Every Trait Counts: Marginal Maximum Likelihood Estimation for Novel Multidimensional Count Data Item Response Models with Rotation or ℓ_1–Regularization for Simple Structure

Marie Beisemann[1], Heinz Holling[2], and Philipp Doebler[1]

[1]TU Dortmund University, Department of Statistics

[2]University of Münster, Institute of Psychology

This is a pre-print (version 1).

Author Note

Correspondence: Correspondence should be addressed to Philipp Doebler, doebler@statistik.tu-dortmund.de.

Acknowledgements: Marie Beisemann would like to thank Paul Bürkner for the helpful discussions and his feedback during the process of working on this project. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded by the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

Funding: This work was supported by research grant DO 1789/7-1 granted by the DFG (Deutsche Forschungsgemeinschaft) to Philipp Doebler. Marie Beisemann's work on this project was funded through this grant.

Conflict of Interest: Heinz Holling is co-author of the *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik*, the intelligence test used in the data example of this study. Otherwise, the authors report no conflicts of interest.

Ethical Approval: This work did not include data collection from human or animal participants. In this work, a pre-existing data set was re-analyzed to illustrate the developed method. The data set was collected by Heinz Holling and colleagues and originally published in Jäger et al. (2006). In Germany, where the data collection was conducted, ethical approval for the study by Jäger et al. (2006) was neither institutionally nor nationally obligatory at the time of data collection and thus no ethical approval was sought for the study at the time.

Data Availability: The algorithms developed in this work have been implemented in the `R` package `countirt` (`https://github.com/mbsmn/countirt/tree/multidimensional`). The R code for the simulation study and the R data files of the simulation results are available on OSF (`https://osf.io/n5792/`). The data set re-analyzed for the example could not be made publicly available as this was guaranteed to participants at the time of data collection during the original study.

Conference Presentations: Some of the computational and software aspects of this work have been included in a presentation on the `countirt` package and its algorithms at the Psychoco Workshop 2023 in Zürich, Switzerland. The corresponding slides have been posted on the conference website (`https://www.psychoco.org/2023/program.html`). An abstract regarding this work was submitted to and accepted for the Methods Retreat for young researchers in the work group methods and evaluation (FGME) of the German psychological society (DGPs) in Kassel, Germany (2022), as well as for the 16th conference of the work group methods and evaluation (FGME) of the German psychological society (DGPs) in Konstanz, Germany (2023), but the work could not be presented at either conference due health reasons.

Abstract

The framework of multidimensional item response theory (MIRT) offers psychometric models for various data settings, most popularly for dichotomous and polytomous data. Less attention has been devoted to count responses. A recent growth in interest in count item response models (CIRM)—perhaps sparked by increased occurrence of psychometric count data, e.g., in the form of process data, clinical symptom frequency, number of ideas or errors in cognitive ability assessment—has focused on unidimensional models. A few recently proposed unidimensional CIRMs rely on the Conway-Maxwell-Poisson distribution as the conditional response distribution which allows to model conditionally over-, under-, and equidispersed responses. In this article, we generalize one of those CIRMs to the multidimensional case, introducing the Multidimensional Two-Parameter Conway-Maxwell-Poisson Model (M2PCMPM) class. Using the Expectation-Maximization (EM) algorithm, we develop marginal maximum likelihood estimation methods, primarily for exploratory M2PCMPMs. The resulting discrimination matrices are rotationally indeterminate. We pursue the goal of obtaining a simple structure for them by (1) rotating and (2) regularizing the discrimination matrix. Recent IRT research has successfully used regularization of the discrimination matrix to obtain a simple structure (i.e., a sparse solution) for dichotomous and polytomous data. We develop an EM algorithm with lasso (ℓ_1) regularization for the M2PCMPM and compare (1) and (2) in a simulation study. We illustrate the proposed model with an empirical example using intelligence test data.

Keywords: Item Response Theory, count data, Conway-Maxwell-Poisson distribution, 2PCMPM, multidimensional IRT, EM algorithm, lasso regularization

Every Trait Counts: Marginal Maximum Likelihood Estimation for Novel Multidimensional Count Data Item Response Models with Rotation or ℓ_1–Regularization for Simple Structure

Multidimensional item response theory (MIRT) provides a framework in which responses to a set of items are explained by the items' relation to a number of latent traits (Reckase, 2009). We assume that person i's response to item j is influenced by L latent traits $\theta_{1i}, \ldots, \theta_{Li}$, where the influence strength is determined by discrimination parameters $\alpha_{j1}, \ldots, \alpha_{jL}$ similar to factor loadings in linear factor analysis. The discrimination parameters for all items and all traits are contained in the discrimination matrix $\boldsymbol{\alpha}$. The assumption of a number of latent traits—rather than just one, as in more traditional unidimensional item response models—is often considered more realistic in psychological research. Psychological constructs are often by definition composed of multiple subcomponents, or response behavior is assumed to be complex and multifactorial.

Multidimensional item response models can be divided into confirmatory and exploratory models, analogous to the factor analytical tradition (McDonald, 1999). While confirmatory models test the fit of a pre-specified item-trait relationship structure to the data, exploratory models aim to determine which items stand in relation to which factors, for instance through rotation of the discrimination (or factor loadings) matrix $\boldsymbol{\alpha}$. A common goal of this popular method is to find a simple structure, that is, an item-trait relationship structure where each item loads primarily onto one factor and not (or only to a small extent) on the remaining factors (Browne, 2001; Thurstone, 1947). An alternative strategy to this end—which has only recently gained popularity in the context of MIRT—is regularization (Cho, Xiao, Wang, & Xu, 2022; Sun, Chen, Liu, Ying, & Xin, 2016). Regularization includes techniques often originally developed for variable selection in (generalized) linear models (Hastie, Tibshirani, & Friedman, 2009). By including a penalty term in the model likelihood, sparse parameter estimates with many zeroes can be enforced. In comparison to unpenalized estimation, parameter values are shrunken towards

0, often improving predictive performance and model interpretation. In the context of MIRT, this leads to more parsimonious estimates of discrimination matrices $\boldsymbol{\alpha}$ by selecting only notable item-trait relationships and shrinking the rest towards 0 (see also Trendafilov, 2014).

Research into regularization as a tool to find simply structured discrimination matrices $\boldsymbol{\alpha}$ in MIRT models has so far focused on models for binary and ordinal response data. But some psychometric tests and self-reports generate another type of response data: counts. For instance, divergent thinking and verbal fluency tasks (Forthmann et al., 2016; Myszkowski & Storme, 2021), or processing speed tasks (Baghaei, Ravand, & Nadri, 2019; Doebler & Holling, 2016). Psychological count responses also occur among self-reports (e.g., in clinical psychology; Magnus & Thissen, 2017; Wang, 2010), or as biometric indicators (e.g., number of fixations in eye-tracking; Man & Harring, 2019). Count data naturally occur in text data analysis (Proksch & Slapin, 2009). Corresponding count data item response models have received increasingly more attention in the psychometric literature in recent years (e.g., Beisemann, 2022; Forthmann, Gühne, & Doebler, 2020; Graßhoff, Holling, & Schwabe, 2020; Man & Harring, 2019).

The simplest count data item response model, Rasch's Poisson Counts Model (RPCM; Rasch, 1960; see also e.g., Holling, Böhning, & Böhning, 2015; Jansen, 1994, 1995; Jansen & van Duijn, 1992; Verhelst & Kamphuis, 2009), models the expected count response μ_{ij} for person i to item j as $\mu_{ij} = \exp(\delta_j + \theta_i)$, where δ_j is the item easiness and θ_i is the sole latent trait.[1] Conditional (upon θ_i) responses are assumed to follow a Poisson distribution. Extensions of the RPCM provided more general models, for example by substituting the log-linear relationship in the RPCM by a sigmoid curve (Doebler, Doebler, & Holling, 2014), or by addressing the conditional equidispersion implied by the Poisson

[1] For consistency and readability, we use a parameterization and notation here which is going to most easily generalize to the multidimensional case in the following sections. The original parameterization by Rasch (1960) is not log-linear but multiplicative.

distribution. Conditional equidispersion leads to the strong assumption that $\mathbb{E}(X_{ij}|\theta_i) = \mathbb{V}\text{ar}(X_{ij}|\theta_i)$. Early extensions of the RPCM allowed overdispersed (i.e., $\mathbb{E}(X_{ij}|\theta_i) < \mathbb{V}\text{ar}(X_{ij}|\theta_i)$) conditional response distributions (e.g., Wang, 2010; Hung, 2012). More recently, models for item-specific conditional equi-, over-, or underdispersion (i.e., $\mathbb{E}(X_{ij}|\theta_i) > \mathbb{V}\text{ar}(X_{ij}|\theta_i)$) were proposed by employing the more general Conway-Maxwell-Poisson (CMP) distribution (Conway & Maxwell, 1962; Huang, 2017; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005). The Conway Maxwell Poisson Model (CMPCM; Forthmann et al., 2020) has no discrimination parameters like a Rasch model, while the Two Parameter Conway Maxwell Poisson Model (2PCMPCM; Beisemann, 2022) includes discrimination parameters. Qiao, Jiao, and He (2023) propose a CMP-based joint modeling approach. Tutz (2022) provides an alternative approach all together for dispersion handling. Regardless of the approach, the adequate consideration of dispersion for count data is important to ensure proper uncertainty quantification, i.e., correct standard errors and model-implied reliability (Forthmann et al., 2020).

These generalizations have focused on unidimensional count item response models. Apart from bidimensional extensions of RPCM (Forthmann, Çelik, Holling, Storme, & Lubart, 2018 for a model without discrimination parameters, and Myszkowski & Storme, 2021 for a two-parameter Poisson model), multidimensional count data models have mostly been developed within the frameworks of generalized linear latent and mixed models (GLLAMM; Skrondal & Rabe-Hesketh, 2004) or factor analysis (Wedel, Böckenholt, & Kamakura, 2003) rather than within MIRT. These works have primarily relied on the Poisson distribution, with Wedel et al. (2003) accomodating some flexibility through truncation of the Poisson distribution leading to underdispersion, and allowing different link functions.

With the present work, we aim to generalize the 2PCMPM (Beisemann, 2022) to a multidimensional count data item response model framework which offers the advantages of multidimensional item response modeling for count data in conjunction with the dispersion flexibility of the CMP distribution. The framework contains a number of existing count data item response models as special cases, allowing for easy testing of assumptions by means of model comparisons. Our goal is further to provide marginal maximum likelihood estimation methods for the framework, with a focus on exploratory models. For these, interpretability of the discrimination matrix $\boldsymbol{\alpha}$ is a crucial goal and is aided by pursuing a simple structure for $\boldsymbol{\alpha}$. To this end, we explore both traditional rotation techniques (Browne, 2001), and more novel regularization approaches (Hastie et al., 2009). The remainder of the paper is structured as follows: In the next section, we introduce and formulate the proposed multidimensional count data item response model framework. We proceed to present marginal maximum likelihood estimation methods for the framework, based on the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). We present both unpenalized and penalized estimation methods. Afterward, we assess the proposed models and algorithms in a simulation study and illustrate the framework with a real-world application example. Finally, a discussion of the presented work is provided.

Multidimensional Two-Parameter Conway-Maxwell-Poisson Models

The tests and self-reports for which methods are developed in this article consist of count data items. Item scores are calculated by counting events or by aggregating across a large number tasks each with a binary score. From each participant $i \in \{1, \ldots, N\}$ we obtain a response x_{ij} to each item $j \in \{1, \ldots, M\}$, where $x_{ij} \in \mathbb{N}_0$, $\forall i \in \{1, \ldots, N\}$, $\forall j \in \{1, \ldots, M\}$. An example of such count data tests in the psychological literature are tests in the creative thinking literature which ask participants for different associations in response to items (e.g., the alternate uses task, AUT, to assess divergent thinking; see e.g., Forthmann et al., 2016, 2020; Myszkowski & Storme, 2021 for psychometric analyses of AUT items). The associations given by each person i to each item j can be counted, resulting in the count response x_{ij}.

To model these count responses in an item response theory framework, we assume that the responses depend on item characteristics and L different latent traits θ_{li} for person

i and trait $l \in \{1, \ldots, L\}$. When $L > 1$, the model is multidimensional. This assumption grants more flexibility as (1) unidimensional models are contained as special cases (for $L = 1$), and (2) the assumption of more than one latent trait is often frequently more realistic and is often empirically supported. An overarching latent trait can be made up of different subdomains which influence item responses differently. Items may also share commonalities beyond the unidimensional trait they measure, violating the local independence assumption in unidimensional models (in the AUT example, this could be different domains the items tap into like figural or verbal; Forthmann et al., 2018; Myszkowski & Storme, 2021). In a multidimensional framework, this can be accounted for by modeling the item domains as different latent traits.

We propose to extend the recently proposed Two-Parameter Conway-Maxwell-Poisson model (2PCMPM; Beisemann, 2022)—which models differing item discriminations and dispersions in a unidimensional model—to the multidimensional case. The proposed Multidimensional Two-Parameter Conway-Maxwell-Poisson Models (M2PCMPM) assumes a log-linear factor model for the expected count response μ_{ij};

$$\mu_{ij} = \exp\left(\alpha_{j1}\theta_{1i} + \cdots + \alpha_{jL}\theta_{Li} + \delta_j\right) = \exp\left(\sum_{l=1}^{L} \alpha_{jl}\theta_{li} + \delta_j\right). \quad (1)$$

In this extension of the slope-intercept parametrized 2PCMPM, we denote by α_{jl} the slope for item j and trait l, which quantifies the extent to which differences in the latent trait l are reflected in the expected responses to item j. The parameter δ_j is the intercept for item j, which is related to—but does not directly correspond to—item j's easiness. Analogously to the 2PCMPM, we then assume that responses follow a Conway-Maxwell-Poisson (CMP) distribution conditional on the L latent traits. We use the mean parameterization of the CMP distribution (Huang, 2017), denoted as CMP_μ. Thus, we assume that

$$P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j) = \text{CMP}_\mu(x_{ij}; \mu_{ij}, \nu_j) = \frac{\lambda(\mu_{ij}, \nu_j)^{x_{ij}}}{(x_{ij}!)^{\nu_j}} \frac{1}{Z(\lambda(\mu_{ij}, \nu_j), \nu_j)}, \quad (2)$$

with $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{Li})^T$ denoting the L latent traits of person i, μ_{ij} as in Equation 1 and ν_j as the item-specific dispersion parameter ($\nu_j < 1$ implies overdispersed, $\nu_j > 1$

underdispersed, and $\nu_j = 1$ equidispersed conditional responses). In Equation 2 the expression $Z(\lambda(\mu_{ij}, \nu_j), \nu_j) = \sum_{x=0}^{\infty} \lambda(\mu_{ij}, \nu_j)^x/(x!)^{\nu_j}$ is the normalizing constant of the CMP_μ distribution (Huang, 2017). For simpler notation, we denote all item parameters α_{jl}, $\forall l$, δ_j, and ν_j, for one item j concatenated in one vector with $\boldsymbol{\zeta}_j$. As Huang (2017) showed, we obtain the rate $\lambda(\mu_{ij}, \nu_j)$ by solving

$$0 = \sum_{x=0}^{\infty} (x - \mu_{ij}) \frac{\lambda^x}{(x!)^{\nu_j}} \tag{3}$$

for $\lambda(\mu_{ij}, \nu_j)$.

With the assumption of conditional independence given all L latent traits, the probability of the response vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})^T$ of person i is the product of Equation 2 for each item. The L latent traits $\boldsymbol{\theta}_i$ for each person i jointly follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}_\theta = \mathbf{0} \in \mathbb{R}^L$ and covariance matrix $\boldsymbol{\Sigma}_\theta$, where $\boldsymbol{\Sigma}_\theta$ is a full rank $L \times L$ matrix with all diagonal entries equal to 1 for model identification purposes (more details on assumptions for $\boldsymbol{\Sigma}_\theta$ follow in section *Latent Trait Covariance Matrix*). Assuming that persons respond independently of each other, we obtain

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) = \prod_{i=1}^{N} \int \cdots \int \prod_{j=1}^{M} P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j) \Psi(\boldsymbol{\theta}_i; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) d\theta_{1i} \ldots d\theta_{Li} \tag{4}$$

as the marginal likelihood for the data $\mathbf{x}$ of all N respondents, where Ψ denotes the density of the multivariate normal distribution and $\boldsymbol{\zeta}$ denotes the item parameters $\{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_M\}$ for all M items.

Special cases

The M2PCMPM contains a number of count data item response models as special cases. For $L = 1$, the M2PCMPM simplifies to the 2PCMPM (Beisemann, 2022) and with the additional constraint that $\alpha_{11} = \cdots = \alpha_{1M}$ the model further simplifies to the Conway-Maxwell-Poisson Counts Model (CMPCM; Forthmann et al., 2020). For $L > 1$, but equal slope parameters across items and traits, the M2PCMPM simplifies to a multidimensional CMPCM. Whenever all item-specific dispersions are fixed to be equal to

1 (i.e., $\forall j \in \{1, \ldots, M\} : \nu_j = 1$), the CMP density simplifies to the Poisson density. Consequently, the M2PCMPM also contains the RPCM (Rasch, 1960), the Two-Parameter Poisson Counts Model (2PPCM; Myszkowski & Storme, 2021), and multidimensional extensions of the RPCM and the 2PPCM (Forthmann et al., 2018; Myszkowski & Storme, 2021). Thereby, the M2CMPM offers the possibility of a comprehensive framework for count data item response modeling which subsumes a number of existing count data item response models.

Model identification

The full M2PCMPM as presented in Equation 1 constitutes an exploratory multidimensional item response model: Any item can be associated by any degree with any latent trait. For this reason, the full M2PCMPM as in Equation 1 is not uniquely identified; it is rotationally indeterminate. To enable estimation, we thus need to impose identification constraints on the discrimination matrix $\boldsymbol{\alpha}$. A common constraint is a triangular $(L-1) \times (L-1)$ submatrix of zeroes in the discrimination matrix (as we believe is for example implemented in the `mirt` package; Chalmers, 2012), i.e., we impose constraints to $L-1$ out of the M items to fix rotational indeterminacy. W.l.o.g., let these be the first $L-1$ items. α_{j1} on the first trait is estimated freely and $\forall \alpha_{jl'} = 0, l' \in \{2, \ldots, L\}$. For the following items $j \in \{2, \ldots, L-1\}$, the first j discriminations are free and we constrain $\forall \alpha_{jl'} = 0, l' \in \{j+1, \ldots, L\}$. In the following, this constraint will be referred to as the upper-triangle identification constraint. See e.g., Sun et al., 2016, for examples of alternative constraints. Note that imposing too strong or empirically insensible constraints may impact the model fit (negatively) (Sun et al., 2016). Identification constraints are imposed upon initial estimation to enable finding a likelihood mode. When rotating the initial solution, constraints are lifted, and the discrimination matrix $\boldsymbol{\alpha}$ is rotated freely.

Marginal Maximum Likelihood Estimation Methods for the M2PCMPM

The goal of (frequentist) model estimation of the M2PCMPM is to maximize the model's marginal likelihood (Equation 4) in terms of item parameters $\boldsymbol{\zeta}$. An elegant and popular approach to marginal likelihood estimation in the context of item response models is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; for an introduction see McLachlan & Krishnan, 2007; see Bock & Aitkin, 1981 for the first IRT application). The expected complete-data likelihood—rather than the observed marginal likelihood—is determined in each Expectation (E) step. It includes unobservable parameters, i.e., the latent traits. The expected complete-data likelihood is maximized in each Maximization (M) step. E and M steps are repeated until a convergence criterion is met.

Expectation-Maximization Algorithm

As the M2PCMPM is an extension of the 2PCMPM, estimation methods for the 2PCMPM can be extended to develop estimation methods for the M2PCMPM. Beisemann (2022) provided an EM algorithm for the 2PCMPM which we use as the basis for proposing EM algorithms for the M2PCMPM. The integral in Equation 4 does not exist in closed form and thus has to be approximated in estimation, for example by Gauss-Hermite quadrature with fixed quadrature points. Relying on such a Gauss-Hermite quadrature for the integral approximation with K^L quadrature points, we generalize the expected complete-data log likelihood of the 2PCMPM (Beisemann, 2022) to $L \geq 1$ latent traits for the expected complete-data log likelihood of the M2PCMPM:

$$\begin{aligned}\mathbb{E}(LL_c) \propto \sum_{k_L=1}^{K} \dots \sum_{k_2=1}^{K} \sum_{k_1=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} \log(\lambda(\mu_{jk_1,\dots,k_L}, \nu_j)) - \nu_j \log(x_{ij}!) \\ - \log(Z(\lambda(\mu_{jk_1,\dots,k_L}, \nu_j), \nu_j)))P(q_{k_1}, \dots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}'),\end{aligned} \tag{5}$$

where LL_c denotes the complete-data log likelihood, and

$$\mu_{jk_1,\dots,k_L} = \exp(\alpha_{j1} q_{1k_1} + \dots + \alpha_{jl} q_{lk_l} + \dots + \alpha_{jL} q_{Lk_L} + \delta_j) \tag{6}$$

with $k_l \in \{1, \dots, K\}$ as the node index for trait l. Here, the joint posterior probability of the multidimensional quadrature point $(q_{k_1}, \dots, q_{k_L})$ is given by

$$P(q_{k_1}, \dots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') = \frac{\prod_{j=1}^{M} \mathrm{CMP}_\mu(x_{ij} | q_{k_1}, \dots, q_{k_L}, \boldsymbol{\zeta}_j') w_{k_1} \dots w_{k_L}}{\sum_{k_1'=1}^{K} \dots \sum_{k_L'=1}^{K} \prod_{j=1}^{M} \mathrm{CMP}_\mu(x_{ij} | q_{k_1'}, \dots, q_{k_L'}, \boldsymbol{\zeta}_j') w_{k_1'} \dots w_{k_L'}}, \tag{7}$$

where w_{k_l}, $k_l \in \{1, \ldots, K\}$, denote the nodes' quadrature weights. The E step consists of computing Equation 7. In the subsequent M step, we maximize Equation 5 iteratively as a function of the item parameters $\boldsymbol{\zeta}$. To this end, we need to take the derivatives of Equation 5 with respect to the item parameters. We optimize in $\log \nu_j$ rather than ν_j to search on an unconstrained parameter space (compare Beisemann, 2022). Similar to the techniques in Beisemann (2022) and Huang (2017), we form derivatives (using some results from Huang, 2017), resulting in gradients

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \alpha_{jl}} = \sum_{k_L=1}^{K} \cdots \sum_{k_1=1}^{K} \sum_{i=1}^{N} \frac{q_{k_l} \mu_{jk_1,\ldots,k_L}}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)} (x_{ij} - \mu_{jk_1,\ldots,k_L}) P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') \tag{8}$$

for slopes α_{jl} (note that q_{k_l} in the numerator of the fraction does not loop over all trait dimensions 1 to L, but instead is specific to dimension $l \in \{1, \ldots, L\}$ for the slope α_{il} we are considering),

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \delta_j} = \sum_{k_L=1}^{K} \cdots \sum_{k_1=1}^{K} \sum_{i=1}^{N} \frac{\mu_{jk_1,\ldots,k_L}}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)} (x_{ij} - \mu_{jk_1,\ldots,k_L}) P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') \tag{9}$$

for intercepts δ_j, and

$$\begin{aligned} \frac{\partial \mathbb{E}(LL_c)}{\partial \log \nu_j} = \sum_{k_L=1}^{K} \cdots \sum_{k_1=1}^{K} \sum_{i=1}^{N} \nu_j \left(A(\mu_{jk_1,\ldots,k_L}, \nu_j) \frac{x_{ij} - \mu_{jk_1,\ldots,k_L}}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)} - (\log(x_{ij}!) - B(\mu_{jk_1,\ldots,k_L}, \nu_j)) \right) \\ \times P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') \end{aligned} \tag{10}$$

for log dispersions $\log \nu_j$, with $A(\mu_{jk_1,\ldots,k_L}, \nu_j) = \mathbb{E}_{X_j}(\log(X_j!)(X_j - \mu_{kj}))$ and $B(\mu_{jk_1,\ldots,k_L}, \nu_j)) = \mathbb{E}_{X_j}(\log(X_j!))$ (Huang, 2017). Furthermore,

$$V(\mu_{jk_1,\ldots,k_L}, \nu_j) = \sum_{x=0}^{\infty} \frac{(x - \mu_{jk_1,\ldots,k_L})^2 \lambda(\mu_{jk_1,\ldots,k_L}, \nu_j)^x}{(x!)^{\nu_j} Z(\lambda(\mu_{jk_1,\ldots,k_L}, \nu_j), \nu_j)} \tag{11}$$

(Huang, 2017) is the variance of the CMP$_\mu$ distribution which depends on $\mu_{jk_1,\ldots,k_L}$ and ν_j.

A known limitation of quadrature is its poor scaling to high dimensions (McLachlan & Krishnan, 2007); that is, in the context of the M2PCMPM, settings with greater numbers of latent traits. However, as illustrated with our example, in count data item response settings a smaller number of latent traits is frequently realistic.

Simple Structure via Rotation

After obtaining an initial solution with the EM algorithm described above, the classical approach for interpretable results is to apply a rotation to the discrimination parameters. Lifting the identification constraints after the initial solution is obtained, we have an infinite number of alternative solutions which can be obtained via rotation (i.e., rotational indeterminancy) (Scharf & Nestler, 2019). That is, there is an infinite number of rotation matrices $V \in \mathbb{R}^{L \times L}$ for which $\boldsymbol{\alpha}\boldsymbol{\Theta}^T = \boldsymbol{\alpha}VV^{-1}\boldsymbol{\Theta}^T = (\boldsymbol{\alpha}V)(V^{-1}\boldsymbol{\Theta}^T)$, where $\boldsymbol{\alpha} \in \mathbb{R}^{M \times L}$ is the discrimination matrix and $\boldsymbol{\Theta} \in \mathbb{R}^{N \times L}$ the latent trait matrix (Scharf & Nestler, 2019; Trendafilov, 2014). A preferred rotation matrix V has to be selected, usually one optimizing a specific criterion such as indicating a simple structure (Browne, 2001; Thurstone, 1947) of $\boldsymbol{\alpha}$ (Scharf & Nestler, 2019). Rotation techniques differ in the employed criterion and in whether they allow latent traits to be correlated (i.e., oblique methods) or not (i.e., orthogonal methods) (Scharf & Nestler, 2019; Trendafilov, 2014). Popular rotation techniques are for instance Varimax (Kaiser, 1958, 1959), which is an orthogonal rotation method, and Oblimin (Carroll, 1957; Clarkson & Jennrich, 1988), which is an oblique rotation method.

Simple Structure via Regularization

Recently, a simple structure has also been obtained with regularization techniques (Cho et al., 2022; Sun et al., 2016; Trendafilov, 2014). A perfect simple structure is a sparse matrix: Each item loads on exactly one latent trait, and the other loadings are zero (Scharf & Nestler, 2019; Trendafilov, 2014). Finding a sparse solution to an optimization problem is one aim of regularization (Hastie et al., 2009). By imposing a penalty term R onto the likelihood, regularization methods shrink parameter estimates toward 0 (Hastie et al., 2009). R is a function of all parameters to be regularized and grows as the absolute value of each parameter estimate grows (Scharf & Nestler, 2019). As a result, only substantial parameters (in our case, loadings or discriminations) remain notably different from 0, essentially encouraging a (more) simple structure of the discrimination matrix $\boldsymbol{\alpha}$

(Scharf & Nestler, 2019). As opposed to rotation methods, which are implemented after finding an initial estimate with the M2PCMPM EM algorithm, regularization methods modify the likelihood and have to be integrated into the EM algorithm. In general, the regularized estimates cannot be rotated without changing the value of R; they are hence rotationally determined in this sense.

As we maximize the expected complete-data log likelihood in each M step, we subtract the penalty term $R \geq 0$ from it, weighted with a hyperparameter η (notation here inspired by Scharf & Nestler, 2019; Sun et al., 2016 and in line with Beisemann, 2022). The penalty term R is a function of all slopes $\alpha_{11}, \ldots, \alpha_{jl}, \ldots, \alpha_{ML}$, as contained in $\boldsymbol{\alpha}$. We aim for a sparse solution specifically for $\boldsymbol{\alpha}$ (ideally a simple structure), which is why we only impose the penalty term over $\boldsymbol{\alpha}$. We obtain

$$\mathbb{E}(LL_c)_{\text{reg}} \propto \sum_{k_L=1}^{K} \cdots \sum_{k_2=1}^{K} \sum_{k_1=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} \log(\lambda(\mu_{jk_1,\ldots,k_L}, \nu_j)) - \nu_j \log(x_{ij}!)$$
$$- \log(Z(\lambda(\mu_{jk_1,\ldots,k_L}, \nu_j), \nu_j)))P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') - \eta R(\boldsymbol{\alpha}), \qquad (12)$$

with $P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}')$ as in Equation 7. We can immediately see that for $\eta = 0$, the unregularized maximum likelihood estimate is optimal. The hyperparameter $\eta \geq 0$ should be tuned, i.e., selected from a grid of possible values to provide the best result in terms of a tuning criterion (Hastie et al., 2009). We are going to return to this point further below.

Depending on the penalty term R, different regularization methods are implemented (for an introduction and an overview, see Hastie et al., 2009). In this work, we employ the lasso (Tibshirani, 1996) penalty,

$$R_{\text{lasso}}(\boldsymbol{\alpha}) = ||\boldsymbol{\alpha}||_1 = \sum_{l=1}^{L} \sum_{j=1}^{M} |\alpha_{jl}|. \qquad (13)$$

For binary and polytomous MIRT models, the lasso penalty has yielded promising results as a method to find a well-fitting discrimination matrix $\boldsymbol{\alpha}$ with a (rather) simple structure (Cho et al., 2022; Sun et al., 2016).

Lasso Penalty

Integrating the lasso penalty (Tibshirani, 1996) into the M2PCMPM EM algorithm requires an extension of the algorithm. We plug Equation 13 into Equation 12 and we observe that the E step of the M2PCMPM algorithm remains unaltered by the penalty term. In the M step, we are confronted with the problem that due to the ℓ_1 norm, the gradient only exists for $\alpha_{jl} \neq 0$. To solve this issue for binary and polytomous MIRT models, Sun et al. (2016) employed the coordinate descent algorithm (Friedman, Hastie, & Tibshirani, 2010) in the M step (see also Cho et al., 2022, for a related approach using variational estimation). Binary and polytomous MIRT models have an estimation advantage over count MIRT models in that they require only the estimation of discrimination and location parameters (e.g., item intercepts or threshold parameters) since the conditional variance is implied by the location parameters. The M2PCMPM additionally requires estimation of the dispersion parameters. A strategy in the context of (generalized) linear mixed models optimizing penalized (fixed) effects in one step, and then optimizing remaining model parameters in another step, alternating the steps until convergence (note that random effects are estimated in yet another step, but this is not of interest to us here; Nestler & Humberg, 2022; Schelldorfer, Meier, & Bühlmann, 2014). Inspired by these approaches, we propose the M2PCMPM lasso-EM algorithm (see Algorithm 1) that—during each M step—first optimizes α's and δ's using item-blockwise coordinate descent, and then optimizes dispersion parameters using Equation 10.

Taking an item-blockwise optimization approach as in Sun et al. (2016), we exploit that the expected complete-data log likelihood decomposes into the sum of the item contributions (immediately observable in Equation 5). During each M step of the EM algorithm, we further assume (as is common in EM algorithms) the posterior probabilities from the previous E step for latent traits to be known (via the quadrature approximation). Thus, the (penalized) optimization problem during each M step and for each item j is that of a generalized linear model (GLM) with intercept δ_j and (penalized) slopes $\boldsymbol{\alpha}_j$. Note that

Algorithm 1 Lasso EM with Blockwise Coordinate Descent during M Step

(0) Choose start values and η value

(1) EM cycle:

while not converged **do** ▷ EM algorithm

(a) E step: Equation 7

(b) M step:

(i) Optimization of slopes $\boldsymbol{\alpha}_j$ and intercept δ_j

for $j = 1, \ldots, M$ **do** ▷ Blockwise cyclic coordinate descent

while not converged **do**

(i') Update δ_j using Equation 14

(ii') Update $\boldsymbol{\alpha}_j$:

for $l = 1, \ldots, L$ **do**

(i*) Update α_{jl} with Equation 15

(ii*) Update $\boldsymbol{\alpha}_j$ with new α_{jl} value

end for

end while

end for

(ii) Optimization for remaining parameters $\boldsymbol{\nu}_j$ with Equation 10

end while

CMP$_\mu$-regression is a "bona fide GLM[...]" (Huang, 2017, p. 365). This allows the use of algorithmic techniques developed for ℓ_1-regularization in GLMs, such as coordinate descent (Friedman et al., 2010).

As we can see in Algorithm 1, we need updating rules for δ_j and the $\boldsymbol{\alpha}_j$ within the blockwise coordinate descent during the M step. To this end, we follow Sun et al. (2016): They approximate the expected complete-data log likelihood for item j (i.e., item-specific increment in Equation 5 in our case) as a univariate function of each item parameter, respectively, with a local quadratic approximation. Using this approximation, the resulting lasso update (with tuning parameter η) takes the following shape (Sun et al., 2016; adapted to our model and parameterization):

$$\hat{\delta}_j = \delta_j' - \frac{\frac{\partial \mathbb{E}(LL_c)_j}{\partial \delta_j}}{\frac{\partial^2 \mathbb{E}(LL_c)_j}{\partial^2 \delta_j}} \tag{14}$$

(Sun et al., 2016) for each δ_j and

$$\hat{\alpha}_{jl} = -\frac{S(-\frac{\partial^2 \mathbb{E}(LL_c)_j}{\partial^2 \alpha_{jl}} \alpha_{jl}' + \frac{\partial \mathbb{E}(LL_c)_j}{\partial \alpha_{jl}}, \eta)}{\frac{\partial^2 \mathbb{E}(LL_c)_j}{\partial^2 \alpha_{jl}}} \tag{15}$$

(Sun et al., 2016) for each α_{jl}.[2] Here, S denotes the soft thresholding operator (Donoho & Johnstone, 1995) which is defined as

$$S(x, \eta) = \text{sign}(x)(|x| - \eta)_+ = \begin{cases} x - \eta, & \text{if } x > 0 \text{ and } \eta < |x|, \\ x + \eta, & \text{if } x < 0 \text{ and } \eta < |x|, \\ 0 & \text{if } \eta \geq |x| \end{cases} \tag{16}$$

(Sun et al., 2016). We substitute the M2PCMPM specific terms. $\partial \mathbb{E}(LL_c)_j / \partial \delta_j$ and $\partial \mathbb{E}(LL_c)_j / \partial \alpha_{jl}$ are given in Equations 8 and 9. Using the second derivatives of the variance $V(\mu_{jk_1,\ldots,k_L}, \nu_j)$ in terms of δ_j and α_{jl} (see Appendix A) and results from Huang (2017), we obtain the following second derivatives in terms of δ_j and α_{jl},

$$\frac{\partial^2 \mathbb{E}(LL_c)_j}{\partial^2 \alpha_{jl}} = \sum_{k_L=1}^{K} \cdots \sum_{k_1=1}^{K} \sum_{i=1}^{N} \frac{q_{k_l}^2 \mu_{jk_1,\ldots,k_L} P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}')}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)^2} C(\mu_{jk_1,\ldots,k_L}, \nu_j) \tag{17}$$

and

$$\frac{\partial^2 \mathbb{E}(LL_c)_j}{\partial^2 \delta_j} = \sum_{k_L=1}^{K} \cdots \sum_{k_1=1}^{K} \sum_{i=1}^{N} \frac{\mu_{jk_1,\ldots,k_L} P(q_{k_1}, \ldots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}')}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)^2} C(\mu_{jk_1,\ldots,k_L}, \nu_j), \tag{18}$$

where

$$\begin{aligned} C(\mu_{jk_1,\ldots,k_L}, \nu_j) &= V(\mu_{jk_1,\ldots,k_L}, \nu_j)(x_{ij} - 2\mu_{jk_1,\ldots,k_L}) \\ &\quad - \mu_{jk_1,\ldots,k_L}(x_{ij} - \mu_{jk_1,\ldots,k_L}) \left(\frac{\mathbb{E}_X(X^3 - \mu_{jk_1,\ldots,k_L} X^2)}{V(\mu_{jk_1,\ldots,k_L}, \nu_j)} - 2\mu_{jk_1,\ldots,k_L} \right). \end{aligned} \tag{19}$$

[2] Following our understanding of the notation in Sun et al. (2016), in each iteration of (1)(b)(i) in Algorithm 1, we update δ_j one-step late in (ii'). That is, we update δ_j in (i') on the basis of the at that point most up-to-date $\boldsymbol{\alpha}_j$, but use the previous δ_j in (ii'). Please compare the appendix in Sun et al. (2016).

Latent Trait Covariance Matrix

In the M2PCMPM EM algorithm (including the regularized variants), we assume the latent trait covariance matrix, $\mathbf{\Sigma}_\theta$, fixed. The diagonal of $\mathbf{\Sigma}_\theta$ is fixed to the canonical value $\mathbf{1} \in \mathbb{R}^L$ for identification purposes in this model with discrimination parameters—this is analogous to the identification assumption made in the unidimensional case in Beisemann (2022). A convenient choice for the off-diagonal is to assume orthogonal latent traits during estimation (i.e., fix all off-diagonal elements of $\mathbf{\Sigma}_\theta$ to 0). If the latent traits are in fact correlated, pronounced double loadings of items can result. For the classical rotation approach, an oblique rotation can find a correlated solution with fewer double loadings.

In the case of strong(er) correlations between latent factors, this may put the regularized approach at a disadvantage as a sparse solution will not fit well when double loadings are required to account for latent factor correlations. Sun et al. (2016) approach this problem by first estimating an unpenalized MIRT model to obtain latent factor correlation estimates from this model, which they plug into $\mathbf{\Sigma}_\theta$ for the respective off-diagonal estimates. We use the same approach in this work, but we obtain the latent factor correlation from oblique rotation of the $\boldsymbol{\alpha}$ matrix. Note that an alternative would be to estimate the latent factor correlations within the EM algorithm, albeit this would require adjustments to the algorithm as well as the model identification constraints (compare Sun et al., 2016).

Confirmatory Models by Imposing Constraints

While not a focus of the present work, we wanted to note that with the M2PCMPM EM algorithm, one can also fit confirmatory multidimensional count data item response models. That is, one can impose constraints on the item parameters (in particular but not exclusively, the slope parameters) and evaluate the specified model's fit to the data. Confirmatory models should be identified by the imposed constraints. For instance, the fit of a perfect simple structure to the data can be evaluated by imposing constraints which imply single loadings of each item onto only one trait l (for a fixed l) of the latent traits, respectively, and $\alpha_{jl'} = 0 \; \forall l' \neq l$.

Computational Aspects

The M2PCMPM EM algorithms are computationally expensive. Thus, we dedicated some effort to improving computational efficiency, as outlined below.

Start Values

In line with the start value approach Beisemann (2022) uses for the 2PCMPM, we set starting values for the M2PCMPM by fitting multi-dimensional two-parameter Poisson models to the data and compute starting values for the dispersion parameters as described in Beisemann (2022). Fitting these Poisson variants first saves computation time as each Poisson iteration of the EM algorithm is much less expensive than a CMP iteration, the obtained start values are already quite close to the CMP solution for the α_{jl} and the δ_j, and therewith reduce the number of required iterations of the M2PCMPM EM algorithm (compare Beisemann, 2022).

Regularization tuning and warm starts

For the lasso-penalized M2PCMPM EM algorithm, the hyperparameter η requires tuning to be optimally chosen. To this end, we use a grid of η values to assess. Values of the grid are chosen equidistantly on the log scale (Hastie et al., 2009). To increase computational efficiency when fitting a penalized M2PCMPM for each η value on the grid, we implemented warm starts (Hastie et al., 2009), that is, we used the model parameter estimates of the previous model as start values for the subsequent model. To select the optimal η, one has to impose a criterion which η has to optimize. Traditionally, one may use cross-validation and optimize the RMSE of model predictions (Hastie et al., 2009). However, due to the high computational cost of the M2PCMPM EM algorithm and in line with prior research (Sun et al., 2016), we opted to use the Bayesian Information Criterion (BIC) as a criterion to optimize instead. Following Sun et al. (2016), for the lasso penalty, we computed the BIC (Schwarz, 1978) dependent on η as

$$\mathrm{BIC}_\eta = p^* \log N - 2LL_m(\hat{\boldsymbol{\zeta}}_{\boldsymbol{\eta}}; \mathbf{x}), \tag{20}$$

where $LL_m(\hat{\boldsymbol{\zeta}}_{\boldsymbol{\eta}}; \mathbf{x})$ is the unpenalized marginal log-likelihood for the penalized model parameter estimates (using hyperparameter value η), and p^* is the number of parameters $\neq 0$, i.e., the number of parameters for which the estimate is neither shrunken to 0 nor constrained to 0. We select the η value minimizing BIC_η.

Implementation

We implemented M2PCMPM EM algorithm (with and without penalities) in the `R` package `countirt` (`https://github.com/mbsmn/countirt`; please consult the package's GitHub page for more information on the implementation and its limitations)[3]. For computational efficiency, the algorithm was implemented in `R` and `C++`, using among others the package `GSL` (Galassi et al., 2010), tied into `R` using `Rcpp` (Eddelbuettel et al., 2011). Multidimensional Gauss-Hermite quadrature was implemented using `MultiGHQuad` (Kroeze, 2016). For efficiency, quadrature grid truncation is used per default (i.e., quadrature points with very low quadrature weights are precluded from the grid).

Simulation Study

In this small simulation study, we aimed to validate the proposed algorithms, and illustrate the viability of their usage in different psychometric settings. The simulation study was run in `R` (R Core Team, 2023), using the package `countirt` to fit the M2PCMPMs. The code for the simulations as well as `rds` files of the saved simulation results are available at `https://osf.io/n5792/`.

[3] At the time of writing this manuscript, the M2PCMPM related algorithms are implemented on `multidimensional` branch: `https://github.com/mbsmn/countirt/tree/multidimensional`. In the future, this branch is going to be merged into the main branch.

Design

In line with previous simulations regarding regularized item response models (Sun et al., 2016), we varied the number of latent traits between $L = 3$ and $L = 4$. Further, we varied the correlation between these latent traits ($\rho = 0$ vs. $\rho = .3$). For the model parameters, we used the same range of δ_j and ν_j values across all conditions. For δ_j, we used values between 1.5 and 3.5, and for $\log \nu_j$, we used values between -0.8 and 0.8 (i.e., implying—not very large—over- and underdispersion of varying degree), assigned randomly to the items. These values are empirically realistic for CMP-based count item response models (but not extreme, cf. Beisemann, 2022; Beisemann, Forthmann, & Doebler, 2024; Forthmann et al., 2020; see also *Application Example*). The true α_j values depended on the simulation condition: Apart from the number of latent traits, we also varied the number number of items per trait ($m = 3$ vs. $m = 5$). To the best of our knowledge, settings with small(er) numbers of items are realistic for count tests, with count tests often being comprised of less items than binary tests. We further varied the type of structure of the $\boldsymbol{\alpha}$ matrix (simple vs. slightly complex). With regard to the $\boldsymbol{\alpha}$ matrix structure, simple implies only single loadings of items on their assigned traits. Slightly complex implies that a quarter of the items for each trait additionally—but to a lesser extent—load onto at least one of the other traits. For the simple structure, non-zero discriminations α_{jl} were chosen between 0.2 and 0.3. For the slightly complex structure, one quarter of zero-elements in the simple structure discrimination matrix of the same dimensions were randomly replaced with values of 0.05 or 0.1 (each with probability $p = .125$). Ranges for the discrimination parameters were again chosen to be empirically realistic (cf. Beisemann, 2022; Beisemann et al., 2024; Forthmann et al., 2020, but not extreme; see also *Application Example*). All true parameter values for the respective conditions can be reproduced from the `R` code on the OSF repository (`https://osf.io/n5792/`). The described design factors were fully crossed to yield 16 simulation conditions. We ran $T = 40$ simulation trials per condition.[4]

Data Generation and Model Fitting

In each trial in each respective condition, we generated (inspired by our application example) $N = 1200$ responses to $M = L \times m$ items under the M2PCMPM with the condition-specific model parameters. With regard to simulating item response data from the CMP distribution, we followed prior simulation studies on CMP-based item response models, using and adapting code from Forthmann et al. (2020) and Beisemann (2022). In each trial, we first fitted an exploratory M2PCMPM with upper-triangle identification constraint. The obtained solution was rotated once using the orthogonal Varimax criterion(Kaiser, 1958, 1959) and once using the oblique Oblimin (Clarkson & Jennrich, 1988), relying on the `GPArotation` package (Bernaards & Jennrich, 2005). Then, we fitted the lasso-penalized M2PCMPMs for hyperparameter tuning with regard to the BIC.[5] We used a 12-value penalization grid of $[0, 1000]$ with values chosen equidistantly on the log scale (compare Hastie et al., 2009). We tuned the lasso-penalized M2PCMPMs once with the orthogonal latent trait assumption and once with a latent trait covariance matrix which incorporates the latent traits correlations obtained from the obliquely rotated M2PCMPM (see *Latent Trait Covariance Matrix*). All M2PCMPMs were fitted using the `countirt` package (see *Computational Aspects*).

We enhanced computational efficiency through several techniques. First, we used

[4] Note that with these models and the hyper parameter tuning for the regularization, each trial is computationally very expensive. For computational feasibility and as we simulated for a large sample of $N = 1200$, we were only able to run 40 simulation trials. This is in line with prior research (e.g., Sun et al., 2016 ran only 50 trials).

[5] Here, we opted for fitting the penalized models for the different η values on the entire data set and selected the best fitting one. This approach is more comparable to the rotated models. However, note that in the machine learning literature, it would be preferred to tune the hyperparameter first on a training data set (i.e., a sub-sample of the sample) and then fit the model with the selected η on the remaining test data set. The latter approach will be less prone to overfitting than the first.

warm starts in tuning η with regard to the BIC for the penalized M2PCMPMs (see *Computational Aspects*). Second, we used the parameter estimates obtained from the unpenalized exploratory M2PCMPM as start values for $\eta = 0$ (which should result in immediate convergence as $\eta = 0$ is the unpenalized case). Third, we adjusted the number of quadrature nodes per trait, in relation to the number of latent traits (with 10 nodes per trait for $L = 3$, and 4 nodes per trait with $L = 4$).

Evaluation Criteria

For the penalized M2PCMPMs, we evaluated the models for the η value selecting during hyperparameter tuning. Following Sun et al. (2016), we evaluated the correct estimation rate (CER) which we adapted to the upper-triangle identification constraint used here. The CER (adapted from Sun et al., 2016) is defined here as

$$\text{CER} = \frac{\sum_{l=1}^{L} \sum_{j=1}^{M} \mathbb{I}(\hat{\lambda}_{jl} = \lambda_{jl}) - c}{L \times M - c}, \tag{21}$$

with c is the number of constraints imposed on $\boldsymbol{\alpha}$ for identification, $L \times M$ the number of elements in $\boldsymbol{\alpha}$, and $\lambda_{jl} = \mathbb{I}(\alpha_{jl} \neq 0)$ and $\hat{\lambda}_{jl} = \mathbb{I}(\hat{\alpha}_{jl} \neq 0)$, where $\mathbb{I}(.)$ denotes the indicator function. Note that we defined the CER slightly differently than Sun et al. (2016) to better accommodate our identification constraint. The CER helps to assess whether the variable selection in the lasso-penalized models worked correctly, or to what extent. Performance of the BIC-based tuning for the lasso-penalized models was assessed by comparing the two ηs selected by minimizing BIC and maximizing CER (Sun et al., 2016).

Further, we assessed bias and RMSE for the intercept and (log-)dispersion parameters, as well as for the multidimensional discrimination parameters. As there are an infinite number of rotated solutions, bias and RMSE on each single discrimination parameter are less meaningful for rotated exploratory item response models. Multidimensional discrimination instead assesses the impact of all factors onto each item j at once. We computed the item-specific multidimensional discrimination as

$$A_j = \sqrt{\sum_{l=1}^{L} \alpha_{jl}^2}. \tag{22}$$

(Reckase & McKinley, 1991).

Results

All trials were completed without any numerical instabilities and the EM algorithm(s) converged for all models in all trials and conditions. Bias and RMSE estimates for the multidimensional discriminations across trials and items are displayed in Figure 1. As the x-axes show, the range of bias and RMSE estimates is rather small for most conditions. Conditions with simple as opposed to more complex $\boldsymbol{\alpha}$ structure showed less bias and RMSE, with less variation between items. Generally, the M2PCMPM EM algorithm in conjunction with rotation performed most often well in terms of bias and RMSE on multidimensional discrimination parameters. In any conditions where the M2PCMPM EM algorithm in conjunction with rotation performed very well, the lasso-regularized M2PCMPM EM algorithm also performed decently in terms of bias and RMSE, albeit slightly less well than the rotation approach. We observed more bias and larger RMSE estimates for conditions with four (as opposed to three) latent traits, more so for five than for three items per trait. This result is likely explained by the number of observations to number of parameters ratio which decreases as the number of parameters grow with L and m, while the number of observations N remained the same in our simulation.

Figure 2 shows the average CER per condition and per method or model used. In the first two rows of Figure 2, we see the results for the simple $\boldsymbol{\alpha}$ structure, and in the last two rows, the results for the complex $\boldsymbol{\alpha}$ structure are displayed. There was a clear difference in performance between the two different $\boldsymbol{\alpha}$ structures. For the simple $\boldsymbol{\alpha}$ structure, in line with expectations, we see poor performance of the rotation methods (which are not able to shrink estimates down to exactly 0, putting them at a disadvantage in general in terms of CER). In conditions with complex $\boldsymbol{\alpha}$ structure, the rotation methods performed better in these conditions as we would expect when there are fewer parameters that require shrinkage to exactly 0. In conditions with correlated latent traits, we can see that only the oblique lasso model showed decent performance (in most but not all conditions) in terms of CER. Especially for correlated latent traits, performance fell off for four latent traits in conjunction with five items per trait, even for the oblique lasso. For $L = 3$ latent traits, more items per trait tended to increase performance (at least for complex $\boldsymbol{\alpha}$ structure), but for $L = 4$ latent traits, more items tended to decrease performance (for both $\boldsymbol{\alpha}$ structures). One can again speculate that these last two observed patterns in the results might be due to the number of observations to number of parameters ratio which is considerably decreased for 4 traits and 5 items per trait.

Figure 3 plots the (condition average) CER for the tuning parameter η selected via the BIC (on the y axis) against the maximum (condition average) CER obtained by any of the models on the η grid, i.e., the model we would have selected based on the CER. Figure 3 shows the two different lasso models in two separate panels. Figure 3 describes how well the BIC performed in terms of parameter tuning (Sun et al., 2016). Ideally, the BIC-selected η is the CER-selected η which would mean that the condition's point in Figure 3 would lie on the diagonal black line. In Figure 3, we can see that this is the case for one condition for the oblique lasso ($L = 4, \rho = 0.3, m = 3$ with simple $\boldsymbol{\alpha}$ structure), and for four conditions for the orthogonal lasso ($L = 3, \rho = 0.3, m = 3$, $L = 3, \rho = 0.3, m = 5$, $L = 4, \rho = 0.3, m = 3$, and $L = 4, \rho = 0.3, m = 5$ with simple $\boldsymbol{\alpha}$ structure, and $L = 3, \rho = 0.3, m = 5$ with complex $\boldsymbol{\alpha}$ structure). For either method, conditions with simple $\boldsymbol{\alpha}$ structure, more items, and/or more traits tended to exhibit better accuracy of BIC-based η tuning with points in proximity of the line. For complex $\boldsymbol{\alpha}$ structure (compared to the other conditions), the CER were lower even when η was selected based on the CER. Figure 3 shows here that for complex $\boldsymbol{\alpha}$ structure (compared to the other conditions), BIC-based tuning works notably better (with points closer to the diagonal line) for more items per trait (and even better if that is in conjunction with more latent traits).

Bias and RMSE estimates for the remaining item parameters (δ_j's and $\log \nu_j$'s) are shown in Tables 1 and 2, respectively. We can see that the intercept parameters can be estimated very well with very little bias (Table 1). For the dispersion parameters, we have slightly larger bias and RMSE estimates (Table 2), but overall still satisfactory performance. In particular for $L = 4$ traits, performance is better for larger m, that is, for more items per trait. Settings with $L = 3$ traits yielded better performance than those with $L = 4$, likely as the number of observations to number of items ratio is smaller in the latter case for constant $N = 1200$.

Application Example

To illustrate the application of an exploratory M2PCMPM together with a comparison of the two regularization based approaches with the traditional rotation based approach, we re-analyze data ($N = 1318$ adolescents, including 434 adolescents diagnosed as highly gifted) from a German intelligence test (*Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik*, BIS-HB; Jäger et al., 2006). The BIS-HB is an operationalization of the Berlin model of intelligence structure (Jäger, 1967, 1982, 1984). In line with this model, the BIS-HB assesses intelligence across four operational abilities (each measured in three content domains: figural, verbal, and numerical): processing capacity, creativity, memory, and processing speed. We re-analyze the responses for the two operational abilities, creativity and processing speed, which generate count responses. Processing speed is assessed using nine items (also re-analyzed in Doebler et al., 2014), creativity (in terms of idea flexibility) with five.

In our re-analysis, we investigate in how far we can recover the theoretical factor structure of two latent traits in an exploratory M2PCMPM. We fit the two variants (i.e., lasso and rotation) of the exploratory two-factor M2PCMPM with the upper-triangle identification constraint to the data and 12 quadrature nodes per trait, using the `countirt` package (see *Computational Aspects*). For the M2PCMPM in conjunction with rotation, we used an orthogonal Varimax (Kaiser, 1958, 1959) and an oblique Oblimin rotation (Clarkson & Jennrich, 1988). For the lasso-penalized M2PCMPM, we fitted one model with a priori orthogonal (i.e., uncorrelated) latent factors and one with a priori oblique (i.e., correlated) latent factors. For the latter, latent factor correlations obtained from the obliquely rotated M2PCMPM were used (compare Sun et al., 2016). We tuned the lasso-penalized M2PCMPMs using a 20-value penalization grid of $[0, 1000]$ with values chosen equidistantly on the log scale (cf. Hastie et al., 2009) and used warm starts in η-tuning (see *Computational Aspects*). As in the simulation study, start values for the first M2PCMPMs on the tuning grid (i.e., for $\eta = 0$) were the parameter estimates from the unpenalized M2PCMPM (before rotation).

The results are shown in Table 3. While we do not obtain a pattern of perfect $\boldsymbol{\alpha}$ simple structure for any of the methods, we can see that in particular for the approaches with oblique latent traits, the estimates for the $\boldsymbol{\alpha}$ matrix align well with theoretical considerations. That is, for the Oblimin-rotated unpenalized M2PCMPM, we can see that the processing speed items load mostly on the first trait (i.e., processing speed), while the creative thinking items load mostly on the second trait (i.e., creative thinking). Only the processing speed items BD and OE load overall rather weakly onto either factor, with a small preference for the processing speed factor. A similar pattern of results emerged for the lasso-penalized M2PCMPM with oblique latent traits, with the penalty-imposed shrinkage amplifying the theoretically implied loading structure further. For the creative thinking items AM and ZF as well as for the processing speed item UW, the discrimination parameters were even shrunken to 0. We can see that the assumption that the latent traits are uncorrelated (i.e., Varimax-rotated unpenalized M2PCMPM and lasso-penalized M2PCMPM with orthogonal latent traits) yielded a less differentiated loading structure, in particular for the creative thinking items which still load highest onto the second trait but also less negligibly onto the first, especially for the lasso-penalized M2PCMPM with orthogonal latent traits. Intercept (δ_j) and log-dispersion ($\log \nu_j$) estimates were—as we would expect—very similar across methods. Note the rotated M2PCMPMs have only one set each as they are both based on the same unpenalized M2PCMPM for which we only rotate the $\boldsymbol{\alpha}$ matrix, leaving the other parameters unchanged. Items exhibited a mix of over- and underdispersion, with some even close to equidispersion (i.e., 0 for $\log \nu_j$ as $\log(1) = 0$), highlighting the strength of the CMP distribution to account for such a variation of dispersion across items.

Discussion

This work proposes a novel multidimensional count item response model with flexible dispersion modeling: the multidimensional two-parameter Conway-Maxwell-Poisson model (M2PCMPM). A number of existing count item response models (Beisemann, 2022; Forthmann et al., 2018, 2020; Myszkowski & Storme, 2021; Rasch, 1960) can be understood as special cases of the M2PCMPM, rendering the M2PCMPM a general overarching model class. The M2PCMPM can be employed in an exploratory manner—which this work primarily focused on—but also in a confirmatory manner by imposing constraints on model parameters. As a consequence, even more special cases of count item response models can be obtained and formulated as well as estimated within the M2PCMPM framework. We derived marginal maximum likelihood estimation methods based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). For exploratory M2PCMPMs, we investigated using rotation methods (e.g., Carroll, 1957; Clarkson & Jennrich, 1988; Kaiser, 1958, 1959) in conjunction with the proposed M2PCMPM-EM algorithm for obtaining a simple structure solution for the discrimination parameter matrix. Alternatively, we developed a ℓ_1-penalized (i.e., lasso-penalized; Tibshirani, 1996) variant of the M2PCMPM-EM algorithm which can be used to the same end. We explored versions of this algorithm with a priori uncorrelated latent traits and with a priori correlated latent traits. In a simulation study and an application example, we assessed and compared the two proposed algorithms for fitting exploratory M2PCMPMs.

Performance Patterns from the Simulation Study

The conducted simulation study showed stable numerical performance for the developed algorithms in the investigated simulation settings. Bias and RMSE on the intercept and (log) dispersion parameters were overall satisfactory, with differences in performance between conditions in line with prior research on CMP-based count item response models (Beisemann, 2022; Beisemann et al., 2024). In conditions with more latent traits, we tended to observe more bias, in particular for the (log) dispersion parameters.

Due to rotational indeterminacy, we assessed bias and RMSE on the discrimination parameters for the multidimensional discriminations. For a number of the conditions, we observed decent performance here, with the rotation approach performing slightly better than the lasso approach. Conditions in which bias and RMSE were more pronounced were those with more traits, especially in conjunction with more items per trait. This pattern also emerged when we assessed the rate of parameters which was correctly estimated to be different from 0 (compare Sun et al., 2016): Even though especially the lasso-penalized M2PCMPM-EM algorithm which accounted for a priori correlated latent traits performed quite well in a number of conditions, performance for it as well as all other variants of the M2PCMPM-EM algorithms decreased for conditions with more traits in conjunction with more items per trait, that is, for conditions with overall larger number of items (and therewith model parameters). This may be a suprising pattern at first glance as regularization may be expected to offer more advantages for larger $\boldsymbol{\alpha}$ matrices.

We speculate that this pattern of results for intercept, (log) dispersion, and discrimination parameters might be explained by the ratio of number of observations to number of model parameters. As the sample size was held constant in the simulation study, this ratio decreased for conditions with more traits and more items per trait, that is, more model parameters. For larger sample sizes where the ratio of number of observations to number of model parameters is similar to conditions with fewer traits in our simulation study, we would hypothesize that performance should be improve for more traits and items per trait. Further, to be able to achieve acceptable (albeit still long) computation times, we used a comparably low number of quadrature nodes per trait for conditions with four latent traits. This may also have affected parameter estimation accuracy.

In terms of BIC-based hyperparameter tuning for the lasso-penalized

M2PCMPM-EM algorithm (with either a priori correlated or a priori uncorrelated latent factors), we found performance differed notably depending on the condition. Assessing tuning performance following Sun et al. (2016), we found that performance was in general better for an underlying simple structure of the $\boldsymbol{\alpha}$ matrix. Unsurprisingly, more complex structures of the $\boldsymbol{\alpha}$ matrix were more challenging as these are less clearly variable selection problems. With more items and/or more traits, the accuracy of the BIC-based hyperparameter tuning tended to improve. Compared to Sun et al. (2016)'s assessment of BIC-based hyperparameter tuning for lasso-penalized binary models, we observed overall (more or less pronounced) worse performance for count models (not just of the BIC tuning, but also of the CER based tuning which is perhaps surprising at first glance). It is worth pointing out that the direct comparison to the models in Sun et al. (2016) is not entirely appropriate as Sun et al. (2016) defined the CER slightly differently to us (see above). The observed pattern may also be confounded with the number of penalized parameters—in our simulation, the smallest setting only included nine items, which leaves (with identification constraints) only six freely estimated, penalized parameters. In this instance, a misclassification equates to a change of $\frac{1}{6}$ in the CER, while in a setting with for example 20 freely estimated, penalized parameters, it would equate to only $\frac{1}{20}$. As Sun et al. (2016) studied settings with far more items—as is realistic for binary data, but not for count data—this means that single or small numbers of misclassifications affected the CER estimates less drastically than in our simulation. As discussed further below, these results suggest that while the BIC-based hyperparameter tuning appears to work decently for some conditions, hyperparameter tuning for the lasso-penalized M2PCMPM-EM algorithm could still be improved by future research. These results also suggest that future research might wish to consider alternatives to the CER for performance evaluation. For example, one could extract the model-implied item covariance matrix and compare it to the observed item covariance matrix using matrix norms.

Limitations and Further Avenues for Future Research

Our simulation study was designed to provide a proof of concept for the proposed model and algorithms. As such, and as guided by previous research (Sun et al., 2016), it focused on scenarios with three or four latent traits. Future research could explore higher dimensional scenarios. In such settings, the Gauss-Hermite quadrature based M2PCMPM EM algorithm is likely going to reach its limitation, as Gauss-Hermite quadrature is known not to scale well to high-dimensional problems (Chalmers, 2012). Thus, future research in this regard could explore alternative integral approximations, such as Monte Carlo based methods. Further, the maximum test length investigated in our simulation study was 20 items. Future research could investigate more extensive tests. An important point to address in corresponding future research would be the ratio of the number of observations to the number of model parameters. With its fixed sample size, the simulation study cannot sufficiently speak to sample size recommendations—albeit observed results patterns suggest that estimation performance may suffer from too low ratios of the number of observations to the number of model parameters.

We implemented the proposed algorithms in `R` and `C++` within the `countirt` package. To this end, we built upon implementations of the 2PCMPM (Beisemann, 2022) and related models (Beisemann et al., 2024) in `countirt`. These implementations all use a naive interpolation-from-grid approach for some of the CMP distribution related quantities to stabilize, facilitate and fasten computations. This approach worked well in our simulation study and its settings, but can be expected to work less well in settings where the data do not align well with the interpolation grid (see `https://github.com/mbsmn/countirt` for details). In a regression framework, Philipson and Huang (2023) developed a sophisticated and theory-based interpolation approach for CMP models which allows not only inter- but also extrapolation from a specifically designed grid. Future research could aim to apply and extend their work to the (multidimensional) IRT context for CMP models.

For comparability with the rotation approach and for computational reasons, we did not tune our lasso penalty term on a training data set. However, for regularization methods that would be the recommended approach (Hastie et al., 2009) and is what we would recommend for high-stakes applications. This approach should prohibit over-fitting to the data more aptly. In general, our tuning for the lasso penalty term simply used a grid with equidistant tuning parameter values on the log-value space (as is typically recommended; Hastie et al., 2009) and was based on the BIC. As we saw in the simulation study results, for certain settings, the selection of the tuning parameter could still be improved. In fact, sometimes the correct estimation rates were even low when they were used to choose the tuning parameter value. Future research might research how parameter tuning can be improved for the M2PCMPM lasso-EM algorithm and what computationally equally economical alternatives to the BIC as a tuning criterion could be used. Further, more investigation of tuning and the tuning grid used could also be interesting and helpful. Such investigations are going to have to face the computation time challenge that these computationally expensive models pose. Other than the warm starts already used in this work, other avenues such as EM algorithm accelerators might be explored (see Beisemann, Wartlick, & Doebler, 2020, for a recent overview of state-of-the-art methods).

Using the lasso penalty in the M2PCMPM not only encourages a sparse solution for the discrimination matrix $\boldsymbol{\alpha}$, but it also imposes a certain degree of shrinkage onto each discrimination estimate in $\boldsymbol{\alpha}$. To avoid shrunken estimates, future research could explore the relaxed lasso (Meinshausen, 2007): The lasso-penalized M2PCMPM can be fitted to the data for model selection, and afterwards an unpenalized M2PCMPM with appropriate constraints (as selected by the lasso) can be fitted to the data for interpretation of the model parameters.

For the penalization, we focused on the lasso (Tibshirani, 1996) which aligns with other research on penalization in item response models (Cho et al., 2022; Sun et al., 2016). However, lasso penalization is known to perform less well in settings with correlated variables (Hastie et al., 2009), which corresponds to latent factor correlations in item response model settings. However, as we can see from our application example, such settings are empirically realistic. Future research could address such limitation by extending the lasso-penalized M2PCMPM EM algorithm to penalties such as the elastic net (Zou & Hastie, 2005) which adaptively combines properties of the lasso and the ridge (Hoerl & Kennard, 1970) penalty. Alternative penalties such as the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001) could also be explored (for an application of SCAD in IRT, see e.g., Robitzsch, 2023). Other ways in which the penalized algorithms themselves could be extended by future research would be for example the incorporation of latent factor correlation estimation into the algorithm, rather than the two-step method by Sun et al. (2016) that we used here to have the algorithm account for a priori expected correlated factors. In the unpenalized M2PMCPM, such extensions would not be as necessary as factor correlations can be accounted for by oblique rotations (e.g, Clarkson & Jennrich, 1988).

Finally, the M2PCMPM framework proposed in this work can also in itself be a stepping stone for future research. That is, the M2PCMPM framework offers researchers the opportunity to propose, fit, and investigate a number of new count item response models that can be accomodated by the M2PCMPM framework as special cases. This can be achieved by exploring the confirmatory side of the M2PCMPM framework which the present work only briefly touched on. Future research could suggest new constraints through which new count item response models can be obtained from the M2PCMPM. Furthermore, for the M2PCMPM framework to be complete and applicable in practice, it needs to be enriched in the future by developing multi-group and differential item functioning extensions within the framework as well as by deriving person parameter estimators, item fit, and person fit measures.

References

Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson counts model. *Perceptual and Motor Skills*, *126*(1), 70–86. doi: 10.1177/0031512518812183

Beisemann, M. (2022). A flexible approach to modeling over-, under-and equidispersed count data in IRT: The two-parameter Conway-Maxwell-Poisson model. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 411–443. doi: 10.1111/bmsp.12273

Beisemann, M., Forthmann, B., & Doebler, P. (2024). Understanding ability and reliability differences measured with count items: The distributional regression test model and the count latent regression model. *Multivariate Behavioral Research*, *Advance online publication*, 1–21. doi: 10.1080/00273171.2023.2288577

Beisemann, M., Wartlick, O., & Doebler, P. (2020). Comparison of recent acceleration techniques for the EM algorithm in one-and two-parameter logistic IRT models. *Psych*, *2*(4), 209–252. doi: 10.3390/psych2040018

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*, 676–696. doi: 10.1177/0013164404272507

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

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. doi: 10.1207/S15327906MBR3601_05

Carroll, J. B. (1957). Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science*, *126*(3283), 1114–1115. doi: 10.1126/science.126.3283.1114

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

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2022). Regularized variational estimation for exploratory item factor analysis. *Psychometrika*, 1–29. doi: 10.1007/s11336-022-09874-6

Clarkson, D. B., & Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, *53*(2), 251–259. doi: 10.1007/BF02294136

Conway, R., & Maxwell, W. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, *12*, 132–136.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, *38*(8), 587–598. doi: 10.1177/0146621614543513

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson counts model. *Learning and Individual Differences*, *52*, 121–128. doi: 10.1016/j.lindif.2015.01.013

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224. doi: 10.1080/01621459.1995.10476626

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., … Bates, D. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. doi: 10.18637/jss.v040.i08

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273

Forthmann, B., Çelik, P., Holling, H., Storme, M., & Lubart, T. (2018). Item response modeling of divergent-thinking tasks: A comparison of Rasch's Poisson model with a two-dimensional model extension. *The International Journal of Creativity & Problem Solving*, *28*(2), 83—95.

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32. doi: 10.1016/j.intell.2016.03.005

Forthmann, B., Gühne, D., & Doebler, P. (2020). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 32–50. doi: 10.1111/bmsp.12184

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1. doi: 10.18637/jss.v033.i01

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., ... Rossi, F. (2010). GNU scientific library reference manual (3rd ed.) [Computer software manual]. Retrieved from `http://www.gnu.org/software/gsl`

Graßhoff, U., Holling, H., & Schwabe, R. (2020). D-optimal design for the Rasch counts model with multiple binary predictors. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 541–555. doi: 10.1111/bmsp.12204

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). New York, NY: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82. doi: 10.2307/1267351

Holling, H., Böhning, W., & Böhning, D. (2015). The covariate-adjusted frequency plot for the Rasch Poisson counts model. *Thailand Statistician*, *13*(1), 67–78.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, *17*(6), 359–380. doi:

10.1177/1471082X17697749

Hung, L.-F. (2012). A negative binomial regression model for accuracy tests. *Applied Psychological Measurement*, *36*(2), 88–103. doi: 10.1177/0146621611429548

Jäger, A. (1967). *Dimensionen der Intelligenz [Dimensions of intelligence].* Göttingen, Germany: Hogrefe.

Jäger, A. (1982). Mehrmodale Klassifikation von Intelligenzleistungen [Multimodal classifications of intelligence achievements]. *Diagnostica*, *28*, 195–225.

Jäger, A. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Intelligence structure research: Competing models, new developments, perspectives]. *Psychologische Rundschau*, *35*, 21–25.

Jäger, A., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., & Beauducel, A. (2006). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik [Berlin intelligence structure test for adolescents: diagnosis of diftedness and high giftedness] (BIS-HB).* Göttingen, Germany: Hogrefe.

Jansen, M. G. (1994). Parameters of the latent distribution in Rasch's Poisson counts model. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* Berlin/Heidelberg, Germany: Springer. doi: 10.1007/978-1-4612-4308-3_23

Jansen, M. G. (1995). The Rasch Poisson Counts Model for incomplete data: An application of the EM algorithm. *Applied Psychological Measurement*, *19*(3), 291–302. doi: 10.1177/014662169501900307

Jansen, M. G., & van Duijn, M. A. (1992). Extensions of Rasch's multiplicative Poisson model. *Psychometrika*, *57*(3), 405–414. doi: 10.1007/BF02295428

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187. doi: 10.1007/BF02289233

Kaiser, H. F. (1959). Computer program for varimax rotation in factor analysis. *Educational and psychological measurement*, *19*(3), 413–420. doi:

10.1177/001316445901900314

Kroeze, K. (2016). MultiGHQuad: Multidimensional Gauss-Hermite quadrature [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=MultiGHQuad` (R package version 1.2.0)

Magnus, B. E., & Thissen, D. (2017). Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *Journal of Educational and Behavioral Statistics*, *42*(5), 531–558. doi: 10.3102/1076998617694878

Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, *79*(4), 617–635. doi: 10.1177/0013164418824148

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. 382). Hoboken, NJ: John Wiley & Sons.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393. doi: 10.1016/j.csda.2006.12.019

Myszkowski, N., & Storme, M. (2021). Accounting for variable task discrimination in divergent thinking fluency measurement: An example of the benefits of a 2-Parameter Poisson Counts Model and its bifactor extension over the Rasch Poisson Counts Model. *The Journal of Creative Behavior*, *55*(3), 800–818. doi: 10.1002/jocb.490

Nestler, S., & Humberg, S. (2022). A lasso and a regression tree mixed-effect model with random effects for the level, the residual variance, and the autocorrelation. *Psychometrika*, *87*(2), 506–532. doi: 10.1007/s11336-021-09787-w

Philipson, P., & Huang, A. (2023). A fast look-up method for Bayesian mean-parameterised Conway–Maxwell–Poisson regression models. *Statistics and Computing*, *33*(4), 81. doi: 10.1007/s11222-023-10244-0

Proksch, S.-O., & Slapin, J. B. (2009). How to avoid pitfalls in statistical analysis of political texts: The case of Germany. *German Politics*, *18*(3), 323–344. doi: 10.1080/09644000903055799

Qiao, X., Jiao, H., & He, Q. (2023). Multiple-group joint modeling of item responses, response times, and action counts with the conway-maxwell-poisson distribution. *Journal of Educational Measurement*, *60*(2), 255–281. doi: 10.1111/jedm.12349

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.* Denmark: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361–373. doi: 10.1177/014662169101500407

Robitzsch, A. (2023). Comparing robust linking and regularized estimation for linking two groups in the 1PL and 2PL models in the presence of sparse uniform differential item functioning. *Stats*, *6*(1), 192–208. doi: 10.3390/stats6010012

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576–590. doi: 10.1080/10705511.2018.1558060

Schelldorfer, J., Meier, L., & Bühlmann, P. (2014). GLMMlasso: An algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics*, *23*(2), 460–477. doi: 10.1080/10618600.2013.773239

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464. Retrieved from `http://www.jstor.org/stable/2958889.`

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(1), 127–142. doi: 10.1111/j.1467-9876.2005.00474.x

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via l_1 regularization. *Psychometrika*, *81*(4), 921–939. doi: 10.1007/s11336-016-9529-6

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: Chicago University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, *29*(3), 431–454. doi: 10.1007/s00180-013-0434-5

Tutz, G. (2022). Flexible item response models for count data: The count thresholds model. *Applied Psychological Measurement*, *46*(8), 643–661. doi: 10.1177/01466216221108124

Verhelst, N. D., & Kamphuis, F. H. (2009). *A Poisson-Gamma model for speed tests* (Cito Measurement and Research Department Reports No. Technical Report 2009-2). Arnhem, The Netherlands: Cito.

Wang, L. (2010). IRT–ZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, *35*(6), 671–692. doi: 10.3102/1076998610375838

Wedel, M., Böckenholt, U., & Kamakura, W. A. (2003). Factor models for multivariate count data. *Journal of Multivariate Analysis*, *87*(2), 356–369. doi: 10.1016/S0047-259X(03)00020-4

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Table 1

Average bias (between-item SD in parentheses) and RMSE (between-item SD in parentheses) on δ_j parameters across all items per condition

Design				Bias (SD)			RMSE (SD)		
L	$\boldsymbol{\alpha}$ structure	ρ	m	Lasso (obli)	Lasso (ortho)	Rotate	Lasso (obli)	Lasso (ortho)	Rotate
3	simple	0	3	0.001 (0.002)	0.001 (0.002)	-0.001 (0.002)	0.011 (0.002)	0.011 (0.002)	0.011 (0.002)
3	simple	0	5	0.002 (0.003)	0.002 (0.003)	-0.001 (0.002)	0.012 (0.004)	0.012 (0.004)	0.012 (0.004)
3	simple	.3	3	0.002 (0.002)	-0.000 (0.001)	-0.001 (0.001)	0.012 (0.003)	0.012 (0.003)	0.012 (0.003)
3	simple	.3	5	0.003 (0.003)	0.000 (0.002)	-0.001 (0.002)	0.014 (0.004)	0.013 (0.004)	0.013 (0.004)
3	complex	0	3	0.002 (0.002)	0.002 (0.002)	0.001 (0.002)	0.011 (0.002)	0.011 (0.002)	0.011 (0.002)
3	complex	0	5	0.000 (0.002)	0.002 (0.002)	-0.000 (0.002)	0.013 (0.004)	0.013 (0.004)	0.013 (0.004)
3	complex	.3	3	0.003 (0.001)	-0.000 (0.001)	-0.001 (0.001)	0.013 (0.003)	0.012 (0.003)	0.012 (0.003)
3	complex	.3	5	0.002 (0.002)	0.000 (0.002)	-0.000 (0.002)	0.014 (0.003)	0.012 (0.003)	0.012 (0.003)
4	simple	0	3	0.006 (0.005)	0.006 (0.004)	0.002 (0.003)	0.013 (0.004)	0.014 (0.003)	0.013 (0.002)
4	simple	0	5	0.006 (0.002)	0.009 (0.003)	0.004 (0.002)	0.014 (0.003)	0.016 (0.003)	0.015 (0.003)
4	simple	.3	3	0.008 (0.006)	0.005 (0.004)	0.003 (0.003)	0.015 (0.005)	0.014 (0.003)	0.013 (0.003)
4	simple	.3	5	0.007 (0.003)	0.006 (0.002)	0.005 (0.002)	0.015 (0.003)	0.014 (0.002)	0.014 (0.002)
4	complex	0	3	0.005 (0.003)	0.005 (0.004)	0.003 (0.003)	0.014 (0.004)	0.013 (0.004)	0.013 (0.004)
4	complex	0	5	0.006 (0.002)	0.008 (0.002)	0.005 (0.002)	0.015 (0.003)	0.016 (0.003)	0.014 (0.002)
4	complex	.3	3	0.007 (0.002)	0.005 (0.003)	0.005 (0.003)	0.015 (0.003)	0.014 (0.003)	0.014 (0.003)
4	complex	.3	5	0.003 (0.003)	0.005 (0.003)	0.004 (0.003)	0.018 (0.003)	0.017 (0.004)	0.016 (0.004)

Notes. Note that rotated models have the same δ_j estimates regardless of rotation methods as those only affect $\hat{\boldsymbol{\alpha}}$. obli = oblique (latent traits are a priori assumed to be correlated). ortho = orthogonal (latent traits are a priori assumed to be orthogonal). L = number of latent traits. ρ = true latent trait correlation. m = number of items per trait.

Table 2

Average bias (SD in parentheses) and RMSE (SD in parentheses) on $\log \nu_j$ *parameters across all items per condition*

Design				Bias (SD)			RMSE (SD)		
L	$\boldsymbol{\alpha}$ structure	ρ	m	Lasso (obli)	Lasso (ortho)	Rotate	Lasso (obli)	Lasso (ortho)	Rotate
3	simple	0	3	-0.007 (0.013)	-0.007 (0.014)	0.007 (0.017)	0.084 (0.029)	0.084 (0.029)	0.082 (0.030)
3	simple	0	5	-0.006 (0.009)	-0.010 (0.027)	-0.004 (0.031)	0.060 (0.018)	0.062 (0.022)	0.061 (0.022)
3	simple	.3	3	-0.007 (0.014)	0.010 (0.012)	0.013 (0.013)	0.071 (0.020)	0.075 (0.025)	0.076 (0.026)
3	simple	.3	5	-0.013 (0.022)	-0.002 (0.020)	-0.001 (0.022)	0.061 (0.023)	0.060 (0.019)	0.060 (0.019)
3	complex	0	3	0.006 (0.013)	0.012 (0.015)	0.015 (0.015)	0.075 (0.023)	0.076 (0.022)	0.077 (0.022)
3	complex	0	5	-0.005 (0.008)	-0.010 (0.021)	-0.005 (0.019)	0.056 (0.013)	0.058 (0.017)	0.055 (0.014)
3	complex	.3	3	-0.007 (0.012)	0.011 (0.010)	0.013 (0.011)	0.068 (0.018)	0.074 (0.025)	0.075 (0.024)
3	complex	.3	5	-0.014 (0.019)	-0.001 (0.012)	-0.000 (0.011)	0.059 (0.018)	0.056 (0.015)	0.056 (0.014)
4	simple	0	3	-0.076 (0.148)	-0.106 (0.214)	-0.071 (0.165)	0.126 (0.134)	0.156 (0.194)	0.132 (0.144)
4	simple	0	5	-0.069 (0.095)	-0.077 (0.104)	-0.068 (0.106)	0.095 (0.087)	0.102 (0.096)	0.098 (0.095)
4	simple	.3	3	-0.077 (0.142)	-0.064 (0.147)	-0.057 (0.138)	0.125 (0.124)	0.122 (0.129)	0.115 (0.120)
4	simple	.3	5	-0.059 (0.098)	-0.049 (0.088)	-0.048 (0.089)	0.093 (0.088)	0.085 (0.075)	0.086 (0.075)
4	complex	0	3	-0.068 (0.135)	-0.073 (0.209)	-0.066 (0.206)	0.120 (0.122)	0.133 (0.186)	0.132 (0.182)
4	complex	0	5	-0.064 (0.093)	-0.065 (0.093)	-0.064 (0.096)	0.097 (0.081)	0.096 (0.081)	0.096 (0.082)
4	complex	.3	3	-0.067 (0.146)	-0.060 (0.173)	-0.060 (0.173)	0.122 (0.131)	0.126 (0.151)	0.126 (0.150)
4	complex	.3	5	-0.059 (0.080)	-0.053 (0.076)	-0.053 (0.076)	0.091 (0.071)	0.086 (0.064)	0.086 (0.064)

Notes. Note that rotated models have the same δ_j estimates regardless of rotation methods as those only affect $\hat{\boldsymbol{\alpha}}$. obli = oblique (latent traits are a priori assumed to be correlated). ortho = orthogonal (latent traits are a priori assumed to be orthogonal). L = number of latent traits. ρ = true latent trait correlation. m = number of items per trait.

Table 3

Results example (Processing speed (P) and creativity (C))

Item Method	 Parameter	RZ (P)	IT (C)	SI (P)	XG (P)	UW (P)	TG (P)	KW (P)	ZS (P)	BD (P)	OE (P)	AM (C)	ZF (C)	EF (C)	OJ (C)
Varimax	α_{j1}	0.262	0.114	0.219	0.258	0.325	0.219	0.157	0.139	0.080	0.091	0.095	0.079	0.104	0.097
	α_{j2}	0.067	0.192	0.099	0.042	0.117	0.064	0.062	0.055	0.049	0.053	0.272	0.212	0.253	0.189
Oblimin	α_{j1}	0.278	0.046	0.212	0.284	0.329	0.228	0.156	0.138	0.071	0.082	-0.013	-0.004	0.006	0.027
	α_{j2}	-0.013	0.193	0.042	-0.041	0.026	-0.000	0.020	0.018	0.031	0.032	0.296	0.228	0.269	0.195
Lasso (ortho)	α_{j1}	0.261	0.154	0.229	0.250	0.335	0.222	0.163	0.145	0.087	0.099	0.155	0.126	0.158	0.138
($\eta = 26.367$)	α_{j2}	0	0.150	0.033	-0.026	0.019	0.000	0.014	0.012	0.023	0.022	0.231	0.175	0.212	0.149
Lasso (obli)	α_{j1}	0.256	0.050	0.208	0.270	0.333	0.225	0.161	0.140	0.072	0.086	0.000	0.000	0.007	0.026
($\eta = 48.329$)	α_{j2}	0	0.172	0.029	-0.040	0.000	-0.009	0.004	0.005	0.024	0.020	0.266	0.213	0.251	0.183
Varimax /	δ_j	2.369	1.780	3.580	2.747	3.069	2.405	3.223	3.460	3.944	3.486	1.313	1.547	1.576	1.491
Oblimin	$\log \nu_j$	0.282	0.815	-1.131	-0.095	-0.436	0.566	0.510	0.009	-0.117	-0.021	0.787	0.909	0.627	0.993
Lasso (ortho)	δ_j	2.370	1.781	3.580	2.748	3.069	2.405	3.223	3.460	3.944	3.486	1.314	1.548	1.577	1.492
($\eta = 26.367$)	$\log \nu_j$	0.271	0.814	-1.134	-0.108	-0.434	0.564	0.515	0.014	-0.119	-0.021	0.793	0.898	0.636	0.988
Lasso (obli)	δ_j	2.358	1.773	3.570	2.737	3.053	2.394	3.215	3.453	3.939	3.481	1.307	1.540	1.569	1.484
($\eta = 48.329$)	$\log \nu_j$	0.264	0.799	-1.141	-0.117	-0.416	0.573	0.540	0.015	-0.120	-0.022	0.770	0.911	0.623	0.998

Notes. Factor correlation from oblique rotation (Oblimin): $r = .611$. Identification constraints are printed in gray. obli = oblique (a priori correlated latent factors). ortho = orthogonal (a priori uncorrelated latent factors).

Figure 1

Distribution of bias (black) and RMSE (gray) estimates across items for each simulation condition. (L = number of latent tarits. r = true correlation between latent traits. m = number of items per trait. simple / complex = type of $\boldsymbol{\alpha}$ *structure. Lasso / Rotate = model variant. ortho = orthogonal. obli = oblique.)*

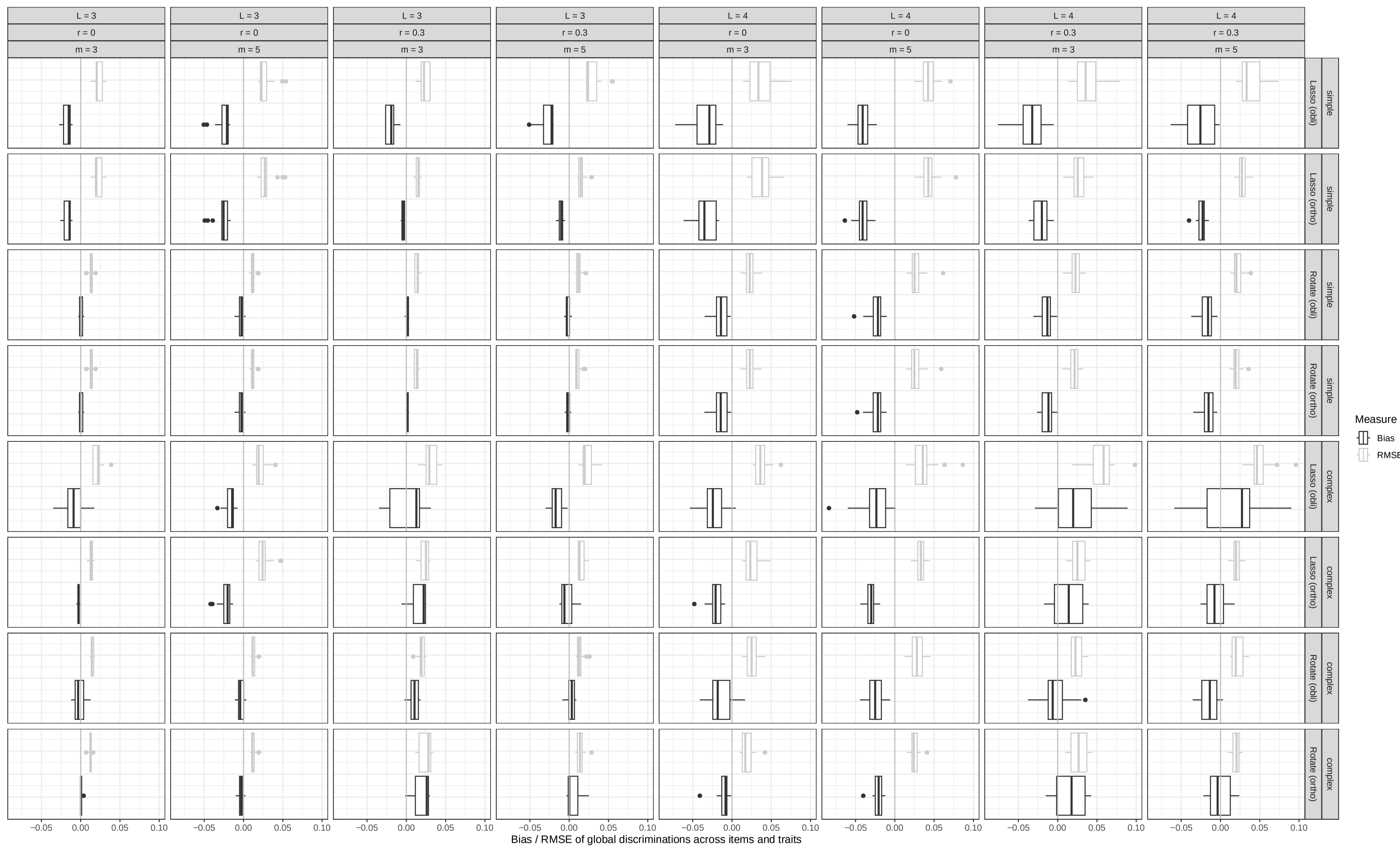


Figure 2

Mean Correct Estimation Rate (CER) estimates for each simulation condition. Estimates for the different model variants are shown on the x-axis and indicated by different shapes as detailed in the legend on the right-hand side. (L = number of latent tarits. r = true correlation between latent traits. m = number of items per trait. simple / complex = type of $\boldsymbol{\alpha}$ *structure. Lasso / Rotate = model variant. ortho = orthogonal. obli = oblique.)*

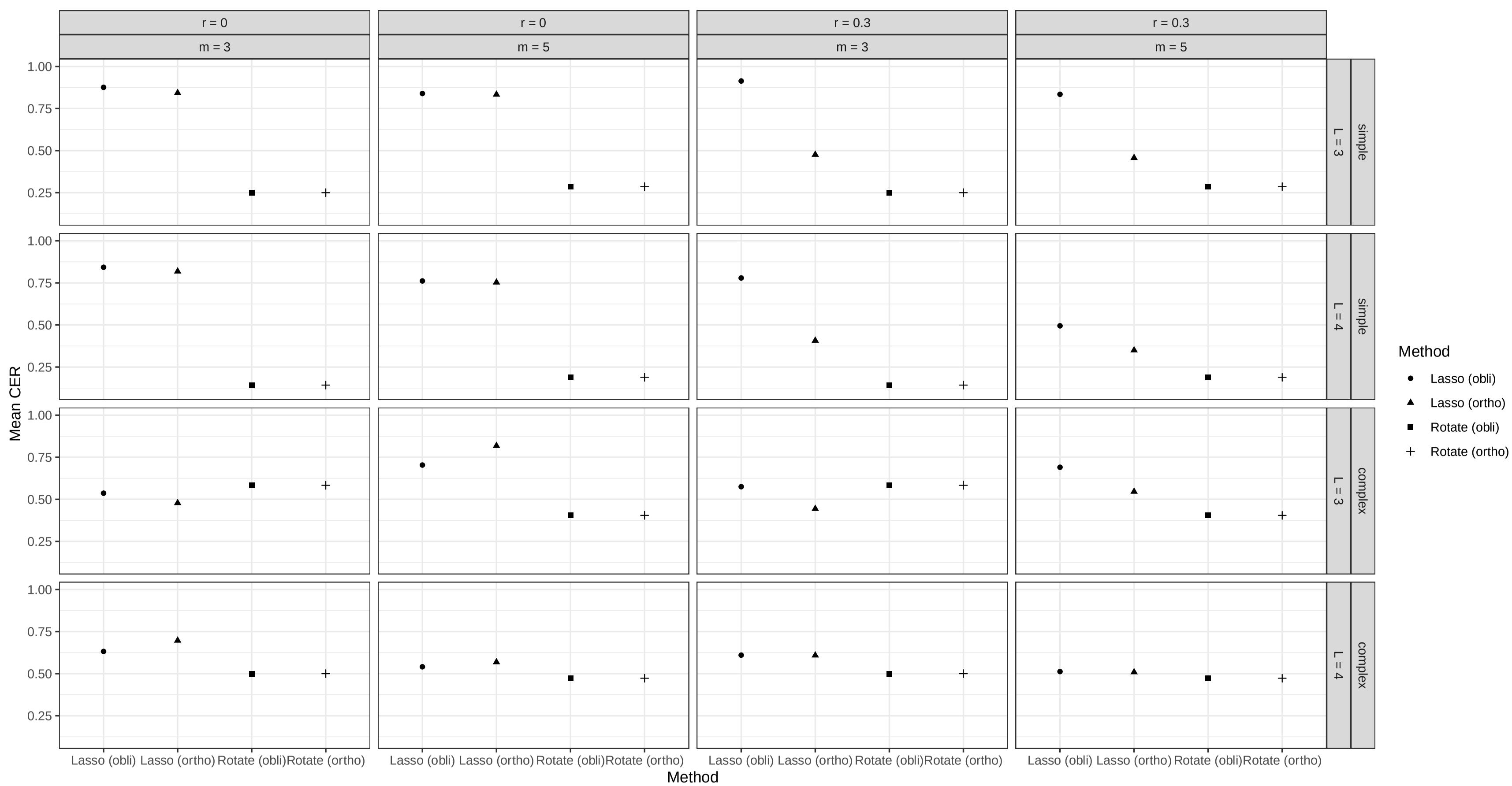
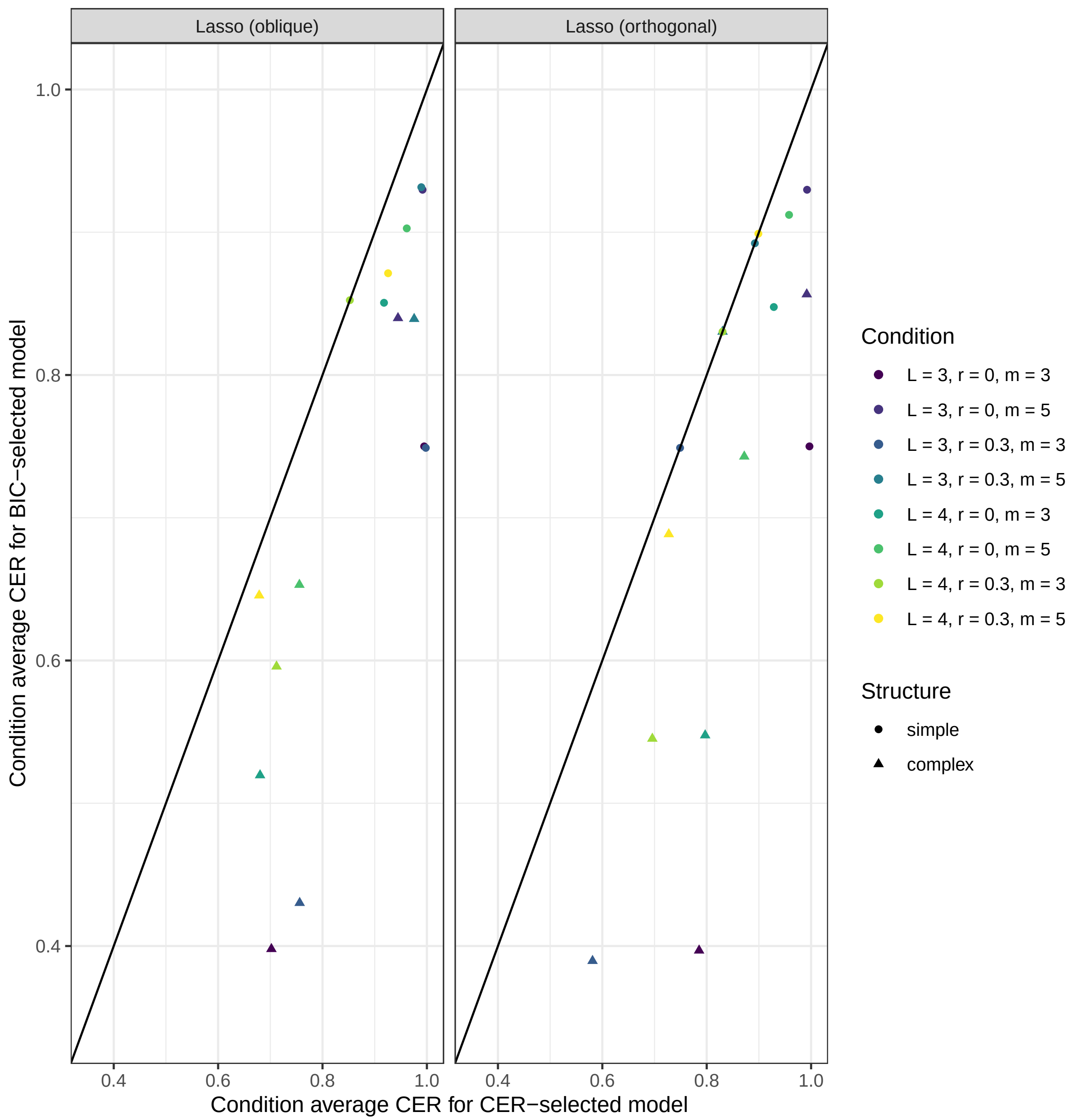


Figure 3

Condition average CER for the BIC-selected model (y-axis) against condition average CER for the CER-selected model (x-axis), shown in two separate panels (lasso with oblique latent covariance matrix on the left and lasso with orthogonal latent covariance matrix on the right). Simulation conditions (in terms of number of latent traits (L), latent factor correlation (r), and number of items per trait (m)) are shown in different colours as indicated by the legend on the right-hand side (under "Condition"). Different $\boldsymbol{\alpha}$ *structures are represented by different shapes as indicated by the legend on the right-hand side (under "Structure").*



Appendix

First derivative of the CMP Variance

For the second derivatives in terms of δ_j and α_{jl} in Equations 17–18, we need the derivative of the variance V in terms of δ_j and α_{jl}. That is,

$$\frac{\partial V(\mu_{jk_1,\dots,k_L}, \nu_j)}{\partial \alpha_{jl}} = \frac{\partial \mathbb{E}_X(X^2)}{\partial \alpha_{jl}} - \frac{\partial \mu^2_{jk_1,\dots,k_L}}{\partial \alpha_{jl}} \tag{A1}$$

$$= \frac{\mu_{jk_1,\dots,k_L} q_{k_l}}{V(\mu_{jk_1,\dots,k_L}, \nu_j)} \mathbb{E}_X(X^3 - \mu_{jk_1,\dots,k_L} X^2) - 2q_{k_l} \mu^2_{jk_1,\dots,k_L}, \tag{A2}$$

and

$$\frac{\partial V(\mu_{jk_1,\dots,k_L}, \nu_j)}{\partial \delta_j} = \frac{\partial \mathbb{E}_X(X^2)}{\partial \delta_j} - \frac{\partial \mu^2_{jk_1,\dots,k_L}}{\partial \delta_j} \tag{A3}$$

$$= \frac{\mu_{jk_1,\dots,k_L}}{V(\mu_{jk_1,\dots,k_L}, \nu_j)} \mathbb{E}_X(X^3 - \mu_{jk_1,\dots,k_L} X^2) - 2\mu^2_{jk_1,\dots,k_L}. \tag{A4}$$

The first equality in both equation holds because for any random variable W it holds that $\mathbb{V}(W) = \mathbb{E}(W^2) - \mathbb{E}(W)^2$. Taking the derivative of $\mu^2_{jk_1,\dots,k_L}$ with regard to α_{jl} and δ_j is trivial. To take the derivative of $\mathbb{E}_X(X^2)$ with regard to α_{jl} and δ_j, we used results provided in Huang (2017) and derivation rules.