Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: www.tandfonline.com/journals/hsem20

Bayesian Hierarchical Moderated Factor Analysis for Testing Measurement Invariance in Multilevel Data: Model Development, Simulation Studies, and Experience Sampling Application

Julian F. Lohmann, Steffen Zitzmann, Martin Hecht, Christoph Niepel & Esther Ulitzsch

To cite this article: Julian F. Lohmann, Steffen Zitzmann, Martin Hecht, Christoph Niepel & Esther Ulitzsch (30 Apr 2026): Bayesian Hierarchical Moderated Factor Analysis for Testing Measurement Invariance in Multilevel Data: Model Development, Simulation Studies, and Experience Sampling Application, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2026.2642783

To link to this article: https://doi.org/10.1080/10705511.2026.2642783

Published online: 30 Apr 2026.

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2026.2642783

Bayesian Hierarchical Moderated Factor Analysis for Testing Measurement Invariance in Multilevel Data: Model Development, Simulation Studies, and Experience Sampling Application

Julian F. Lohmann[a], Steffen Zitzmann[b], Martin Hecht[c], Christoph Niepel[d], and Esther Ulitzsch[e]

[a]Leibniz Institute for Science and Mathematics Education; [b]Medical School Hamburg; [c]Helmut Schmidt University; [d]University of Luxembourg; [e]University of Oslo

ABSTRACT

Moderated Nonlinear Latent Factor Analysis (MNLFA) has been introduced as a flexible approach for testing measurement invariance among categorical and continuous covariates. Equipped with Bayesian shrinkage priors, MNLFA can handle large numbers of covariates and potentially invariant item parameters. The present study extends the capabilities of the Bayesian MNLFA to multilevel and longitudinal confirmatory factor analysis. We show how a Bayesian hierarchical MNLFA (BH-MNLFA) can be implemented and provide two simulation studies to demonstrate its functionality. Focusing on invariance explorations in experience sampling data as a potential use case in the context of longitudinal data analysis, we showcase the utility of BH-MNLFA with data from educational psychology, and test invariance of state self-concepts measures across time and school subjects.

KEYWORDS

Bayesian modeling; confirmatory factor analysis; measurement invariance; moderated nonlinear latent factor analysis; multilevel modeling

1. Introduction

The measurement of latent variables, such as intelligence, personality, motivation, and emotion, using multiple indicators (i.e., test or questionnaire items) is ubiquitous in psychological research. Latent factor models are typically employed to relate item responses to the latent variable of interest that itself is not directly observable. The (estimated) item parameters of latent factor models reflect how items relate to the latent variable. However, in heterogeneous populations, the interrelation between specific items and the latent factor can vary across individuals. This phenomenon is well known under the terms *measurement non-invariance* or *differential item functioning* (e.g., Meredith, 1993). Several statistical techniques for investigating measurement invariance (MI) have been proposed (e.g., Bauer, 2017; Millsap, 2012). These approaches explore whether item parameters vary as a function of relevant covariates such as group memberships. Among those, moderated nonlinear latent factor analysis (MNLFA) stands out as a flexible and powerful approach for examining MI that overcomes several limitations of other MI approaches (see Bauer, 2017). In essence, MNLFA allows researchers to test whether and how the relationships between observed variables and latent factors vary across different levels of a set of covariates. Thereby, MNLFA makes it possible to test MI not only as a function of categorical but also continuous covariates. What is more, in MNLFA all parameters of a measurement model (item loadings, item intercepts, residual variances, latent mean, and variance) can be moderated by these covariates, going beyond the capabilities of existing approaches. To additionally overcome identification concerns that require preselecting model parameters as being invariant (so-called *anchor items*), it has recently been proposed to combine the MNLFA approach with (Bayesian) regularization techniques (Brandt et al., 2025; Chen et al., 2022, and see also Bauer et al., 2020 for a non-Bayesian approach). These regularization techniques allow for the estimation of weakly identified models and, thus, make typically unjustified assumptions of MI for some preselected model parameters unnecessary (Bauer, 2023; Brandt et al., 2025; Robitzsch, 2022).

In the present study, we expand the capabilities of MNLFA to the hierarchical data structures often encountered in psychological research, for example, when students are nested in classes or measurement repetitions are nested within persons. We do so by combining MNLFA with multilevel confirmatory factor analysis (MLCFA; Muthén, 1991; Rabe-Hesketh et al., 2004). In addition to supporting MI investigations in hierarchical data scenarios, this newly proposed approach will also allow researchers to consider potential moderators at both the within- and between-level. Thereby, this model extension can provide a more nuanced understanding of measurement invariance across different hierarchical levels. For instance, in a longitudinal study, where measurements are nested within persons, person characteristics (e.g., age,

CONTACT Julian F. Lohmann lohmann@leibniz-ipn.de Leibniz Institute for Science and Mathematics Education, Kiel, Olshausenstrasse 62, 24118 Kiel, Germany.

gender) on the between-level as well as situational characteristics (e.g., day time, social context) on the within-level could moderate item parameters. In an educational study with students nested within classes, school and classroom characteristics (e.g., class size, school track) on the between-level and student-specific factors (e.g., gender, socioeconomic background) on the within-level could be examined as moderators of item functioning.

To facilitate the exposition of BH-MNLFA, we focus on hierarchical data structures arising from intensive longitudinal data with repeated measurements—specifically, its application to Ecological Momentary Assessments (EMA; Stone & Shiffman, 1994; Trull & Ebner-Priemer, 2013). EMA has attracted growing interest in psychological research in recent years. A defining feature of EMA studies is that participants repeatedly respond to questionnaires at multiple time points over a period of days or weeks throughout their daily lives. These assessments are often conducted in a variety of naturally occurring contexts (e.g., at work, in class, at home). Such data can provide a rich basis for nuanced analyses of within-person processes and interindividual differences (see, e.g., Jongerling et al., 2015; Lohmann et al., 2024; Nesselroade, 1991), but they also raise important questions about the quality and comparability of measurements across time points, contexts, and individuals, thereby highlighting the need for formal investigation of MI (e.g., Adolf et al., 2014; Horstmann & Ziegler, 2020; McNeish et al., 2021; Ulitzsch et al., 2025; Vogelsmeier, et al., 2019; 2024). However, examining measurement invariance across time, situational settings, and person characteristics is not straightforward, and even simpler MI explorations are still often lacking in EMA studies (e.g., Moeller et al., 2024; Vogelsmeier et al., 2024). We believe that BH-MNLFA provides a powerful, yet simple tool to address this gap. The remainder of this article is organized as follows. We begin with a brief introduction to MNLFA and MLCFA. Based on this, we then develop BH-MNLFA. In the third and fourth section, we provide two simulation studies demonstrating the functionality of our approach. In the fifth sections, we present an extensive empirical real data example from an experience sampling study. Finally, we discuss the limitations of our work and directions for future research.

1.1. MNLFA

The MNLFA approach is based on a standard CFA model with $i = 1, 2, ..., I$ observed indicator items measured for $j = 1, 2, ..., J$ subjects (hereafter referred to as Level-1 *units*). For the sake of simplicity—and without loss of generality—we focus here on a unidimensional factor model, in which all items load on a single latent construct. This unidimensional model can be expressed by e.g., Rabe-Hesketh et al. (2007)

$$y_{ij} = \alpha_i + \lambda_i \eta_j + \varepsilon_{ij}, \quad \text{with } \eta_j \sim N(0,1), \varepsilon_{ij} \sim N(0, \omega_i), \tag{1}$$

where y_{ij} denotes the observed response of unit j on item i, and η_j represents the latent factor score for unit j. The parameter α_i is the intercept of item i, λ_i is the item-specific loading relating observations to the latent factor, and ε_{ij} is the item- and unit-specific residual with expected value zero and variance ω_i. For model identification, it can be assumed that the latent factor has mean zero and a variance of one.

In MNLFA, the item parameters as well as the mean and variance of the latent factor are allowed to vary as functions of a given number of observed covariates $r = 1, 2, ..., R$, i.e., the parameters are moderated by this set of covariates (Bauer, 2017). Let z_j denote an $r \times 1$ unit-specific vector of observed continuous and/or categorical covariates. To investigate how these are related to CFA model parameters, Equation (1) is extended by moderation effects given by (see, Bauer, 2017; Brandt et al., 2025; Kolbe et al., 2024)

$$y_{ij} = \alpha_{ij} + \lambda_{ij}\eta_j + \varepsilon_{ij}, \quad \text{with} \quad \eta_j \sim N\left(\mu_j, \psi_j\right), \varepsilon_{ij} \sim N(0, \omega_{ij}) \tag{2}$$

with:

$$\alpha_{ij} = \alpha_0 + \boldsymbol{\kappa_i z_j} \tag{3}$$

$$\lambda_{ij} = \lambda_0 + \boldsymbol{\nu_i z_j} \tag{4}$$

$$\omega_{ij} = \omega_0 \exp\left(\boldsymbol{\xi_i z_j}\right) \tag{5}$$

$$\mu_j = \boldsymbol{\gamma z_j} \tag{6}$$

$$\psi_j = \exp(\boldsymbol{\rho z_j}). \tag{7}$$

Here, $\boldsymbol{\kappa_i}$, $\boldsymbol{\nu_i}$, $\boldsymbol{\xi_i}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\rho}$ are $1 \times r$ matrices of moderation effects, indicating the effect of covariates z on the respective moderated item parameters α_{ij}, λ_{ij}, and ω_{ij} or the factor parameters μ_j and ψ_j. Non-zero elements in $\boldsymbol{\kappa_i}$, $\boldsymbol{\nu_i}$, $\boldsymbol{\xi_i}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\rho}$ indicate non-invariance. The parameters α_0, λ_0, and ω_0, represent the mean baseline item parameters, when all covariates $\boldsymbol{z_j}$ are zero.

Frequentist approaches for estimating such MNLFA models have preselected one or more items as so-called anchor item(s) (e.g., Kolbe et al., 2024). The parameters of these anchor items are constrained to be fixed across covariate values z_i (i.e., to be invariant), thereby ensuring model identification. However, selecting anchor items a priori is often difficult to justify in practical applications, as true invariance is rarely known in advance. To address this limitation of MNLFA, Brandt et al. (2025) proposed a Bayesian estimation procedure for MNLFA that incorporates shrinkage priors for all moderation effects. The shrinkage priors serve as regularization factors allowing for the estimation of MNLFA as weakly identified models with potential moderation effects for all item and factor parameters. Similar regularization-based methods have also been developed within the frequentist framework (Bauer et al., 2020; Belzak & Bauer, 2020, 2024; Robitzsch, 2023).

Different shrinkage priors have been proposed and evaluated for regularization in Bayesian estimation (Brandt et al., 2018; van Erp et al., 2019), with the Bayesian adaptive Lasso (BaLasso) prior (Leng et al., 2014) tending to achieve among the most promising results in the context of testing MI with MNLFA (e.g., Brandt et al., 2025). BaLasso uses a parameter-specific penalty, which is handled as a

hyperparameter that itself is sampled from a distribution (e.g., van Erp et al., 2019). The BaLasso can be represented as

$$\beta_j|\varphi_j \sim Double\ Exponential\left(0, \frac{1}{\varphi_j{}^2}\right), \quad (8)$$

$$\text{with the penalty hyperprior } \varphi_j{}^2|a, b^2 \sim Gamma(a, b^2),$$

where β_j is the coefficient to be regularized, and the double-exponential (Laplace) prior has a location parameter fixed at zero and a scale determined by the shrinkage parameter $\varphi_j{}^2$. The hyperprior on $\varphi_j{}^2$ allows for data-driven estimation of the shrinkage intensity for each coefficient. The fixed parameters a and b^2 define the shape and scale of the Gamma distribution, thereby determining the average degree and variability of potential shrinkage, respectively. For Bayesian estimation of the MNLFA, the BaLasso can be employed for the moderation effects (see Brandt et al., 2025), i.e., for the parameter vectors $\boldsymbol{\kappa}$, $\boldsymbol{\nu}$, $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\rho}$.

1.2. Multilevel CFA

Extending the single-level factor model given in Equation (1) to a two-level MLCFA with two separate level-specific latent factors for the within-level and between-level, respectively, we introduce another subscript $k = 1, 2, \ldots, K$ denoting the cluster membership of unit j. In the context of longitudinal data from EMA studies, for example, K refers to the number of participants and j indicates a specific measurement for person k. Furthermore, Equation (1) is extended by the measurement part for Level-2 denoting the within part via superscript $^{(W)}$ and between part via superscript $^{(B)}$ (e.g., Muthén, 1991; Rabe-Hesketh et al., 2007):

$$y_{ijk} = \alpha_i + \lambda_i{}^{(B)}\eta_k{}^{(B)} + \varepsilon_{ik}{}^{(B)} + \lambda_i{}^{(W)}\eta_{jk}{}^{(W)} + \varepsilon_{ijk}{}^{(W)} \quad (9)$$

$$\text{with } \eta_k{}^{(B)} \sim N(\mu, \psi^{(B)}), \eta_{jk}{}^{(W)} \sim N\left(0, \psi^{(W)}\right),$$

$$\varepsilon_{ik}{}^{(B)} \sim N\left(0, \omega_i{}^{(B)}\right), \varepsilon_{ijk}{}^{(W)} \sim N\left(0, \omega_i{}^{(W)}\right),$$

where $\lambda_i{}^{(B)}$ and $\lambda_i{}^{(W)}$ are the factor loadings for indicator i relating observations to the within-level and between-level common factor, respectively, $\eta_k{}^{(B)}$ is the unobserved between-level factor score for cluster k and $\eta_{jk}{}^{(W)}$ is the unobserved within-level factor score for unit j in cluster k. In the context of EMA, $\eta_k{}^{(B)}$ would represent stable trait-like between-person differences in the construct under investigation while $\eta_{jk}{}^{(W)}$ contains person- and measurement-occasion-specific deviations from the individual trait-like average η_k. Furthermore, $\varepsilon_{ik}{}^{(B)}$ and $\varepsilon_{ijk}{}^{(W)}$ are the level-specific residuals. Figure 1 presents a path model for a two-level MLCFA with four indicator items.

1.3. A Bayesian Hierarchical MNLFA

To extend the flexibility of MNLFA to settings with clustered data, we propose combining Bayesian MNLFA with MLCFA. In two-level data structures, covariates can be classified as within-level—characteristics of the individual unit—or between-level—characteristics of the cluster. In EMA studies, within-level covariates often capture time-specific situational

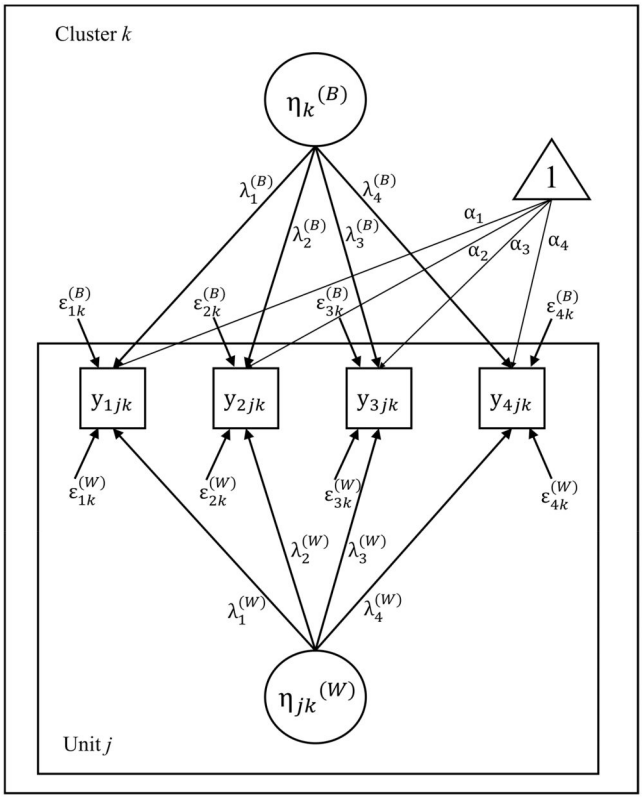

Figure 1. Path diagram of a unidimensional MLCFA with four indicator items.

factors such as time of day or social context. Between-level covariates, by contrast, represent stable person-level characteristics, including age, gender, or personality traits.

Integrating MNLFA and MLCFA enables researchers to examine how these different types of covariates are associated with the measurement properties and the latent construct being assessed—for example, with a self-report scale measuring test anxiety or situational interest. An educational researcher might find that specific item intercepts and loadings vary across different school subjects or as a function of time of day. At the between-level, covariates such as cognitive ability or socio-economic background could explain systematic differences in item functioning—for instance, why some students consistently score higher on particular items or why certain items relate less strongly to the latent factor for students from disadvantaged backgrounds. Researchers may also explore effects on the latent factor itself, such as whether within-person interest decreases over the course of the day or whether socio-economic background accounts for stable between-person differences in test anxiety.

Therefore, the item and latent factor parameters on both levels can be subject of moderation effects: within-level parameters can be functions of within-level covariates $\boldsymbol{z_{jk}}$ and between-level parameters can be functions of between-level covariates $\boldsymbol{z_k}$. Starting from the MLCFA as given in Equation (9) we develop the hierarchical (two-level) MNLFA that reads as follows:

$$y_{ijk} = \alpha_{ik}{}^{(B)} + \lambda_{ik}{}^{(B)}\eta_k{}^{(B)} + \varepsilon_{ik}{}^{(B)} + \alpha_{ikj}{}^{(W)} + \lambda_{ikj}{}^{(W)}\eta_{jk}{}^{(W)} + \varepsilon_{ijk}{}^{(W)} \quad (10)$$

with

$$\eta_k^{(B)} \sim N\left(\mu_k^{(B)}, \psi^{(B)}\right), \eta_{jk}^{(W)} \sim N\left(\mu_{jk}^{(W)}, \psi^{(W)}\right), \varepsilon_{ik}^{(B)} \sim N\left(0, \omega_i^{(B)}\right), \varepsilon_{ijk}^{(W)} \sim N\left(0, \omega_i^{(W)}\right)$$

and between-level moderations:

$$\alpha_{ik}^{(B)} = \alpha_{i0}^{(B)} + \boldsymbol{\kappa_i}^{(B)} \boldsymbol{z_k} \quad (11)$$

$$\lambda_{ik}^{(B)} = \lambda_{i0}^{(B)} + \boldsymbol{\nu_i}^{(B)} \boldsymbol{z_k} \quad (12)$$

$$\omega_i^{(B)} = \omega_0^{(B)} \exp\left(\boldsymbol{\xi_i}^{(B)} \boldsymbol{z_k}\right) \quad (13)$$

$$\mu_k^{(B)} = \boldsymbol{\gamma}^{(B)} \boldsymbol{z_k} \quad (14)$$

$$\psi_k^{(B)} = \psi_0^{(B)} \exp\left(\boldsymbol{\rho}^{(B)} \boldsymbol{z_k}\right). \quad (15)$$

and within-level moderations:

$$\alpha_{ijk}^{(W)} = \boldsymbol{\kappa_i}^{(W)} \boldsymbol{z_{jk}} \quad (16)$$

$$\lambda_{ijk}^{(W)} = \lambda_0^{(W)} + \boldsymbol{\nu_i}^{(W)} \boldsymbol{z_{jk}} \quad (17)$$

$$\omega_{ijk}^{(W)} = \omega_0^{(W)} \exp\left(\boldsymbol{\xi_i}^{(W)} \boldsymbol{z_{jk}}\right) \quad (18)$$

$$\mu_{jk}^{(W)} = \boldsymbol{\gamma}^{(W)} \boldsymbol{z_{jk}} \quad (19)$$

$$\psi_{jk}^{(W)} = \psi_0^{(W)} \exp\left(\boldsymbol{\rho}^{(W)} \boldsymbol{z_{jk}}\right), \quad (20)$$

where all parameters have the same meaning as described above but (W) and (B) distinguish within-level from between-level parameters. Residual variances were modeled separately at both levels, assuming independence between within-level and between-level residual components, consistent with standard two-level CFA formulations. The exponential parameterization ensures that residual variances and latent variances remain strictly positive across all moderator values. For identification, the latent factor means were fixed to zero and the latent variances to one at both levels when all moderators were set to zero. For the moderation effects in $\boldsymbol{\kappa_i}^{(B)}, \boldsymbol{\nu_i}^{(B)}, \boldsymbol{\xi_i}^{(B)}, \boldsymbol{\gamma}^{(B)}$, and $\boldsymbol{\rho}^{(B)}$ as well as $\boldsymbol{\kappa_i}^{(W)}$, $\boldsymbol{\nu_i}^{(W)}$, $\boldsymbol{\xi_i}^{(W)}$, $\boldsymbol{\gamma}^{(W)}$, and $\boldsymbol{\rho}^{(W)}$, the BaLasso introduced in Equation (8) can be employed to achieve model identification, with each moderation effect receiving its own shrinkage parameter τ drawn from the same hyperprior. The BaLasso prior induces shrinkage of moderation effects toward zero, concentrating the posterior of weak or unsupported effects near zero, whereas nonzero effects are retained when supported by the data. For the other model parameters, standard prior distribution as typically used in hierarchical Bayesian factor models can be specified (see, e.g., Depaoli & Clifton, 2015; van Erp & Browne, 2021; Zitzmann et al., 2016, 2020). The means and variances of the within and between-level factor can be set to be zero and one, respectively, when all potential moderators have value zero. The path model of this BH-MNLFA is shown in Figure 2.

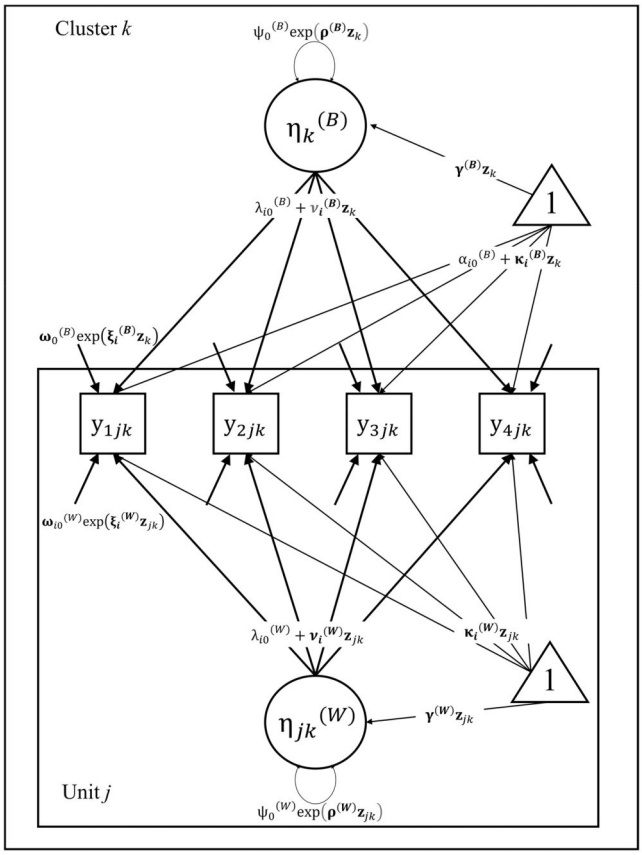

Figure 2. Path diagram of the BH-MNLFA.

2. Simulation Study 1: Level-1 Moderators

A key feature of the proposed hierarchical extension of the Bayesian MNLFA is its ability to distinguish between within-person and between-person levels, allowing for moderation effects at both levels. This first simulation study focuses on Level-1 moderators (continuous and dichotomous) in the BH-MNLFA and the power to detect considerable DIF effects.

2.1. Method

We conducted a simulation study using nine conditions with 250 replications per condition. The data were simulated in R and analyzed using Stan via RStan (Stan Development Team, 2024). The R code is available on our Open Science Framework (OSF) repository (https://osf.io/hd5bw).

The data-generating model in Simulation Study 1 was a unidimensional BH-MLCFA model with five indicator items, reflecting relatively short self-report scales commonly used in psychological research in general and in EMA studies in particular. Additionally, a Level-1 continuous covariate (scaled and mean centered) and dichotomous covariate (effect coded) were simulated. These covariates could act as potential moderators of item parameters. The analysis model was identical to the data-generating model.

Across all simulation conditions, this design yielded 2,500 potential moderation effects per parameter type (250 datasets x 5 items x 2 moderators)—for item loadings, intercepts, and residual variances. In each simulated dataset, for each covariate, two or three of the five item loadings, intercepts, and residual variances were randomly selected to be moderated, resulting in a minimum of 12 and a maximum of 18 moderated item parameters per dataset. Both the number of moderated parameters and the specific item

parameters subject to moderation were randomly determined in each dataset. The latent mean and variance were not allowed to vary as a function of the covariates.

The item-level reliability was assumed to be .5, which can be considered a typical value in psychological research and was also used in the simulation study by Brandt et al. (2025). Thus, standardized factor loadings had the value $\sqrt{0.5} \cong 0.71$. The intraclass correlation coefficient was set to .5, indicating that 50% of the variance in each item was attributable to within-level variance and 50% to between-level variance. The latent mean was set to 0, and latent variances on the within and between level were set to 1. The item-specific means were randomly sampled from value range [2, 2.5, 3, 3.5, 4], reflecting typical response patterns in Likert-type scales commonly used in psychological research, which often yield scale averages within this range (see, e.g., Furr & Bacharach, 2013; OECD, 2024; and also, our empirical example later on).

2.1.1. Simulation Conditions

Two design factors varied across simulation conditions: (1) size of DIF and (2) Level-2 sample size. DIF was set to 0.1 (small), 0.2 (medium), and 0.3 (large). These values were chosen to span a practically relevant range from minor to more pronounced deviations from measurement invariance and are broadly in line with effect magnitudes considered in recent simulation studies on measurement invariance (e.g., Bauer et al., 2020; Brandt et al., 2025). In the data-generating model, these coefficients represent standardized linear moderation of item intercepts and factor loadings per one standard-deviation change in the moderator. Moderation of residual variances was implemented via a log-link (i.e., an exponential transformation). Accordingly, the same numerical DIF coefficient does not imply identical practical impact across different item parameters (see, e.g., Bauer, 2017; Millsap, 2012).

Level-2 sample size was set to $N_{Level\text{-}2} = 25$, 50, and 100. Fully crossing the two design factors resulted in nine simulation conditions.[1] The Level-1 sample size for each cluster was held constant at $N_{Level\text{-}1} = 20$ across conditions.[2] However, because Level-1 sample size is also expected to influence the power to detect DIF in BH-MNLFA, we conducted a supplementary simulation in which Level-1 sample size was varied, too. The results of this study can be found in the Online Supplemental Material A on OSF (https://osf.io/hd5bw).

2.1.2. Model Specification and Estimation

Weakly informative normal distributions were used as priors for item intercepts $\mu_i \sim N(0, 10)$ and exponentially transformed item loadings $\tilde{\lambda}_i \sim N(0, 0.5)$. The exponential transformation ensured that the item loadings were always positive ($\lambda_i = \exp(\tilde{\lambda}_i)$). For residual variances, we specified weakly informative half-cauchy priors $\omega_i \sim HC(0, 2)$. Identical priors were used for the within- and the between-level.

BaLasso priors with a hyperprior $\phi_i \sim Gamma(10, 1)$ were employed for the potential moderation effects of the item parameters and factor mean and variance—a typical choice in other MNLFA applications (e.g., Chen et al., 2022).[3] We used Markov Chain Monte Carlo (MCMC) estimation with 5,000 iterations in Stan (Carpenter et al., 2017; Stan Development Team, 2024). The first 1,000 iterations were discarded as warm-up. Convergence was assessed using the potential scale reduction (PSR) measure (Gelman et al., 2013), which should be below 1.1 for all model parameters (see Brandt et al., 2025).

Table 1. Convergence rates and accuracy for detecting DIF.

	Convergence			Accuracy		
	DIF = 0.1	DIF = 0.2	DIF = 0.3	DIF = 0.1	DIF = 0.2	DIF = 0.3
$N_{Level\text{-}2} = 25$	92%	96%	93%	60%	84%	96%
$N_{Level\text{-}2} = 50$	86%	89%	92%	65%	95%	99%
$N_{Level\text{-}2} = 100$	86%	83%	89%	75%	98%	99%

2.2. Results

Table 1 presents the model convergence rates and the overall accuracy of detecting moderation effects of item parameters for each simulation condition. The convergence rates were slightly higher for the smaller sample size conditions. The accuracy of detecting moderation effects increased with higher DIF and larger sample sizes.

Figure 3 displays the item parameter-specific true and false positive rates. Because continuous and dichotomous moderator showed almost identical results across conditions, we report the average results in the following. However, moderator-specific results can be found on OSF (https://osf.io/hd5bw). The true positive rate refers to the probability of detecting a moderation effect when it is present, while the false positive rate refers to the probability of detecting a moderation effect when it is absent. We based statistical inferences on Bayesian 95% credible intervals (i.e., computed as the 2.5th and 97.5th percentiles of the posterior distribution), corresponding to an alpha level of .05. A moderation effect was classified as detected when its 95% Credible Interval did not include zero. The false positive rates were consistently below 5% across all sample size and DIF conditions, indicating that the Type I error rate was well-controlled.

In contrast, the true positive rates varied substantially across simulation conditions. As could have been expected, higher DIF and larger sample sizes were associated with higher true positive rates. This pattern was consistent across the three item parameters. In psychological research, a power of 0.8 (i.e., 80%) is typically considered sufficient.

[1] DIF was varied in magnitude but not in structural pattern (e.g., intercept-only, loading-only, or residual-only DIF). Isolating these patterns was beyond the scope of this study but may affect detection performance.

[2] This decision was made to restrict the number of design factors.

[3] In the simulation studies, we did not vary the shrinkage hyperprior. However, in the empirical application later on, we introduce a sensitivity approach to assess how different hyperparameter settings affect inferences regarding DIF.

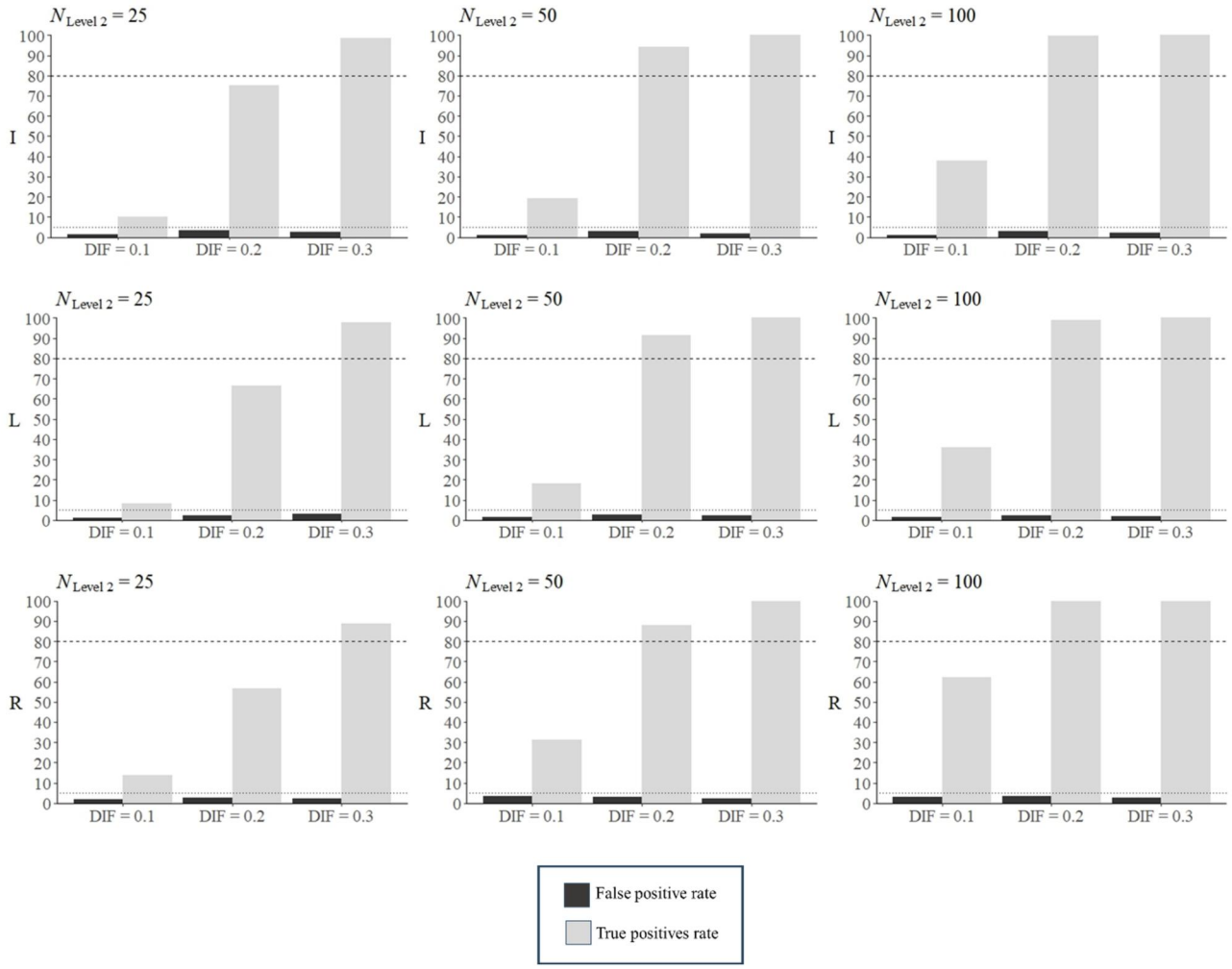

Figure 3. True and false positive rates for detecting moderation effects on item intercepts (I), loadings (L), and residuals (R) at the within-level.

However, for the small DIF condition (DIF = 0.1), the true positive rates were below 80%, indicating that it was challenging to detect moderation effects under these conditions reliably.

In the medium DIF condition, the detection rates for moderation effects were below the desired threshold only in the small sample size condition. In contrast, in the large DIF condition, the true positive rates were above 80% for all three item parameters when the sample size was moderate ($N_{Level\text{-}2} = 50$) or large ($N_{Level\text{-}2} = 100$). These findings suggest that the power to detect moderation effects was generally adequate when the size of DIF and the sample size were both at least medium.

2.3. Discussion

The first simulation study exclusively focused on Level-1 moderation effects in a multilevel CFA and demonstrated that BH-MNLFA allows for the detection of moderation effects among all item parameters with respect to continuous as well as dichotomous moderators. As was to be expected, however, the results also highlight that the accuracy of detecting these effects depends on sample size and the magnitude of the DIF effect. When sample sizes are small or DIF effects are weak, the power to detect moderation effects may be limited. Results from an additional simulation study (Online Supplemental Material A) further demonstrate that the Level-1 sample size (i.e., the number of observations per cluster) substantially affects Level-1 DIF detection. Holding DIF magnitude and the Level-2 sample size constant, larger Level-1 sample sizes yielded higher DIF detection rates.

3. Simulation Study 2: Level-2 Moderators

In the second simulation study, we examined the performance of BH-MNLFA in detecting moderation effects at the cluster level (Level 2). We expected that power to detect such cluster-level moderation would depend primarily on the Level-2 sample size and would benefit much less from increases in Level-1 sample size, in contrast to the within-level moderation design in Simulation Study 1 (see additional simulation in Online Supplemental Material A). Nevertheless, we were also interested in whether Level-1 sample size might still influence the detection of DIF at the between level, even though the moderators were purely between-cluster variables in the second simulation study. We therefore explicitly incorporated variation in Level-1

Table 2. Convergence rates and accuracy for detecting DIF.

		Convergence			Accuracy		
$N_{Level\text{-}2}$	$N_{Level\text{-}1}$	DIF = 0.1	DIF = 0.2	DIF = 0.3	DIF = 0.1	DIF = 0.2	DIF = 0.3
400	5	98%	98%	99%	58%	71%	88%
	20	90%	93%	92%	59%	78%	94%
800	5	95%	95%	92%	61%	87%	98%
	20	82%	79%	82%	64%	93%	99%

sample size into the design of Simulation Study 2. To identify suitable sample size conditions for detecting cluster-level moderation effects, we based the following design choices broadly on the findings of Brandt et al. (2025), who evaluated the performance of single-level Bayesian MNLFA.

3.1. Method

The data-generating model was again a unidimensional MLCFA model with five indicator items. A cluster-level continuous and a dichotomous covariate were simulated, acting as potential moderators of item parameters on the between-level. Again, there were 2500 potential moderation effects (250 datasets x 5 items x 2 moderators) for item loadings, intercepts, and residual variances, respectively. In each simulated dataset, two or three of the five item loadings, intercepts, and residual variances were randomly selected to be moderated per covariate, resulting in a minimum of 12 and a maximum of 18 moderation effects per dataset. In line with Simulation Study 1, the latent mean and variance were never the subject of moderation effects.

In Simulation Study 2, three design factors were varied across simulation conditions: Level-2 sample size ($N_{Level\text{-}2}$ = 400 vs. 800, Level-1 sample size ($N_{Level\text{-}1}$ = 5 vs. 20) and size of DIF (DIF = 0.1 vs. 0.2 vs. 0.3). All other parameters and prior specifications were identical to those used in Simulation Study 1.

3.2. Results

Table 2 presents the convergence rates and the accuracy of detecting item parameter moderations. Again, convergence rates were higher for the small sample size condition. Overall accuracy of detecting whether there was a considerable moderation effect or not was higher for larger sample sizes and stronger DIF effects. Notably, higher Level-1 sample size also improved detection rates, indicating that within-cluster sample size contributes to identifying between-level moderation effects.

Figure 4 presents the true positive and false positive rates for detecting DIF in Level-2 item intercepts, loadings, and residual variances across the two moderators (moderator-specific results are available on OSF). As expected, detection rates increased with larger DIF and larger sample sizes. For all Level-2 item parameters, higher Level-1 sample size was also associated with higher detection rates. However, it became apparent that only for the largest DIF (0.3) and the largest sample size condition (i.e., $N_{Level\text{-}2}$ = 800, $N_{Level\text{-}1}$ = 20) did detection rates for all three Level-2 item parameters exceed the desired 80% threshold. Detection rates for intercepts and loadings were also above this threshold for all DIF = 0.3 conditions across sample size combinations. For DIF = 0.2, moderation of Level-2 intercepts and loadings was only reliably detected when the Level-2 sample size was high (i.e., 800).

3.3. Discussion

The second simulation study demonstrated that the proposed BH-MNLFA approach can detect moderation effects on between-level item parameters. As expected, power was driven primarily by Level-2 sample size and DIF magnitude, but Level-1 sample size also contributed: larger within-cluster samples yielded higher detection rates for all Level-2 item parameters. However, satisfactory performance was obtained only under relatively favorable conditions, indicating that substantially larger samples are required for reliably detecting item parameter moderations on between-level compared to within-level moderation in practice.

4. Empirical Application

In this empirical application, we investigated MI for different state self-concepts using a large dataset from educational psychology (Niepel et al., 2022, 2025). In this study, we drew on data from $N_{\text{Level-2}}$ = 355 students, who participated in a three-week EMA conducted in German schools. The students repeatedly completed a questionnaire regarding several situational variables after each German, English, Mathematics, and Physics lesson. The target variables of the present application were students' German, English, Mathematics, and Physics state self-concepts. Each of these state self-concepts was assessed with three items after each lesson (Niepel et al., 2022). Importantly, the latter implies that for each subject (e.g., English, Mathematics), students' self-concepts were assessed after every lesson regardless of the subject being taught. For instance, Physics self-concepts were not only assessed after Physics lessons but after English, German, and Mathematics as well. The items were adapted from the Self-Description Questionnaire (Marsh et al., 1983) to assess situational expressions of self-concepts (i.e., state self-concepts). The item wordings can be found in Online Supplemental Material B on OSF (https://osf.io/hd5bw), and further details can be found in Niepel et al. (2022). We only used data from measurement occasions at which a student responded to all 3 items of a given self-concept scale.

We employed BH-MNLFA to answer the following two research questions involving one continuous and one categorical Level-1 moderator:

1. How do the measurement properties of state self-concepts change over the period of the EMA study (continuous moderator)?
2. How do the measurement properties of subject-specific state self-concepts differ across the school subjects Mathematics, Physics, German, and English (categorical moderator)?

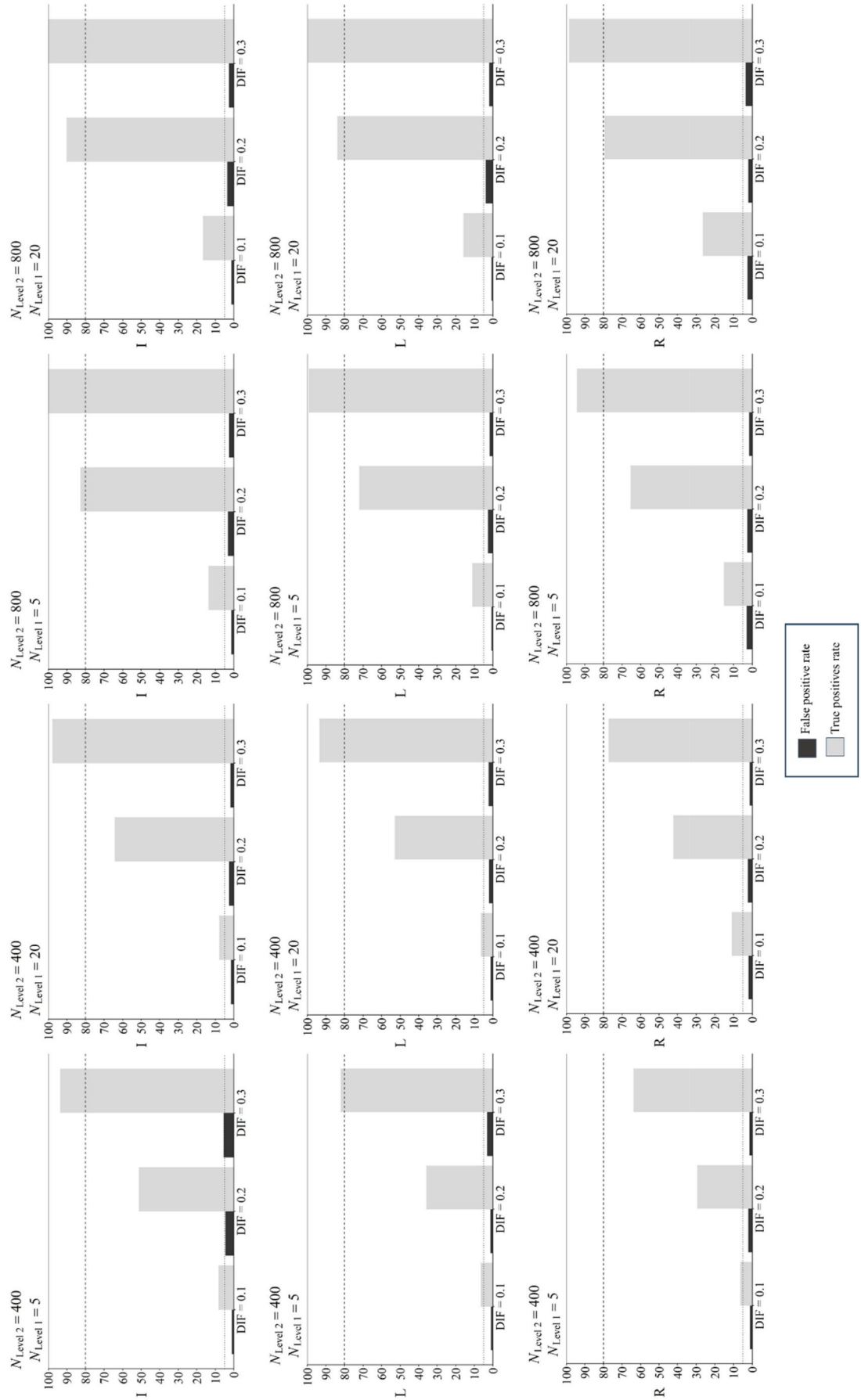

Figure 4. True and false positive rates for detecting moderation effects on item intercepts (I), loadings (L), and residuals (R) at the between level.

4.1. Method

4.1.1. Data Analysis

We ran a separate BH-MNLFA for each of the four state self-concepts (i.e., four unidimensional BH-MNLFA models). To address research question 1, time in hours—as tracked from the beginning of the EMA period onwards—was introduced as a continuous Level-1 moderator in each of these models. To address research question 2, we used three Level-1 dummy variables to represent which subject was taught in a given lesson. It is important to note that each of the self-concepts were assessed after each lesson type. The particular subject currently under study served as the reference category. For example, when the Mathematics state-self-concept was under study, we introduced the three dummies "German lesson", "English lesson", and "Physics lesson". If one of these dummies significantly moderates the item parameters, for example, "English lesson", this would indicate that the measurement properties of Mathematics self-concept differ when measured after English compared to Mathematics lessons.

4.1.2. Model Estimation and Sensitivity Analysis

Each model was estimated using MCMC with three parallel chains and 10,000 iterations per chain. We used the same weakly informative prior setting introduced in Simulation Study 1. The shrinkage hyperprior was set $\tau_j^2 \sim Gamma(a, 1)$.

Although the BaLasso automatically adaptively determines the appropriate amount of shrinkage during model estimation, the choice of hyperprior nevertheless influences the range and strength of possible shrinkage effects (see Chen et al., 2022). This leads to an inevitable tradeoff: A hyperprior that is overly restrictive may incorrectly shrink meaningful moderation effects toward zero, while a too-weak prior may lead to poor convergence. Thus, it is crucial to assess whether inferences are robust to varying assumptions about the degree of shrinkage. To identify hyperprior settings that best balance this tradeoff, we conducted a sensitivity analysis that systematically varied the hyperprior settings and evaluated their impact on both model convergence and parameter estimates. We explored six different hyperprior values setting the shape parameter a to 400, 200, 100, 50, 25, and 10 and re-estimating each model under each setting. These values represent a gradient from stronger to weaker regularization. Lower values of the shape parameter place less prior mass on large penalty values and increase the relative variability of the penalty parameters, thereby weakening the average shrinkage and allowing regression coefficients to deviate more freely from zero. For each setting, we assessed model convergence using the PSR, which should be below 1.05, and an effective size of at least 400 for all model parameters.[4]

[4]Following the recommendations of Zitzmann and Hecht (2019), we applied stricter convergence criteria for the empirical application, whereas the simulation studies required a more pragmatic balance between convergence diagnostics and computational efficiency.

4.2. Results

Figure 5 depicts the 95% Bayesian Credible Intervals for the moderation effects in the four analyses of the subject-specific state self-concept scales.

4.2.1. Time as a Moderator of Item Parameters

Across the various self-concept scales, Figure 5 shows that only 2 of the 12 item loadings—and only 1 of the 12 item intercepts—exhibited a moderator effect of time that differed substantially from zero. Overall, this pattern indicates that both metric and scalar measurement invariance were largely supported for the state self-concept scales by our analyses. In addition, the results revealed several consistent trends in how time moderated the parameters across the self-concept measures. Most notably, three item residuals were consistently negatively moderated, suggesting that measurement error tended to decrease over time. This pattern can be interpreted as an instance of what the literature on repeated self-report measurement refers to as the *Socrates effect* (e.g., Jagodzinski, 1986)—a specific form of the mere-measurement effect (Long et al., 2025). The term describes the phenomenon whereby repeated self-reflection prompted by questionnaires or interviews leads to shifts in response behavior, greater self-awareness, and in some cases even personal change over time (e.g., Godin et al., 2008). Recent work also suggests that such processes may be reflected in the psychometric properties of self-report scales themselves (McNeish et al., 2021). A reduction in measurement error suggests that the students may increasingly form a clearer internal representation of the latent construct being assessed, thereby aligning their responses across items more consistently.

Four further significant moderation effects were present with respect to the moderator time. For the math state self-concept, one item intercept and the latent mean were positively moderated. Although just credibly different from zero, these positive moderation effects—with only the latent mean moderation indicating true change in the underlying construct—suggest a small overall increase in students' math state self-concept over the study period. In addition, for English and Physics self-concept, there was a negative moderation of the second item loading by the time covariate, implying that the second item slightly became less related to the latent factor over time.

4.2.2. School Subjects as Moderators of Item Parameters

Again, with only 5 of the 36 possible item loading moderations and 2 of the 36 possible item intercept moderations yielding Credible intervals not including zero, the findings for school subjects largely support metric and scalar invariance of the state self-concept scales. Overall, 26 of the 132 possible moderation effects across the four subject-specific analyses were credibly different from zero. All substantial moderation effects related to school subjects were consistently negative, and most of them concerned variance components—namely, latent factor variance and item residual variance. Specifically, the latent variance showed 7 significant moderations, and the item residuals showed

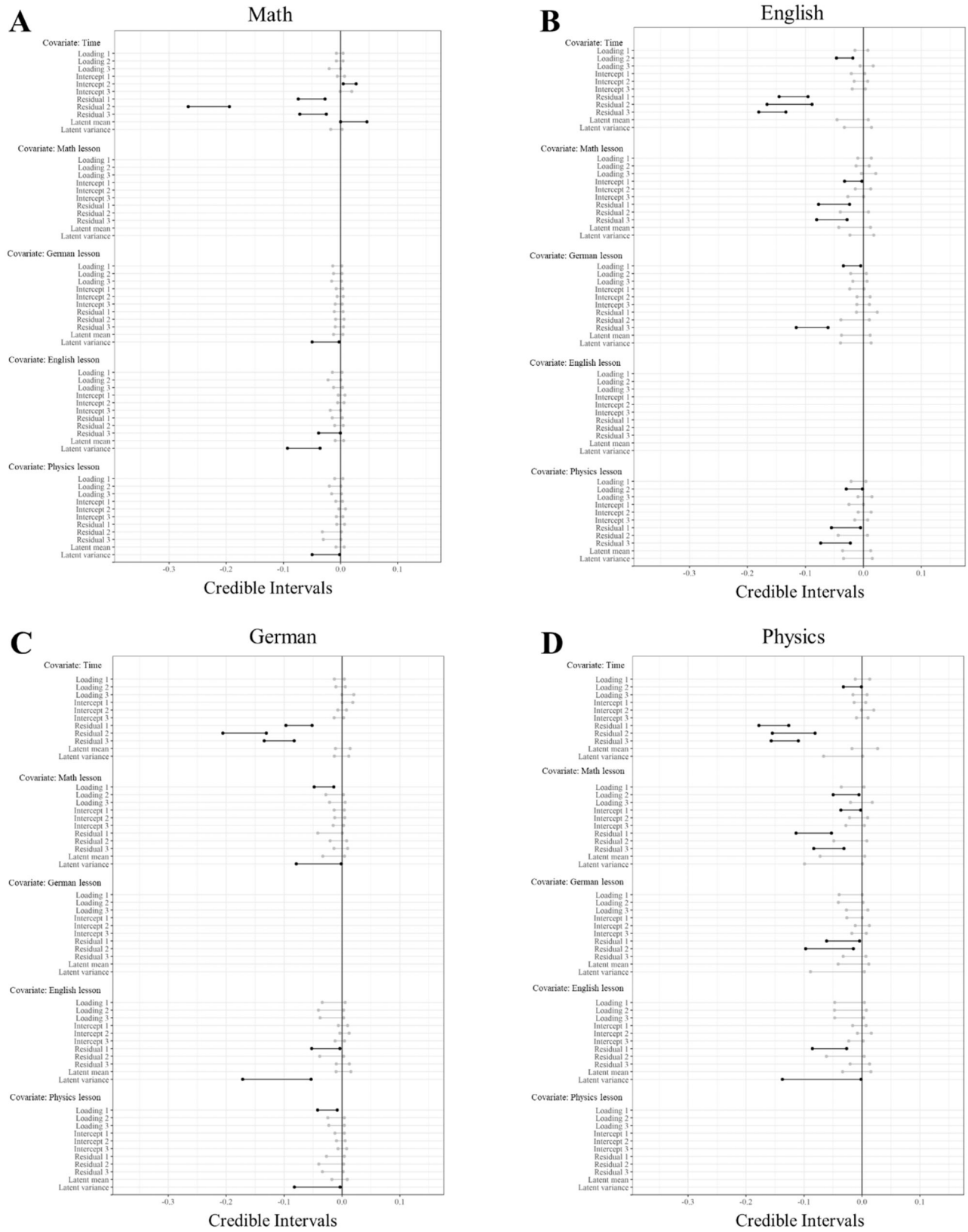

Figure 5. Bayesian 95% credible intervals for the moderation effects of the covariates time and type of lesson across math, german, english, and physics state self-concepts.

12. These results suggest that state self-concepts tend to be less variable after lessons in non-corresponding subjects than after lessons in the corresponding subject. In addition, the findings indicate that state self-concept indicators tend to contain less measurement noise following lessons in non-corresponding subjects.

Furthermore, concerning item loadings, five significant moderation effects were observed. These effects indicated

Table 3. Sensitivity analysis of moderation effects on physics state self-concept under varying shrinkage prior settings.

Moderation effect	*Gamma* (400,1)	*Gamma* (200,1)	*Gamma* (100,1)	*Gamma* (50,1)	*Gamma* (25,1)	*Gamma* (10,1)
Item 1						
Intercept						
Time	−0.002 (0.003)	−0.003 (0.004)	−0.003 (0.005)	0 (0.008)	−0.001 (0.012)	−0.013 (0.01)
English	−0.007 (0.004)	**−0.014 (0.006)***	**−0.019 (0.009)***	**−0.026 (0.012)***	−0.024 (0.017)	−0.017 (0.023)
Physics	0.001 (0.003)	−0.001 (0.004)	−0.004 (0.006)	−0.01 (0.009)	−0.009 (0.013)	0.012 (0.024)
German	−0.004 (0.004)	−0.008 (0.005)	−0.012 (0.007)	−0.016 (0.009)	−0.017 (0.014)	−0.015 (0.036)
Loading						
Time	−0.001 (0.003)	−0.001 (0.004)	0.001 (0.006)	0.146 (0.204)	0.008 (0.014)	−0.188 (0.141)
English	−0.005 (0.005)	−0.013 (0.008)	−0.015 (0.01)	−0.007 (0.016)	−0.019 (0.029)	−0.019 (0.199)
Physics	−0.008 (0.006)	−0.012 (0.01)	−0.018 (0.014)	−0.02 (0.016)	−0.026 (0.022)	−0.087 (0.114)
German	−0.004 (0.004)	−0.014 (0.008)	−0.019 (0.011)	−0.026 (0.013)	−0.021 (0.021)	0.24 (0.398)
Residual						
Time	**−0.134 (0.013)***	**−0.146 (0.013)***	**−0.152 (0.013)***	**−0.155 (0.013)***	**−0.157 (0.013)***	**−0.14 (0.012)***
English	**−0.02 (0.011)***	**−0.054 (0.014)***	**−0.083 (0.016)***	**−0.095 (0.016)***	**−0.102 (0.017)***	**−0.098 (0.019)***
Physics	−0.005 (0.006)	**−0.029 (0.013)***	**−0.056 (0.015)***	**−0.069 (0.015)***	**−0.072 (0.016)***	**−0.084 (0.019)***
German	−0.001 (0.004)	−0.011 (0.01)	**−0.032 (0.015)***	**−0.048 (0.017)***	**−0.048 (0.015)***	**−0.058 (0.017)***
Item 2						
Intercept						
Time	0.005 (0.004)	0.008 (0.005)	0.01 (0.006)	0.014 (0.009)	0.013 (0.013)	0 (0.011)
English	−0.001 (0.003)	−0.003 (0.005)	−0.004 (0.008)	−0.01 (0.012)	−0.007 (0.018)	0.001 (0.025)
Physics	0.002 (0.003)	0.003 (0.004)	0.003 (0.006)	0 (0.009)	0.002 (0.014)	0.025 (0.027)
German	0.001 (0.003)	0.001 (0.004)	0 (0.006)	−0.001 (0.008)	−0.001 (0.015)	0.001 (0.04)
Loading						
Time	**−0.023 (0.007)***	**−0.019 (0.007)***	**−0.017 (0.008)***	0.142 (0.222)	−0.008 (0.015)	−0.205 (0.153)
English	**−0.012 (0.006)***	**−0.024 (0.008)***	**−0.028 (0.012)***	−0.015 (0.019)	−0.029 (0.03)	−0.018 (0.211)
Physics	−0.005 (0.005)	−0.01 (0.01)	−0.016 (0.015)	−0.019 (0.017)	−0.019 (0.024)	−0.096 (0.12)
German	−0.008 (0.005)	**−0.017 (0.008)***	−0.019 (0.011)	−0.023 (0.013)	−0.016 (0.023)	0.28 (0.442)
Residual						
Time	**−0.07 (0.022)***	**−0.104 (0.019)***	**−0.118 (0.019)***	**−0.136 (0.024)***	**−0.129 (0.019)***	**−0.128 (0.025)***
English	−0.002 (0.004)	−0.004 (0.008)	−0.014 (0.015)	−0.046 (0.027)	**−0.046 (0.021)***	**−0.069 (0.025)***
Physics	−0.001 (0.004)	−0.007 (0.009)	−0.023 (0.017)	**−0.05 (0.021)***	**−0.063 (0.024)***	**−0.064 (0.025)***
German	−0.005 (0.006)	**−0.025 (0.015)***	**−0.054 (0.021)***	**−0.083 (0.022)***	**−0.096 (0.025)***	**−0.106 (0.026)***
Item 3						
Intercept						
Time	0 (0.003)	0 (0.004)	0.001 (0.005)	0.004 (0.008)	0.003 (0.013)	−0.009 (0.01)
English	−0.002 (0.003)	−0.006 (0.005)	−0.01 (0.008)	−0.018 (0.012)	−0.016 (0.017)	−0.009 (0.023)
Physics	−0.003 (0.003)	−0.006 (0.005)	−0.01 (0.006)	−0.016 (0.009)	−0.015 (0.012)	0.004 (0.023)
German	0 (0.003)	−0.001 (0.004)	−0.004 (0.006)	−0.007 (0.008)	−0.008 (0.014)	−0.007 (0.037)
Loading						
Time	−0.002 (0.003)	−0.004 (0.005)	−0.003 (0.006)	0.145 (0.209)	0.003 (0.013)	−0.194 (0.144)
English	0 (0.003)	−0.001 (0.005)	−0.001 (0.009)	0.004 (0.013)	−0.004 (0.029)	−0.013 (0.193)
Physics	−0.008 (0.006)	−0.011 (0.009)	−0.019 (0.014)	−0.017 (0.02)	−0.028 (0.021)	−0.09 (0.113)
German	0 (0.003)	−0.003 (0.005)	−0.007 (0.009)	−0.014 (0.013)	−0.01 (0.021)	0.254 (0.404)
Residual						
Time	**−0.12 (0.012)***	**−0.129 (0.012)***	**−0.133 (0.012)***	**−0.136 (0.012)***	**−0.136 (0.012)***	**−0.133 (0.012)***
English	**−0.017 (0.01)***	**−0.042 (0.012)***	**−0.057 (0.013)***	**−0.057 (0.018)***	**−0.07 (0.015)***	**−0.059 (0.018)***
Physics	0.001 (0.003)	0 (0.005)	−0.003 (0.008)	−0.005 (0.011)	−0.007 (0.013)	−0.011 (0.014)
German	−0.001 (0.003)	−0.005 (0.007)	−0.01 (0.01)	−0.014 (0.012)	−0.017 (0.014)	−0.018 (0.015)
Latent Mean						
Time	0.001 (0.003)	0.002 (0.006)	0.003 (0.011)	−0.011 (0.025)	−0.001 (0.033)	0.049 (0.035)
English	−0.005 (0.006)	−0.014 (0.012)	−0.028 (0.021)	−0.026 (0.032)	−0.05 (0.044)	−0.052 (0.064)
Physics	0 (0.003)	−0.001 (0.006)	−0.006 (0.012)	−0.004 (0.023)	−0.019 (0.034)	−0.008 (0.113)
German	−0.001 (0.003)	−0.003 (0.007)	−0.01 (0.013)	−0.012 (0.023)	−0.03 (0.036)	−0.022 (0.05)
Latent Variance						
Time	−0.007 (0.007)	−0.018 (0.013)	−0.029 (0.018)	−0.148 (0.16)	−0.053 (0.037)	0.208 (0.191)
English	−0.006 (0.007)	−0.023 (0.017)	−0.045 (0.027)	−0.081 (0.04)	−0.057 (0.074)	0.063 (0.187)
Physics	**−0.027 (0.015)***	**−0.065 (0.026)***	**−0.076 (0.038)***	−0.095 (0.051)	−0.078 (0.059)	−0.029 (0.19)
German	−0.005 (0.006)	−0.021 (0.016)	−0.038 (0.026)	−0.067 (0.044)	−0.056 (0.056)	−0.357 (0.454)

Note. Columns in light grey show the parameter estimates from the models that did not meet the convergence criterion.
*Bayesian 95% Credible Interval does not include zero.

that only the first and second items exhibited school-subject–specific variation. For the item intercepts, only two parameters showed moderation effects by the school subject factor. More specifically, for the English and physics self-concept scales, students tended to rate Item 1 lower after math lessons compared to after lessons in English and physics, respectively. Notably, no significant moderation effects were found for the latent means in any of the analyses.

4.2.3. Sensitivity Analysis

To illustrate the results of the sensitivity analyses, Table 3 depicts the estimated moderation effects for Physics state self-concepts comparing six models ran with different settings for the shrinkage hyperprior. The three models with the highest shrinkage effect fulfilled the convergence criteria and were thus candidates for final inference. The parameter estimates of the models that do not fulfill the convergence criteria and are thus impractical for final inference are

printed in grey. For the report of final results in the previous sections, we used the least strong shrinkage effect that still resulted in model convergence (i.e., *Gamma*(100, 1)).

From Table 3, it can be inferred that parameter estimates for the moderation effects are higher when the shrinkage effect is smaller. It also becomes apparent that some moderation effects are not statistically significantly different from zero when the shrinkage factor is high, but become significant in the models with lower shrinkage factors (e.g., the moderation of the residual of item one by the Physics lesson dummy). It is also observable that the standard deviation of the parameter estimates also increases when the shrinkage factor decreases. Sensitivity analyses for the other self-concept scales can be found in Online Supplemental Material C on OSF (https://osf.io/hd5bw). Overall, the hyperprior *Gamma*(100,1) achieved the best balance for the given scenario, with one exception—the math self-concept scale.

4.3. Discussion

Across the subject-specific analyses, the BH-MNLFA results indicate that the state self-concept scales demonstrated largely robust metric and scalar invariance over time and across school subjects. Only a few item loadings and intercepts showed significant moderation. This suggests that, overall, the scales measured the same constructs consistently across repeated assessments and instructional contexts.

At the same time, a consistent pattern of negative moderation effects for item residuals emerged over time. This reduction in measurement error may be due to a *Socrates effect*, whereby repeated self-reporting leads participants to develop an internal representation of the latent construct over time and provide more consistent responses across items.

Regarding school-subject effects, significant moderation was found primarily for variance components, with both latent variances and item residuals tending to be lower after non-corresponding lessons. This pattern suggests fewer intraindividual fluctuations and less measurement noise when the immediate instructional context is not directly tied to the domain being assessed. As a result, students' responses in these situations appeared more homogeneous and trait-like.

Taken together, the empirical application highlights how BH-MNLFA can be used for examining how measurement properties and structural parameters change across time and contexts.

5. General Discussion

In the present study, we introduced BH-MNLFA for testing measurement invariance in clustered data with continuous and categorical moderators at both the within- and between-cluster levels. This approach combines multilevel CFA, moderated nonlinear factor analysis, Bayesian estimation, and Bayesian regularization via shrinkage priors. Two simulation studies demonstrated the functionality of BH-MNLFA and provided information on power to detect non-invariance, offering practical guidance for researchers considering its use. Although flexibly applicable to a range of hierarchical data structures, we illustrated the method using repeated EMA measurements and showed how BH-MNLFA can be used to investigate contextual correlates and shifts in measurement invariance over time. A sensitivity analysis further illustrated how to navigate the tradeoff between model convergence and shrinkage of moderation effects. In our OSF repository, we also make editable code available to run the BH-MNLFA with Stan (Carpenter et al., 2017).

5.1. Comparison of BH-MNLFA with Alternative Approaches for Testing MI in EMA Data

While BH-MNLFA can be applied to various types of hierarchical data, in this study, we highlighted its possible application for testing MI in data obtained from EMA. In this section, we therefore contextualize BH-MNLFA among existing methodological approaches designed to address MI in the context of EMA data.

In recent years, several methods have been proposed and discussed for testing or exploring MI with EMA data. Most of them deal with the MI among the two essential dimensions of EMA data—persons and time. For instance, Adolf et al. (2014) proposed to explore MI by iteratively imposing equality constraints on item parameters in a CFA model. The procedure begins with an unconstrained model, where each parameter is estimated freely for each person and each time point. Next, researchers can sequentially impose equality constraints—for instance, by first holding item loadings equal across time points within persons (i.e., weak invariance over time) and secondly also holding item loadings equal across persons (i.e., weak invariance across persons). Fit measures such as the Bayesian information criterion or likelihood ratio test statistics can be used as indicators of whether MI holds. The procedure proceeds until strict invariance is established across measurement occasions and persons or rejected by poor model fit.

McNeish et al. (2021) addressed the two-level nesting structure of EMA data by proposing a cross-classified multilevel CFA model in which Level-1 item parameters are treated as random effects across both time and individuals. These random effects are assumed to follow a normal distribution with a population mean parameter and a corresponding variance component quantifying the heterogeneity of a given item parameter along a given dimension (e.g., heterogeneity across persons or across measurement occasions). This model specification acknowledges that MI violations can occur in a nonsystematic (random) fashion across either dimension and quantifies the extent of measurement variability by estimating variance components for item parameters directly. If random effects variances are near zero, this suggests approximate invariance; higher variance indicates more substantial MI violations. Again, model comparisons and fit measures can be used to test the necessity of a given variance parameter.

Vogelsmeier et al. (2019) proposed a Latent Markov Factor Analysis (LMFA) for investigating MI in EMA. This method assumes the existence of a small number of latent measurement models, or "measurement regimes," within a given dataset. Participants are allowed to transition between these regimes over time, reflecting potential qualitative shifts in the measurement structure. Transitions between regimes are governed by a first-order Markov process, which models the probability of moving from one latent state (i.e., measurement model) to another across adjacent measurement occasions. Notably, LMFA is not limited to a confirmatory framework. Instead, it also supports an exploratory investigation of factor structures and loading patterns across different latent states (Vogelsmeier et al., 2025). This makes it particularly well-suited for settings where researchers suspect discrete, regime-like changes in how constructs are measured—such as changes in how constructs manifest during different psychological or physiological states—rather than gradual or moderated variations.

BH-MNLFA adds a new method to the existing approaches for investigating MI in EMA contexts that focusses on covariate effects. It accounts for the nested structure of EMA data, in which repeated measurements are nested within individuals. However, unlike cross-classified approaches (e.g., Kim et al., 2023; McNeish et al., 2021), BH-MNLFA does not assume a second clustering along measurement occasions. This aspect reflects the fact that measurement occasions in EMA data are often highly individualized and exhibit no clear structure across persons, as in panel data. However, BH-MNLFA allows for systematic and flexible testing of measurement moderation by observed covariates on the person level (person characteristics as moderators of the measurement model) and time-point-specific level (situational factors or time itself moderating measurement properties).

5.2. Limitations and Future Research

The BH-MNLFA, as described in this article, is not without its limitations and could be further extended and refined in future work.

First, the proposed BH-MNLFA is a complex model. Estimating these models using MCMC can be time-consuming when datasets are large. For instance, the runtime for a model in the empirical example was up to 24 hours, even running chains in parallel. Future studies could explore alternative estimation procedures, such as Bayesian Penalized Maximum Likelihood (e.g., Lüdtke et al., 2021), to accelerate estimation. However, a critical aspect in this context is how to derive valid standard errors for statistical inference on moderation effects under Penalized Maximum Likelihood (Casella et al., 2010).

Second, the proposed model does not quantify between-cluster differences in item parameters or DIF effects. In this sense, our specification can be viewed as a fixed-effects model. However, in the literature on measurement models with intensive longitudinal data, the possibility to quantify and test between-person differences in measurement properties was recently introduced and discussed (e.g., McNeish et al., 2021; Schuurman & Hamaker, 2019). Future extensions of BH-MNLFA could therefore incorporate random effects for item parameters—and potentially also for moderator effects—to capture heterogeneity in measurement characteristics across clusters.

Third, our proposed sensitivity analysis procedure to find appropriate regularization parameters for a given dataset considers the tradeoff between model convergence and the detection of substantial moderation effects. Other approaches also consider model fit as another quantity to be considered in selecting shrinkage factors (Orzek et al., 2023; Robitzsch, 2023). These approaches, for instance, use typical fit measures, such as the Bayesian information criterion (BIC), and search for the shrinkage factor that results in the best fit indices. Future approaches could also rely on cross-validation and out-of-sample predictions to optimize this procedure and adapt it for the BH-MNLFA. However, it will be a challenge to reconcile such approaches with an acceptable runtime.

Fourth, other Bayesian shrinkage priors have been developed and recently applied and evaluated in the context of MNLFA (Brandt et al., 2018, 2025; Chen et al., 2022; van Erp et al., 2019). The results of extensive simulation studies implied that lasso priors provide among the best results (e.g., Brandt et al., 2025). Nevertheless, potential disadvantages of the BaLasso, such as that coefficients are not shrunken exactly to zero when using posterior means, should be considered (for details see Brandt et al., 2018; van Erp et al., 2019). This property might be less important when investigating MI but alternative priors could be tested in the context of the BH-MNLFA in the future.

Fifth, in the EMA application, we considered only linear changes in item parameters over time. In practice, however, other functional forms—such as quadratic, piecewise, or even nonparametric trajectories—may be theoretically plausible and substantively important. The BH-MNLFA framework can, in principle, be extended to accommodate such alternative specifications of moderation effects. Moreover, researchers may be interested not only in change across the entire study period but also in changes within specific temporal windows (e.g., time of day) or person-specific indices of assessment history (e.g., the consecutive number of completed prompts). BH-MNLFA is well suited to address these types of questions, as long as the time-related covariates are appropriately defined and coded.

Sixth, the presented model implementation is limited to two-level data. However, three or more level data might be of interest for applied researchers (e.g., Parrisius et al., 2022). In addition, similar to recent implementations and recommendations concerning the (single-level) MNLFA, we implemented our BH-MNLFA as a unidimensional model. However, sometimes it might be of interest to consider more complex factor models (bivariate or G-factor models). Future work could therefore extend the BH-MNLFA for these scenarios.

Finally, missing data are common in (intensive) longitudinal and multilevel designs and can affect both estimation

and the sensitivity to detect violations of measurement invariance. Our BH-MNLFA implementation assumes data in long format and uses all available information on the item responses at Level 1. Thus, individuals contribute information from the measurements they provided, which is typically appropriate under ignorable missingness assumptions (i.e., MCAR or MAR; Enders, 2022). Nevertheless, higher proportions of missing item responses reduce the Level-1 sample size and will therefore lower power for detecting DIF. In contrast, missingness in moderators is more consequential for the current implementation. The present model treats moderators as fully observed, such that missing covariate values may lead to listwise or casewise deletion for the affected occasions, classes, or persons (depending on the covariate's level). Although in many applications moderators are background or contextual variables collected via separate mechanisms (e.g., baseline questionnaires or administrative sources) and therefore typically exhibit relatively low missingness—as in our empirical example—this need not hold generally. Therefore, missing moderators in BH-MNLFA is an important direction for future studies. Further model developments could extend BH-MNLFA, for instance, by including Bayesian approaches for addressing missing data (Gelman et al., 2013) or by applying multiple imputation as a prior modeling step. These approaches would allow the model to retain information from partially observed covariates and may reduce bias under MAR.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a grant from the Luxembourg National Research Fund (C16/SC/11333571) to Christoph Niepel.

ORCID

Julian F. Lohmann http://orcid.org/0000-0002-5864-9692
Steffen Zitzmann http://orcid.org/0000-0002-7595-4736
Martin Hecht http://orcid.org/0000-0002-5168-4911
Christoph Niepel http://orcid.org/0000-0001-6376-7901
Esther Ulitzsch http://orcid.org/0000-0002-9267-8542

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., & Dolan, C. V. (2014). Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra- and inter-individual model structures. *Frontiers in Psychology*, *5*, 883. https://doi.org/10.3389/fpsyg.2014.00883

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *The British Journal of Mathematical and Statistical Psychology*, *76*, 435–461. https://doi.org/10.1111/bmsp.12316

Bauer, D. J., Belzak, W. C. M., & Cole, V. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized Moderated Nonlinear Factor Analysis to detect differential item functioning. *Structural Equation Modeling: a Multidisciplinary Journal*, *27*, 43–55. https://doi.org/10.1080/10705511.2019.1642754

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*, 673–690. https://doi.org/10.1037/met0000253

Belzak, W. C. M., & Bauer, D. J. (2024). Using regularization to identify measurement bias across multiple background characteristics: A penalized Expectation–Maximization Algorithm. *Journal of Educational and Behavioral Statistics*, *49*Article, 976–1012. 10769986231226439. Advance online publication. https://doi.org/10.3102/10769986231226439

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive Bayesian Lasso approach with spike-and-slab priors to identify multiple linear and nonlinear effects in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 946–960. https://doi.org/10.1080/10705511.2018.1474114

Brandt, H., Chen, S. M., & Bauer, D. J. (2025). Bayesian penalty methods for evaluating measurement invariance in moderated nonlinear factor analysis. *Psychological Methods*, *30*, 482–512. Advance online publication. https://doi.org/10.1037/met0000552

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J. [., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*, 369–411. https://doi.org/10.1214/10-BA607

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist Lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 122–139. https://doi.org/10.1080/10705511.2021.1948335

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 327–351. https://doi.org/10.1080/10705511.2014.937849

Enders, C. K. (2022). *Applied missing data*. The Guilford Press.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. SAGE publications.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Godin, G., Sheeran, P., Conner, M., & Germain, M. (2008). Asking questions changes behavior: Mere measurement effects on frequency of blood donation. *Health Psychology: official Journal of the Division of Health Psychology, American Psychological Association*, *27*, 179–184. https://doi.org/10.1037/0278-6133.27.2.179

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*, *34*, 1037–1059. https://doi.org/10.1002/per.2266

Jagodzinski, W. (1986). *Black and White statt LISREL? Wie groß ist der Anteil von "Zufallsantworten" beim Postmaterialismusindex?* [How large is the proportion of "random answers. " in the post-materialism index? *ZA-Information/Zentralarchiv für Empirische Sozialforschung, 19*, 30–51. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-205447

Jongerling, J., Laurenceau, J. -P., & Hamaker, E. L. (2015). A Multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*, 334–349. https://doi.org/10.1080/00273171.2014.1003772

Kim, E., Cao, C., Liu, S., Wang, Y., & Dedrick, R. (2023). Testing measurement invariance over time with intensive longitudinal data and identifying a source of non-invariance. *Structural Equation*

Modeling: A Multidisciplinary Journal, *30*, 393–411. https://doi.org/10.1080/10705511.2022.2130331

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2024). Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, *29*, 388–406. Advance online publication. https://doi.org/10.1037/met0000501

Leng, C., Tran, M., N., & Nott, D. - (2014). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, *66*, 221–244. https://doi.org/10.1007/s10463-013-0429-6

Lohmann, J. F., Zitzmann, S., & Hecht, M. (2024). Studying between-subject differences in trends and dynamics: Introducing the random coefficients Continuous-Time Latent Curve Model with Structured Residuals. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*, 151–164. https://doi.org/10.1080/10705511.2023.2192889

Long, P. A., Huberts, A. S., Di Torrero, A. N., Otto, L. R., Rogge, A. A., Ritschl, V., & Stamm, T. A. (2025). The mere-measurement effect of patient-reported outcomes: A systematic review and meta-analysis. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *34*, 1211–1220. https://doi.org/10.1007/s11136-025-03909-y

Lüdtke, O., Ulitzsch, E., & Robitzsch, A. (2021). A comparison of penalized maximum likelihood estimation and Markov chain Monte Carlo techniques for estimating confirmatory factor analysis models with small sample sizes. *Frontiers in Psychology*, *12*, 615162. https://doi.org/10.3389/fpsyg.2021.615162

Marsh, H. W., Relich, J. D., & Smith, I. D. (1983). Self-concept: The construct validity of interpretations based upon the SDQ. *Journal of Personality and Social Psychology*, *45*, 173–187. https://doi.org/10.1037/0022-3514.45.1.173

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: a Multidisciplinary Journal*, *28*, 807–822. https://doi.org/10.1080/10705511.2021.1915788

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Moeller, J., Dietrich, J., & Baars, J. (2024). The experience sampling method in the research on achievement-related emotions and motivation. In G. Hagenauer, R. Lazarides, & H. Järvenoja (Eds.), *Motivation and emotion in learning and teaching across educational contexts: Theoretical and methodological perspectives and empirical insights*. Routledge.

Muthén, B. O. (1991). Multilevel Factor Analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354. https://doi.org/10.1111/j.1745-3984.1991.tb00363.x

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* (pp. 92–105). American Psychological Association. https://doi.org/10.1037/10099-006

Niepel, C., Hausen, J. E., Weber, A. M., & Möller, J. (2025). Understanding mean-level and intraindividual variability in state academic self-concept: The role of students' trait expectancies and values. *Journal of Educational Psychology*, *117*, 772–788. https://doi.org/10.1037/edu0000946

Niepel, C., Marsh, H. W., Guo, J., Pekrun, R., & Möller, J.] (2022). Revealing dynamic relations between mathematics self-concept and perceived achievement from lesson to lesson: An experience-sampling study. *Journal of Educational Psychology*, *114*, 1380–1393. https://doi.org/10.1037/edu0000716

OECD. (2024). *PISA 2022 technical report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

Orzek, J. H., Arnold, M., & Voelkle, M. C. (2023). Striving for sparsity: On exact and approximate solutions in Regularized Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 956–973. https://doi.org/10.1080/10705511.2023.2189070

Parrisius, C., Gaspard, H., Zitzmann, S., Trautwein, U., & Nagengast, B. (2022). The "situative nature" of competence and value beliefs and the predictive power of autonomy support: A multilevel investigation of repeated observations. *Journal of Educational Psychology*, *114*, 791–814. https://doi.org/10.1037/edu0000680

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190. https://doi.org/10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. *In Handbook of computing and statistics with applications. Handbook of latent variable and related models* (Vol. 1, pp. 209–227). Elsevier. https://doi.org/10.1016/S1871-0301(06)01010-9

Robitzsch, A. (2022). Estimation methods of the multiple-group one-dimensional factor model: Implied identification constraints in the violation of measurement invariance. *Axioms*, *11*, 119. https://doi.org/10.3390/axioms11030119

Robitzsch, A. (2023). Implementation aspects in regularized structural equation models. *Algorithms*, *16*, 446. https://doi.org/10.3390/a16090446

Stan Development Team. (2024). *RStan: The R interface to Stan* (Version 2.32.6) [Computer software]. https://mc-stan.org/

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, *16*, 199–202. https://doi.org/10.1093/abm/16.3.199

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Ulitzsch, E., Viechtbauer, W., Lüdtke, O., Myin-Germeys, I., Nagy, G., Nestler, S., & Eisele, G. V. (2025). Investigating the effect of experience sampling study design on careless and insufficient effort responding identified with a screen-time-based mixture model. *Psychological Assessment*, *37*, 347–359. Advance online publication. https://doi.org/10.1037/pas0001379

van Erp, S., & Browne, W. J. (2021). Bayesian Multilevel Structural Equation Modeling: An investigation into robust prior distributions for the Doubly Latent Categorical Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 875–893. https://doi.org/10.1080/10705511.2021.1915146

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *33*, 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

Vogelsmeier, L. V. D. E., Jongerling, J., & Ulitzsch, E. (2025). Accounting for measurement invariance violations in careless responding detection in intensive longitudinal data: Exploratory vs. Partially constrained latent markov factor analysis. *Multivariate Behavioral Research*, *60*, 878–897. https://doi.org/10.1080/00273171.2025.2492016

Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., Roover., & K., de. (2019). Latent Markov Factor Analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 557–575. https://doi.org/10.1080/10705511.2018.1554445

Zitzmann, S., Helm, C., & Hecht, M. (2020). Prior specification for more stable Bayesian estimation of multilevel latent variable models in small samples: A comparative investigation of two different approaches. *Frontiers in Psychology*, *11*, 611267. https://doi.org/10.3389/fpsyg.2020.611267

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 661–679. https://doi.org/10.1080/10705511.2016.1207179