two distinct scenarios based on the probability distribution of observing each category of the covariate. In the first scenario, each category had an equal probability of .2, leading to a uniform distribution across the five possible values. In the second scenario, we implemented a nonuniform distribution, characterized by the relative frequencies of the categories as follows: .1 for Categories 1 and 2, .35 for Categories 3 and 4, and .1 for Category 5. If parameter heterogeneity was present, the item parameters differed between persons with a covariate category of 3 or less and persons with a category of 4 or higher. The other settings were identical to those of the first two simulation studies.

For each setting of the three simulation studies, we applied two separate score-based tests to test for parameter heterogeneity in α^{nm} and α^e, leading to two p-values in each simulated data set. These tests considered only the individual score contributions for the respective item parameters and thus implemented the approach outlined in the introduction, where test statistics of score-based tests are only calculated for a subset of the item parameters. We evaluated the rate of p-values below .05 for each setting, which led to an evaluation of the Type I error and the power of both statistical tests under each setting.

In the second and third simulation studies, we further investigated the accuracy of the algorithm used for the split point detection. In this evaluation, we first determined the split point based on the algorithm outlined in the previous section if true parameter heterogeneity was detected by the corresponding statistical test. The result of the second statistical test was ignored for this evaluation. For each condition with simulated parameter heterogeneity effects and positive statistical tests, we calculated the mean and the standard deviation of the split points and compared it to the true value of 40 in the second simulation study and 3 in the third simulation study.

In all simulation studies, the model parameters were estimated using the Metropolis–Hastings Robbins–Monro algorithm (Cai, 2010) using the default settings for this algorithm in `mirt`.[1] This algorithm is computationally more efficient for parameter estimation in high-dimensional IRT models, such as IRTree models, than classical methods such as marginal maximum likelihood estimation based on an expectation-maximization algorithm (Bock & Aitkin, 1981). For scaling the item and person parameters, this algorithm uses a normal prior with mean 0 for the person parameters of the multidimensional IRT model. The variance covariance matrix of the latent variables was freely estimated in all simulations.

---

[1]At the time of writing this paper, these default settings use 150 burn-in cycles in Stage 1 of the algorithm, 100 SEM cycles in Stage 2, and 5 Metropolis–Hastings draws at each iteration. The estimation algorithm uses a maximum of 2000 iterations and a convergence threshold of .001.

5 | RESULTS

We first discuss results on the power and the Type I error rate, and then we discuss the accuracy of the split point detection. Additional plots are presented in the files provided as supporting information for this article.

5.1 | Type I error and power

If parameter heterogeneity was absent in the factor loading tested for invariance, the Type I error was slightly increased, but usually close to the nominal alpha level of .05. This overall result was found independently of which factor loading was tested for changes and also independently of whether heterogeneity was present in the other factor loading. As an illustration, we present the results for conditions with a metric covariate in the following Figure 2 when testing for parameter changes in α^e under conditions with no parameter changes in both factor loadings. For conditions investigating categorical covariates with the DM statistic, the Type I error was usually close to the nominal alpha level. Under conditions with an ordinal person covariate, the score-based tests based on the test statistics WDMo and maxLMo usually had a slightly increased Type I error, in particular when the distribution of the covariate categories was skewed.

When a parameter change of .5 was present in a factor loading tested for heterogeneity, the power of the corresponding test was usually close to 1. We therefore focus on the results of tests in conditions with a parameter change of .2, which are summarized in the following Figures 3 and 4. As can be seen, the tests had considerable power to detect this type of effect. The power to detect parameter changes in α^{nm} was slightly larger than that to detect parameter changes in α^e. This difference was due to the hierarchical nature of the IRTree model, where all rating responses provide information for the pseudo-items

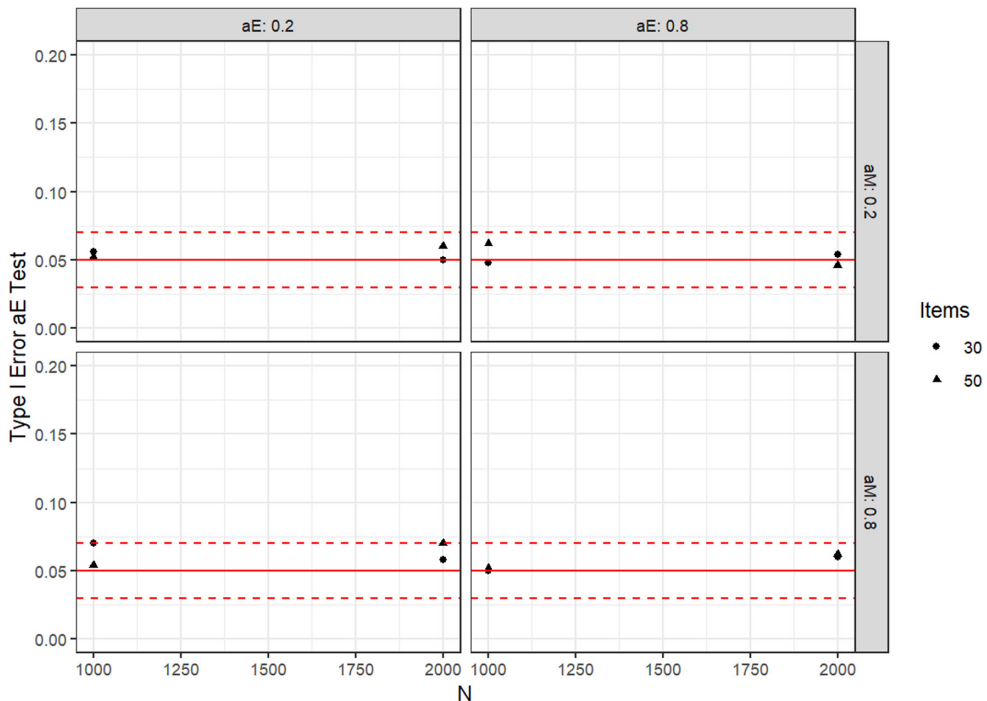**FIGURE 2** Type I error when testing for parameter changes in α^e when all parameters are invariant. Red lines indicate nominal alpha level of .05 and approximate 95% confidence interval for the hit rates.
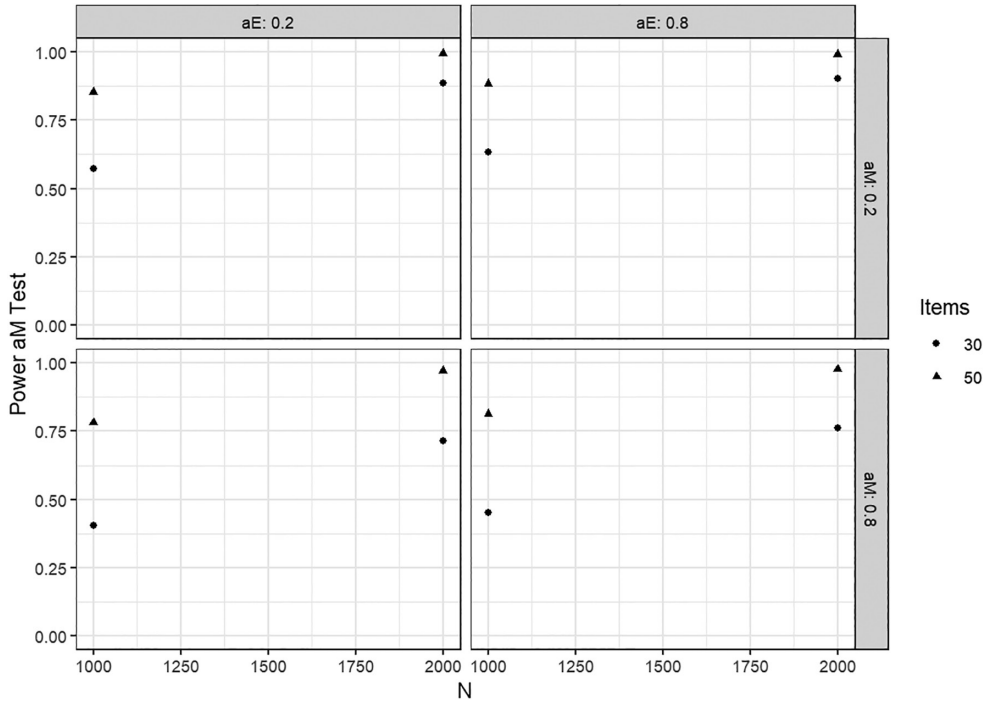
FIGURE 3 Power when testing for parameter changes in α^{nm} when this parameter changed by +.2.
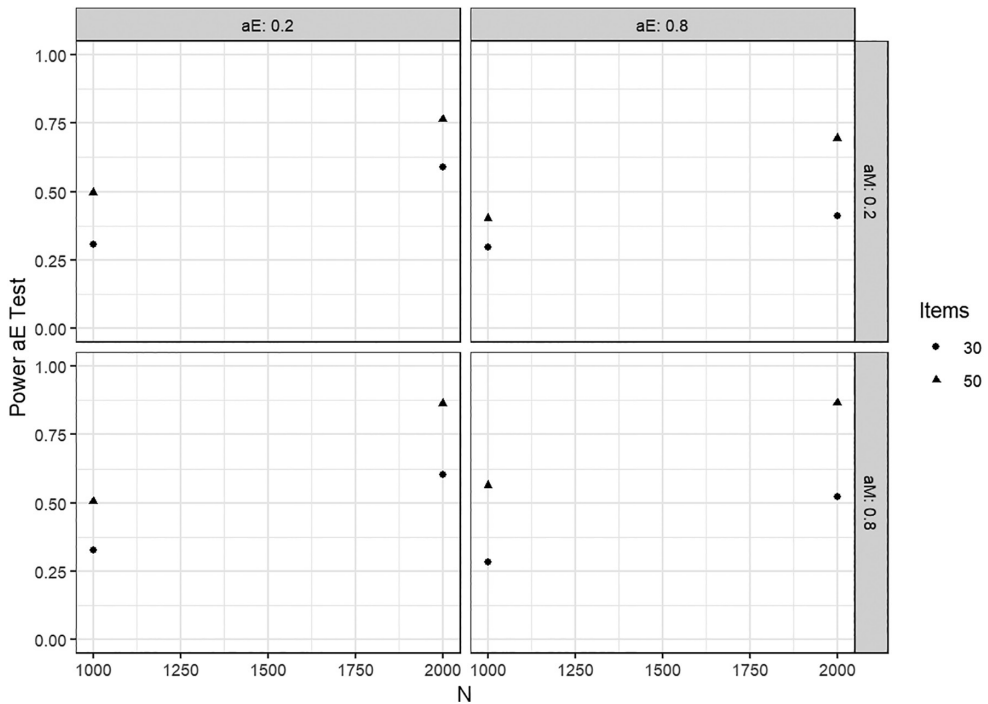

FIGURE 4 Power when testing for parameter changes in α^{e} when this parameter changed by +.2.

of nonmoderate responding, but only a subset of responses provides information for the pseudo-items of extreme responding (Figure 1). Under conditions with a larger sample size and test length, the power was slightly increased overall. Again, very similar results were found in the simulation studies with categorical and ordinal covariates, and we omit details for brevity.

5.2 | Cutpoint selection

Under all conditions, the mean cutpoint proposed for splitting the groups affected by parameter heterogeneity was larger than the true value of 40, but lower than the value of 50, which would split the sample into two groups of the same size. This bias decreased with the sample size and with the size of the actual parameter change. As an illustration, Figures 5 and 6 present detailed results for the cutpoint selection for parameter heterogeneity tests for α^e when this parameter was affected by a parameter change of +.2 and +.5, respectively. Analogous results were obtained for changes in α^{nm} and in the simulation study with an ordinal covariate.

6 | EMPIRICAL EXAMPLE

Eid and Rauber (2000) analyzed data from an organizational survey with a six-item questionnaire on employees' satisfaction with their superiors. Responses were observed on a six-point rating scale, and the authors analyzed the data with a mixture distribution partial credit model (Rost, 1991). Model selection via the Bayesian information criterion revealed a two-class solution, where the class-specific threshold parameters indicated a response tendency to use nonextreme response categories for the larger class and a tendency toward extreme categories for the smaller class. Moreover,
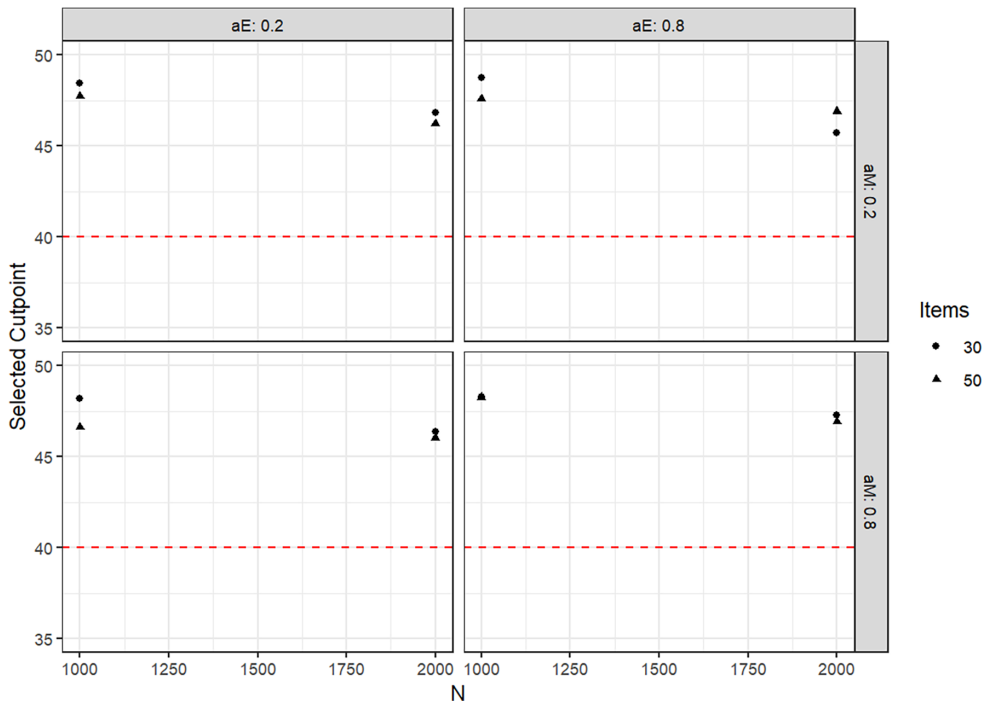

FIGURE 5 Mean proposed cutpoint under different conditions of sample size, test length, and factor loadings when α^e was affected by a parameter change of +.2. The red line denotes the true cutpoint of 40.
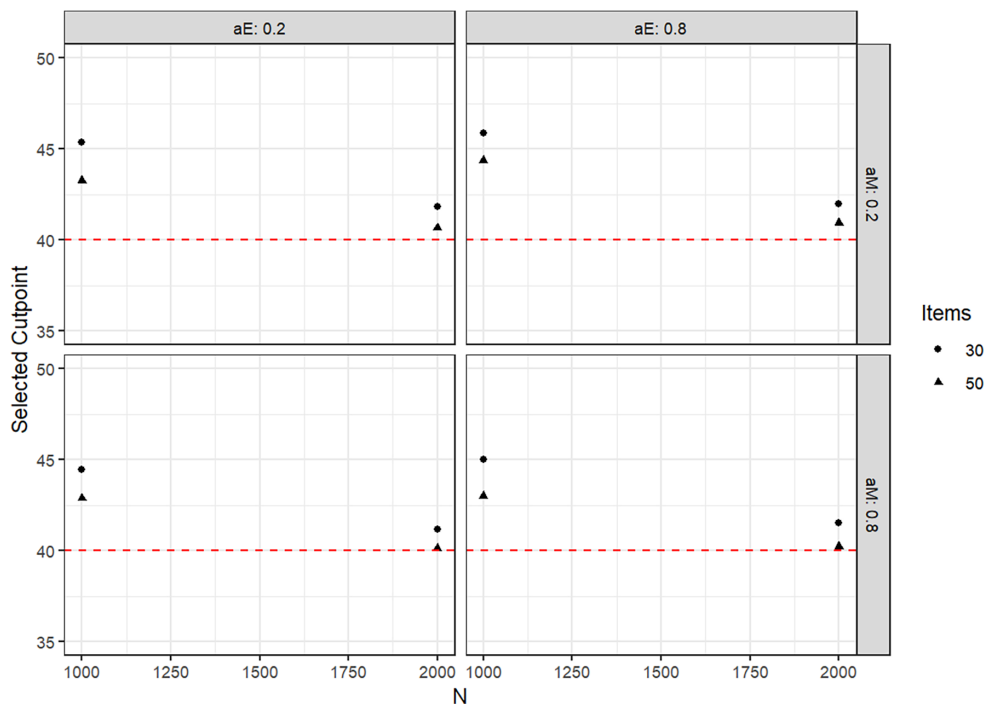
FIGURE 6 Mean proposed cutpoint under different conditions of sample size, test length, and factor loadings when α^e was affected by a parameter change of +.5. The red line denotes the true cutpoint of 40.

the authors found relationships of class assignment with demographic and job-related person co-variates, including the covariate "length of service on the same position" with five ordinal categories ($\leq$6, 7–12 months, 1–5, 5–10, >10 years).

Here we reanalyze the data from Eid and Rauber (2000) with the IRTree model in Figure 1 and test for parameter heterogeneity using the new score-based partitioning method.[2] Unlike the mixed partial credit model in the original analysis, the IRTree model accommodates nonmoderate and extreme response styles as continuous dimensions, so that parameter heterogeneity is disentangled from individual response styles. A sample of $N = 3964$ respondents was available with observed values in the six items and in the covariate "length of service on the same position." Parameter esti-mates for the IRTree model in Figure 1 were obtained with the Metropolis–Hastings Robbins–Monro algorithm of the R package mirt using default convergence settings (Chalmers et al., 2023), that is, a convergence threshold of .001. The joint distrbution of the person parameter vector was specified as $(\theta, \eta^{nm}, \eta^e)' \sim N(0, \Sigma)$ with freely estimated variance covariance matrix Σ. Parameter estimation of the IRTree model in Figure 1 showed loadings for the trait θ in the pseudo-items of nonmoderate responding (Equations 2 and 3) and extreme responding (Equations 4 and 5) of $\widehat{\alpha}^{nm} = .761$ and $\widehat{\alpha}^e = .613$. These loadings refer to the common trait θ in the model equations for the agreement, nonmoderate, and extreme nodes of the IRTree and can be interpreted in relation to the loading of 1.0 in the agreement node (Equation 1). The results therefore replicate earlier findings that the trait has a smaller weight for gradual choices among (dis)agreement categories relative to the pseudo-item of general agreement versus disagreement (Meiser et al., 2019).

Score-based tests for parameter heterogeneity in α^{nm} and α^e with the person covariate "length of ser-vice on the same position" and the test statistic LM_{uo} in Equation (14) indicated significant heterogeneity

---

[2]The original data can be retrieved from: https://www.hogrefe.com/us/ejpa/special-features.

for α^{nm}, $f(efp) = 9.80$, $p = .007$, but not for α^{e}, $f(efp) = .29$, $p = .961$. Analogously to the tests evaluated in the simulation studies, the score-based tests used for the empirical data examined parameter heterogeneity in individual loading parameters. As a next step, we applied the partitioning algorithm to find respondent groups for which the item parameter vector of the IRTree model is invariant. The partitioning algorithm located the cutoff point at length of service up to 5 years as opposed to longer than 5 years. Subsequent score-based tests within the two subgroups did not reveal further significant heterogeneity in α^{nm} for respondents with length of service up to 5 years, $f(efp) = 1.01$, $p = .482$, or for respondents with length of service longer than 5 years, $f(efp) = 3.23$, $p = .071$.

The resulting subgroups showed estimated trait loadings of $\widehat{\alpha}^{nm} = .777$ and $\widehat{\alpha}^{nm} = .754$ for respondents up to 5 years and more than 5 years of service on the same position, respectively. The estimates suggest that respondents with a shorter period on the same position engage in a stronger trait-based response process in their choices of nonmoderate versus moderate (dis)agreement categories relative to their general (dis)agreement decision compared to respondents with a longer period on the same position. Although the score-based test was only significant for the loading parameter α^{nm}, descriptively we found the same pattern for α^{e} in the choices between extreme versus nonextreme (dis)agreement categories, with $\widehat{\alpha}^{e} = .653$ and $\widehat{\alpha}^{e} = .587$ for respondents up to 5 years or more than 5 years on the same position.

In addition to heterogeneity in the factor loadings, the partition of the sample also captures heterogeneity in the intercept parameters of the pseudo-items. In contrast to differences in the loading parameters, which reflect the impact of individual differences in θ on the observed responses, differences in the intercept parameters indicate general tendencies to select or avoid certain kinds of categories that characterize subgroups of respondents. For the agreement node, the mean estimated intercept parameter δ_{1i} across the six rating items was 2.100 for the first subgroup and 1.800 for the second subgroup, indicating that respondents with up to 5 years of service had a stronger tendency to endorse the positive statements about their superiors compared to respondents with longer periods of service. The mean intercepts of the nonmoderate disagreement and agreement nodes δ_{2i} and δ_{3i} were -2.473 and $.335$ for the first subgroup and -2.376 and $.160$ for the second group, suggesting that respondents with shorter service on the same position had stronger overall tendencies to avoid clear-cut disagreement and to show clear-cut agreement than respondents who had served longer on the same position. Finally, the mean intercepts for extreme disagreement and agreement decisions δ_{4i} and δ_{5i} were -3.782 and -1.788 for the first subgroup and -2.959 and -2.283 for the second subgroup, which signals a stricter tendency to avoid extreme disagreement and a less strict tendency to avoid extreme agreement categories for respondents with shorter service on the same position. Together the differences in the various intercept parameters reveal that respondents with up to 5 years of service were more inclined to avoid negative ratings and to give favorable ratings than respondents with more than 5 years of service.

It has recently been demonstrated that differences in structural item parameters along the branches of sequential IRTree decision nodes may reflect changes in the underlying person distribution (Lyo et al., 2023), which, among other factors, could be due to sample selection effects (Meiser & Reiber, 2023). Likewise, in the present context of model-based partitioning, differences in the item parameters between subgroups can mirror variations in the person distribution, as is obvious for the reported differences in the intercept parameters: Inasmuch as the differences consistently indicated a tendency to less negative and more positive ratings for the first subgroup relative to the second subgroup, the differences in the intercept parameters may be driven by higher average values for the trait θ in the first group. That is, because the expectation of θ is fixed to zero in either group for technical reasons of identifiability, group differences in θ may manifest themselves in shifts of threshold parameters. To demonstrate this point, and to safeguard our interpretation of the differences in loading parameters against alternative interpretations as mere reflections of alterations in the person distribution, we conducted an auxiliary multigroup analysis. In this analysis, we fixed the expectation vector of the person parameters to zero only for one subgroup and constrained the intercept parameters to be equal across subgroups while leaving the variance covariance matrix free between groups. This specification allowed us to guarantee parameter identification while at the same time accounting for potential differences in

the person distribution. In particular, the expectation of θ was estimated as a free parameter for the subgroup of respondents with up to 5 years of service on the same position. The estimated expectation of θ for this subgroup increased to .324 in the multigroup analysis, thereby showing the greater satisfaction of respondents on the trait level. The higher trait level may indicate a change effect with greater satisfaction at a more recently acquired position that deteriorates over time, or it may be due to a selection effect such that employees with higher satisfaction are more likely to be promoted to the next position within a period of 5 years. At the same time, the expectations of nonmoderate and extreme response styles η^{nm} and η^e were estimated at $-.240$ and $-.438$ in the subgroup of respondents with up to 5 years of service, indicating more moderate and less extreme response behavior presumably due to less judgment confidence given the shorter experience with superiors. Importantly, however, the direction of the differences in the loading parameters α^{nm} and α^e on the trait θ was replicated in the modified model specification, with $\hat{\alpha}^{nm} = .810$ and $\hat{\alpha}^{nm} = .737$ and $\hat{\alpha}^e = .705$ and $\hat{\alpha}^e = .565$ in the subgroups with up to 5 years versus more than 5 years of service.

Finally, we used a model with item-specific loading parameters on the substantive trait θ to demonstrate the score-based testing and partitioning approach in a generalized framework. For this purpose, the variance of the trait dimension θ was fixed at 1.0, and item-specific loading parameters on this dimension were specified for the pseudo-items in Equations (1)–(5). Score-based tests were applied to the subsets of loading parameters referring to the pseudo-items of nonmoderate responding and extreme responding, respectively. Replicating the previous findings, the tests revealed significant heterogeneity along the person covariate for the loading parameters in the pseudo-items for nonmoderate responding, $f(\textit{efp}) = 17.26$, $p = .031$, but not for extreme responding, $f(\textit{efp}) = 13.68$, $p = .118$. The partitioning algorithm identified the cutoff at the same point as before, that is, between up to 5 years as opposed to more than 5 years of service on the same position. Mean estimated loading parameters for the trait dimension θ across the six rating items were 3.001, 2.216, and 1.776 for the agree, nonmoderate, and extreme pseudo-items in the first subgroup and 3.369, 2.340, and 1.800 in the second subgroup. These results indicate a larger differentiatedness of general agreement responses in the second subgroup with a longer period of service, presumably due to more experience with their superiors and, thus, higher judgment certainty. In line with the previous results, however, the trait loadings showed a steeper decrease from the superordinate agreement node to the nonmoderate and extreme decision nodes in the second subgroup with more than 5 years of service compared to the first subgroup with up to 5 years of service on the same position.

7 | DISCUSSION

IRTree models can accommodate quantitative individual differences in response styles, such as nonmoderate or extreme response tendencies, and thereby allow researchers to disentangle the measured traits of interest from additional response processes (Böckenholt, 2017; Böckenholt & Meiser, 2017; Khorramdel & von Davier, 2014; Meiser et al., 2019; Plieninger & Meiser, 2014). Like other IRT model families, however, IRTree models are based on the premise that the structural item parameters are homogeneous across the entire population of respondents, so that the loading weights of the latent dimensions are assumed invariant over subgroups. As a consequence, standard IRTree models cannot account for heterogeneity in the composition or weighting of response processes across respondents. Such heterogeneity can be essential, however, for theoretical and practical reasons. First, existing sources of heterogeneity are of interest for analyzing the nature of response processes that underlie observed rating responses in psychological assessment and surveys. Second, accounting for heterogeneity in the parameters affords a tailored measurement model for the target traits that optimally fits the individual response sequence.

In this research, we introduced and evaluated a partitioning approach for IRTree models that captures parameter heterogeneity in pseudo-items as a function of observed person covariates. The proposed partitioning method builds on score-based tests that have been used for tests of parameter invariance

in unidimensional IRT models with maximum likelihood estimation (Debelak & Strobl, 2019; Komboz et al., 2018; Schneider et al., 2022; Strobl et al., 2015). Here we extended this method to IRTree models with a multidimensional latent space across and within pseudo-items and to parameter estimation with a Metropolis–Hastings Robbins–Monro algorithm that is recommended for complex models of high dimensionality (Cai, 2010) and that can be used in the R package mirt (Chalmers, 2012; Chalmers et al., 2023). In contrast to earlier studies on score-based tests and model-based recursive partitioning (e.g., Debelak & Strobl, 2019), our study also focused on tests that were sensitive to heterogeneity in individual item parameters. This possibility was previously mentioned by Schneider et al. (2022) but has not yet been systematically evaluated.

The simulation studies generally demonstrated accurate Type I error rates and sufficient power of the score-based tests for invariance of the factor loadings, which formed the focus of the present research. We only found a handful of conditions where the Type I error was slightly increased, for instance, when parameter heterogeneity with respect to an ordinal person covariate was investigated. This finding might be related to the use of the Metropolis–Hastings Robbins–Monro algorithm that might cause numerical inaccuracies in the resulting score-based tests. An anonymous reviewer pointed us to the possibility that additional fine-tuning of this algorithm (Ju & Falk, 2019; Monroe & Cai, 2014) and other options for estimating the covariance matrix of the individual score contributions (Falk & Monroe, 2018; Guastadisegni et al., 2022; Paek & Cai, 2014) might also help to obtain accurate Type I error rates under these conditions. Importantly, our simulation studies demonstrated that the score-based tests were able to distinguish between parameter heterogeneity in α^{nm} and α^{e}. The empirical example illustrated that the partitioning algorithm could reveal heterogeneity between subgroups in the factor loadings for real data, such that variation was detected in the weighting of the substantive trait for gradual category choices of disagreement and agreement categories. In addition, general tendencies to prefer or avoid specific categories were reflected in the estimated intercept parameters of the subgroups and may indicate differences in the latent person distributions. Unlike the mixture distribution models proposed to capture heterogeneity in IRTree models (e.g., Alagöz & Meiser, 2023; Kim & Bolt, 2021), the model-based partitioning approach allows that both substantive traits and response styles contribute to the decisions in the pseudo-items for all subgroups. Moreover, because the partitioning algorithm exploits extraneous person covariates, it facilitates the analysis of sources or concomitants of parameter heterogeneity, and it circumvents common problems of mixture distribution models like local maxima and extensive sample size requirements.

In this paper, we introduced score-based tests to analyze heterogeneity in the factor loadings of substantive traits by means of an IRTree model that accounts for nonmoderate and extreme response styles in six-point rating responses (Böckenholt, 2017; Meiser et al., 2019). The proposed score-based testing approach and corresponding partitioning algorithm can easily be adapted to other IRTree model structures and parameters, however, like IRTree models for four-, five-, or seven-point rating scales and models with estimated loadings for both trait and response style dimensions (Khorramdel & von Davier, 2014; Merhof & Meiser, 2023; Plieninger & Meiser, 2014). Over and above generalizing the score-based partitioning algorithm to different IRTree model structures and model specifications, future research could extend the systematic investigation of its statistical properties like Type I error rate and power for tests of invariance of other structural parameters in IRTree modeling. While our focus was on heterogeneity in the factor loadings of the substantive trait, future analysis might focus on heterogeneity in the intercept parameters or in the loading parameters for other response processes.

To conclude, this research introduced a new approach to testing for parameter heterogeneity in multidimensional IRTree models. The new method has proven to be a promising tool that can be used to investigate the nature and sources of differences in response processes and to account for such differences in the measurement of psychological traits.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
The simulated data and the code that support the findings of this study are openly available in OSF at https://osf.io/fpt7e/?view_only=e47933453ded45b0b5434ceb37409b75. The data of the empirical example can be obtained from https://www.hogrefe.com/us/ejpa/special-features.

COMPUTATIONAL DETAILS
All results were obtained using the R system for statistical computing version 4.4.0 (R Core Team, 2024), employing the add-on packages (in alphabetical order) `mirt` version 1.41 (Chalmers, 2012), `SimDesign` version 2.15 (Chalmers & Adkins, 2020), and `strucchange` version 1.5-3 (Zeileis et al., 2002) in the simulation study. The figures were created using `ggplot2` (Wickham, 2016) version 3.5.1.

ORCID
Rudolf Debelak https://orcid.org/0000-0001-8900-2106

REFERENCES
Alagöz, Ö. E. C., & Meiser, T. (2023). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, *84*(5), 957–993. https://doi.org/10.1177/00131644231206765

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*(4), 821–856. https://doi.org/10.2307/2951764

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

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83. https://doi.org/10.1037/met0000106

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541. https://doi.org/10.1037/met0000016

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

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

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

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chalmers, R. P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2023). *mirt: Multidimensional item response theory (R package)*. https://CRAN.R-project.org/package=mirt

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. https://doi.org/10.18637/jss.v048.c01

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. https://doi.org/10.1111/bmsp.12275

Debelak, R., & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement*, *79*(2), 385–398. https://doi.org/10.1177/0013164418777784

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20–30. https://doi.org/10.1027/1015-5759.16.1.20

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. https://doi.org/10.1037/met0000059

Falk, C. F., & Monroe, S. (2018). On Lagrange multiplier tests in multidimensional item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement*, *78*(4), 653–678. https://doi.org/10.1177/0013164417714506

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*(3), 647–667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*(3), 273–294. https://doi.org/10.1007/BF02294296

Glas, C. A. W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M. A. van Duijn, & T. A. Sniders (Eds.), *Essays on item response theory* (pp. 131–148). Springer.

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 269–288). Springer.

Glas, C. A. W., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. https://doi.org/10.1177/0146621602250530

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 603–626. https://doi.org/10.1348/000711009X481360

Guastadisegni, L., Cagnone, S., Moustaki, I., & Vasdekis, V. (2022). Use of the lagrange multiplier test for assessing measurement invariance under model misspecification. *Educational and Psychological Measurement*, *82*(2), 254–280. https://doi.org/10.1177/00131644211020355

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, *25*(5), 560–576. https://doi.org/10.1037/met0000249

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, *14*(1–2), 113–132. https://doi.org/10.1080/10485250211394

Jeon, M., & De Boeck, P. (2016). A general item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138. https://doi.org/10.1177/0013164413498876

Ju, U., & Falk, C. F. (2019). Modeling response styles in cross-country self-reports: An application of a multilevel multidimensional nominal response model. *Journal of Educational Measurement*, *56*(1), 169–191. https://doi.org/10.1111/jedm.12205

Khorramdel, L., Pokropek, A., & von Davier, M. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, *72*, 538–559. https://doi.org/10.1111/bmsp.12179

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*(2), 161–177. https://doi.org/10.1080/00273171.2013.866536

Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

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

Lyo, W., Bolt, D. M., & Westby, S. (2023). Exploring the effects of item-specific factors in sequential and IRTree models. *Psychometrika*, *88*(3), 745–775. https://doi.org/10.1007/s11336-023-09912-x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 501–516. https://doi.org/10.1111/bmsp.12158

Meiser, T., & Reiber, F. (2023). Item-specific factors in IRTree models: When they matter and when they don't (commentary on Lyu, Bolt & Westby, 2023). *Psychometrika*, *88*(3), 739–744. https://doi.org/10.1007/s11336-023-09916-7

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. Eleven International.

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, *88*(4), 1354–1380. https://doi.org/10.1007/s11336-023-09901-0

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., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods*, *21*(2), 151–163. https://doi.org/10.1037/met0000038

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

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343–369. https://doi.org/10.1177/0013164413499344

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, *74*(1), 58–76. https://doi.org/10.1177/0013164413500277

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

R Core Team. (2024). *R: A language and environment for statistical computing [computer software manual]*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, C. R. (2005). Score test: Historical review and recent developments. In N. Balakrishnan, H. N. Nagaraja, & N. Kannan (Eds.), *Advances in ranking and selection, multiple comparisons, and reliability: Methodology and applications* (pp. 3–20). Birkhäuser Boston. https://doi.org/10.1007/0-8176-4422-9_1

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Reckase, M. (2009). *Multidimensional item response theory*. Springer.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75–92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131–151. https://doi.org/10.1007/BF02294453

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. https://doi.org/10.3758/s13428-021-01689-0

Schoenmakers, M., Tijmstra, J., Vermunt, J., & Bolsinova, M. (2024). Correcting for extreme response style: Model choice matters. *Educational and Psychological Measurement*, *84*(1), 145–170. https://doi.org/10.1177/00131644231155838

Sterner, P., & Goretzko, D. (2023). Exploratory factor analysis trees: Evaluating measurement invariance between multiple covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 871–886. https://doi.org/10.1080/10705511.2023.2188573

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., 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. https://doi.org/10.5282/ubm/epub.10588

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Wang, T., & Merkle, E. (2023). *Self-normalized score-based tests to detect parameter heterogeneity for mixed models*. https://arxiv.org/abs/2302.14275

Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Frontiers in Psychology*, *5*, 438. https://doi.org/10.3389/fpsyg.2014.00438

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, *48*(4), 441–456. https://doi.org/10.1111/j.1745-3984.2011.00154.x

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

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 and Graphical Statistics*, *17*(2), 492–514.

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

SUPPORTING INFORMATION

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

How to cite this article: 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*, *00*, 1–20. https://doi.org/10.1111/bmsp.12367