technische universität
dortmund

fakultät
statistik

Item Response Models for Count Data: Generalizations and Estimation Algorithms

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund

Vorgelegt von

Marie Beisemann

geboren in Witten

Dortmund, März 2024

Amtierender Dekan:
Prof. Dr. Philipp Doebler

Gutachter:
Prof. Dr. Philipp Doebler (Technische Universität Dortmund)
Prof. Dr. Andreas Groll (Technische Universität Dortmund)

Tag der Prüfung:
23. Mai 2024

Für meine Großmutter Margarete Beisemann

Abstract

Item response theory (IRT) represents a statistical framework within which responses to psychological tests can be modelled. A psychological test consists of a set of items (e.g., tasks to solve or statements to rate) to which a person taking the test responds. IRT assumes that responses are influenced by respondents' latent traits (e.g., personality traits or cognitive abilities) as well as by items' characteristics (e.g., difficulty). IRT models exist for a variety of different response types; the focus of this thesis lies on count responses. These can for example be generated by cognitive tests measuring idea fluency (counts: number of ideas), as process data during test taking (counts: number of clicks), or by reading proficiency assessments (counts: number of errors). Previously comparatively understudied, the field of count item response theory (CIRT) has witnessed a steady increase in interest in recent years. As a result, a number of new CIRT models have been proposed that address limitations of previously existing CIRT models, broadening the empirical applicability of CIRT.

An important concern regarding modelling of counts is their dispersion: The most common distribution for counts, the Poisson distribution, assumes its mean equals its variance (so called equidispersion). By relying on the Poisson distribution, prominent CIRT models assume such equidispersion for responses (conditional on the latent trait(s)). Research has found this assumption empirically violated for some tests. A recently introduced unidimensional CIRT model using the Conway-Maxwell-Poisson (CMP) distribution instead, accommodates over- and underdispersed conditional responses as well. Nonetheless, the model maintains some of the restricting assumptions of previous models. Thus, even with new model proposals, CIRT still offers less modelling flexibility than IRT for other response types (such as binary responses).

The present cumulative thesis aims to address three such gaps in the CIRT landscape. In the first article, I propose a unidimensional CIRT model with a conditional CMP response distribution which extends a previously proposed model through the inclusion of another item parameter (i.e., a discrimination parameter). As such a model has previously not been computable with existing estimation methods, I derive a maximum likelihood estimation procedure to this end, using the Expectation-Maximization (EM) algorithm. In the second article, we propose two extensions of this model which allow

the inclusion of item- and person-specific covariates, respectively. Therewith, we allow to investigate explanations for differences between items and participants, respectively. Again, we provide corresponding estimation methods. In the third article, we generalize the unidimensional CIRT model proposed in the first article to a multidimensional count item response model framework, with a focus on exploratory models. We provide a respective estimation procedure, of which we additionally develop a lasso-penalized variant. The articles in this thesis are accompanied by the development of an `R` package that implements the proposed models and estimation methods.

Zusammenfassung

Item-Response-Theorie (IRT) stellt ein statistisches Framework dar, in welchem Antworten aus psychologischen Tests modelliert werden können. Ein psychologischer Test besteht aus einer Menge an Items (z.B. Aufgaben, die gelöst werden müssen oder Aussagen, die bewertet werden müssen), auf die eine den Test bearbeitende Person antworten muss. IRT trifft die Annahme, dass die Antworten durch latente Merkmale der getesteten Personen (z.B. Persönlichkeitsmerkmale oder kognitive Fähigkeiten) sowie durch Eigenschaften der Items (z.B. Schwierigkeit) beeinflusst werden. IRT-Modelle sind für eine Reihe verschiedener Antwortarten verfügbar; der Schwerpunkt dieser Dissertation liegt auf Zähldaten. Diese entstehen zum Beispiel aus kognitiven Tests, die Ideenflüssigkeit messen (Zähldaten: Anzahl der Ideen), aus Prozessdaten während der Testdurchführung (Zähldaten: Anzahl von Klicks) oder aus Erfassungen der Lesefähigkeit (Zähldaten: Anzahl der Fehler). Das zuvor vergleichsweise weniger entwickelte Feld der Zähldaten-Item-Response-Theorie (ZIRT) hat in den letzten Jahren einen stetigen Anstieg an Interesse erlebt. In der Folge wurden eine Reihe neuer ZIRT-Modelle vorgeschlagen, die Limitationen der vorangegangenen ZIRT-Modelle adressieren und damit die empirische Anwendbarkeit der ZIRT ausweiten.

Ein wichtiger Punkt in der Modellierung von Zähldaten ist deren Dispersion: Die am häufigsten verwendete Zähldatenverteilung, die Poisson-Verteilung, trifft die Annahme, dass der Erwartungswert der Varianz entspricht (sog. Equidispersion). Durch die Verwendung der Poisson-Verteilung treffen bekannte ZIRT-Modelle ebenso die Annahme equidispersierter Antworten (bedingt auf das oder die latente(n) Merkmal(e)). Forschung hat gezeigt, dass diese Annahme empirisch durch manche Tests verletzt wird. Ein kürzlich vorgeschlagenes eindimensionales ZIRT-Modell, welches die Conway-Maxwell-Poisson- (CMP-) anstelle der Poisson-Verteilung verwendet, kann zusätzlich über- und unterdispersierte bedingte Antworten modellieren. Nichtsdestotrotz behält auch dieses Modell manche der einschränkenden Annahmen der vorherigen Modelle bei. Folglich bietet ZIRT, trotz neuer Modellentwicklungen, immer noch weniger Modellierungsflexibilität als IRT für andere Antworttypen (wie z.B. binäre Antworten).

Die vorliegende kumulative Dissertation verfolgt das Ziele drei Forschungslücken in der ZIRT-Landschaft zu schließen. Im ersten Artikel der Arbeit schlage ich ein eindi-

mensionales ZIRT-Modell mit einer bedingten CMP-Verteilung der Antworten vor, welches ein zuvor in der Literatur vorgeschlagenes Modell erweitert durch die Integrierung eines weiteren Itemparameters (nämlich eines Diskriminationsparameters). Ein solches Modell ist zuvor nicht berechenbar gewesen mit existierenden Schätzmethoden, daher entwickele ich Maximum-Likelihood-Schätzmethoden hierfür, basierend auf dem Expectation-Maximization-Algorithmus (EM-Algorithmus). Im zweiten Artikel schlagen wir zwei Erweiterungen dieses Modells vor, die es erlauben, jeweils item- oder personenspezifische Kovariaten in das Modell aufzunehmen. Damit erlauben wir es, Erklärungen für Unterschiede zwischen Items und Personen zu untersuchen. Auch hier entwickeln wir entsprechende Schätzmethoden. Im dritten Artikel generalisieren wir das im ersten Artikel vorgeschlagene eindimensionale ZIRT-Modell zu einem multidimensionalen ZIRT-Framework, mit einem Fokus auf exploratoriven Modellen. Wir entwickeln entsprechende Schätzmethoden, von denen wir zudem eine Lasso-penalisierte Variante erarbeiten. Die Artikel dieser kumulativen Dissertation werden durch die Entwicklung eines `R`-Pakets begleitet, in welchem die vorgeschlagenen Modelle und Schätzmethoden implementiert sind.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor and first reviewer of this thesis, Prof. Dr. Philipp Doebler. I am very grateful for his help and feedback with this thesis, his support of my research, for everything he has taught me, and for the freedom and the opportunities he has given me in my work. I would also like to thank my second reviewer, Prof. Dr. Andreas Groll, as well as Prof. Dr. Jörg Rahnenführer and Dr. Uwe Ligges who formed my PhD committee. I greatly appreciate their support, their expertise, and their time.

I thank my co-authors Dr. Boris Forthmann and Prof. Dr. Heinz Holling for each working with me on one of the projects included in this thesis, respectively, and sharing their ideas and knowledge with me.

My position at the department was (partially) funded through research grant DO 1789/7-1 (from Deutsche Forschungsgemeinschaft, DFG, to Philipp Doebler) for the last three years, for which I am very grateful. Some of the simulations carried out in this thesis were run 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. I am grateful for access to the cluster. I would like to thank Prof. Dr. Paul Silvia and Dr. Simone Heine, for (publicly) sharing their data so that it could be re-analysed for some of the empirical examples in this thesis' articles. Thank you to Dr. Thilo Welz and Dr. Lena Schmid for sharing their dissertation LaTeX templates with me.

Thank you to Sofie Beisemann, Philip Buczak, and Paul Bürkner for proofreading (parts of) this thesis.

I thank my colleagues and my dear friends at the Department of Statistics. Working with and along side them has been a privilege and a pleasure. I am very grateful for the wonderful friendships I have forged here.

I thank my wonderful friends for their love, their support, and for laughing with me through everything. One truly could not be blessed with better friends than mine. I am grateful for every single one of my friends, but only have the space to thank a few by name (for which I apologize): Thank you to my best friend Johanna Rehder, to Stella Mercedes Fingas, to Benedikt Schuler, to Ina Dormuth, and to Julia Duda. Thank you to Paul Bürkner for conversations we have had about this work and the continued support and friendship.

Last but certainly not least, I would like to thank my family. Thank you to Philip Buczak for everything. It is said too often, but I mean it with all my heart: I really, truly could not have done this without you. Thank you to my parents, Britta and Martin Beisemann, and to my sister, Sofie Beisemann. No words can ever adequately express my love and gratitude for you. I love you to the moon and back.

List of Publications

This cumulative thesis is based on the following three manuscripts:

Article 1: **Beisemann, M.** (2022). A flexible approach to modelling 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. https://doi.org/10.1111/bmsp.12273

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution-NonCommercial License.

Article 2: **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. https://doi.org/10.1080/00273171.2023.2288577

Contribution of the thesis author: The author of this thesis formulated the proposed models based on a rough sketch by Prof. Dr. Doebler. She developed the estimation procedures for the models and implemented them in `R` and `C++`. She designed and implemented the simulation studies under the supervision of Prof. Dr. Doebler. She carried out the illustrative data analyses together with Dr. Forthmann. She wrote the initial draft of the manuscript, with input from Prof. Dr. Doebler and Dr. Forthmann.

Article 3: **Beisemann, M.**, Holling, H., & Doebler, P. (2024). Every trait counts: Marginal maximum likelihood estimation for novel multidimensional count data item response models with rotation or ℓ_1–regularization for simple structure. *PsyArXiV pre-print, version 1*. https://doi.org/10.31234/osf.io/fqyjs

Under review with Psychometrika at the time of publishing the thesis.

Contribution of the thesis author: The author of this thesis formulated the proposed model based on a rough sketch by Prof. Dr. Doebler. She developed the estimation procedures for the model and implemented them in `R` and `C++`. She

designed and implemented the simulation study under the supervision of Prof. Dr. Doebler. She carried out the illustrative data analysis under the supervision of Prof. Dr. Doebler and Prof. Dr. Holling. She wrote the initial draft of the manuscript, with input from Prof. Dr. Doebler and Prof. Dr. Holling.

In the course of these projects, the author of this thesis wrote the following `R` package:

(1) `countirt` *R package for item response models for count data*
https://github.com/mbsmn/countirt

Notation

This overview lists the most important notation used throughout this thesis. Any additional notation is explained upon first introduction.

$\mathbb{N}$	Natural numbers
$\mathbb{R}$	Real numbers
N	Number of participants (persons)
M	Number of items
L	Number of latent traits
K	Number of quadrature nodes (unidimensional model) or number of quadrature nodes per trait (multidimensional model) in Gauss-Hermite quadrature
i	Person index
j	Item index
l	Trait index
k	Quadrature node index
X_{ij}	Response for person i to item j (random variable)
x_{ij}	Response for person i to item j (realization)
U_{jc}	cth item covariate for item j (random variable)
u_{jc}	cth item covariate for item j (realization)
T_{ip}	pth person covariate for person i (random variable)
t_{ip}	pth person covariate for person i (realization)
θ_i	Latent trait for person i (unidimensional model)
θ_{li}	lth latent trait for person i (multidimensional model)
α_j	Slope parameter for item j (unidimensional model)

α_{jl}	Slope parameter for item j and trait l (multidimensional model)
δ_j	Intercept parameter for item j
$\boldsymbol{\zeta}_j$	Set of all item parameters for item j
q_k	kth quadrature node
w_k	kth quadrature weight

Vectors and matrices are represented by bold symbols without the indices. E.g., responses for all persons to all items are a $N \times M$ matrix $\mathbf{X}$ (random variable) or $\mathbf{x}$ (realization). Responses to all items for a person i are a vector of length M, $\boldsymbol{X}_i$ (random variable) or $\mathbf{x}_i$ (realization). Responses to one item j for all persons are a vector of length N, $\boldsymbol{X}_j$ (random variable) or $\mathbf{x}_j$ (realization).

Contents

Part I

Summary of Thesis Work

1 Introduction

Item response theory (IRT) is a theoretical framework within the field of psychometrics (i.e., the study of psychological testing and measurement) which describes participants' response behaviour in an assessment situation (e.g., a cognitive test or a personality questionnaire) as a function of the participant's latent trait(s) and characteristics of the assessment tool (Embretson & Reise, 2000). With a test or a self-report, researchers typically intend to measure one or multiple latent traits, i.e., unobservable psychological constructs such as cognitive abilities or personality traits that influence observable behaviour. Examples of such latent traits could be intelligence, processing speed, or creative thinking, to only name a handful.

A psychological test or self-report consists of a set of items which are stimuli to which participants respond, e.g., tasks for participants to solve or statements to which participants express their agreement. IRT assumes that each item has specific characteristics, for example each item in a test has its specific difficulty. These item characteristics are assumed to influence participants' responses along with their latent trait(s) the test or self-report is intended to measure. Statistical IRT models predict participants' responses to each item as a function of these item characteristics and these latent trait(s).

The most widely known IRT models are perhaps those for binary data (see e.g., Baker & Kim, 2004; Embretson & Reise, 2000). This type of data is for example generated by intelligence tests, where items can be answered either correctly or incorrectly. Binary item response models predict the probability to answer the item correctly. Another well known and used class of IRT models in psychology are ordinal IRT models (see e.g., Embretson & Reise, 2000). They can be applied to self-report data, where responses are given on rating scales, for instance. Ordinal item response models predict the probability to answer in specific rating categories. Unsurprisingly in view of their popularity among applied researchers, binary and ordinal IRT models are important foci in the psychometric research landscape. However, binary and ordinal response data are not the only types of data generated by psychological assessment tools. Count data constitute another possible psychometric response type.

Count responses have recently received increasing attention in the psychometric literature (e.g., Forthmann, Gühne, & Doebler, 2020b; Man & Harring, 2019; Qiao, Jiao,

& He, 2023). Apart from more traditional but ever relevant sources of count responses, for example reading assessments (Rasch, 1960; Verhelst & Kamphuis, 2009), count responses that have only more recently become subject of item response analyses include process data, such as fixations during eye-tracking while working on a test (Man & Harring, 2019; Man, Harring, & Zhan, 2022) or action counts during computer-based assessments (Qiao et al., 2023). With the increased interest in and new forms of generation of count responses in psychometric assessment, interest in and the relevance of corresponding statistical methods is also growing.

1.1 Psychometric Count Data

Count data responses arise from psychological tests and self-reports where answers can be summarized into a count. A common source for psychometric count data are cognitive (ability) tests, such as tests to assess processing speed (Baghaei, Ravand, & Nadri, 2019; Doebler & Holling, 2016), intelligence (Ogasawara, 1996), and verbal fluency or divergent thinking (which are related to creative thinking; Forthmann, Holling, Çelik, Storme, & Lubart, 2017; Forthmann, Çelik, Holling, Storme, & Lubart, 2018; Myszkowski & Storme, 2021). Responses to (cognitive) ability tests are often counts, for example, in reading assessments (Rasch, 1960; Verhelst & Kamphuis, 2009) or language proficiency tests (Forthmann, Grotjahn, Doebler, & Baghaei, 2020a). Self-reports can also be sources of count data responses, for instance in clinical psychology (e.g., depressive symptoms or drug use; Magnus & Thissen, 2017; Wang, 2010). Count data responses that have received increasingly more attention in psychometrics are process data during (computer-based) assessments (e.g., eye-tracking fixations or action counts; Man & Harring, 2019; Man et al., 2022; Qiao et al., 2023). Count data which can be structured analogously to participants' responses to items also occur in other fields, for example, in political science in the shape of text data (Proksch & Slapin, 2009) or in researchers' scores on bibliometric indicators in researcher performance (Forthmann & Doebler, 2021; Mutz & Daniel, 2018). As long as the data can be thought of as a responses-to-items structure and one or more latent dimensions are assumed to influence the responses, psychometric count item response models can be applied to such data (e.g., Forthmann & Doebler, 2021). In such settings, statistical models which have strong commonalities with item response models are also frequently applied (e.g., Jentsch, Lee, & Mammen, 2021).

To illustrate, we are going to briefly take a look at a type of count-data generating task that frequently features as an example in the works included in this thesis. These are divergent thinking or fluency tasks (e.g., Forthmann et al., 2016; Forthmann et al.,

2017; Forthmann et al., 2018). An example for such a task would be an alternate uses task (Guilford, 1967), where participants are instructed to name as many different (or as creative; Forthmann et al., 2016; Nusbaum, Silvia, & Beaty, 2014) alternative uses as possible for an everyday object, such as a brick. A participant answers with a list of uses, such as *bookend, door stop, tread, weight, weapon*. This list can be summarized into a count of 5 which this particular participant answered to this particular item. In a test of idea fluency, we would have several different items like this and a sample of participants to answer them. For N participants and M items, we obtain a data set of counts of $N \times M$ dimensions. This is illustrated in Table 1.1. The counts in Table 1.1 could be numbers of ideas, but also any of the above mentioned examples of count data, such as the number of eye-tracking fixations per task, the number of mistakes per text read, etc.

Table 1.1: Illustration of a fictional psychometric count data set for $N = 1000$ participants and $M = 5$ items

Participant	Item 1	Item 2	Item 3	Item 4	Item 5
1	5	10	6	7	9
2	3	6	4	4	5
3	4	9	5	9	6
4	6	12	9	6	9
...	...	...	...	...	...
1000	4	11	3	7	7

The goals of applying a count item response model to psychometric count data (or of applying any item response model to any psychometric data) can be manifold: In general, (count) item response models are applied when (count) responses to multiple items are supposed to be modelled as underlaid by one or more latent traits. The assumption is that differences in this or these latent trait(s) can be – to a certain extent – captured in the observable responses, thus, they in turn can be used to learn about differences on these latent traits. The measurement of unobservable latent traits is typically one of the core interests of applied psychometricians, and item response models can be used to this end (Baker & Kim, 2004). Related are also questions of measurement precision, that is, how well latent traits can be measured by psychological tests or self-report instruments. In this context, psychometricians are often interested in a test's reliability, which in psychological classical test theory is defined (for a unidimensional construct) as the ratio of latent trait variance to manifest response variance (Lord & Novick, 1968). In IRT, corresponding concepts exist and can be investigated with the help of item response models (see e.g., Baker & Kim, 2004). To be able to pursue measurement aims, item response models need to be calibrated, that is, the item properties, also referred to

as item parameters, need to be estimated (Van der Linden, 2018).

1.2 Item Response Models for Count Data

The first count item response model was introduced by Rasch (1960): With Rasch's Poisson Counts Model (RPCM), Rasch (1960) modelled psychometric count responses as a function of a unidimensional latent trait and an item-specific difficulty parameter. The RPCM can be understood as a log-linear model and a special case of a generalized linear mixed model (Baghaei & Doebler, 2019). Since its proposal, the RPCM has been the subject of numerous psychometric works, working on estimation methods (e.g., Jansen, 1986, 1994, 1995; Verhelst & Kamphuis, 2009), model diagnostics (e.g., Holling, Böhning, & Böhning, 2015) or extensions, such as different link functions (Doebler, Doebler, & Holling, 2014), explanatory versions of the RPCM where item or person characteristics are included in the model (e.g., Graßhoff, Holling, & Schwabe, 2013, 2020; Jansen, 2003; Ogasawara, 1996, see Chapter 3.2 for more details on exploratory models), or other extensions (Jansen, 1995; Jansen & van Duijn, 1992; Verhelst & Kamphuis, 2009).

Conditional on the latent trait, the RPCM assumes count responses are Poisson distributed. As a result of a property of the Poisson distribution (i.e., equidispersion; see e.g., Fahrmeir, Heumann, Künstler, Pigeot, & Tutz, 2016), this assumption implies that the conditional count responses are equidispersed, that is, the conditional mean equals the conditional variance of the count responses (see Chapter 2 for more details). Empirically, this is a rather strong assumption which has been found to be violated in real-life applications (e.g., Forthmann et al., 2020b; Forthmann & Doebler, 2021). Violation of the equidispersion assumption can take the shape of overdispersion, that is, the conditional distribution of count responses has a variance larger than its mean. A violation in this direction can cause liberal standard errors and upwards-biased model-based reliability (Forthmann et al., 2020b; Hilbe, 2011). To account for overdispersed conditional count responses, several different approaches have been proposed in the psychometric literature: For instance, Hung (2012) proposed to rely on the negative binomial rather than the Poisson distribution. The negative binomial distribution is a count distribution which allows for overdispersion. Overdispersion can sometimes also be caused by an excess of zero counts, which can be accounted for using the zero-inflated Poisson distribution, as used in the count item response model by Wang (2010). Other approaches have been proposed for example by Verhelst and Kamphuis (2009), Mutz and Daniel (2018).

The equidispersion assumption for the conditional count responses cannot only be violated in the direction of overdispersion but also in the direction of underdispersion (Forthmann et al., 2020b). For underdispersed conditional count responses, the conditional distribution's variance is smaller than its mean (see Chapter 2 for more details). Such a violation may lead to conservative standard errors and downwards-biased model-based reliability (Faddy & Bosch, 2001; Forthmann et al., 2020b). Statistically, this violation is more difficult to address, with the first count item response model able to flexibly account for over-, equi- and underdispersion of conditional count responses having been only recently proposed by Forthmann et al. (2020b). The Conway-Maxwell-Poisson Counts Model (CMPCM; Forthmann et al., 2020b) is a direct generalization of the RPCM as it replaces the Poisson distribution with the Conway-Maxwell-Poisson (CMP) distribution (Conway & Maxwell, 1962; Huang, 2017; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005), which generalizes the Poisson distribution through the addition of a dispersion parameter. Adequately modelling dispersion of conditional responses is approached from a different angle in Tutz (2022).

Apart from equidispersion, the RPCM (Rasch, 1960) makes another assumption which may prove empirically unrealistic: The RPCM implicitly assumes that all items capture the latent trait equally well, that is, they all have the same discrimination. In IRT, item discrimination is the property of an item (potentially included in IRT models as an item parameter) that describes to what extent differences on the latent trait can be depicted in the predicted count responses, akin to the concept of factor loadings. Empirically, this assumption may be violated unless the item construction process was specifically aimed towards equally discriminant items (Myszkowski & Storme, 2021). The CMPCM retains this assumption. Thus, available count item response models which include item-specific discrimination parameters do not at the same time accommodate underdispersion. There is for example a direct extension of the RPCM (also using the Poisson distribution) including a discrimination parameter (Myszkowski & Storme, 2021), a special case within the Generalized Linear Latent Mixed Models (GLAMM) framework (Skrondal & Rabe-Hesketh, 2004). An approach that allows for overdispersed conditional count responses is for example provided by Wang (2010).

The RPCM is a unidimensional count item response model. The accompanying assumption is that only one trait underlies the count responses. This can (but must not necessarily) be a strong assumption as psychological constructs are often complex, decomposing into several subconstructs, and response behaviour can be complexly determined through more than one construct at a time. Multidimensional item response models (MIRM; for an introduction, see e.g., Embretson & Reise, 2000, and Chapter 3.3 for more details) accommodate this on occasion required empirical complexity

by allowing responses to be a function of item characteristics and multiple latent traits. Unidimensional item response models constitute special cases of MIRM. While popular in IRT for binary and ordinal responses (Embretson & Reise, 2000; Chalmers, 2012), the research landscape for multidimensional count item response models is comparably sparse. Apart from some bivariate extensions of the RPCM, with (Myszkowski & Storme, 2021) or without (e.g., Forthmann et al., 2018) discrimination parameters, multidimensional count data models are often embedded in other frameworks, such as GLAMM (Skrondal & Rabe-Hesketh, 2004) or count data factor analysis (Wedel, Böckenholt, & Kamakura, 2003). While item-specific discrimination can be accommodated in these frameworks through model parameters such as factor loadings, dispersion flexibility is typically quite limited due to use of the Poisson distribution, albeit Wedel et al. (2003) allow dispersion flexibility to a certain extent via truncation and also allow for different link functions.

1.3 The Present Thesis

This cumulative thesis consists of three articles which extend the existing count item response model landscape to address some of the outlined restrictions and limitations of previously available count item response models. In the first article (Beisemann, 2022, Chapter 3.1), a unidimensional count item response model is proposed that combines the dispersion flexibility of the CMP distribution with item-specific discriminations. To allow for better understanding of test and item properties as well as driving factors of latent trait differences in the model proposed in Beisemann (2022), two explanatory extensions of this model are proposed in the second article (Beisemann, Forthmann, & Doebler, 2024a, Chapter 3.2). In the third article (Beisemann, Holling, & Doebler, 2024b, Chapter 3.3), the unidimensional count item response model proposed in Beisemann (2022) is further generalized to a multidimensional count item response model framework, with a focus on exploratory models.

For each (set of) new count item response model(s) proposed in this thesis, maximum likelihood estimation procedures are derived and evaluated in simulation studies. In IRT, we typically have two types of estimation tasks: IRT models first have to be calibrated, that is, item parameters have to be estimated, and can then be used for measurement purposes, that is, to estimate participants' latent trait(s) scores (Van der Linden, 2018). The focus of the present work is going to be on calibration. In the first article of this thesis (Beisemann, 2022), an Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is proposed which is extended and generalized in the second and third article, respectively.

The overall aim of this thesis consists of extending item response models for count data, allowing accommodation of previously unaddressed empirical challenges. The estimation procedures were implemented in an `R` package (Chapter 4) to allow for these extensions to be available for use by applied researchers. Impact, challenges, and limitations of the present work as well as avenues for future research are discussed in Chapter 5.

2 Statistical Methods and Background

In this section, statistical and psychometric constructs central to the works of this thesis are going to be introduced and explained to give an overview of the statistical research landscape in which the works of this thesis are embedded.

In the following, let X denote a random (count, unless stated otherwise) variable with realisation x. Let i be a person index, running from 1 to N, and j an item index, running from 1 to M.

2.1 The Conway-Maxwell-Poisson Distribution

For the count item response models in the works of this thesis, we are going to assume conditional responses follow a Conway-Maxwell-Poisson (CMP) distribution. The CMP distribution constitutes a generalization of the well known discrete Poisson distribution for count data, for which the count density is defined as

$$\text{Pois}(x;\lambda) = P(X = x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda), \tag{2.1}$$

where $\lambda \in \mathbb{R}^+$ and $x \in \mathbb{N}_0$ (Fahrmeir et al., 2016). The rate parameter λ determines both the expectation and the variance of the Poisson distribution, that is, $\mathbb{E}(X) = \mathbb{V}\text{ar}(X) = \lambda$, which is referred to as the equidispersion assumption of the Poisson distribution (Fahrmeir et al., 2016). Figure 2.1 shows the Poisson count densities for three different values of λ, illustrating how λ determines both the location and the spread of the distribution (i.e., equidispersion).

In empirical applications, such as count data item response modeling in this thesis, assuming equidispersion can be limiting and unrealistic (Shmueli et al., 2005). When data exhibit more variance than the mean would imply under the Poisson distribution (i.e., $\mathbb{E}(X) < \mathbb{V}\text{ar}(X)$), data are overdispersed. When the opposite is the case (i.e., $\mathbb{E}(X) > \mathbb{V}\text{ar}(X)$), data are underdispersed. In these instances, the Poisson distribution is no longer an appropriate choice to model such data (Shmueli et al., 2005).

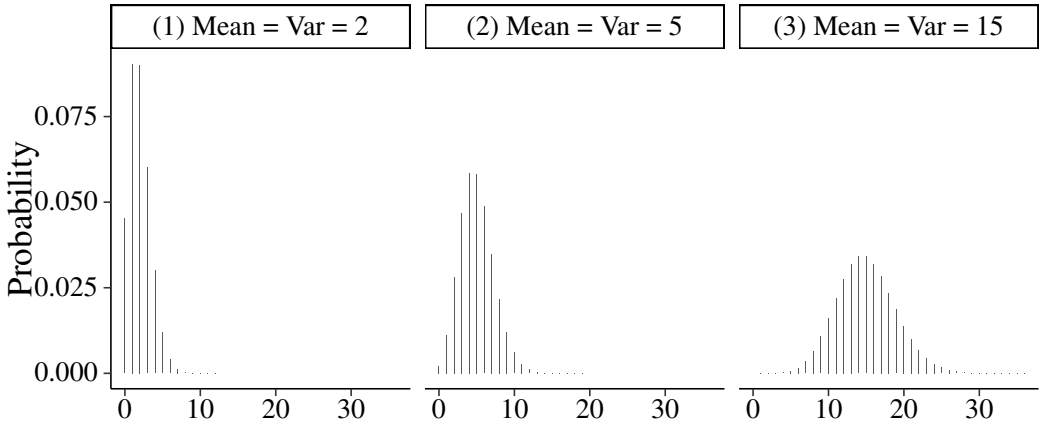


Figure 2.1: Poisson count densities for $\lambda = 2$ (Panel (1)), $\lambda = 5$ (Panel(2)), and $\lambda = 15$ (Panel (3)) illustrating equidispersion.

A more appropriate distribution is the generalization of the Poisson distribution: the Conway-Maxwell-Poisson (CMP) distribution (Conway & Maxwell, 1962). The CMP distribution generalizes the Poisson distribution by including an additional parameter, the dispersion parameter $\nu \in \mathbb{R}_0^+$. The count density is defined as

$$\text{CMP}(x; \lambda, \nu) = P(X = x; \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)} \tag{2.2}$$

(Shmueli et al., 2005), where

$$Z(\lambda, \nu) = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x!)^\nu} \tag{2.3}$$

(Shmueli et al., 2005) is a normalizing constant. Comparing Equations 2.1 and 2.2, one can immediately see that for $\nu = 1$, the CMP distribution simplifies to the Poisson distribution. Further, the Bernoulli (when $\nu \to \infty$, $Z(\lambda, \nu) \to 1 + \lambda$) and the geometric distribution (when $\nu = 0, \lambda < 1$) constitute border cases of the CMP distribution (Shmueli et al., 2005). The CMP distribution is undefined for $\nu = 0$ and $\lambda \geq 1$ (Shmueli et al., 2005). The CMP distribution is a member of the exponential family with sufficient statistics $S_1 = \sum_{i=1}^n x_i$ and $S_2 = \sum_{i=1}^n \log(x_i!)$ for n i.i.d. observations $x_1, \ldots, x_n$ (Shmueli et al., 2005).

For $\nu \neq 1$, the mean and variance of the CMP distribution do not coincide with the rate λ (Shmueli et al., 2005). Huang (2017) argues that for applications where count

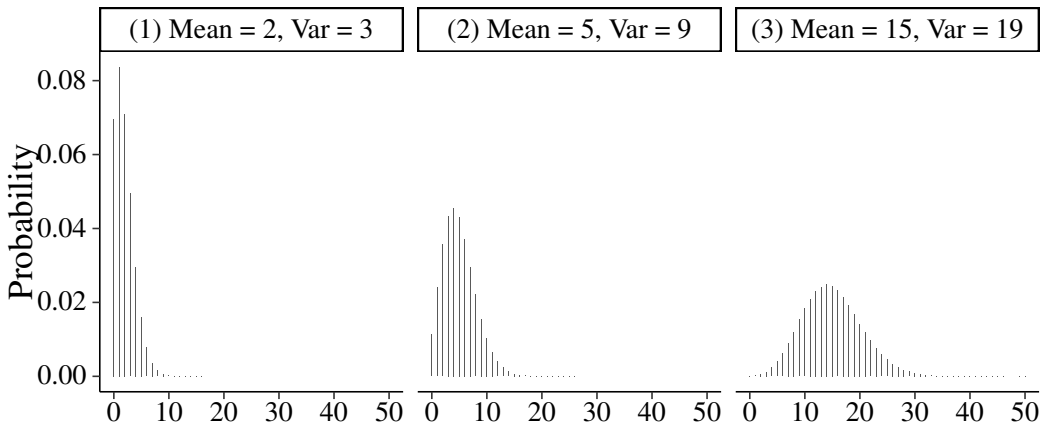


Figure 2.2: CMP$_\mu$ count densities for $\mu = 2$ (Panel (1)), $\mu = 5$ (Panel (2)), and $\mu = 15$ (Panel (3)), all with $\nu = 0.5$, illustrating overdispersion (variance indicated in the panels is rounded to the nearest integer).

responses are predicted, such as regression, a parameterization of the CMP distribution in terms of its mean is beneficial in terms of interpretation. He introduces the mean-parameterized CMP distribution, denoted in the following as CMP$_\mu$. He defines the count density as

$$\text{CMP}_\mu(x; \mu, \nu) = P(X = x; \mu, \nu) = \frac{\lambda(\mu, \nu)^x}{(x!)^\nu} \frac{1}{Z(\lambda(\mu, \nu), \nu)} \tag{2.4}$$

(Huang, 2017), where $\mu \in \mathbb{R}_0^+$ denotes the mean, and with the rate $\lambda(\mu, \nu)$ as a function of μ and ν implicitly defined through the root to

$$0 = \sum_{x=0}^{\infty} (x - \mu) \frac{\lambda^x}{(x!)^\nu} \tag{2.5}$$

(Huang, 2017). Given the data, the different CMP parameterizations are equivalent (Huang, 2017). If $\nu > 1$, the distribution is underdispersed, if $\nu < 1$, the distribution is overdispersed, and – as mentioned above – for $\nu = 1$, the distribution simplifies to the Poisson distribution and is equidispersed (Huang, 2017). This is illustrated in Figure 2.2 for overdispersion and in Figure 2.3 for underdispersion. The means of the distributions are the same in the respective panels (1) – (3) as in Figure 2.1, but the variances are larger in Figure 2.2 compared to Figure 2.1 and smaller in Figure 2.3 compared to Figure 2.1.

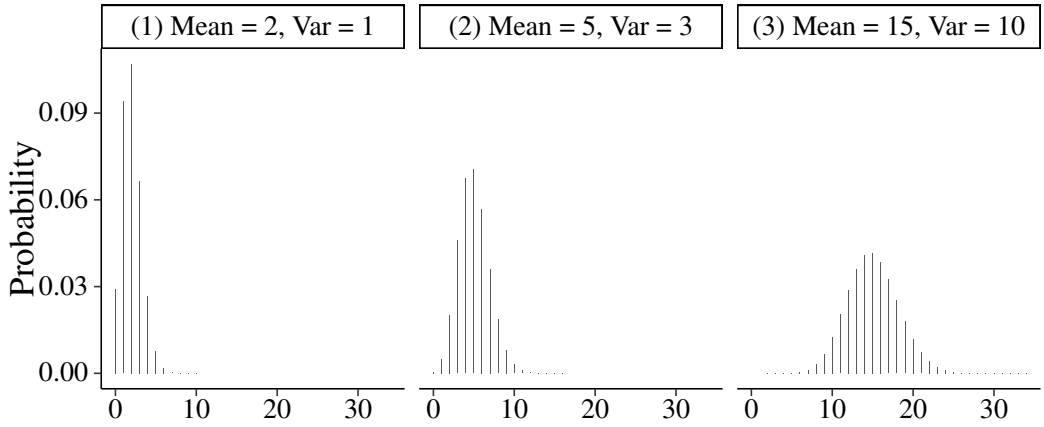


Figure 2.3: CMP_μ count densities for $\mu = 2$ (Panel (1)), $\mu = 5$ (Panel (2)), and $\mu = 15$ (Panel (3)), all with $\nu = 1.5$, illustrating underdispersion (variance indicated in the panels is rounded to the nearest integer).

For the CMP_μ distribution, the expectation $\mathbb{E}(X) = \mu$ and the variance is given by

$$\mathbb{V}\text{ar}(X) = \sum_{x=0}^{\infty} \frac{(x-\mu)^2 \lambda(\mu,\nu)^x}{(x!)^\nu Z(\lambda(\mu,\nu),\nu)} \tag{2.6}$$

(Huang, 2017).

Huang (2017) shows that in the CMP_μ distribution, the parameters μ and ν are orthogonal. Huang (2017) proves that the CMP_μ distribution is a member of the two-parameter exponential family as well as a member of the one-parameter exponential family if ν is fixed. Thus, the CMP_μ distribution lends itself well to generalized linear regression modelling (Huang, 2017), as for example implemented in the `R` package `glmmTMB` (Brooks et al., 2017).

For further properties of the rate- and mean-parameterized CMP distribution, see Shmueli et al. (2005) and Huang (2017), respectively.

2.2 Count Item Response Models

Item Response Theory (IRT) is a popluar statistical framework in psychometrics. For a general introduction, see Van der Linden (2018). Among IRT models, those for count data are less well known than for example the very popular IRT models for binary data

(cf. Birnbaum, 1968; Rasch, 1960). The first count IRT model was introduced by Rasch (1960) and is going to be described in detail in Section 2.2.2. Van der Linden (2018) includes a chapter on some specific count IRT models.

As sources for general count IRT models are scarce beyond single chapters on specific count IRT models in IRT textbooks, I describe a general count IRT model in the following based on work on specific count IRT models (Baghaei et al., 2019; Forthmann et al., 2020b) and textbooks on dichotomous and polytomous IRT models (Baker & Kim, 2004; Reckase, 2009).

Generally spoken, item response models formulate a relationship between person i's latent trait(s), $\boldsymbol{\theta}_i$, and item j's properties, $\boldsymbol{\zeta}_j$, to predict the person's answer to the item. Additionally, one may include person- or item-specific covariates, which I denote with $\boldsymbol{u}_j$ and $\boldsymbol{t}_i$, respectively. IRT models which include item- and/or person-specific covariates are referred to explanatory item response models (for an introduction, see De Boeck & Wilson, 2004). We assume that answers X_{ij} are – conditional on the latent trait(s) $\boldsymbol{\theta}_i$ – distributed according to a count distribution $\tau(\mu, \boldsymbol{\nu})$, where μ is the mean of the distribution and $\boldsymbol{\nu}$ contains the remaining parameters of the distribution, that is, $X_{ij}|\boldsymbol{\theta}_i \sim \tau(\mu_{ij}, \boldsymbol{\nu}_j)$. With this, we assume that the mean of that count distribution, μ_{ij}, depends on the person-specific latent trait(s) and item-specific properties (as well as person and/or item covariates, if applicable), while the remaining parameters $\boldsymbol{\nu}_j$ are only item-specific (albeit in theory, this can also be extended to allow additional dependence on the person-specific latent trait(s)). The mean μ_{ij} is modeled as

$$\mu_{ij} = g(f(\boldsymbol{\theta}_i, \boldsymbol{\zeta}_j, \boldsymbol{u}_j, \boldsymbol{t}_i)), \tag{2.7}$$

where $g(.)$ is an inverse-link function and $f(.)$ models the relationship between $\boldsymbol{\theta}_i$ and $\boldsymbol{\zeta}_j$. IRT models assume local independence (Van der Linden, 2018): When conditioning on the latent trait(s), the items are independent of one another.

2.2.1 The Psychometric Concept of Reliability

Psychometric reliability is a (psychological) test goodness criterion. It quantifies how precisely a test measures the latent trait. The classical test theory (CTT) tradition of psychometrics assumes that test scores Y additively decompose into the true latent trait values T and an error term E, that is, $Y = T + E$ (Lord & Novick, 1968). CTT defines the reliability for a test with test scores Y as $\text{Rel}(Y) = \frac{\mathbb{V}\text{ar}(T)}{\mathbb{V}\text{ar}(Y)}$ (Lord & Novick, 1968), where T are the true latent trait values. With this definition, CTT obtains one reliability value for a test across all possible true latent trait values.

In IRT, reliability is typically defined in relation to or directly as test information. Test information is additively composed of the item information for each item $j = 1, \ldots, M$, that is,

$$\mathcal{I}(\theta) = \sum_{j=1}^{M} \mathcal{I}_j(\theta), \tag{2.8}$$

where $\mathcal{I}_j(\theta)$ is the item information for item j (described for dichotomous and polytomous IRT models e.g., in Baker & Kim, 2004, and analogous for count IRT models). The item information is the Fisher information for that item with respect to the latent trait parameter θ. The item as well as the test information are functions of θ, implying that reliability can vary across the latent trait range in an IRT conceptualization of reliability (described for dichotomous and polytomous IRT models e.g., in Baker & Kim, 2004, and analogous for count IRT models). For a study of the item information under the RPCM, see for example Doebler et al. (2014), Graßhoff et al. (2013, 2020).

An alternative is the empirical (marginal) reliability, which is defined as

$$\text{Rel}_{\text{emp}} = 1 - \frac{\overline{\text{SE}(\theta)^2}}{\mathbb{V}\text{ar}(\theta)} \tag{2.9}$$

(Brown & Croudace, 2014; Green, Bock, Humphreys, Linn, & Reckase, 1984). When estimating the empirical reliability, one can use the latent trait's standard error and variance estimates obtained from one's IRT model. This approach was for example used by Forthmann et al. (2020b) for their count IRT model (the CMPCM, see below). A characteristic of empirical reliability is that this approach only provides one reliability for the test across the whole range of the latent trait. Forthmann et al. (2020b) found that specifically for count IRT models, reliability estimates are biased if equidispersion is assumed by the count IRT model, but over- or underdispersion is present in the data.

2.2.2 Unidimensional Count Item Response Models

This subsection introduces two unidimensional count item response models which were generalized in the works of this thesis. They are important previous works for this thesis which is why they receive special attention in this section. For an extensive overview of other existing unidimensional count item response models, the reader is referred back to Chapter 1. The models are introduced in the following using a parameterization and notation that is consistent with the parameterization and notation of the models

proposed in the works of this thesis as this is more convenient for the reader. Please note that in the original and related publications, in part, other parameterizations and notation were used.

Rasch's Poisson Counts Model

The first count data IRT model was proposed by Rasch (1960): the Rasch Poisson Counts Model (RPCM; Rasch, 1960) which models the expected count μ_{ij} as

$$\mu_{ij} = \exp(\theta_i + \delta_j). \tag{2.10}$$

Referring back to Equation 2.7, the RPCM uses the exponential function for g, an additive relationship between the uni-dimensional latent trait θ_i and one item-specific parameter δ_j (i.e., for an item j, $\boldsymbol{\zeta}_j = \{\delta_j\}$) as f. The latent abilities θ_i are not directly observable, they are latent variables. Depending on the specific estimation approach, we handle them differently. In the context of this thesis, we use the assumption that $\theta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, N$ (as an alternative, one can e.g., assume a gamma distribution for the θ_i; Jansen & van Duijn, 1992). The latent mean is fixed to 0 for identification. In this formulation of the RPCM, the latent variance σ^2 can be estimated (see e.g., Baghaei & Doebler, 2019). In the parameterization used here, δ_j is an item-specific intercept which can be interpreted as the item-specific easiness: The higher δ_j, the higher the expected count μ_{ij}, regardless of the person's latent trait θ_i. For example, if one models the number of ideas, this means that in response to an easier item, all participants, even the less creative ones, are able to generate higher number of ideas compared to a more difficult item. In the original RPCM, there are no item or person covariates (i.e., no $\boldsymbol{u}_j$ or $\boldsymbol{t}_i$), but the RPCM has since been extended to include those (see e.g., Graßhoff et al., 2013, 2020; Jansen, 2003; Ogasawara, 1996).

The RPCM assumes for the conditional response distribution τ the Poisson distribution which has no further parameters (we set $\lambda = \mu_{ij}$). Relying on the Poisson distribution as the conditional response distribution implies that the RPCM assumes equidispersed conditional responses.

Several estimation approaches have been developed for the RPCM, such as conditional maximum likelihood (e.g., Jansen, 1995; Rasch, 1960), and marginal maximum likelihood (MML) using for instance an Expectation-Maximization algorithm (e.g., Jansen, 1995) but also other MML approaches (e.g., Jansen & van Duijn, 1992). In the formulation used in Equation 2.10, it is easy to see that the RPCM also constitutes a special case of a generalized linear mixed model (GLMM; see e.g., McCulloch & Searle,

2004), with the θ_i modelled as a random intercept (see e.g., Baghaei & Doebler, 2019). As such, the RPCM can be estimated in any GLMM estimation framework, such as the `R` package `lme4` (Bates, Mächler, Bolker, & Walker, 2015), as explained in detail in Baghaei and Doebler (2019).

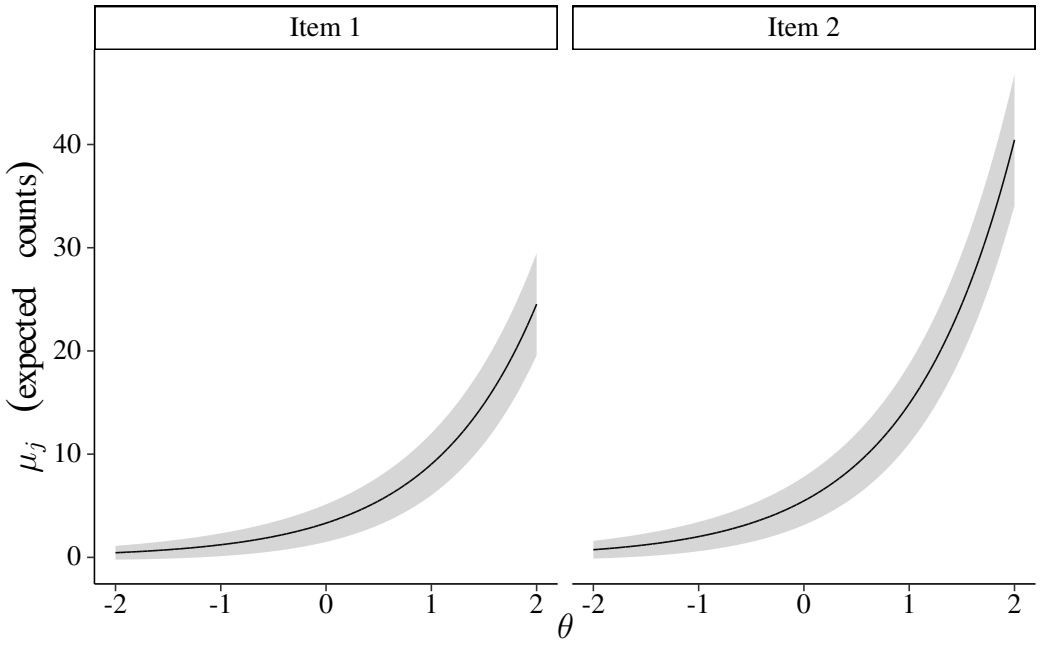


Figure 2.4: Item response curves under the RPCM for an item with easiness $\delta_1 = 1.2$ (Item 1) and an item with easiness $\delta_2 = 1.7$ (Item 2) (shaded ribbons indicate ± 1 conditional SD).

Figure 2.4 illustrates the expected responses μ_j per item j as a function of the latent trait θ for two different easiness values (item 1 with easiness $\delta_1 = 1.2$ and item 2 with easiness $\delta_2 = 1.7$). These plots are called item response curves. We can see that under the RPCM, the item easiness determines both where the item response curve crosses the y-axis for $\theta = 0$ (i.e., what number of counts we expect from a person of average ability) as well as – due to the inverse log-link function – the steepness of the curve. The traditional psychometric understanding – stemming from IRT models for binary responses – associates the steepness of item response curves with the item's capability to differentiate between persons of different ability. It is slightly more complex for count IRT items, nonetheless, psychometricians are often interested in estimating a separate parameter for this property. This parameter is usually referred to as the item discrimination (or simply slope, depending on the parameterization; Baker & Kim, 2004).

Figure 2.4 further illustrates the equidispersion assumption of the RPCM: The conditional standard deviation is fully determined by the conditional mean, that is, by the item's easiness and the person's latent trait level. Again, psychometrically, we might wish to have a separate parameter which allows to modulate the mean-implied conditional variance. This latter limitation of the RPCM was recently addressed by Forthmann et al. (2020b) with the introduction of a new count IRT model, the Conway-Maxwell-Poisson Counts Model.

Conway-Maxwell-Poisson Counts Model

The Conway-Maxwell-Poisson Counts Model (CMPCM; Forthmann et al., 2020b) constitutes a generalization of the RPCM as it uses the same model formulation given in Equation 2.10 but assumes the CMP_μ distribution as the conditional response distribution τ rather than the Poisson distribution. With the use of the CMP_μ distribution, the CMPCM has an additional item-specific (or global, if constrained equal across items) dispersion parameter, ν_j (but note that in the original model formulation, Forthmann et al., 2020b, use the inverse of ν_j as the dispersion parameter). As illustrated in Figure 2.5, the dispersion parameter allows to model overdispersion, that is, more conditional variance than the mean implies (see items 3 and 4 which have the same conditional mean as items 1 and 2 in Figure 2.4, but more conditional variance), and underdispersion, that is, less conditional variance than the mean implies (see items 5 and 6 which have the same conditional mean as items 1 and 2 in Figure 2.4 but less conditional variance).

Analogously to the RPCM, the CMPCM can be understood as a special case within the GLMM framework (Forthmann et al., 2020b; Huang, 2017). It can therefore be estimated with corresponding GLMM software as long as the software implements the CMP_μ distribution as a response distribution. This is the case for the `glmmTMB` package (Brooks et al., 2017) in `R`. Forthmann et al. (2020b) described how the CMPCM can be estimated with the help of `glmmTMB`.

In IRT, models such as the CMPCM and the RPCM with are typically referred to as one-parameter models. Two-parameter IRT models additionally include another item-specific parameter, (depending on the parameterization) a discrimination or slope parameter (see e.g., Baker & Kim, 2004, for an introduction to one- and two-parameter IRT models for binary data and their estimation routines)[1]. This parameter allows to modulate the response curve steepness beyond the steepness that is implied by the item

[1] Arguably the most well known IRT models have been developed for binary data, an introduction to which is beyond the scope of this thesis.

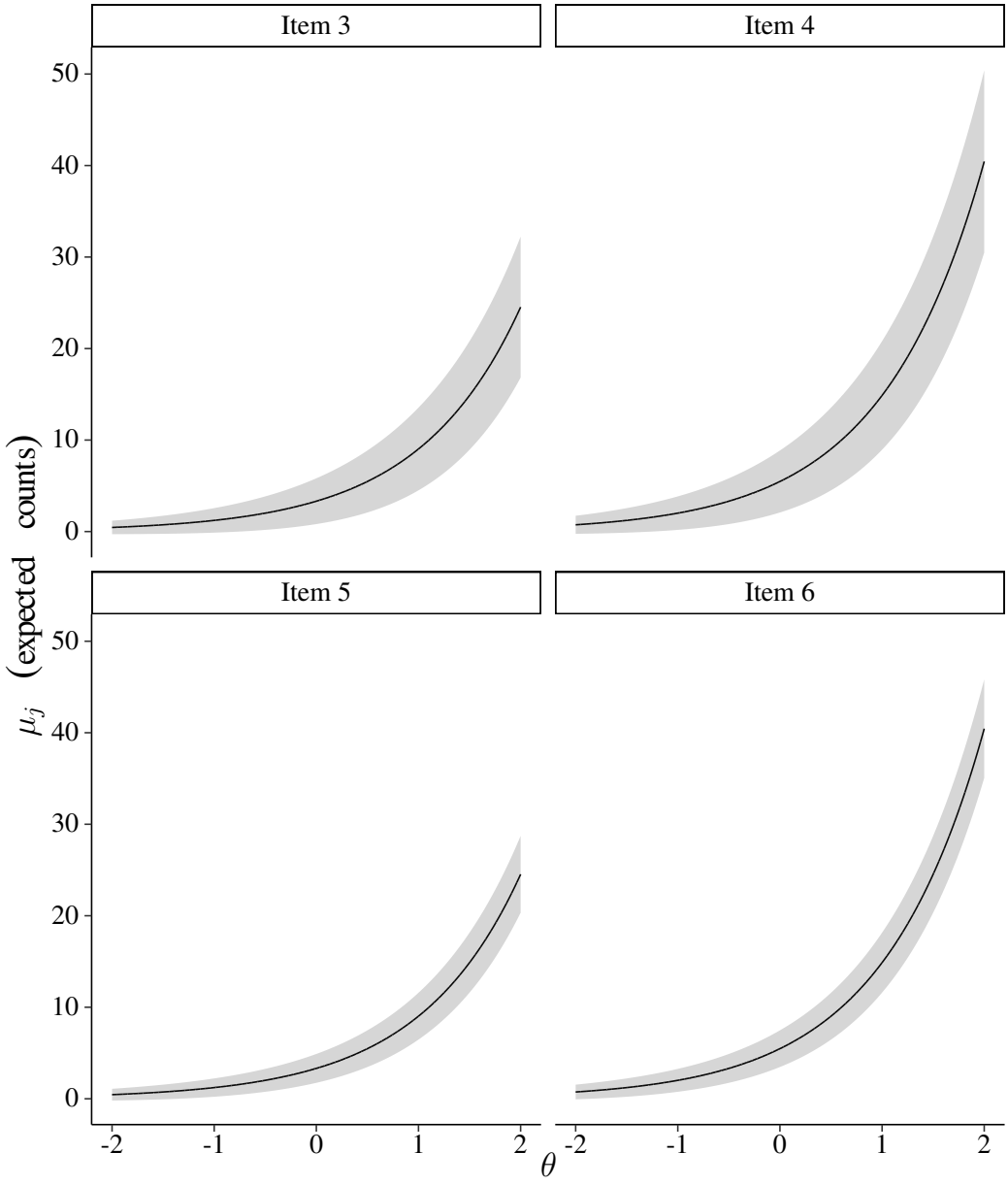


Figure 2.5: Item response curves under the CMPCM for items with easiness $\delta_3 = \delta_5 = 1.2$ (Items 3 and 5) and items with easiness $\delta_4 = \delta_6 = 1.7$ (Items 4 and 6), items 3 and 4 exhibit conditional overdispersion ($\nu_3 = \nu_4 = 0.4$) and items 5 and 6 exhibit conditional underdispersion ($\nu_5 = \nu_6 = 1.4$) (shaded ribbons indicate ± 1 conditional SD).

easiness in count IRT models with an inverse log-link function. The formulation for the RPCM and the CMPCM in Equation 2.10 implicitly assumes the discriminations or slopes for all items to be fixed to 1, and allows to estimate the latent trait variance σ^2. This formulation aligns with estimation within a GLMM framework and highlights how the RPCM and the CMPCM can be understood as special cases in the GLMM framework (Baghaei & Doebler, 2019; Forthmann et al., 2020b). Alternatively, we could formulate the prediction of the expected counts μ_{ij} in the RPCM and the CMPCM as $\mu_{ij} = \exp(\alpha\theta_i + \delta_j)$, assuming $\theta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, for $i = 1, \ldots, N$, that is, estimating one slope parameter α that is equal across all items and fixing the latent trait variance to 1 for model identification. I explain this alternative formulation here as this generalizes more organically to the two-parameter count IRT model using the CMP_μ distribution I proposed in the first article of this thesis (see Chapter 3.1).

2.2.3 Multidimensional Item Response Models

While multidimensional IRT (MIRT) models have become a staple in psychometric analyses for binary and ordinal data (Chalmers, 2012), MIRT models for count data are comparably underdeveloped. Multidimensional approaches for count data have predominantly been developed in other frameworks than IRT, such as factor analysis (Wedel et al., 2003) or generalized linear additive mixed models (Skrondal & Rabe-Hesketh, 2004), which both have parallels to IRT. In the third article of this thesis (see Chapter 3.3), we proposed a MIRT model for count data, filling this gap in the literature landscape.

As the literature on specifically MIRT models for count data is scarce, I will briefly give some background on MIRT models for binary data. A popular unidimensional model for binary data is the two-parameter logistic (2PL) model (Birnbaum, 1968) which models the binary responses X_{ij} as conditionally Bernoulli distributed with success probability π_{ij}. In the multidimensional extension of the 2PL model, the success probability π_{ij} for person i and item j is modelled as

$$\pi_{ij} = \text{expit}\left(\sum_{l=1}^{L} \alpha_{lj}\theta_{li} + \delta_j\right) \tag{2.11}$$

(Reckase, 2009), where we have L latent trait θ_{li} with a trait $\times$ item-specific slope α_{jl} (which are also often written in matrix notation as $\boldsymbol{\alpha} \in \mathbb{R}^{M\times L}$, in this thesis referred to as the discrimination matrix) and an item-specific intercept δ_j. Note that due to the different inverse link function, the units for these parameters are different to what

they are in count IRT models, as would their interpretation and item response curves (or surfaces, in the multidimensional case) be. The expit function is the inverse logit function, defined as $\text{expit}(x) = \exp(x)/(1 + \exp(x))$. This type of MIRT model is also called a compensatory MIRT model (Reckase, 2009). One frequently assumes a multivariate normal distribution as the joint distribution for the latent traits (Chalmers, 2012), where one typically assumes latent trait means to be 0 and fixes the diagonal of the covariance matrix to 1 for identification.

As I will describe in more detail in Section 2.3, a very popular method for estimating (M)IRT models is the Expectation-Maximization algorithm (Bock & Aitkin, 1981; Dempster et al., 1977). A range of popular MIRT models for binary, polynomous, and ordinal data are implemented in the `R` package `mirt` (Chalmers, 2012).

Confirmatory and Exploratory Models

The MIRT literature usually distinguishes between exploratory and confirmatory MIRT models; a distinction also made in the factor analysis literature (Chalmers, 2012; McDonald, 1999). The discrimination matrix $\boldsymbol{\alpha}$ can be regarded as analogous to a factor loading matrix in factor analysis (e.g., McDonald, 2000, for MIRT model for binary data). A confirmatory MIRT model imposes constraints on the discrimination matrix $\boldsymbol{\alpha}$ so that specific relationships between each item and the factors are modelled. By fitting the confirmatory MIRT model to the data and assessing the model fit, one can assess whether the theoretical assumptions that informed the imposed constraints find empirical support (McDonald, 2000).

Exploratory MIRT models aim to estimate the whole discrimination matrix $\boldsymbol{\alpha}$ from the data. Thereby, relationships between items and factors are not modelled according to substantive considerations but estimated empirically from the data (McDonald, 2000). Due to neither the α_{jl} nor the θ_{li} being fixed and their linear relationship, there are an infinite number of possible solutions for $\boldsymbol{\alpha}$ – a phenomenon referred to as rotational indeterminancy (Scharf & Nestler, 2019). There are different criteria for choosing a preferred solution for $\boldsymbol{\alpha}$ (Scharf & Nestler, 2019). Popular ones are those which imply a simple structure for the $\boldsymbol{\alpha}$ matrix (Thurstone, 1947), in simple terms that is, a discrimination matrix in which each item is associated with predominantly one latent trait (see e.g., Trendafilov, 2014, for a more specific definition). For a setting with two latent traits and six items, a perfect simple structure would for example be given by

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{11} & 0 \\ \alpha_{21} & 0 \\ \alpha_{31} & 0 \\ 0 & \alpha_{42} \\ 0 & \alpha_{52} \\ 0 & \alpha_{62} \end{pmatrix}. \tag{2.12}$$

Rotation for Simple Structure

To obtain a simple structure solution for $\boldsymbol{\alpha}$, one first estimates the MIRT model with sufficient identification constraints and rotates the resulting $\boldsymbol{\alpha}$ matrix after dropping the identification constraints. Denote the $N \times L$ matrix of latent traits with $\boldsymbol{\Theta}$. The $\boldsymbol{\alpha}$ matrix is rotated by multiplying with a rotation matrix $\boldsymbol{V} \in \mathbb{R}^{L \times L}$ which satisfies that

$$\boldsymbol{\alpha}\boldsymbol{\Theta}^T = \boldsymbol{\alpha}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{\Theta}^T \tag{2.13}$$

(Scharf & Nestler, 2019; Trendafilov, 2014). The rotation matrix $\boldsymbol{V}$ is selected according to different criteria depending on the specific rotation method (Scharf & Nestler, 2019). Rotation methods can be classified into orthogonal and oblique approaches (Trendafilov, 2014). An example for an orthogonal rotation technique, which assumes uncorrelated latent traits, is Varimax (Kaiser, 1958, 1959). Denote the elements of the rotated discrimination matrix ($\boldsymbol{\alpha}^* = \boldsymbol{\alpha}\boldsymbol{V}$) as α^*_{jl} for the element in the jth row and lth column. The Varimax approach chooses the orthogonal rotation matrix $\boldsymbol{V}$ that maximizes the sum of the column-wise variances of discriminations (analogous to factor loadings) $\mathbb{V}\text{ar}(\boldsymbol{\alpha}^*_l)$, where

$$\mathbb{V}\text{ar}(\boldsymbol{\alpha}^*_l) = \sum_{j=1}^{M} \alpha^{*4}_{jl} - \frac{1}{M}\left(\sum_{j=1}^{M} \alpha^{*2}_{jl}\right)^2 \tag{2.14}$$

(Trendafilov, 2014). An example for an oblique rotation technique, which allows correlations between the latent traits, is Oblimin (Carroll, 1957; Clarkson & Jennrich, 1988). For the Oblimin family (Carroll, 1957; Clarkson & Jennrich, 1988), the goal is to find the oblique rotation matrix V that minimizes the function

$$f(\boldsymbol{\alpha}^*) = \sum_{l \neq l^*}\left(M\sum_{j=1}^{M} \alpha^{*2}_{jl}\alpha^{*2}_{jl^*} - \kappa\sum_{j=1}^{M} \alpha^{*2}_{jl}\sum_{j=1}^{M} \alpha^{*2}_{jl^*}\right) \tag{2.15}$$

(Clarkson & Jennrich, 1988). The value κ is chosen depending on the specific variant; it must hold that $0 \leq \kappa \leq 1$.

A range of orthogonal and oblique rotation criteria, including Varimax and Oblimin, are implemented in the `R` package `GPArotation` (Bernaards & Jennrich, 2005) which implements the gradient projection algorithm (GPA; Jennrich, 2001, 2002, 2004).

2.3 Marginal Maximum Likelihood Estimation in IRT

Apart from the articles on maximum likelihood approaches for specific count IRT models (e.g., Jansen, 1995, 2003, for the RPCM), there is little literature on (marginal) maximum likelihood (MML) estimation for count IRT models in general. There is however ample good literature on (M)ML estimation for dichotomous and polytomous IRT models (e.g., Baker & Kim, 2004; Reckase, 2009). I have written the following general section on MML for count IRT models based on these works (Baker & Kim, 2004; Reckase, 2009) and adapted them to the generic count IRT model introduced in Section 2.2.

For any count IRT model, let $P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j, \boldsymbol{u}_j, \boldsymbol{t}_i)$ denote the probability of observing x_{ij} counts for person i responding to item j. With the assumption of local independence, the probability for the response vector $\boldsymbol{x}_i$ for person i to all M items can be obtained as

$$P(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\zeta}, \boldsymbol{u}, \boldsymbol{t}_i) = \prod_{j=1}^{M} P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j, \boldsymbol{u}_j, \boldsymbol{t}_i). \tag{2.16}$$

If estimated with a marginal maximum likelihood approach, one further places a distribution assumption on the latent trait(s) which often is a uni- or multivariate normal distribution (but see also e.g., Jansen & van Duijn, 1992), depending on whether the model is uni- or multidimensional. Let Ψ generically denote the density of the distribution assumed for the latent trait(s), which depends on parameters $\boldsymbol{\xi}$. The probability of response vector $\boldsymbol{x}_i$ for person i, marginalized over $\boldsymbol{\theta}_i$, is

$$P(\boldsymbol{x}_i; \boldsymbol{\zeta}, \boldsymbol{u}, \boldsymbol{t}_i) = \int_{\boldsymbol{\theta}_i} P(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\zeta}, \boldsymbol{u}, \boldsymbol{t}_i) \Psi(\boldsymbol{\theta}_i; \boldsymbol{\xi}) d\boldsymbol{\theta}_i. \tag{2.17}$$

Further assuming the N persons have been sampled independently from one another, the marginal likelihood for $\boldsymbol{\zeta}$ given the data $\boldsymbol{x} \in \mathbb{N}_0^{N \times M}$ is

$$L_m(\boldsymbol{\zeta}; \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{t}) = \prod_{i=1}^{N} \int_{\boldsymbol{\theta}_i} P(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\zeta}, \boldsymbol{u}, \boldsymbol{t}_i) \Psi(\boldsymbol{\theta}_i; \boldsymbol{\xi}) d\boldsymbol{\theta}_i. \tag{2.18}$$

Dropping the possible item and person covariates in the following for readability, the marginal likelihood can be written more succinctly as

$$L_m(\boldsymbol{\zeta}; \boldsymbol{x}) = \prod_{i=1}^{N} \int_{\boldsymbol{\theta}_i} P(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\zeta}) \Psi(\boldsymbol{\theta}_i; \boldsymbol{\xi}) d\boldsymbol{\theta}_i. \tag{2.19}$$

Usually, one takes the logarithm of the marginal likelihood for optimization, obtaining the marginal log-likelihood $LL_m(\boldsymbol{\zeta}; \boldsymbol{x}) = \log(L_m(\boldsymbol{\zeta}; \boldsymbol{x}))$. Maximizing (the logarithm of) Equation 2.19 in terms of item-specific model parameters $\boldsymbol{\zeta}$ – in the sense of estimating the item-specific model parameters – constitutes the goal of calibration of IRT models. For specific count IRT models, such as the RPCM, estimation can also be tackled via other approaches, for example using a conditional maximum likelihood approach (Jansen, 1995). Outside of these cases, estimation usually takes a marginal maximum likelihood (MML) approach. MML is preferable to joint maximum likelihood (JML; i.e., assuming and estimating fixed latent abilities for the set of observed persons) in this instance, as in JML, the number of model parameters grows with the number of persons in the sample grows, potentially negatively impacting the estimator's consistency. MML instead integrates over the latent abilities, eliminating the problem. However, MML faces the challenge that the integral in Equation 2.19 is analytically intractable. It is beyond the scope of this thesis to give a comprehensive overview over all the possible approaches of how this challenge can be addressed. Instead, in the following, I am only going to briefly give an introduction to the methods we used in the articles of this thesis to this end.

2.3.1 The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a popular algorithm for marginal maximum likelihood estimation problems such as the one presented above. For a general introduction, see McLachlan and Krishnan (2007) and for an IRT-specific discussion, see Baker and Kim (2004). The EM algorithm is a powerful tool in ML estimation whenever the estimation problem can be re-formulated as an incomplete-data problem (McLachlan & Krishnan, 2007). For instance, in the IRT context, we can regard the observed responses $\boldsymbol{x}$ as the incomplete data, lacking the (unobservable) latent trait(s) $\boldsymbol{\theta}$, which together would make up the complete data $(\boldsymbol{x}, \boldsymbol{\theta})$.

One can formulate the (log-)likelihood for the complete data (McLachlan & Krishnan, 2007), which I denote here with $LL_c(\boldsymbol{\zeta}; \boldsymbol{x}, \boldsymbol{\theta})$. In many cases, the complete-data likelihood is easier to maximize than the incomplete-data marginal likelihood (McLachlan & Krishnan, 2007). The EM algorithm takes advantage thereof as well as of the result that the same estimates $\hat{\boldsymbol{\zeta}}$ maximize the complete-data likelihood and the incomplete-data marginal likelihood (McLachlan & Krishnan, 2007). Directly maximizing $LL_c(\boldsymbol{\zeta}; \boldsymbol{x}, \boldsymbol{\theta})$ is not possible, as the $\boldsymbol{\theta}$ have not been observed. Instead, the EM algorithm takes an iterative approach. Starting with initial values $\boldsymbol{\zeta}^{(0)}$ for $\boldsymbol{\zeta}$, which are assumed known for the time being, the EM algorithm computes the conditional (on $\boldsymbol{x}$) (posterior) expectation of $LL_c(\boldsymbol{\zeta}; \boldsymbol{x}, \boldsymbol{\theta})$ in its initial Expectation (E) step, that is,

$$Q(\boldsymbol{\zeta}; \boldsymbol{\zeta}^{(0)}) = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\zeta}^{(0)}}(LL_c(\boldsymbol{\zeta}; \boldsymbol{x}, \boldsymbol{\theta})|\boldsymbol{x}) \tag{2.20}$$

(McLachlan & Krishnan, 2007). In its first Maximization (M) step, the EM algorithm proceeds to maximize (the assumed given) $Q(\boldsymbol{\zeta}; \boldsymbol{\zeta}^{(0)})$ in terms of $\boldsymbol{\zeta}$, obtaining a new set of current item parameters, $\boldsymbol{\zeta}^{(1)}$ (McLachlan & Krishnan, 2007). Subsequently, the EM algorithm iterates between E and M step, where $\boldsymbol{\zeta}^{(0)}$ is replaced by the respective last obtained item parameter estimates (McLachlan & Krishnan, 2007). That is, in the ith iteration of the EM algorithm, the E step computes $Q(\boldsymbol{\zeta}; \boldsymbol{\zeta}^{(i-1)})$ and the M step maximizes $Q(\boldsymbol{\zeta}; \boldsymbol{\zeta}^{(i-1)})$, until a criterion of convergence is met (McLachlan & Krishnan, 2007). A helpful property of the EM algorithm is that each EM iteration either increases or maintains (but does not decrease) the log-likelihood, which guarantees convergence as long as the log-likelihood is of such a shape that it can be maximized (Dempster et al., 1977; McLachlan & Krishnan, 2007).

Evaluating Equation 2.20 requires computing an expectation, i.e., evaluating an integral. Often, this integral is analytically intractable, requiring a type of numerical approximation. A popular method to this end is Gauss-Hermite (GH) quadrature (Baker & Kim, 2004). GH quadrature approximates the integral over a continuous variable by a sum over discrete points which are weighted according to the distribution of the continuous variable (Baker & Kim, 2004). The discrete points are called quadrature nodes and the associated weights, quadrature weights. Gauss-Hermite quadrature is for example implemented in `R` in the package `fastGHQuad` (Blocker, 2018).

For unidimensional logistic IRT models (i.e., the 2PL model and a further generalization including a third item parameter), Bock and Aitkin (1981) provided an EM algorithm which simplifies the E step by computing sufficient statistics for the (posterior) expected complete-data log-likelihood (Baker & Kim, 2004). The E step then consists of computing these sufficient statistics (Baker & Kim, 2004). Please see Baker and Kim

(2004) for further details and the specific algorithm.

2.3.2 Penalized MML Estimation in IRT

A simple structure for the discrimination matrix $\boldsymbol{\alpha}$ cannot only be obtained through rotation (see Sections 2.2.3 and 2.2.3). Recently, research in factor analysis (see e.g., Trendafilov, 2014, for a review) as well as dichotomous and polytomous IRT (e.g., Cho, Xiao, Wang, & Xu, 2022; Robitzsch, 2023; Sun, Chen, Liu, Ying, & Xin, 2016) has explored an alternative: Obtaining a simple structure for the $\boldsymbol{\alpha}$ via regularization. Regularization methods were originally developed within the context of variable selection problems in (generalized) linear models (Hastie, Tibshirani, & Friedman, 2009). Obtaining a simple structure for the $\boldsymbol{\alpha}$ matrix in IRT can also be viewed as a variable selection problem: The aim is a sparse discrimination matrix in which only certain parameters are different from 0 (Scharf & Nestler, 2019; Trendafilov, 2014).

Regularization is a broad research field in the statistical literature, a summary of which would be beyond the scope of this thesis. For a general introduction, see for example Hastie et al. (2009). In regularized estimation approaches, one imposes a penalty onto the likelihood that is to be maximized. Such a penalty is usually a function of all or a subset of the model parameters and it penalizes larger estimates for these parameters (Hastie et al., 2009). Thereby, it imposes shrinkage on these parameters: Estimates are gradually shrunken towards 0. Depending on the specific penalty imposed, parameters can even be shrunken to exactly 0 (e.g., for the least absolute shrinkage and selection operator (lasso) penalty; Hastie et al., 2009; Tibshirani, 1996, see also below). Thus, regularization can serve as a means of variable selection. Originally developed for regression, regularization methods such as the lasso have grown to be popular in a wider array of contexts, including generalized linear mixed models (e.g., Groll & Tutz, 2014; Schelldorfer, Meier, & Bühlmann, 2014) and extensions (e.g., Nestler & Humberg, 2022), factor analysis (e.g., Trendafilov, 2014), or structural equation modeling (e.g., Jacobucci, Grimm, & McArdle, 2016). In the following, I am going to focus on prior work in lasso-regularized IRT estimation that specifically informed the method development of this thesis.

For the third article in this thesis, an important prior work is the ℓ_1-penalized EM algorithm for dichotomous and polytomous IRT models developed by Sun et al. (2016). I am going to briefly outline their work with a focus on their algorithm rather than the models, as those are for binary and ordinal rather than count data. Thus, I am going to start the description by looking at the marginal log-likelihood for a multidimensional 2PL model (Equation 2.11; Birnbaum, 1968; Reckase, 2009) which I denote $LL_m(\boldsymbol{\zeta}; \boldsymbol{x})$

in the following. I adapted the notation to align with the notation used throughout this thesis. Imposing a lasso penalty (Tibshirani, 1996) on the discrimination matrix $\boldsymbol{\alpha}$ ($\subset \boldsymbol{\zeta}$), the goal of penalized maximum likelihood estimation is to maximize

$$LL_m(\boldsymbol{\zeta}; \boldsymbol{x}) - \eta||\boldsymbol{\alpha}||_1 \tag{2.21}$$

in terms of $\boldsymbol{\zeta}$ (Sun et al., 2016). The hyperparameter $\eta > 0$ modulates how strongly shrinkage is imposed onto the $\boldsymbol{\alpha}$ matrix (Sun et al., 2016). Sun et al. (2016) tune η using the Bayesian information criterion (BIC; Schwarz, 1978). Note that for $\eta = 0$, the unpenalized ML estimator is obtained (Sun et al., 2016). The lasso penalty $||\boldsymbol{\alpha}||_1$ is

$$||\boldsymbol{\alpha}||_1 = \sum_{j=1}^{M} \sum_{l=1}^{L} |\alpha_{jl}| \tag{2.22}$$

(Sun et al., 2016). For estimation, Sun et al. (2016) employ an EM algorithm (Dempster et al., 1977) in combination with the coordinate descent algorithm (Friedman, Hastie, & Tibshirani, 2010) during the M step to implement the penalization. During the E step, they numerically approximate the (posterior) expectation of the complete-data (i.e., $(\boldsymbol{x}, \boldsymbol{\theta})$) log-likelihood through quadrature, while assuming the item parameters from the previous M step, $\boldsymbol{\zeta}'$, known (Sun et al., 2016). Denote the approximated posterior expectation as $\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}')$ (Sun et al., 2016). During the M step, $\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}')$ is penalized by substracting the lasso penality, i.e.,

$$\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}') - \eta||\boldsymbol{\alpha}||_1, \tag{2.23}$$

and this penalized (approximated) expected log-likelihood is in turn maximized with regard to $\boldsymbol{\zeta}$ (Sun et al., 2016). Sun et al. (2016) carry out the maximization per item, as $\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}')$ decomposes into each item's contribution, and employ Friedman et al. (2010)'s cyclic coordinate descent algorithm. There are $L + 1$ parameters associated with each item. Using a discrimination-difficulty parameterization rather than an intercept-slope parameterization in the IRT model, Sun et al. (2016) have a difficulty parameter d_j and L discrimination (= slope) parameters $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jL})^T$ to optimize. Within the (item-wise) cyclic coordinate descent, they do so by iteratively updating the parameters through the following updating rules: For d_j update through

$$\hat{d}_j = d_j - \frac{\partial_{d_j} \hat{Q}(\boldsymbol{\zeta}_j|\boldsymbol{\zeta}'_j)}{\partial^2_{d_j} \hat{Q}(\boldsymbol{\zeta}_j|\boldsymbol{\zeta}'_j)} \tag{2.24}$$

(Sun et al., 2016), and for each α_{jl} update through

$$\hat{\alpha}_{jl} = -\frac{S(-\partial^2_{\alpha_{jl}}\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}')\alpha_{jl} + \partial_{\alpha_{jl}}\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}'), \eta)}{\partial^2_{\alpha_{jl}}\hat{Q}(\boldsymbol{\zeta}|\boldsymbol{\zeta}')} \tag{2.25}$$

(Sun et al., 2016). The elements in $\boldsymbol{\alpha}_j$ are updated successively, with each element previous to the currently updated element already in its updated form and each subsequent element still in its previous form (Sun et al., 2016). In Equation 2.25, S is defined as

$$S(x, \eta) = \text{sign}(x)(|x| - \eta)_+ = \begin{cases} x - \eta, & \text{if } x > 0 \text{ and } \eta < |x|, \\ x + \eta, & \text{if } x < 0 \text{ and } \eta < |x|, \\ 0 & \text{if } \eta \geq |x| \end{cases} \tag{2.26}$$

(Sun et al., 2016), and is called the soft threshold operator (Donoho & Johnstone, 1995) through which shrinkage can be imposed on the penalized item parameters. Further details on how Sun et al. (2016) derive these updating rules are described in the appendix of their paper.

3 Summary of the Articles

3.1 Article 1: The Two-Parameter Conway-Maxwell-Poisson Model (BJSMP, 2022)

The article summarized in the following is published in *British Journal of Mathematical and Statistical Psychology*: **Beisemann, M.** (2022). A flexible approach to modelling 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. https://doi.org/10.1111/bmsp.12273

3.1.1 Motivation

Existing count item response models were limited in their ability to accommodate empirical data settings in which conditional response distributions were underdispersed, at least for a subset of items (Forthmann et al., 2020b), as Chapter 1 outlined. The introduction of the Conway-Maxwell-Poisson Counts Model (CMPCM; Forthmann et al., 2020b, see Chapter 2.2.2) allowed to account for item-specific over-, equi-, and also underdispersion for the first time. A limitation of the CMPCM is that it assumes equal discrimination or slope parameters across all items. This assumption is empirically sometimes violated (Myszkowski & Storme, 2021) and should at least be tested. The aim of this first article was the extension of the CMPCM to allow for modelling item-specific discrimination or slope parameters. This required the proposal of a corresponding estimation technique, as existing estimation procedures and software implementations could not accomodate such an extension (Forthmann et al., 2020b). Importantly, with the addition of discrimination or slope parameters, we leave the GLMM context and can no longer estimate such a count IRT model with GLMM software.

3.1.2 The Two-Parameter Conway-Maxwell-Poisson Model

The proposed Two-Parameter Conway-Maxwell-Poisson Model (2PCMPM) extends the CMPCM through the inclusion of item-specific discriminations, or in the param-

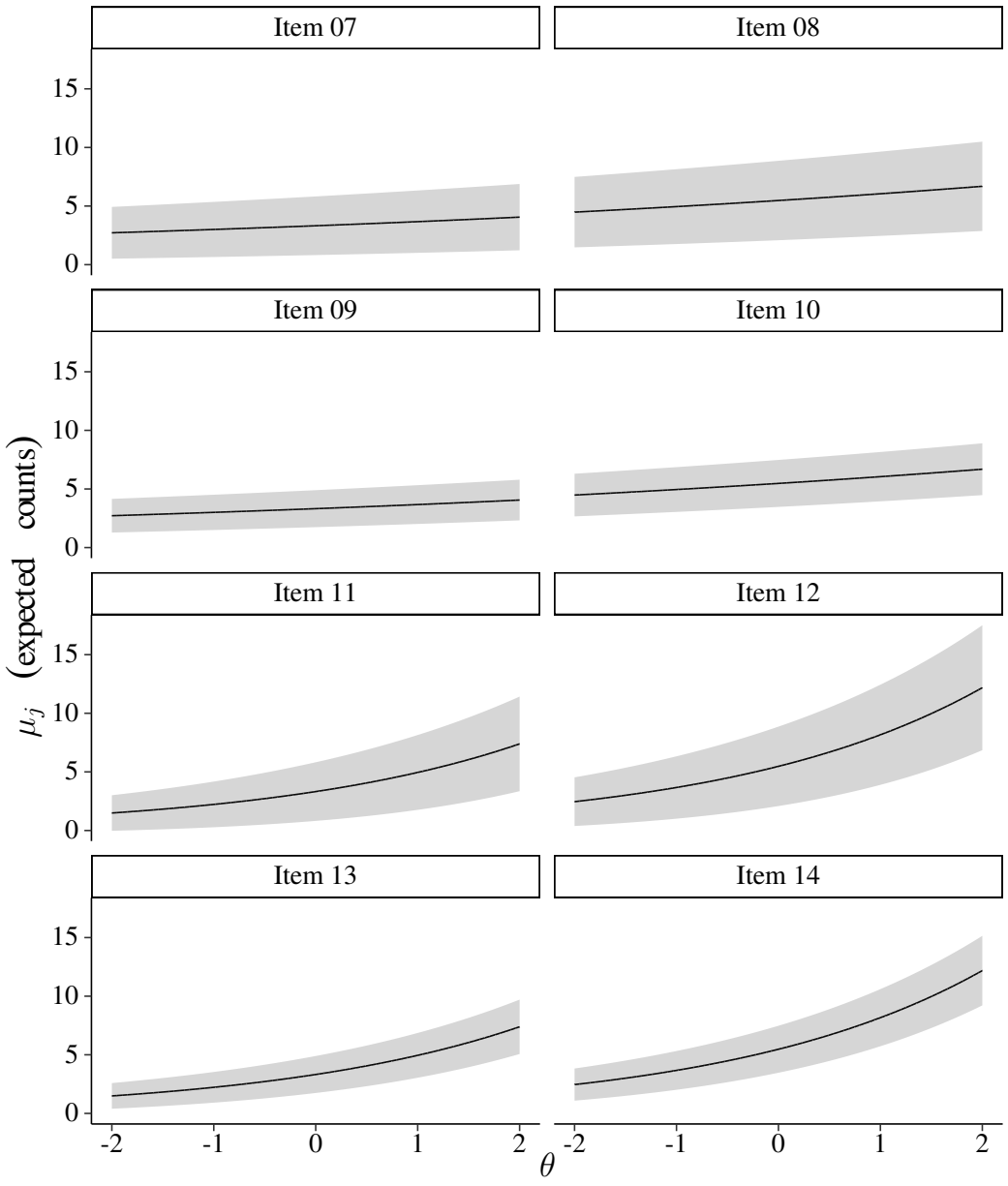


Figure 3.1: Item response curves under the 2PCMPM for items with easiness $\delta_7 = \delta_9 = \delta_{11} = \delta_{13} = 1.2$ (Items 7, 9, 11, and 13) and items with easiness $\delta_8 = \delta_{10} = \delta_{12} = \delta_{14} = 1.7$ (Items 8, 10, 12, and 14); items 7, 8, 11, and 12 exhibit conditional overdispersion ($\nu_7 = \nu_8 = \nu_{11} = \nu_{12} = 0.4$) and items 9, 10, 13, and 14 exhibit conditional underdispersion ($\nu_9 = \nu_{10} = \nu_{13} = \nu_{14} = 1.4$) (shaded ribbons indicate ± 1 conditional SD); items 7 – 10 have a very gentle slope of $\alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10} = 0.1$ and items 11 – 14 have a steeper slope of $\alpha_{11} = \alpha_{12} = \alpha_{13} = \alpha_{14} = 0.4$.

eterization used here (see Beisemann, 2022, or Chapter 2.2.2 for details), slope parameters α_j, for items $j = 1, \ldots, M$. The 2PCMPM assumes that count responses $X_{ij}|\theta_i \sim \text{CMP}_\mu(\mu_{ij}, \nu_j)$ for person $i \in \{1, \ldots, N\}$ and item $j \in \{1, \ldots, M\}$, with item-specific dispersion parameter ν_j. Further, the expected count response μ_{ij} for person i to item j is modelled as

$$\mu_{ij} = \exp\left(\alpha_j \theta_i + \delta_j\right) \tag{3.1}$$

(Beisemann, 2022). The intercept δ_j represents the expected log counts for a person of average ability (i.e., $\theta_i = 0$) answering item j. Note that with the introduction of the slope parameter, the intercept δ_j is no longer directly equivalent to the item easiness. The slope α_j represents the degree to which differences on the latent trait are depicted in differences on the count responses (similar to a factor loading). Analogously to the count IRT models described in Chapter 2, the item response curves under the 2PCMPM for items with varying slopes, intercepts, and dispersions are illustrated in Figure 3.1. We can see that the 2PCMPM can model item-specific dispersion in the same manner as the CMPCM (Figure 2.5) and is additionally able to model item-specific differences in the item's capability to differentiate between latent trait levels through the added slope parameter (beyond what is implied through the item intercept).

For the latent trait θ_i, assume that $\theta_i \sim \mathcal{N}(0, 1)$, for $i = 1, \ldots, N$. The latent trait variance has to be fixed to 1 for identification purposes when estimating slope parameters in IRT models (Baker & Kim, 2004). With the local independence assumption, the probability for the response vector $\mathbf{x}_i$ containing the responses of person i to all M items under the 2PCMPM is

$$P(\mathbf{x}_i|\theta_i, \boldsymbol{\zeta}) = \prod_{j=1}^{M} \text{CMP}_\mu(x_{ij}; \mu_{ij}, \nu_j) \tag{3.2}$$

for $i \in \{1, \ldots, N\}$ (Beisemann, 2022). Denote the density of the standard normal distribution as ϕ. The marginal likelihood for the data $\mathbf{x} \in \mathbb{N}_0^{N \times M}$ (i.e., the responses of all N participants to all M items) under the 2PCMPM is given by

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) = \prod_{i=1}^{N} \int P(\mathbf{x}_i|\theta_i, \boldsymbol{\zeta}) \phi(\theta_i) d\theta_i \tag{3.3}$$

(Beisemann, 2022).

3.1.3 Estimation via Expectation-Maximization Algorithm

The integral in Equation 3.3 is analytically intractable, calling for numeric methods of integration. In this work, I relied on (fixed) Gauss-Hermite quadrature to this end, as it is a common approach in IRT (Baker & Kim, 2004). Let q_k denote the kth quadrature node and w_k the kth quadrature weight. Rewriting Equation 3.3 in quadrature notation, using $K \in \mathbb{N}$ quadrature nodes, one obtains

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{x}_i | q_k, \boldsymbol{\zeta}) w_k \tag{3.4}$$

(Beisemann, 2022). Directly optimizing Equation 3.4 in terms of $\boldsymbol{\zeta}$ is still challenging. We can conceive of the estimation problem as an incomplete-data problem: We have the observed responses to the items, $\mathbf{x}$, but also the latent traits $\boldsymbol{\theta}$, or in quadrature notation, $\mathbf{q}$, which can be thought of as unobservable and therefore missing data. The complete data would consist of both, that is, $(\mathbf{x}, \boldsymbol{\theta})$ or $(\mathbf{x}, \mathbf{q})$. In the article, I show that the expected complete-data log-likelihood can be expressed as

$$\mathbb{E}(LL_c) \propto \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[(x_{ij} \log(\lambda_{jk}) - \nu_j \log(x_{ij}!) - \log(Z_{jk}) \, P(q_k | \mathbf{x}_i, \boldsymbol{\zeta}') \right]. \tag{3.5}$$

with $Z_{jk} = Z(\lambda(\mu_{jk}, \nu_j), \nu_j))$ and $\lambda_{jk} = \lambda(\mu_{jk}, \nu_j)$ for easier readability, and where

$$P(q_k | \mathbf{x}_i, \boldsymbol{\zeta}') = \frac{\prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_k, \boldsymbol{\zeta}_j') w_k}{\sum_{k'=1}^{K} \prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_{k'}, \boldsymbol{\zeta}_j') w_{k'}} \tag{3.6}$$

are the posterior probabilities for q_k, $k = 1, \ldots, K$ (Beisemann, 2022). The latter are computed during each E step using given item parameters $\boldsymbol{\zeta}_j'$ from the previous M step (or using start values in the first iteration). In each subsequent M step, the expected complete-data log-likelihood is maximized with respect to $\boldsymbol{\zeta}$, given the posterior probabilities (3.6) from the previous E step (Beisemann, 2022). In the article, I provide the derivatives required for the M step. E and M steps are iteratively repeated until a criterion of convergence is met. I implemented the 2PCMPM EM algorithm in `R` (R Core Team, 2023) and `C++` (tied into `R` using `rcpp`; Eddelbuettel et al., 2011) in the `R` package `countirt` (see Chapter 4 for details).

In the article, I obtain standard errors for the model parameters using a numerical approximation technique to Oake's identity (Oakes, 1999; Chalmers, 2018; Pritikin, 2017). With regard to measurement, I discuss how expected a-posteriori (EAP) latent

trait estimates can be easily and computationally inexpensively obtained from the EM algorithm (Beisemann, 2022).

3.1.4 Evaluation in Simulation Studies and Application

In a first simulation study, parameter recovery and reliability of the proposed model and algorithm was assessed under 32 different simulation conditions, with varying sample sizes, numbers of items, quadrature nodes, and types of underlying item-specific dispersions (i.e., all items with conditional equidispersion, all items with conditional overdispersion, all items with conditional underdispersion, and a set of items in which each type of dispersion was represented with at least one item). Only a small number of trials in some of the conditions experienced numerical instabilities; reasons for which are discussed in the article (Beisemann, 2022). The results in terms of parameter recovery were overall satisfactory, with greater number of items and larger samples sizes unsurprisingly yielding better results. Increasing the quadrature node number above 121 did not improve results notably (Beisemann, 2022). In a second simulation study, the 2PCMPM was compared to previously proposed count item response models. The 2PCMPM is the most general of the models compared, with all other models constituting special cases of the 2PCMPM. The comparison was conducted under conditions where in truth, a 2PCMPM holds, and all existing count item response models were faced with some or multiple kinds of violations of their assumptions. The comparison highlights that in particular in terms of uncertainty quantification, that is, in terms of standard errors and model-implied reliability estimates, the 2PCMPM can alleviate problems caused for the other models due to assumption violations (Beisemann, 2022). However, it is also noteworthy to observe that in terms of point estimates for the latent trait, differences between all compared models were minimal, especially for the CMPCM and the 2PCMPM (Beisemann, 2022). This as well as other emerging patterns are discussed in the article.

For an empirical illustration of the proposed model, a 2PCMPM was fitted to data from $M = 6$ divergent thinking fluency tasks (Silvia, 2008a, 2008b; Silvia et al., 2008; Silvia, 2013). The resulting parameters indicated that the items varied in their item-specific discrimination or slope parameters as well as in their item-specific dispersion parameters (Beisemann, 2022). These patterns were corroborated by likelihood ratio tests testing (1) the constraint of equal dispersion parameters across items (against item-specific dispersions), and (2) the constraint of equal slope parameters across items (against item-specific slopes). In both cases, the constraints were rejected, indicating that the full 2PCMPM fit better to the data than (previously existing) special cases of

the 2PCMPM (Beisemann, 2022). The example illustrates how special cases, such as the CMPCM, can easily be obtained by introducing constraints in the 2PCMPM, and how this allows for testing of the special cases' assumptions.

3.2 Article 2: Explanatory Extensions of the 2PCMPM (MBR, 2024)

The article summarized in the following has been published as an advance online article in *Multivariate Behavioral Research* at the time of submitting this thesis: **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. https://doi.org/10.1080/00273171.2023.2288577

3.2.1 Motivation

Item response models yield estimates for item-specific characteristics and can be used to obtain latent trait estimates. Differences between items in their characteristics and between test takers in their latent trait values can thus be observed. A specific group of item response models aims to explain these differences: Explanatory item response models (see e.g., De Boeck & Wilson, 2004), or explanatory extensions of item response models, allow the inclusion of item and / or person covariates that might explain (part of) the differences between items and / or test takers, respectively. Popular explanatory item response models include the Log-Linear Test Model (LLTM; Fischer, 1973) and the Latent Regression Model (LRM; Zwinderman, 1991), both explanatory extensions of the Rasch model (a one-parameter item response model for binary data; Rasch, 1960), the former by inclusion of item and the latter by inclusion of person covariates. Explanatory extensions of count item response models have received comparatively less attention. Prior research has focused on explanatory extensions of the RPCM (e.g., Ogasawara, 1996; Graßhoff et al., 2013, 2020) which – as discussed in the previous chapters – assumes discriminations to be equal across items and all items' conditional response distribution to be equidispersed. Differences in item parameters can thus only be modeled and explained for the item-specific easiness (inverse difficulty) parameters (compare Chapter 2.2.2). From an implementation stand point, incorporating covariates into the RPCM has the advantage that such extensions still remain within the GLMM framework (De Boeck et al., 2011).

The 2PCMPM (Beisemann, 2022), proposed in the first article of this thesis, additionally allows for modeling of item-specific discrimination or slope parameters and item-specific dispersion parameter through which the 2PCMPM can account for conditional over-, equi-, and underdispersion. These allow for differences between items among three groups (rather than one) of item-specific parameters (Beisemann et al.,

2024a). Explaining such differences can help to understand reliability differences between items (as reliability in item response models depends on the item parameters), can inform item construction (if certain item properties can empirically be shown to influence item parameters in a certain way, they can be chosen purposefully), and guide item selection (e.g., if certain items empirically demonstrate low discrimination, they are only capable of capturing latent trait differences poorly and thus are less suitable for a test assessing the respective latent trait) (Beisemann et al., 2024a). Between-person differences in latent trait values are commonly a focus of substantive research on the measured constructs (Beisemann et al., 2024a). To study these differences under the 2PCMPM allows to account for empirically present complexity of data and thus enables more accurate uncertainty quantification (Beisemann, 2022). These considerations motivated the aim of this second article: The development of two explanatory item response models, one to include item and one to include person covariates, as well as corresponding estimation procedures based on the EM algorithm proposed for the 2PCMPM in the first article. Note that when basing explanatory CIRT models on the 2PCMPM, we leave the GLMM framework.

3.2.2 The Distributional Regression Test Model

In the second article of this thesis, we proposed the Distributional Regression Test Model (DRTM) which extends the 2PCMPM (Beisemann, 2022) through the inclusion of item covariates on all three (or a subset of) the item parameters in the 2PCMPM. For simpler notation, we formulated the following equations using the same I item covariates to explain differences on all three types of item parameters (but the DRTM also allows different item covariates for each item parameter type, or only including item covariates on one or two item parameter groups; for details see Beisemann et al., 2024a). With $u_{j1}, \dots, u_{jI}$, $j \in \{1, \dots, M\}$, denoting the realizations of the I item covariates, we model

$$\alpha_j = \alpha + \sum_{c=1}^{I} \beta_{\alpha_c} u_{jc} \tag{3.7}$$

$$\delta_j = \delta + \sum_{c=1}^{I} \beta_{\delta_c} u_{jc}, \tag{3.8}$$

under the DRTM, with β_{α_c} and β_{δ_c} for the cth covariate weight in the model for α_j and δ_j, respectively (Beisemann et al., 2024a). The expected counts μ_{ij} are in turn modelled

as

$$\mu_{ij} = \exp\left(\delta + \sum_{c=1}^{I} \beta_{\delta_c} u_{jc} + \left(\alpha + \sum_{c=1}^{I} \beta_{\alpha_c} u_{jc}\right)\theta_i\right), \tag{3.9}$$

under the DRTM (Beisemann et al., 2024a), as opposed to as in Equation 3.1.2 under the 2PCMPM. The dispersion parameters ν_j are modelled as

$$\nu_j = \exp(\tilde{\nu} + \sum_{c=1}^{I} \beta_{\nu_c} u_{jc}). \tag{3.10}$$

under the DRTM, with β_{ν_c} for the cth covariate weight for ν_j (Beisemann et al., 2024a). Note the different scales of $\tilde{\nu}$ and ν_j (due to the log link), highlighted in the notation by writing $\tilde{\nu}$ rather than simply ν (as in Equations 3.7 and 3.8). With Equations 3.7–3.10, the DRTM, as an extension of the 2PCMPM, assumes count responses $X_{ij}|\theta_i \sim \text{CMP}_\mu(\mu_{ij}, \nu_j)$ for person $i \in \{1, \ldots, N\}$ and item $j \in \{1, \ldots, M\}$ (Beisemann et al., 2024a).

3.2.3 The Count Latent Regression Model

We further propose the Count Latent Regression Model (CLRM) in the second article of this thesis. The CLRM extends the 2PCMPM through the inclusion of P person covariates to explain differences in the latent trait values. With $t_{i1}, \ldots, t_{iP}$, $i \in \{1, \ldots, N\}$, denoting the realizations of the P person covariates, we model

$$\theta_i = \theta_i^* + \sum_{p=1}^{P} \gamma_p t_{ip}, \tag{3.11}$$

where γ_p, $p = 1, \ldots, P$ represent the regression weights for the person covariates (Beisemann et al., 2024a). We can interpret the term θ_i^* as a random intercept with $\theta_i^* \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. We fix the variance to 1 for identification purposes. The expected counts μ_{ij} are modelled as

$$\mu_{ij} = \exp\left(\delta_j + \alpha_j\left(\theta_i^* + \sum_{p=1}^{P} \gamma_p t_{ip}\right)\right) \tag{3.12}$$

under the CLRM (Beisemann et al., 2024a). The CLRM assumes count responses $X_{ij}|\theta_i \sim \text{CMP}_\mu(\mu_{ij}, \nu_j)$ for person $i \in \{1, \ldots, N\}$ and item $j \in \{1, \ldots, M\}$, with μ_{ij} as in Equation 3.12 and item-specific dispersion parameters ν_j (as in the 2PCMPM).

As introduced in Equation 3.11, the CLRM can theoretically include continuous and categorical covariates, but in the article, we discuss the computational challenges of continuous covariates and how they can be alleviated when including only categorical covariates. Due to the challenges associated with including continuous covariates, our implementation and evaluation of the CLRM includes only categorical person covariates (see Chapters 4 and 5 for a further discussion) (Beisemann et al., 2024a).

3.2.4 Estimation via Expectation-Maximization Algorithm

In the second article of this thesis, we outline how the DRTM and the CLRM can be estimated using adaptations of the 2PCMPM EM algorithm presented in the first article (Beisemann, 2022). With μ_{ij} and ν_j as in Equations 3.9–3.10 and in Equation 3.12 with item-specific ν_j for the DRTM and the CLRM, respectively, the marginal likelihood for the DRTM and the CLRM remains as given in Equation 3.3 for the 2PCMPM. Following, the EM algorithm to estimate the DRTM and CLRM remains as outlined in Chapter 3.1.3 for the 2PCMPM, but with the respective specifications for μ_{ij} and ν_j and consequently different gradients for the M step which we provide in the second article of this thesis (Beisemann et al., 2024a). The EM algorithms for the DRTM and the CLRM were again implemented in `R` (R Core Team, 2023) and `C++` (tied into `R` using `rcpp`; Eddelbuettel et al., 2011) in the `R` package `countirt` (see Chapter 4 for details, also regarding the computational challenge of continuous person covariates in the CLRM and how computational efficiency was improved for categorical person covariates). As in the first article, we obtained standard errors for the model parameters using a numerical approximation technique to Oake's identity (Oakes, 1999; Chalmers, 2018; Pritikin, 2017).

3.2.5 Evaluation in Simulation Studies and Application

We conducted two simulation studies to assess (I) parameter recovery in the DRTM and the CLRM, and (II) the power and type I error of the Wald tests for covariate effects. In the first simulation study, we systematically varied the sample size, the number of items, and the number and types of covariates, resulting in 28 different simulation conditions. Numerical instabilities were only encountered in a small number of trials in a few simulation conditions, and results in terms of parameter recovery were mostly satisfactory. More challenging conditions (particularly for slope parameter estimation) involved continuous item covariates and smaller numbers of items. Potential reasons for these and other observed result patterns are discussed in the article. In the second

simulation study, we varied sample size, number of items, the type of covariate in the DRTM, and the size of effect, in models with one (either item or person) covariate. In the resulting 68 simulation conditions, we observed numerical instabilities only in some trials in one condition. Results in terms of type I error rate and power for covariate detection in the DRTM and the CLRM were overall quite promising but not perfect. We discuss the result pattern in the second article (Beisemann et al., 2024a).

In two empirical examples, we illustrated how the proposed models can be applied to real data sets. We used the DRTM to re-analyze divergent thinking task data (Forthmann et al., 2016), investigating to what extent item parameter differences under the 2PCMPM can be explained by the instruction used in conjunction with the items, word frequency of the prompt word in the item, and their interaction. We illustrated how likelihood ratio tests and information criteria can be used to compare different DRTMs. We re-analysed language proficiency test data (Forthmann et al., 2020a; Grotjahn, Schlak, & Aguado, 2010; Heine, 2017) with the CLRM and investigated to what extent latent trait differences can be explained by gender, age, and whether the person was a native speaker. Limitations of this analysis and resulting avenues for future method development are discussed in the article (see also Chapter 5) (Beisemann et al., 2024a).

3.3 Article 3: Multidimensional Count Data Item Response Models (PsyArXiv, 2024)

The article summarized in the following has been submitted to and is under review with *Psychometrika* at the time of publishing this thesis and has been published as a pre-print on PsyArXiv: **Beisemann, M.**, Holling, H., & Doebler, P. (2024). Every trait counts: Marginal maximum likelihood estimation for novel multidimensional count data item response models with rotation or ℓ_1–regularization for simple structure. *PsyArXiv pre-print, version 1*. https://doi.org/10.31234/osf.io/fqyjs

3.3.1 Motivation

The proposal of the 2PCMPM (Beisemann, 2022) in the first article of this thesis allows to model count responses to items with varying dispersions and varying discriminations (or slopes). However, the 2PCMPM remains limited in terms of the assumption that only one latent trait, θ, underlies the count responses. Empirically, psychological tests and self reports may often require responses to be influenced by more than one latent trait, for example by construction (i.e., when measured constructs are decomposed into several subfacets) or by test administration (i.e., when internal and external factors additionally influence responses) (Beisemann et al., 2024b).

In item response theory, the framework of multidimensional IRT (MIRT) allows to model responses as influenced by $L \in \mathbb{N}$ latent traits (see e.g., Reckase, 2009, for an introduction; compare also Chapter 2). For binary and ordinal responses, MIRT models enjoy great popularity (Chalmers, 2012). For count data, MIRT models have received comparably less attention (but see Forthmann et al., 2018; Myszkowski & Storme, 2021). Count models with multiple latent traits have been developed in a factor analytical setting (Wedel et al., 2003), using a Poisson response distribution with different link functions and truncating the Poisson distribution, which accommodates some extent of underdispersion. In a different tradition, the generalized linear latent and mixed models (GLLAMM) framework by Skrondal and Rabe-Hesketh (2004) also allows to fit multidimensional count models. To the best of my knowledge, prior to this third article of the thesis, no multidimensional model taking advantage of the dispersion flexibility of the CMP distribution existed.

The third article of this thesis (Beisemann et al., 2024b) extended the 2PCMPM to multidimensional count IRT models: the class of multidimensional two-parameter Conway-Maxwell-Poisson models (M2PCMPM). The focus of the article is the exploratory ver-

sion of M2PCMPMs. We developed a marginal maximum likelihood estimation technique based on the EM algorithm provided in the first article of this thesis (Beisemann, 2022). For obtaining a simple structure (Thurstone, 1947) for the discrimination matrix $\boldsymbol{\alpha}$, we used traditional rotation as well as regularization techniques. For the latter, we developed a penalized version of the M2PCMPM and a corresponding estimation routine inspired by prior work by (Sun et al., 2016), using the lasso penalty (Tibshirani, 1996).

3.3.2 Multidimensional Two-Parameter CMP Models

The proposed class of multidimensional two-parameter Conway-Maxwell-Poisson models (M2PCMPM) generalizes the 2PCMPM from the first article of this thesis to the multidimensional case with $L \in \mathbb{N}$ latent traits. For latent traits $\theta_{1i}, \ldots, \theta_{li}, \ldots, \theta_{Li}$, $i = 1, \ldots, N$, we model the expected count response μ_{ij} for one person i to one item j as (Beisemann et al., 2024b)

$$\mu_{ij} = \exp\left(\sum_{l=1}^{L} \alpha_{jl}\theta_{li} + \delta_j\right), \tag{3.13}$$

where α_{jl} denotes the discrimination or slope for the jth item and lth trait and δ_j denotes the intercept for the jth item (as in the 2PCMPM). It is immediately clear that the 2PCMPM is contained within the M2PCMPM as a special case for $L = 1$. All special cases of the 2PCMPM in turn constitute special cases of the M2PCMPM. With the assumptions (1) that the latent traits are jointly multivariately normal distributed (with density ψ, mean vector $\boldsymbol{\mu}_\theta = \mathbf{0} \in \mathbb{R}^L$, and covariance matrix $\boldsymbol{\Sigma}_\theta \in \mathbb{R}^{L\times L}$, with a diagonal of 1's for identification and either with orthogonal traits or pre-estimated latent trait correlations, as discussed in the article), (2) that the responses X_{ij} are conditionally CMP distributed (i.e., $X_{ij}|\boldsymbol{\theta}_i \sim \text{CMP}_\mu(\mu_{ij}, \nu_j)$ with μ_{ij} as in Equation 3.13 and ν_j estimated item-specifically), and (3) that the items are locally independent, we obtained the marginal likelihood for the M2PCMPM as

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) = \prod_{i=1}^{N} \int \cdots \int \prod_{j=1}^{M} P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j)\Psi(\boldsymbol{\theta}_i; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) d\theta_{1i} \ldots d\theta_{Li}, \tag{3.14}$$

(Beisemann et al., 2024b) where $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{Li})^T$, for $i = 1, \ldots, N$. Analogously to the notation in the summaries of the previous two articles, the probability for a count response $X_{ij} = x_{ij}$ under the M2PCMPM is given by $P(x_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\zeta}_j)$, where $\boldsymbol{\zeta}_j$ denotes the vector of all item parameters for item j in the model. It is $\boldsymbol{\zeta} = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_M\}$. In

the article, we discussed through which constraints on the discrimination matrix $\boldsymbol{\alpha}$ the exploratory M2PCMPM can be identified (Beisemann et al., 2024b).

3.3.3 Estimation via Expectation-Maximization Algorithm

In the third article of this thesis (Beisemann et al., 2024b), we provided a marginal maximum likelihood estimation routine for the M2PCMPMs based on the EM algorithm for the 2PCMPM developed in the first article (Beisemann, 2022). As in Beisemann (2022), we relied on (in this case multidimensional) Gauss-Hermite quadrature in estimation. We discussed in the article that this choice is only going to work well for small numbers of latent traits, as multidimensional quadrature is known not to scale well to higher numbers of dimensions (Baker & Kim, 2004). Generalizing Equation 3.5, we obtained

$$\begin{aligned}\mathbb{E}(LL_c) \propto \sum_{k_L=1}^{K} \dots \sum_{k_2=1}^{K} \sum_{k_1=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} \log(\lambda(\mu_{jk_1,\dots,k_L}, \nu_j)) - \nu_j \log(x_{ij}!) \\ - \log(Z(\lambda(\mu_{jk_1,\dots,k_L}, \nu_j), \nu_j))) P(q_{k_1}, \dots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') \end{aligned} \tag{3.15}$$

as the expected complete-data log likelihood for the M2PCMPs, where $\mu_{jk_1,\dots,k_L} = \exp(\sum_{l=1}^{L} \alpha_{jl}\theta_{li} + \delta_j)$ with node index $k_l \in \{1, \dots, K\}$ for trait l and $(q_{k_1}, \dots, q_{k_L})$ denote a specific combination of quadrature nodes for which the joint posterior probability is

$$P(q_{k_1}, \dots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') = \frac{\prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_{k_1}, \dots, q_{k_L}, \boldsymbol{\zeta}_j') w_{k_1} \dots w_{k_L}}{\sum_{k_1'=1}^{K} \dots \sum_{k_L'=1}^{K} \prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_{k_1'}, \dots, q_{k_L'}, \boldsymbol{\zeta}_j') w_{k_1'} \dots w_{k_L'}} \tag{3.16}$$

(Beisemann et al., 2024b). In each E step of the M2PCMPM EM algorithm, Equation 3.16 is evaluated using the item parameters $\boldsymbol{\zeta}'$ determined in the previous M step. With the posterior probabilities determined in the E step considered given, the M step consists of determining the item parameters $\boldsymbol{\zeta}$ which maximize Equation 3.15. To this end, we provided the first derivatives of Equation 3.15 in terms of the different item parameters (Beisemann et al., 2024b). The article also briefly discussed how confirmatory M2PCMPMs can be fitted with the M2PCMPM EM algorithm.

The EM algorithm for the M2PCMPM was again implemented in `R` (R Core Team, 2023) and `C++` (tied into `R` using `rcpp`; Eddelbuettel et al., 2011) in the `R` package `countirt`. Note that at the time of writing this thesis, the M2PCMPM related implementations are only available on the `multidimensional` branch of `countirt`

(see Chapter 4 for details).

In the paper, we explored orthogonal (Varimax; Kaiser, 1958, 1959) and oblique (Oblimin; Clarkson & Jennrich, 1988) rotation methods to select the rotation matrix V that yields the simplest structure for $\boldsymbol{\alpha}$ (see Chapter 2 for details).

Following previous research (Cho et al., 2022; Sun et al., 2016), we investigated regularization techniques as an alternative to rotation methods to obtain as simple a structure as possible for $\boldsymbol{\alpha}$. In the third article of this thesis (Beisemann et al., 2024b), we developed a regularized EM algrithm for the M2PCMPM inspired by Sun et al. (2016). With $P(q_{k_1}, \ldots, q_{k_L}|\mathbf{x}_i, \boldsymbol{\zeta}')$ remaining as in Equation 3.16, for the regularized expected complete-data log likelihood we obtained

$$\begin{aligned}\mathbb{E}_{\text{reg}}(LL_c) \propto \sum_{k_L=1}^{K} \ldots \sum_{k_2=1}^{K} \sum_{k_1=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} \log(\lambda(\mu_{jk_1,\ldots,k_L}, \nu_j)) - \nu_j \log(x_{ij}!) \\ - \log(Z(\lambda(\mu_{jk_1,\ldots,k_L}, \nu_j), \nu_j)))P(q_{k_1}, \ldots, q_{k_L}|\mathbf{x}_i, \boldsymbol{\zeta}') - \eta R(\boldsymbol{\alpha})\end{aligned} \tag{3.17}$$

(Beisemann et al., 2024b). The penalty term $R(\boldsymbol{\alpha}) \geq 0$ is a function of the parameters we intend to regularize and is weighted by η, a hyperparameter which determines the strength of regularization (for $\eta = 0$, no regularization is applied). Following Sun et al. (2016), we used the lasso penalty (Tibshirani, 1996) for the penalty term, that is,

$$R_{\text{lasso}} = ||\boldsymbol{\alpha}||_1 = \sum_{l=1}^{L} \sum_{j=1}^{M} |\alpha_{jl}|. \tag{3.18}$$

In each E step of the regularized M2PCMPM EM algorithm, we compute Equation 3.16 (given the item parameters $\boldsymbol{\zeta}'$ from the previous M step) to be able to optimize Equation 3.17 in the subsequent M step considering the posterior probabilities given. Inspired by previous work (Nestler & Humberg, 2022; Schelldorfer et al., 2014; Sun et al., 2016), we proposed that in each M step, we optimize for the parameter α_{jl}'s and δ_j's on the mean first using (item-blockwise) coordinate descent (Friedman et al., 2010) and then use the $\log \nu_j$ derivative calculated for (unregularized) M2PCMPM EM algorithm to optimize for the $\log \nu_j$'s. The regularized M2PCMPM EM algorithm was also implemented in `countirt`, together with a function that allows to tune the regularization hyperparameter η based on a user-provided grid and using the BIC as the tuning criterion. We used warm starts (for details, see Hastie et al., 2009) to increase computational efficiency (Beisemann et al., 2024b).

3.3.4 Evaluation in Simulation Studies and Application

We conducted a small simulation study to assess the model's and the algorithms' viability (Beisemann et al., 2024b). Generating data under the MCIRT model, we studied 16 different simulation conditions, resulting from fully crossing varied numbers of latent traits, varied latent-trait correlations, varied numbers of items per trait, and varied true structures of the discrimination matrix. In each condition, we fitted (1) an unpenalized MCIRT model to be rotated (a) orthogonally and (b) obliquely as well as (2) a lasso-penalized MCIRT with (a) a priori uncorrelated latent traits and (b) a priori correlated traits as estimated based on the obliquely rotated unpenalized MCIRT model (an approach suggested by Sun et al., 2016). Due to the high computational cost, we only ran a comparatively small number of simulation trials in each condition. We assessed bias and RMSE for the items' intercepts, log dispersions, and multi-dimensional discrimination, as well as the correct estimation rate (CER) used by Sun et al. (2016), which expresses the fraction of correctly freely estimated (as opposed to shrunken to 0) discrimination parameters.

We found negligible bias for the intercept parameters and – in line with our previous studies (Beisemann, 2022; Beisemann et al., 2024a) – slightly more bias for the dispersion parameters. Bias on the multidimensional item discriminations was overall not too large, with the penalized models usually performing slightly worse than the unpenalized models. For conditions with overall more model parameters (i.e., as the number of traits and items per trait grew), parameter recovery declined which we speculated might have been due to decreasing obervations-to-paramters ratios for these conditions (Beisemann et al., 2024b). A similar pattern emerged for the CER, although here, it was the penalized methods that overall performed better than the unpenalized methods (Beisemann et al., 2024b). In terms of hyperparameter tuning (based on the BIC) for the regularized models, we found that tuning worked best for more items, more traits, and when the underlying $\boldsymbol{\alpha}$ structure was simple (Beisemann et al., 2024b). However, we also observed that hyperparameter tuning on the whole could still be improved upon in the future.

We illustrated the developed models with a re-analysis of an intelligence test for adolescents. We fitted (1) an unpenalized MCIRT model to be rotated (a) orthogonally and (b) obliquely as well as (2) a lasso-penalized MCIRT with (a) a priori uncorrelated latent traits and (b) a priori correlated traits as estimated based on the obliquely rotated unpenalized MCIRT model (an approach suggested by Sun et al., 2016). We observed that the models estimated $\boldsymbol{\alpha}$ matrices which aligned with theoretical expectations (Beisemann et al., 2024b).

4 Computational Implementation

The estimation algorithms for the models developed in this thesis were implemented in the `R` package `countirt` (at the writing of this thesis, available via GitHub: https://github.com/mbsmn/countirt), with parts of the package written in `C++` for computational efficiency (tied into `R` with the help of `rcpp`; Eddelbuettel et al., 2011). The implementation of these algorithms was associated with certain challenges that arose from employing the CMP distribution. The first section of this chapter outlines what these challenges were and how they were addressed in `countirt`. In the second section of this chapter, I provide a brief overview of the user interface `countirt` offers to specify count item response models in `R`.

4.1 Computational Challenges

For all EM algorithms discussed in Chapter 3 and developed in the three articles of this thesis, we are confronted with computational challenges due to the CMP distribution. As we can see in Chapter 2.1, for the CMP_μ density, the normalizing constant $Z(\lambda(\mu,\nu),\nu)$ is an infinite sum, and the rate $\lambda(\mu,\nu)$ is implicitly defined, as the solution to an equation also involving another infinite sum. The articles of this thesis (Beisemann, 2022; Beisemann et al., 2024a; Beisemann et al., 2024b) further include gradients of the expected complete-data log likelihood required for the M step of the respective EM algorithms. These gradients include the variance of the CMP_μ distribution, as an expectation over the count distribution also an infinite sum, as well as other infinite sums (depending on the specific algorithm, see Beisemann, 2022; Beisemann et al., 2024a; Beisemann et al., 2024b). Any of the EM algorithms require numerous evaluation of all of these quantities, as they have to be evaluated for each node and item combination in each E step and multiple times for each node and item combination for each item parameter as their respective gradients are optimized numerically in the M step. As the EM algorithm will also evaluate them for every possible node value (combination, in the multidimensional case), these quantities also have to be evaluated in part for extreme resulting μ_{ij}'s. Evaluating these infinite sums directly as (truncated) finite sums based on sufficiently many summands at each of the numerous times required

would be computationally costly and at times might even lead to numerical instabilities in unfortunate cases, especially for the more extreme node values.

To alleviate this problem, we opted to use an interpolation-from-grid approach in the `countirt` package for some of the quantities associated with the CMP distribution. The approach is naive and ad-hoc, intended to stabilize and accelerate the estimation for the settings we studied in the articles of this thesis. This was done to enable the implementation of the algorithms developed in the works of this thesis, with the implementation not being the focus of the work (for more sophisticated and systematic work specifically on using interpolation in a CMP regression setting, see Philipson & Huang, 2023). To this end, we tabled different μ_{jk} and ν_j combinations using internal functions we customized from the `glmmTMB R` package (Brooks et al., 2017) for the normalizing constant $Z(\lambda(\mu_{jk}, \nu_j), \nu_j)$, the rate $\lambda(\mu_{jk}, \nu_j)$, and the variance $\mathbb{V}\text{ar}(\mu_{jk}, \nu_j)$. Other quantities (also infinite sums) required in some of the gradients used were not tabled but computed in each function evaluation based on the tabled and interpolated quantities (which suffices for stabilization in our implementation). For the infinite sum computation, we implemented functions that were inspired by the series summation in `glmmTMB` (Brooks et al., 2017) or rather `TMB` (Kristensen, Nielsen, Berg, Skaug, & Bell, 2016): We start the summation at the mode of the series and add increments to the left and the right until further increments fall beneath a threshold.

In each E step and in each evaluation of the gradient during the M step (in each M step, the gradient is evaluated multiple times during numerical optimization) in any of the EM algorithms developed in this thesis, the interpolation from the grid (implemented in `C++`) is carried out as follows: We determine the μ_{jk} and ν_j values for all node $\times$ item combinations. We cap μ_{jk} values below a minimum (0.001) and a maximum μ_{jk} (200), to make sure that we stay within the boundaries of the interpolation table (see Chapter 5 for a discussions of the therewith associated limitations). We then interpolate the corresponding $\mathbb{V}\text{ar}(\mu_{jk}, \nu_j)$, $\log Z(\lambda(\mu_{jk}, \nu_j), \nu_j)$ and $\log \lambda(\mu_{jk}, \nu_j)$ for all μ_{jk} and ν_j values from the grid using bicubic interpolation in `GSL` (Galassi et al., 2010). With these values, we then compute the posterior probabilities and gradients for the E and M step, respectively.

This approach is still not without limitations. Critically, for the CLRM (Beisemann et al., 2024a), the number of μ_{jk} values for interpolation grows additionally with the number of persons (N). That is, we need to evaluate K (nodes) $\times M$ (items) $\times N$ (persons) combinations for μ_{ijk}. This renders each EM iteration extremely computationally costly, even using the interpolation-from-grid approach, to the point where computation is too slow for any practical purposes. In Beisemann et al. (2024a), we used a computational trick to speed up the algorithm for categorical person covariates: For categorical

covariates, there is only a limited number of possible value patterns across covariates, with each person exhibiting one of these patterns. It thus suffices to compute μ_{ijk} for this limited number of patterns, reducing the number of times each CMP quantity has to be interpolated. For each person, we can then simply match their covariate pattern to the list of covariate patterns and use the respectively interpolated values. This way, person covariates can be included as long as they are categorical, either naturally or through artificial categorization of continuous covariates.

Even using the interpolation-from-grid approach (and not considering continuous person covariates in the CLRM), each EM iteration still remains relatively (to very, depending on the model) costly. An approach to reducing computation times is to reduce the number of necessary EM iterations. This can for example be achieved if the start values are chosen well. To this end, start values for slopes and intercepts in Beisemann (2022), Beisemann et al. (2024a), Beisemann et al. (2024b) are determined by first fitting a Poisson version of the respective model to the data, with the Poisson density allowing for much faster EM iterations. With these start values (and start values for $\log \nu_j$ as described in Beisemann, 2022), the number of EM iterations of the CMP models can usually be substantially decreased (but note that for the – in particular, exploratory – multidimensional models, quite a high number of iterations are still required, leading to long computation times for the exploratory M2PCMPM).

Limitations of our computational approach are discussed further in Chapter 5.

4.2 User Interface of the `countirt` package

For unidimensional models (Beisemann, 2022; Beisemann et al., 2024a), `countirt` provides the model fitting function `cirt` which allows the user to fit a 2PCMPM, DRTM, or CLRM as developed in Beisemann (2022) and Beisemann et al. (2024a), as well as to fit a CMPCM (Forthmann et al., 2020b), 2PPCM (Myszkowski & Storme, 2021), RPCM (Rasch, 1960), or explanatory extensions of these models. The `cirt` function takes the model specification via the `model` argument. The user passes a string to the `model` argument that specifies the model using syntax that is inspired (but not exactly the same and by no means as flexible) by the `lavaan` package (Rosseel, 2012). If one were to have a data set containing responses to five items, stored in a `data.frame` with five columns, named `item1`, `item2`, etc., one can specify a 2PCMPM with the following syntax:

```
R> model_2pcmpm <- "theta=~item1+item2+item3+item4+item5;"
```

`countirt` uses "=∼" to indicate "is measured by", as `lavaan` does. In unidimensional item response models, we measure one latent trait which is commonly denoted as θ. The model specification in `countirt` requires the user to write "`theta=∼`"—that is, users cannot arbitrarily name latent variables as they can in `lavaan` syntax. On the right to "=∼", the user specify the variable names containing the item responses, separated by "+". Each line of model specification in `countirt` has to end in "`;`".

The model specification and the data set, named for example `df` in this illustration, are passed to the `cirt` function for model fitting. The user further needs to specify the count distribution to be used through the `family` argument. Currently, the available options are "cmp" or "poisson". The `cirt` function allows to control model fitting parameters via the `control` argument which takes a list produced by the function `cirt_control`. The user fits a 2PCMPM using the following code:

```
R> fit_2pcmpm <- cirt(
        model = model_2pcmpm,
        data = df,
        family = "cmp")
```

Changing the `family` argument to "poisson" would result in fitting a 2PPCM instead. Models that are special cases of the 2PCMPM in other ways than simplifying from the CMP to the Poisson distribution (e.g., the CMPCM) can be expressed by imposing constraints in the model specification. For the CMPCM, for instance, the constraint is equal slope parameters across items. The user can specify this constraint using the following syntax:

```
R> model_cmpcm <- "theta=~item1+item2+item3+item4+item5;
                   alphas ~ 1;"
```

The syntax "`alphas ∼ 1`" indicates that all α parameters should be constrained to be equal, with one global α estimated by the model. This model specification can be passed to the `cirt` function as described above. Analogously, the user can specify the dispersion parameters to be equal across items by specifying "`log_nus ∼ 1`". Alternatively, the user could fix the slopes to specific values, by specifying e.g., "`alpha1 ∼ 0.3, alpha2 ∼ 0.2, ..., alpha5 ∼ 0.4;`" (replacing "`...`" by the specifications for the omitted items). Analogously, the user could fix the values of the log dispersion parameters (using `log_nu1`, `log_nu2`, etc. instead of `alpha1`, `alpha2`, etc.). Note that the constraints are imposed on the log dispersions so that fixed values need to be on the log scale. To date, not all combination of constraints

are supported in countirt (the documentation of the cirt functions gives further information). Currently, no constraints on the intercept parameters are supported.

After obtaining a cirtfit model object with the cirt function, the user can apply a set of post-processing functions. If the user is interested in inference, standard errors, confidence intervals, and z- and p-values can be added to the cirtfit model object using the add_inference function:

```
R> fit_2pcmpm <- add_inference(fit_2pcmpm)
```

This was implemented as a separate function as standard error computation for the CMP models can be costly and users might not want to wait the extra time if they are not interested in obtaining standard errors and significance tests. The add_inference function currently only supports cirtfit models of family "cmp".

The user can obtain a summary of the results with the summary function and can compare two nested models using the anova function. EAP latent trait estimates (as described in Beisemann, 2022) can be obtained with the abilities function. Item response curves for the items can be plotted using the item_curves function – either all in one plot (argument grid=FALSE) or in separate panels for each item (argument grid=TRUE).

The explanatory extensions of the 2PCMPM, i.e., the CLRM and the DRTM (Beisemann et al., 2024a), can be obtained by adding either person or item covariates, respectively, to the model specification (for computational reasons similar to those discussed for person-by-item covariates in Beisemann et al., 2024a, and Chapter 3.2, a combination of the two is not available). For the CLRM, the user amends the 2PCMPM model specification as follows:

```
R> model_clrm <- "theta=~item1+item2+item3+item4+item5;
                  thetas~1+covariate1+covariate2;"
```

Using this amended model specification, the CLRM can be fit using the same cirt function call as the 2PCMPM. The cirt function will expect to find two columns named covariate1 and covariate2 in the supplied data frame containing each person's covariate values. The covariate columns can be arbitrarily named in the data frame, the user only has to give their correct names in the model specification. As explained in Chapter 3.2 and in Beisemann et al. (2024a), the implementation of the CLRM in countirt is only recommended for categorical covariates. The user only has to ensure that the covariate columns in the data frame are of class factor and

countirt will handle them accordingly and utilize pattern matching for better efficiency (Beisemann et al., 2024a).

To fit the DRTM in countirt, the user needs to restructure the data frame into long format, with all item responses in an arbitrarily named column, e.g., counts, and item labels in another arbitrarily named column, e.g., item. The model specification takes a slightly different form to accommodate this long format structure:

```
R> model_drtm <- "theta=~counts(item::item1)+
                        counts(item::item2)+
                        counts(item::item3)+
                        counts(item::item4)+
                        counts(item::item5);
                  alphas~1+covariate1+covariate2;
                  deltas~1+covariate1;
                  log_nus~1+covariate2;"
```

The items which measure θ need to be named according to the scheme: column name of the item column with item responses followed by parentheses in which first, the column name of the column with item labels is given, followed by two colons, followed by the item label of the respective item. Item covariates for each item parameter are specified as shown, analogously to how person covariates are specified. The user can specify item covariates for any subset or all of the item parameters. The item covariates may differ for the different parameter types. Currently, interaction terms need to be added manually to the data frame in their own column. Adding item covariates on some item parameters can be combined with constraints on other item parameters, albeit not all combinations are supported yet (the cirt documentation gives more information). The DRTM can be fitted using the cirt function and only differs from the function call for the 2PCMPM and CLRM by additionally setting the data_long argument to TRUE (the arguments defaults to FALSE for wide data sets such as needed for the 2PCMPM and the CLRM) and providing the name of a column with person identifiers to the person_id argument. In this example, the data frame in long format is called df_long and the column with person identifiers within that data frame is called ID. The code for fitting the DRTM is:

```
R> fit_drtm <- cirt(
        model = model_drtm,
        data = df_long,
        family = "cmp",
        data_long = TRUE,
```

```
6         person_id = "ID")
```

The post-processing functions described above do not support any of the explanatory models yet.

For multidimensional models (Beisemann et al., 2024b) – the most recent addition to the countirt package[1] – the user interface is not as far developed at the time of writing this thesis. An idea for how the user interface can be extended for convenient specification and model fitting of multidimensional models is discussed in Chapter 5. So far, exploratory multidimensional count item response models can be specified using the mcirt_explore function which takes the arguments nfactors (number of latent traits), data (data frame with (only) the item responses in columns), family (count data distribution, can be either "cmp" or "poisson"), penalize (indicates if and how the slope matrix should be penalized, can be either "none" or "lasso"[2]), alpha_constraints (a matrix with as many rows as nfactors and as many columns as the number of items; it imposes constraints on the slope matrix of the same shape, currently allowing to specify which slopes should be fixed to certain values, as indicated in the matrix, and which should be estimated freely, as indicated by a NA entry in the matrix), and control (which takes a list of control parameters for estimation as returned by mcirt_control). Currently, subsequent rotation needs to be coded manually, for example using the GPArotation package (Bernaards & Jennrich, 2005). Lasso-penalized MCIRT models can be tuned using the mcirt_tune_lasso function which takes the arguments nfactors, data, family, alpha_constraints, and control (just as mcirt_explore) as well as the additional arguments penalize_grid (a vector of values to use for the lasso tuning parameter, i.e., η in Equation 3.17) and tuning_crit (the tuning criterion used to tune the lasso tuning parameter, can be either "BIC" or "AIC"). As of now, there are no post-processing functions implemented for the multidimensional models, i.e., functions such as summary or plotting functions are only currently available for unidimensional models.

[1] At the time of writing this thesis, the multidimensional models are available only on the *multdimensional* branch of the countirt package, available here: https://github.com/mbsmn/countirt/tree/multidimensional

[2] And potentially "ridge", but the implementation for ridge penalization is not as well developed and tested so far.

5 Discussion

Count item response models have received increasingly more attention in psychometric model development in the recent years (e.g., Forthmann et al., 2020b; Qiao et al., 2023; Man & Harring, 2019; Man et al., 2022; Tutz, 2022). The challenge of conditionally underdispersed count responses was first addressed by Forthmann et al. (2020b) with a generalization of Rasch's Poisson Counts Model (RPCM; Rasch, 1960), relying on the Conway-Maxwell-Poisson distribution (CMP; Shmueli et al., 2005; Conway & Maxwell, 1962; Huang, 2017) rather than the Poisson distribution for conditional responses. The CMP distribution allows to estimate the degree and the type (over-, equi-, or underdispersion) of dispersion in the conditional response distribution. In the three articles that make up the present thesis, the Conway-Maxwell-Poisson Counts Model (CMPCM) proposed by Forthmann et al. (2020b) is further generalized (1) to include an additional item-specific discrimination (or slope) parameter (Beisemann, 2022), (2) to allow for the inclusion of item or person covariates in (1) to obtain explanatory count item response models (Beisemann et al., 2024a), and (3) to extend (1) to the multidimensional case with more than one latent trait (Beisemann et al., 2024b). A major focus of these articles is the maximum likelihood estimation of these proposed models.

5.1 The Two-Parameter Conway-Maxwell-Poisson Model

In the first article of this thesis (Beisemann, 2022), I proposed the Two-Parameter Conway-Maxwell-Poisson Model (2PCMPM), which extends the CMPCM (Forthmann et al., 2020b) through item-specific discrimination (or slope) parameters, and developed a marginal maximum likelihood estimation method based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

The proposal of a two-parameter count IRT model, which takes advantage of the dispersion flexibility the CMP distribution offers in the same way as Forthmann et al. (2020b)'s CMPCM does, allows to account for count data settings which IRT previously had not been able to adequately account for. Furthermore, as we can see in the two subsequent articles of this thesis, the work on an estimation method using the EM

algorithm lays the groundwork for more methodological developments and extensions in count IRT.

Two simulation studies assessed the proposed model and estimation algorithm and found overall satisfactory performance (Beisemann, 2022). In terms of computational efficiency, the simulations showed that the employed approach to choosing start values (see Chapter 4) helped to achieve comparably small numbers of EM algorithm iterations and thereby shorter computation times (Beisemann, 2022). The idea to take advantage of the Poisson distribution's greater computational convenience to obtain good start values for the parameters on the mean of the CMP model is not specific to the 2PCMPM and could also be employed in estimation routines for other CMP based models, such as regression and hierarchical regression models.

In terms of parameter recovery, the simulation studies showed that – as one would expect – increased sample size improved parameter estimation accuracy (Beisemann, 2022). Out of the item parameters, the log dispersion parameters appeared to be the hardest to estimate, exhibiting more bias in more simulation conditions than the other two types of item parameters (Beisemann, 2022). Coverage of 95% confidence intervals (CI) was mostly (but not always) satisfactory and better for larger samples (Beisemann, 2022). As the method for standard error computation (Chalmers, 2012; Pritikin, 2017) was chosen mostly for computational convenience and CIs simply relied on a normal approximation, this was quite a promising result which likely could be further improved upon by trying out different approaches. In the same vein, results in terms of person parameter estimation (a brief exploration of the measurement perspective which has overall taken a backseat throughout this dissertation work) were overall satisfactory despite using the computationally easiest approach of EAP estimation based on the EM algorithm (Beisemann, 2022).

In the second simulation study, the 2PCMPM was used as the data-generating model and compared to previously suggested models which are misspecified in that setting. The results showed showed that parameter point estimation was only slightly improved upon by the 2PCMPM as opposed to the other models (and for person parameters, correlations between point estimates were very high, as is not uncommon for one- and two-parameter IRT models, see e.g., Bürkner, 2020; Loken & Rulison, 2010). However, uncertainty quantification (i.e., 95% CI coverage, model-implied reliability) suffered from the misspecification (in particular, coverage of the slope parameters) and was improved by using the correctly specified 2PCMPM (Beisemann, 2022). This showcased that for settings with item-specific discrimination and dispersion, using the (in this case, correctly specified) 2PCMPM is actually advantageous over using previously existing models.

More discussion regarding the 2PCMPM and the results from the simulation studies assessing the 2PCMPM is provided in Beisemann (2022).

5.1.1 Limitations

The first article (Beisemann, 2022) discusses several limitations and ideas for future research based on the work presented in the article. To summarize the most important ones, a limitation with regard to the simulation studies was the relatively low number of simulation trials due to the computational expensiveness of the 2PCMPM as well as the limited comparison to other count models (Beisemann, 2022). As discussed in Beisemann (2022), approaches to further improve computation times, such as EM accelerators (for a recent IRT-specific overview, see e.g., Beisemann, Wartlick, & Doebler, 2020), could be explored in future research to remedy this limitation.

Some numerical instabilities occurred in a small number of simulation conditions in the first simulation study (Beisemann, 2022). In some cases, these could have been due to unfortunate start values, which would be easily fixed in practical applications by manually adjusting the start values (Beisemann, 2022). An even more sophisticated solution could be to implement a type of screening and if necessary adjusting start values within the algorithm. Another source of numerical instabilities might be (by chance, not by design) more extreme data conditions, for example with very high counts. The algorithm and especially its implementation (see Chapter 4) were not designed for settings such as these, leaving them untested at best and unsuitable at worst for these settings (see also below for a further discussion of this issue). Future research could investigate the algorithm's and implementation's behaviour in more extreme data settings and develop adjustments to adequately handle such settings.

Further limitations are discussed and ideas for future research are suggested in Beisemann (2022).

5.2 Explanatory Extensions

The second article of this thesis (Beisemann et al., 2024a) provided two extensions to the 2PCMPM (Beisemann, 2022): We extended the 2PCMPM to include item covariates on any or all item parameters in the Distributional Regression Test Model (DRTM) and to include (categorical) person covariates on the latent person parameter in the Count Latent Regression Model (CLRM). We developed a marginal maximum likeli-

hood estimation procedure to which end we used the EM algorithm (Dempster et al., 1977), building upon the work of the first article of this thesis (Beisemann, 2022).

The introduction of two new explanatory IRT models for count data enables researchers to better understand existing count data generating tests and could inform construction of new tests. For example, item dispersion has only been flexibly accounted for with the recent introduction of the CMPCM (Forthmann et al., 2020b). After correctly modelling a mixture of item-specific over-, under-, and equidispersion in a given test, substantive researchers are most likely interested in understanding what leads to a specific type of dispersion. With exploratory IRT models, such as the ones we proposed in the second article of this thesis (Beisemann et al., 2024a), researchers can investigate questions like these. We illustrated two possible applications in the paper with real-data examples (Beisemann et al., 2024a).

In Beisemann et al. (2024a), we discussed that the simulation studies revealed overall satisfactory parameter recovery, with dispersion parameters displaying the most bias (paralleling previous findings in Beisemann, 2022, for the 2PCMPM). For the DRTM, it was additionally parameters on the slope that exhibited more bias, in particular in smaller data sets. For the DRTM, only longer tests displayed satisfactory power across all parameters (at least in some conditions). For the CLRM, we only observed power drops for smaller effects in a small sample. Both of which are unsurprising results patterns. In terms of type I error rate, both models yielded mostly satisfactory results, but did exhibit some slightly liberal type I error rates in certain conditions. This could have been caused by the Wald approximations we used for inference. However, we are also estimating power and type I error rates with a certain amount of imprecision as we only ran 250 simulation trials due to the computational expense of the models (Beisemann et al., 2024a).

We discussed the results from the simulation study further in Beisemann et al. (2024a).

5.2.1 Limitations

With regard to the simulation studies, the limitations concerning the data settings studied and the numerical instabilities from the first article of this thesis (Beisemann, 2022) could be re-iterated again for the second article (compare Chapter 5.1.1 and Beisemann et al., 2024a). In the article, we also discuss some limitations to our simulation design specific to the exploratory extensions, such as the number of covariates included or the correlation between them (which we did not vary systematically; Beisemann et al., 2024a).

An important limitation to our work in the second article (Beisemann et al., 2024a) is a computational issue (see also Chapter 4 and Section 5.4.1 below): The CLRM requires interpolating K (number of quadrature nodes) $\times M$ (number of items) $\times N$ (number of persons) values in each E step and each iteration of each M step, quickly leading to very large numbers of values that are interpolated which causes computation times to increase beyond practical feasibility. Our solution to focus on categorical person covariates with few categories (as we can then use a method of pattern matching, see Chapter 4) leads to a loss of information (Beisemann et al., 2024a). The same challenge arises when one considers a future extension of the 2PCMPM to an explanatory IRT model with person-by-item covariates, as we encounter in differential item functioning (DIF) models (see e.g., De Boeck & Wilson, 2004, for an introduction). Finding solutions to this issue will enable the development of such DIF models which are popular among substantive researchers.

We investigated power and type I error rates in covariate detection in our second simulation study (Beisemann et al., 2024a). We used the same method (Chalmers, 2018; Pritikin, 2017) to compute standard errors as I used in Beisemann (2022) and relied on Wald approximations for inference. This approach is easy to implement but resulting power and type I error rates could probably be improved upon by investigating more specifically what type of standard error computation methods works best for these models, and in combination with what type of confidence interval (Beisemann et al., 2024a). This could be an interesting idea for future research, not just for the DRTM and CLRM, but also more generally for the 2PCMPM. The topic would both be interesting from a more mathematical standpoint as well as from a software development standpoint, so that we could give more options for inference in `countirt` (see also below).

The second article of this thesis (Beisemann et al., 2024a) discusses further limitations and ideas for future research.

5.3 Multidimensional Extensions and More General Framework

The third article (Beisemann et al., 2024b) of this thesis extended the 2PCMPM (Beisemann, 2022) to a multidimensional count IRT framework, subsumed under the general multidimensional two-parameter Conway-Maxwell-Poisson model (M2PCMPM). The M2PCMPM models count responses to items as dependent on one or more latent traits, while allowing for item-trait-specific discriminations (conceptually comparable to factor loadings) and flexible item-specific dispersion modeling through the CMP distribution. The M2PCMPM contains existing models, such as the 2PCMPM (Beisemann,

2022), CMPCM (Forthmann et al., 2020b), RPCM (Rasch, 1960), and related count IRT models (e.g., Myszkowski & Storme, 2021), as well as potential new models (e.g., with more than one latent trait) as special cases. In the third article (Beisemann et al., 2024b), we concentrate mostly on the M2PCMPM as an exploratory count IRT model, allowing to exploratorily (i.e., without prior assumptions) estimate the association structure between items and latent traits. Based on the work from the first article of this thesis (Beisemann, 2022) and similarly relying on the EM algorithm (Dempster et al., 1977), we developed marginal maximum likelihood estimation methods for the M2PCMPM. To achieve the in exploratory IRT often pursued goal of a simple structure of the discrimination matrix (Browne, 2001; Thurstone, 1947), we combined the developed EM algorithm with traditional rotation methods (Carroll, 1957; Clarkson & Jennrich, 1988; Kaiser, 1958, 1959) as well as developed a penalized EM algorithm using a lasso penalty (Tibshirani, 1996), inspired by (Sun et al., 2016).

The development of a multidimensional count IRT framework enables researchers to propose a number of new (confirmatory) count IRT models that can be expressed in and estimated with the framework (Beisemann et al., 2024b). Relying on the CMP distribution helps the framework to flexibly model item-specific dispersions, as previous research has shown to often be necessary (e.g., Forthmann et al., 2020b). The exploratory approach predominantly considered in this work is especially helpful in test construction.

We evaluated the developed exploratory methods, both rotational and lasso-penalized, in a small simulation study (Beisemann et al., 2024b). As we have come to expect from the previous two articles of this thesis (Beisemann, 2022; Beisemann et al., 2024a), the (log) dispersion parameters displayed more bias, while the intercept parameters were overall estimated quite well (but differences between conditions were observed and discussed) (Beisemann et al., 2024b). Assessing parameter recovery for the discrimination parameters was more difficult for the multidimensional models, as there is no unique solution for the discrimination matrix. Thus, we opted for evaluating parameter recovery for the multidimensional discriminations. We observed satisfactory performance in several conditions. The rotational approaches tended to perform better than the penalized approaches (Beisemann et al., 2024b). Conditions in which parameter recovery performance and performance in terms of correct estimation rate (i.e., correctly estimating parameters freely as opposed to shrinking them to 0 per regularization) dropped, were conditions with more items and more traits (Beisemann et al., 2024b). As we hypothesized and discussed in Beisemann et al. (2024b), we might have observed this performance pattern because our simulation study held the sample size constant rather than increasing it with the number of model parameters (which would allow to hold the

observations-to-model-parameters ratio constant). The simulation study further showed that tuning the hyperparameter for the lasso-penalized models worked better in some settings than others, but overall could still be improved in future research (Beisemann et al., 2024b).

More discussion regarding the M2PCMPM and the results from the simulation study is provided in Beisemann et al. (2024b).

5.3.1 Limitations

The results from the simulation study are limited in the sense that they can only speak to the algorithms' and models' behaviour in the investigated settings. For the third article (Beisemann et al., 2024b), the simulation study's goal was to provide a proof of concept for our developed methods, we would recommend that future research test the methods more extensively in more exhaustive settings. For example, a simulation study that enables required sample size recommendations as well as tests our hypotheses about performance drops in settings with small parameters-to-sample-size ratios would be interesting and important (Beisemann et al., 2024b).

The computational burden of the multidimensional CMP models is very high, even for comparatively small models with only few items and few latent traits in comparably small samples. This imposed limitations on our simulations, not least in terms of the number of replications we were able to run, but it is likely also going to be a relevant limitation in practical applications. For the models with four latent traits in our simulation study, we decreased the number of quadrature nodes per trait to achieve more manageable computation times. While this helps in terms of computation times, it likely impairs accuracy at least a little bit. In line with ideas suggested for the first and second article of this thesis, future research could investigate how computation times can be improved (Beisemann et al., 2024b).

For higher numbers of latent traits, Gauss-Hermite (GH) quadrature – upon which our EM algorithms are based – is known to work less well (Chalmers, 2012), as computation times drastically increase with the number of traits. An alternative to GH quadrature is Monte Carlo (MC) integration for which the proposed EM algorithms could quite easily be adapted, as I have already roughly sketched out and crudely implemented to test. A challenge which arises when switching to MC integration is how to choose the convergence criterion. This is a general challenge for MC-EM algorithms (McLachlan & Krishnan, 2007), but a possibly even bigger challenge for the M2PCMPM-EM algorithm in our implementation which uses interpolation for some of the CMP quantities,

adding another layer of possible inaccuracies.

Our work in this article did not include an approach to calculating standard errors for the multidimensional count IRT models, as the development of the estimation framework, including the penalized extension, was already quite work intensive, leaving standard errors beyond the scope of this work. However, this would be a worthwhile endeavour for future research, in particular if future research compared different approaches to recommend the best one for multidimensional count IRT models.

As we compared the developed penalized approaches to the rotation methods for the developed unpenalized model, we opted for tuning the hyperparameter in the penalized models on the whole data set (Beisemann et al., 2024b). For hyperparameter tuning – generally in machine learning approaches, and also specifically for lasso – it is at least recommended to divide the data into a training data set upon which the hyperparameter is tuned and then to fit the model again on the test data set with the selected hyperparameter (Hastie et al., 2009). If we had done this here, then our lasso-penalized models would have been fit to a subset of the data while our rotated models would have been fit to the whole data set, making them difficult to compare. However, the approach we have taken here is going to be prone to overfitting – fitting the data at hand too well, while fitting new data more poorly. Alternatively, we could have carried out the comparison for all approaches on the test data set, which could be explored for future comparisons. In general, we observed that hyperparameter could still be improved upon, for example by choosing different tuning grids or other tuning criteria (Beisemann et al., 2024b). Such endeavours will also be subject to the challenge described above: the computational burden of the multidimensional CMP models. Even as it is, hyperparameter tuning (already using warm starts to increase computational efficiency) was computationally very expensive. Thus, I would recommend future research address the computational challenges first to have more freedom in improving hyperparameter tuning.

We discussed these and further limitations as well as ideas for future research in more detail in the third article of this thesis (Beisemann et al., 2024b).

5.4 The `R` package `countirt`

The `R` package `countirt` implements the EM algorithms for the count item response models proposed and developed in this thesis as well as for Poisson variants of these models. The package is written in `R` and `C++`, tied into `R` using `Rcpp` (Eddelbuettel et al., 2011). The main focus in terms of the package development during the time

working on these thesis projects was to develop reasonably stable and computationally efficient estimation algorithms (for details, see Chapter 4).

5.4.1 Limitations

Chapter 4 describes and discusses computational challenges related to the CMP_μ distribution and how they are addressed in `countirt` using a naive interpolation-from-grid approach. For the settings studied in the simulations in Beisemann (2022), Beisemann et al. (2024a), Beisemann et al. (2024b) this approach yielded quite satisfactory results. However, it is worth noting that even with this approach, the MCIRT models (Beisemann et al., 2024b) still showed considerable computation times. Further, as person covariates add another index dimension to the quantites that need to be interpolated, the interpolation-from-grid approach implemented in `countirt` only allows for categorical person covariates (Beisemann et al., 2024a) and this also poses a challenge to the development of DIF models in the future. As discussed in Beisemann et al. (2024a), considering different estimation approaches to the EM algorithm, such as Laplace approximations (for other IRT models, see e.g., Andersson & Xin, 2021), might be a possible avenue for future research which aims to increase computational speed.

In future research, the performance of the `countirt` package could be assessed in further simulations to make recommendations for settings other than examined in Beisemann (2022), Beisemann et al. (2024a), Beisemann et al. (2024b). In particular, the implementation was designed for settings such as in Beisemann (2022), Beisemann et al. (2024a) with small to moderate counts. Large counts are likely going to be less accurately accounted for with this implementation, as the algorithm in `countirt` truncates expected μ values above a maximum value (currently, 200) and sets them to the maximum value. This is necessary to stay within the bounds of the interpolation grid as only inter- and no extrapolation is implemented in `countirt`. It is natural that this will affect parameter estimation accuracy for larger counts. Similarly, more extreme parameter values, such as large slope values (which will also naturally lead to larger counts), might also lead to larger expected μ values, at least for some quadrature nodes, which will likely again affect parameter estimation accuracy. For other data settings than those that inspired the models proposed in this thesis and were examined in Beisemann (2022), Beisemann et al. (2024a), Beisemann et al. (2024b), one might always opt to extend the interpolation grid. A drawback of this – and the reason why it was not chosen larger when the data settings examined here did not require it – is that interpolation and therefore computation time all over will increase.

In general, the interpolation-from-grid approach used in `countirt` is a naive and

ad-hoc implementation used to ease the computational burden of the proposed, computationally expensive CMP-based IRT models developed in this thesis. Philipson and Huang (2023) proposed an interpolation approach for CMP regression models. Their sophisticated approach uses a specifically developed grid from which they not only inter- but also extrapolate values. For future development of `countirt`, one might consider extending and applying the Philipson and Huang (2023) interpolation approach to CMP-based IRT models and implementing this in `countirt`.

The current capabilities and limitations of the user interface for model fitting in the `countirt` package are described in Chapter 4. Important limitations include post processing for explanatory models and the user interface for multidimensional models. How these can be addressed by possible future developments of the package is outlined in the following section.

5.4.2 Ideas for Future Development

The `countirt` package is currently only available from GitHub. The goal of future developments should be to extend the package to the point where it is more feature complete, at what point, it can also be submitted to CRAN. The major steps to take towards that direction are outlined in the following.

The first major idea for future development of `countirt` concerns the most recent implementation of the multidimensional count item response models (Beisemann et al., 2024b). The currently exported multidimensional model fitting functions do not yet align well with the `cirt` model fitting function for unidimensional models[1]. Future package development could focus on developing a `mcirt` model fitting function which would work analogously to the `cirt` function for unidimensional models, including extended model specification syntax which would allow more than one latent trait. Either the model specification or the model fitting function would need to allow users to specify whether they wish to fit an exploratory or confirmatory model.

The post-processing options available in `countirt` – as described in Chapter 4 – are furthest developed for the 2PCMPM and constrained versions thereof. In a first step, these available post-processing options should be extended to support the explanatory as well as the multidimensional models. In a second step, the post processing available in `countirt` could be extended further. For example, functions to calculate model and item fit statistics could be provided, the options for latent trait estimation and standard

[1] This is also the reason why the implementation of the multidimensional models has not yet been merged into the main `countirt` branch on GitHub.

error computation as well as model parameter inference tests could be extended, or functions to calculate and plot item and test information as well as reliability could be added (see also Section 5.5).

On a more fine-grained level, the flexibility of the `cirt` and a potentially added `mcirt` function in terms of model specification could be extended. For instance, the `cirt` function could be extended to allow constraints on the intercept parameters as well as allow even more combinations of constraints on different parameters. The underlying implementation for constraints in a potentially added `mcirt` is already present on the multidimensional branch of `countirt` on GitHub. Currently, only constraints in the shape of fixing parameters to certain values are available. In the future, this could be extended to also allow equality constraints and combinations of the two.

5.5 General Limitations and Avenues for Future Research

In the following, I am going to discuss limitations to the thesis and the work as a whole, beyond the more specific limitations discussed in the context of each individual article, and outline some ideas for how they could be addressed in future research.

With the focus of this work on model calibration, i.e., the estimation of item parameters in the four different models proposed in Beisemann (2022), Beisemann et al. (2024a), Beisemann et al. (2024b), a detailed perspective on measurement in count item response models was beyond the scope of this thesis. This might constitute an interesting avenue for future research, especially as measurement and the related concept of (test and item) information pose interesting challenges in the context of count item response models. As we briefly touched on in the discussion in Beisemann et al. (2024a), item information (i.e., the Fisher information with regard to the latent variable θ) for count item response models (with a log link function) usually grows infinitely with increasing θ. This is challenging to align with the common psychometric understanding of measurement from a substantial perspective: An infinitely large item information for an infinitely large θ would mean that the higher the latent ability, the more precisely it can be measured. This is at odds with the substantial intuition that it is likely difficult to capture the upper end of the ability spectrum and to construct items which are able to do so. Future count item response theory research could investigate the item and test information for different count item response models, including the 2PCMPM (Beisemann, 2022), and provide an integration with substantial psychometric theory.

The data settings for which the models in this thesis were developed were comparably "well behaved", that is, they did not pose extreme challenges to model estimation. As

I discussed above in Chapter 5.4.1, the developed estimation methods are going to be challenged in – or in the worst case, unsuitable for – more extreme data settings. This may include data with large counts (which are not accounted for by the current interpolation grid in `countirt`) or settings in which the latent trait has a notably skewed distribution. Future research may investigate how the proposed models and estimation algorithms could be adapted or extended to account for different more extreme data settings. Ideally, future research could start by reviewing different count data settings in psychometrics and deriving a set of possible more extreme data patterns which in turn could be addressed accordingly in method development.

5.6 Conclusion

In summary, the three works that make up the present thesis extended the count item response theory landscape by (1) introducing a two-parameter count IRT model which allows for flexible dispersion modelling, along with an estimation method for this model which was previously not available, and going on to extend the proposed two-parameter count IRT model (2) to include person or item covariates. We built upon my work in the first article of this thesis to develop estimation methods for the models proposed in the second article of this thesis. Finally, the third article contributed (3) a multidimensional count IRT framework, again building on work from the first article. Along with these methodological developments, this thesis included (4) the development of an `R` package that implements the proposed models and developed estimation methods. With these contributions, this thesis provides new methods to substantive researchers that help to analyze their data, and opens up new avenues for future methodological developments in count IRT.

Bibliography

Andersson, B. & Xin, T. (2021). Estimation of latent regression item response theory models using a second-order Laplace approximation. *Journal of Educational and Behavioral Statistics*, *46*(2), 244–265. doi:10.3102/1076998620945199

Baghaei, P. & Doebler, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. doi:10.1177/0033294118797577

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

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., 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

Beisemann, M., Forthmann, B., & Doebler, P. (2024a). 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., Holling, H., & Doebler, P. (2024b). Every trait counts: Marginal maximum likelihood estimation for novel multidimensional count data item response models with rotation or l1–regularization for simple structure. *PsyArXiV preprint*. doi:10.31234/osf.io/fqyjs

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Boston, IL: Addison-Wesley.

Blocker, A. W. (2018). *Fastghquad: Fast 'rcpp' implementation of gauss-hermite quadrature*. R package version 1.0. Retrieved from https://CRAN.R-project.org/package=fastGHQuad

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

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. doi:10.3929/ethz-b-000240890

Brown, A. & Croudace, T. J. (2014). Scoring and estimating score precision using multidimensional irt models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). New York, NY: Routledge.

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

Bürkner, P.-C. (2020). Analysing standard progressive matrices (spm-ls) with bayesian item response models. *Journal of Intelligence*, *8*(1). doi:10.3390/jintelligence8010005

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

Chalmers, R. P. (2018). Numerical approximation of the observed information matrix with Oakes' identity. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 415–436. doi:10.1111/bmsp.12127

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.

De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, *39*, 1–28. doi:10.18637/jss.v039.i12

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. & 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

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

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

Embretson, S. E. & Reise, S. P. (2000). *Psychometric methods: Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Faddy, M. & Bosch, R. (2001). Likelihood-based modeling and analysis of data underdispersed relative to the Poisson distribution. *Biometrics*, *57*(2), 620–624. doi:10.1111/j.0006-341X.2001.00620.x

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*. Berlin/Heidelberg, Germany: Springer.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. doi:10.1016/0001-6918(73)90003-6

Forthmann, B. & Doebler, P. (2021). Reliability of researcher capacity estimates and count data dispersion: A comparison of Poisson, negative binomial, and Conway-

Maxwell-Poisson models. *Scientometrics*, *126*(4), 3337–3354. doi:10.1007/s11192-021-03864-8

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., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*(3), 257–269. doi:10.1080/10400419.2017.1360059

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., Grotjahn, R., Doebler, P., & Baghaei, P. (2020a). A comparison of different item response theory models for scaling speeded C-tests. *Journal of Psychoeducational Assessment*, *38*(6), 692–705. doi:10.1177/0734282919889262

Forthmann, B., Gühne, D., & Doebler, P. (2020b). 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. Retrieved from http://www.gnu.org/software/gsl

Graßhoff, U., Holling, H., & Schwabe, R. (2013). Optimal design for count data with binary predictors in item response theory. In *Moda 10–advances in model-oriented design and analysis* (pp. 117–124). Berlin/Heidelberg, Germany: Springer. doi:10.1007/978-3-319-00218-7_14

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

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347–360. doi:10.1111/j.1745-3984.1984.tb01039.x

Groll, A. & Tutz, G. (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, *24*, 137–154. doi:10.1007/s11222-012-9359-z

Grotjahn, R, Schlak, T, & Aguado, K. (2010). S-C-Tests: Messung automatisierter sprachlicher Kompetenzen anhand von C-Tests mit massiver textspezifischer Zeitlimitierung [S-C-tests: Measurement of automated language competencies with C-tests with a strict text-specific time limit]. In R Grotjahn (Ed.), *Der C-test: Beiträge aus der aktuellen Forschung [The C-test: Contributions from current research]* (297—319). Frankfurt a.M., Germany: Lang.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

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

Heine, S. (2017). *Fremd-und Zweitsprachenerfolg und seine Erklärung durch Erwerbsalter, kognitive, affektiv-motivationale und sozio-kulturelle Variablen: Eine empirische Studie*. Germany: Kassel University Press GmbH.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge, UK: Cambridge University Press.

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

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. doi:10.1080/10705511.2016.1154793

Jansen, M. G. (1986). A Bayesian version of Rasch's multiplicative Poisson model for the number of errors of an achievement test. *Journal of Educational Statistics*, *11*(2), 147–160. doi:10.3102/10769986011002147

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. (2003). Estimating the parameters of a structural model for the latent traits in Rasch's model for speed tests. *Applied Psychological Measurement*, *27*(2), 138–151. doi:10.1177/0146621602250536

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

Jennrich, R. I. (2001). A simple general procedure for orthogonal rotation. *Psychometrika*, *66*, 289–306. doi:10.1007/BF02294840

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika*, *67*, 7–19. doi:10.1007/BF02294706

Jennrich, R. I. (2004). Derivative free gradient projection algorithms for rotation. *Psychometrika*, *69*, 475–480. doi:10.1007/BF02295647

Jentsch, C., Lee, E. R., & Mammen, E. (2021). Poisson reduced-rank models with an application to political text data. *Biometrika*, *108*(2), 455–468. doi:10.1093/biomet/asaa063

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

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, *70*(5), 1—21. doi:10.18637/jss.v070.i05

Loken, E. & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. doi:10.1348/000711009X474502

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, IL: Addison-Wesley.

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

Man, K., Harring, J. R., & Zhan, P. (2022). Bridging models of biometric and psychometric assessment: A three-way joint modeling approach of item responses, response times, and gaze fixation counts. *Applied Psychological Measurement*, *46*(5), 361–381. doi:10.1177/01466216221089344

McCulloch, C. E. & Searle, S. R. (2004). *Generalized, linear, and mixed models*. Hoboken, NJ: John Wiley & Sons.

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

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114. doi:10.1177/01466210022031552

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

Mutz, R. & Daniel, H.-D. (2018). The bibliometric quotient (BQ), or how to measure a researcher's performance capacity: A Bayesian Poisson Rasch model. *Journal of Informetrics*, *12*(4), 1282–1295. doi:10.1016/j.joi.2018.10.006

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

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, set, create: What instructing people to "be creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 423. doi:10.1037/a0036549

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 479–482. doi:10.1111/1467-9868.00188

Ogasawara, H. (1996). Rasch's multiplicative Poisson model with covariates. *Psychometrika*, *61*(1), 73–92. doi:10.1007/BF02296959

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

Pritikin, J. N. (2017). A comparison of parameter covariance estimation methods for item response models in an expectation-maximization framework. *Cogent Psychology*, *4*(1), 1279435. doi:10.1080/23311908.2017.1279435

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*. R Foundation for Statistical Computing. 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.

Robitzsch, A. (2023). Regularized generalized logistic item response model. *Information*, *14*(6), 306. doi:10.3390/info14060306

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. doi:10.18637/jss.v048.i02

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

Silvia, P. J. (2008a). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual Differences*, *44*(4), 1012–1021. doi:10.1016/j.paid.2007.10.027

Silvia, P. J. (2008b). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139–146. doi:10.1037/1931-3896.2.3.139

Silvia, P. J. (2013). Assessing creativity with divergent thinking tasks (Silvia et al., 2008, Study 2, Psychology of Aesthetics, Creativity, and the Arts). OSF. Retrieved from osf.io/8vrck

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. doi:10.1037/1931-3896.2.2.68

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 $l1$ 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

Van der Linden, W. J. (2018). *Handbook of item response theory: Three volume set*. Boca Raton, FL: CRC Press.

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). Cito. Arnhem, The Netherlands.

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

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589–600. doi:10.1007/BF02294492

Part II

Articles

Article 1

Beisemann, M. (2022). A flexible approach to modelling 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. https://doi.org/10.1111/bmsp.12273

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2022), 75, 411–443

published by John Wiley & Sons Ltd
on behalf of British Psychological Society.

www.wileyonlinelibrary.com

A flexible approach to modelling over-, under- and equidispersed count data in IRT: The Two-Parameter Conway–Maxwell–Poisson Model

Marie Beisemann

Department of Statistics, TU Dortmund University, Germany

Several psychometric tests and self-reports generate count data (e.g., divergent thinking tasks). The most prominent count data item response theory model, the Rasch Poisson Counts Model (RPCM), is limited in applicability by two restrictive assumptions: equal item discriminations and equidispersion (conditional mean equal to conditional variance). Violations of these assumptions lead to impaired reliability and standard error estimates. Previous work generalized the RPCM but maintained some limitations. The two-parameter Poisson counts model allows for varying discriminations but retains the equidispersion assumption. The Conway–Maxwell–Poisson Counts Model allows for modelling over- and underdispersion (conditional mean less than and greater than conditional variance, respectively) but still assumes constant discriminations. The present work introduces the Two-Parameter Conway–Maxwell–Poisson (2PCMP) model which generalizes these three models to allow for varying discriminations and dispersions within one model, helping to better accommodate data from count data tests and self-reports. A marginal maximum likelihood method based on the EM algorithm is derived. An implementation of the 2PCMP model in R and C++ is provided. Two simulation studies examine the model's statistical properties and compare the 2PCMP model to established models. Data from divergent thinking tasks are reanalysed with the 2PCMP model to illustrate the model's flexibility and ability to test assumptions of special cases.

The Rasch Poisson Counts Model (RPCM; Rasch, 1960) is a one-parameter Item Response Theory (IRT) model for count data. Several different types of psychometric test generate count data, for instance reading tests (Rasch, 1960; Verhelst & Kamphuis, 2009). Other examples include but are not limited to processing speed tasks (Baghaei, Ravand, & Nadri, 2019; Doebler & Holling, 2016), language tests in the form of C-tests (Forthmann, Grotjahn, Doebler, & Baghaei, 2020; Forthmann, Gühne, & Doebler, 2020), intelligence tests (Ogasawara, 1996), verbal fluency tasks and fluency measurement in divergent thinking tasks (Forthmann, Çelik, Holling, Storme, & Lubart, 2018; Forthmann, Holling, Çelik, Storme, & Lubart, 2017; Myszkowski & Storme, 2021). Psychometric count data can also arise from self-reports, for instance of drug use (Wang, 2010) or frequency of

Correspondence should be addressed to Marie Beisemann, Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany (email: beisemann@statistik.tu-dortmund.de).

DOI:10.1111/bmsp.12273

depressive symptoms (Magnus & Thissen, 2017). Another application of count data IRT models is the field of text data analysis (Proksch & Slapin, 2009) or the analysis of bibliometric indicators to assess researchers' performance (Forthmann & Doebler, 2021; Mutz & Daniel, 2018). To analyse the properties of these psychometric tests within the framework of IRT, appropriate models for count data are required. Recent advances have generalized the RPCM in different directions to address limits imposed by the model's assumptions (Forthmann, Gühne, et al. 2020; Myszkowski & Storme, 2021). As the proposed models each only address one assumption, they remain restricted with regard to other assumptions, limiting their flexibility in count data IRT modelling. The present work aims to fill this gap by generalizing previous work (Forthmann, Gühne, & Doebler, 2020; Myszkowski & Storme, 2021) further and introducing the Two-Parameter Conway–Maxwell–Poisson (2PCMP) model.

1.1. Prior research: The RPCM and other count IRT models

The RPCM – for an introduction see, for example, Baghaei and Doebler (2019) or Verhelst and Kamphuis (2009) – models a participant's response on their latent ability and an item's difficulty. Different estimation methods and extensions have been developed for the RPCM (*e.g.*, Jansen, 1995, 1997; Jansen & van Duijn, 1992; Ogasawara, 1996; Verhelst & Kamphuis, 2009). The RPCM assumes that for each item, the distribution of responses (conditional on a person's latent ability) follows a Poisson distribution with rate λ. The rate is modelled to depend on difficulty and latent ability θ and determines both the location and the spread of the conditional distribution of responses X, so that $\mathbb{E}(X|\theta) = \mathbb{V}\text{ar}(X|\theta)$ (equidispersion assumption). Conceptually, the spread of the conditional distribution of responses is linked to an item's measurement precision. But as the same parameter determines both location and spread, the RPCM links an item's difficulty deterministically with its measurement precision (for constant ability). This is empirically not always a plausible assumption. For instance, Forthmann, Gühne, et al. (2020) found that divergent thinking tasks showed over- and underdispersion depending on the item, and Forthmann and Doebler (2021) found similar phenomena for items measuring researchers' capacity. A violation of the equidispersion assumption results in impaired standard error and model-implied reliability estimation (Forthmann, Gühne, et al., 2020). If $\mathbb{E}(X|\theta) < \mathbb{V}\text{ar}(X|\theta)$ the conditional response distribution exhibits overdispersion, and if $\mathbb{E}(X|\theta) > \mathbb{V}\text{ar}(X|\theta)$ it is underdispersed, with overdispersion leading to liberal and underdispersion to conservative standard errors (Faddy & Bosch, 2001; Forthmann, Gühne, et al., 2020; Hilbe, 2011). Different extensions of the RPCM have been proposed that are able to account for overdispersion – for example, a negative binomial regression model (NBRM; Hung, 2012), a Poisson mixture model (Verhelst & Kamphuis, 2009), a Bayesian Poisson Rasch model (Mutz & Daniel, 2018), a zero-inflated Poisson model (IRT-ZIP; Wang, 2010) and the ICC Poisson counts model (Doebler, Doebler, & Holling, 2014). The recently proposed Conway–Maxwell–Poisson Counts Model (CMPCM; Forthmann, Gühne, et al., 2020), based on the Conway–Maxwell–Poisson (CMP) distribution (Huang, 2017; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005), is the only count data IRT model as of yet which is able to account for both over- and underdispersion. Just as the Poisson distribution is a special case of the CMP distribution, so the RPCM is a special case of the CMPCM.

The CMPCM – like the RPCM – assumes all items to be equally discriminant of the underlying latent ability. That is, each item is assumed to reflect differences in latent ability equally well in the responses to the item. Unless a test has been explicitly constructed to

satisfy this assumption, which is not necessarily very common for count data generating tasks, it is likely to be violated (Myszkowski & Storme, 2021). This limits the applicability of the CMPCM. Take, for instance, the example used in this work, divergent thinking tasks (see also Section 5). The ability to think divergently (*i.e.*, to generate many different ideas in response to a stimulus; Guilford, 1967) can be measured, for example, with items that ask participants to give alternative uses for everyday objects or with items where participants have to imagine many different consequences of a change in everyday life. There is no guarantee that these two types of tasks discriminate equally well between participants. In any case, it is at least desirable to be able to test that assumption, especially for existing count data tasks which were not developed to be analysed within an IRT framework. Further, estimating item discriminations can help to inform item selection. Previous research has laid the ground work to include discrimination parameters in the RPCM – for example, in a count data factor analysis framework (Wedel, Böckenholt, & Kamakura, 2003), or within the generalized linear latent and mixed models (GLLAMM) framework as Poisson GLAMM (Skrondal & Rabe-Hesketh, 2004) – leading to recent work on the Poisson GLAMM special case in an IRT context with the Two-Parameter Poisson Counts Model (2PPCM; Myszkowski & Storme, 2021). As an extension of the RPCM, the 2PPCM contains the former as a special case. Work on including discrimination parameters in count IRT models without the equidispersion assumption is limited to models able to account for overdispersion (Doebler et al., 2014; Wang, 2010). This limits the applicability of two-parameter count IRT models as psychometric tasks might produce not only equi- or overdispersed but also underdispersed data (Forthmann, Gühne, et al., 2020).

1.2. The present work

The present work introduces a model that is a natural extension of both the 2PPCM and the CMPCM: the Two-Parameter Conway–Maxwell–Poisson (2PCMP) model. It models item-specific discrimination as well as item-specific dispersion parameters (the latter allow for modelling underdispersion as well as over- and equidispersion). The 2PCMP model contains the 2PPCM and the CMPCM as special cases, allowing for easy testing and loosening of their assumptions. The 2PCMP model is thus able to address two major limitations of the RPCM within the same model, which has previously not been possible. A limiting factor for the introduction of a model like the 2PCMP model has been a lack of appropriate estimators (Forthmann, Gühne, et al., 2020). The present work fills this gap by deriving a marginal maximum likelihood estimation method for the 2PCMP model based on the expectation–maximization (EM) algorithm. The paper is accompanied by an R implementation of the 2PCMP model. The 2PCMP model's statistical properties are examined and compared to those of established models in two simulation studies. I further reanalyse a divergent thinking fluency task data set with the 2PCMP model to give an empirical illustration of the model.

2. The two-parameter Conway–Maxwell–Poisson model

Under the 2PCMP model (as under the 2PPCM; Myszkowski & Storme, 2021), one assumes that the expected number of counts μ_{ij} given by person i in response to an item j depends on the item parameters α_j and δ_j and the person's latent ability θ_{ij} (all on the log scale) as follows:

$$\mu_{ij} = \exp\left(\alpha_j \theta_i + \delta_j\right). \tag{1}$$

The parameterization in Equation (1) is referred to as the intercept–slope parameterization, with α_j as the slope and δ_j as the intercept. It is often used in IRT for its computational advantages (Baker & Kim, 2004). An alternative common parameterization is the discrimination–difficulty parameterization (*i.e.*, $\mu_{ij} = \exp\left(a_j\left(\theta_i - d_j\right)\right)$), which can be obtained by substituting $a_j = \alpha_j$ for the discrimination and $d_j = -\delta_j/\alpha_j$ for the difficulty in Equation (1) and rearranging (the computational disadvantage of this parameterization is caused by the multiplicative association between a_j and d_j, resulting in a trade-off between the two parameters in estimation). Under typical distributional assumptions for the latent ability θ_i (*i.e.*, $\mathbb{E}(\theta_i) = 0$), the intercept δ_j indicates the log counts one would expect from a person of average ability (*i.e.*, $\theta_i = 0$). With a decrease in the difficulty d_j, a person of the same ability is expected to respond with a larger number of counts, that is, the item is easier. The slope quantifies how strongly a person's latent ability influences the expected response for them. A larger α_j indicates that a person's response to an item is more representative of their latent ability. Figure 1, as an illustration, shows the item response curves (expected responses μ_{ij} plotted against different latent abilities θ_i) under the 2PCMP model for six divergent thinking items (see Section 5 for more details). One can see that the item response curves differ in their steepness, which indicates differences in the slopes α_j. Items which differentiate better between persons with regard to their latent ability (*e.g.*, item 5) have steeper curves, indicating that the same difference in θ_i (x-axis) leads to greater differences in the expected response μ_{ij} (y-axis) compared to items with less discriminatory power and flatter response curves (*e.g.*, items 3 and 6). This information about items can be helpful to know for researchers in terms of item selection and in terms of weighting items to build a total score that best measures the latent ability.

As the 2PCMP model predicts the expected number of counts, that is, the mean of the corresponding probability distribution, the model requires a parameterization of said distribution in terms of its mean. For a long time, such a parameterization of the CMP distribution was not available. Recently, Huang (2017) provided a mean parameterization of the CMP distribution which also builds on the foundation of the CMPCM (Forthmann, Gühne, et al., 2020). The CMPCM is contained in the 2PCMP model as a special case by imposing the constraint that the slopes are equal across items, $\alpha_1 = \ldots = \alpha_M$. The density function for the mean parameterization of the CMP distribution is denoted by CMP_μ in the following and is given by

$$\text{CMP}_\mu(x; \mu, \nu) = \frac{\lambda(\mu, \nu)^x}{(x!)^\nu} \frac{1}{Z(\lambda(\mu, \nu), \nu)}, \tag{2}$$

where $\mu \in (-\infty, \infty)$ is the mean of the distribution and $\nu \in [0, \infty)$ is the dispersion parameter which controls the spread of the distribution. $Z(\lambda(\mu, \nu), \nu) = \sum_{x=0}^{\infty} \lambda(\mu, \nu)^x/(x!)^\nu$ is a normalizing constant (Huang, 2017). The rate $\lambda(\mu, \nu)$ is a function of μ and ν, given by the solution to (Huang, 2017)

$$0 = \sum_{x=0}^{\infty} (x - \mu) \frac{\lambda^x}{(x!)^\nu}. \tag{3}$$

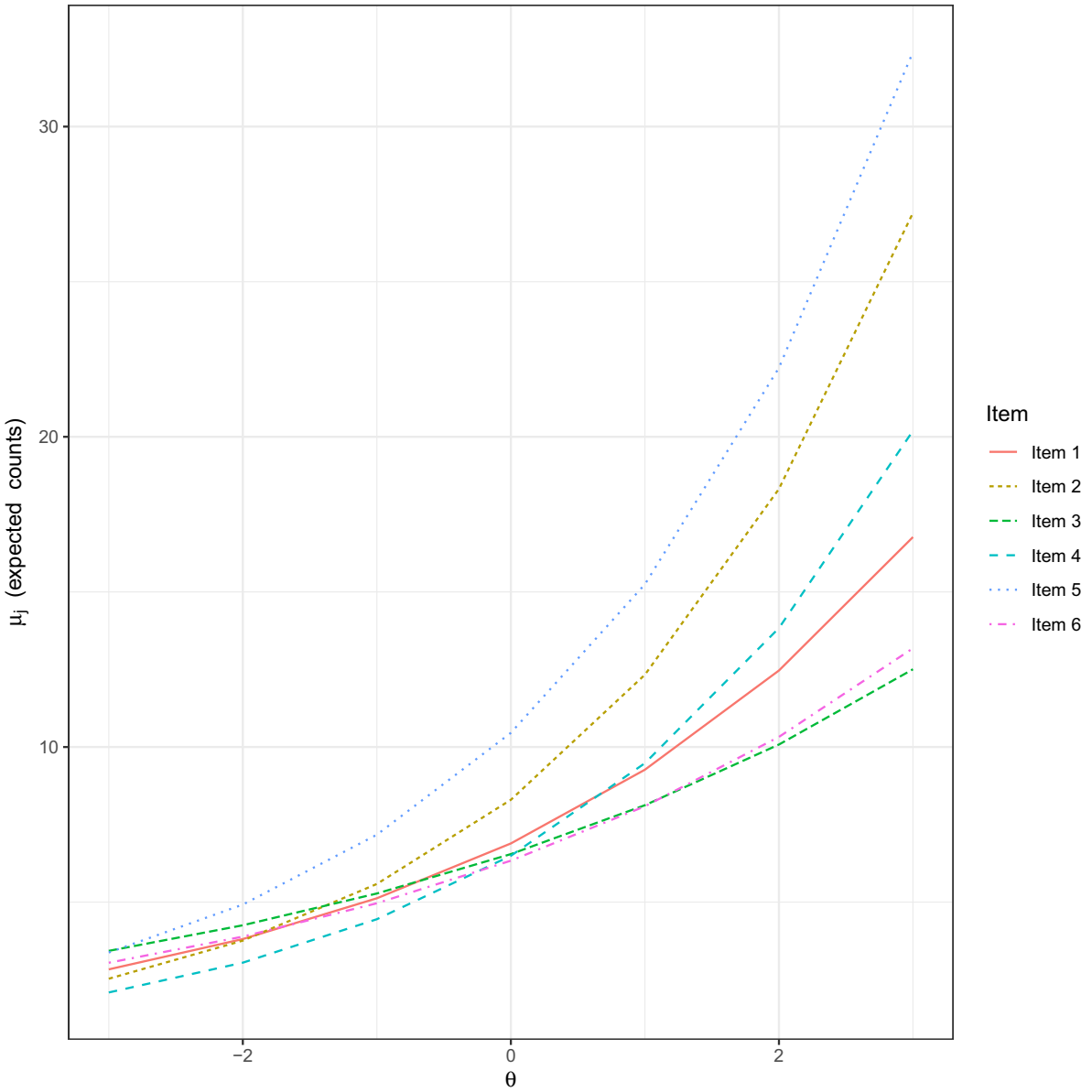


Figure 1. Item response functions (*i.e.*, plotting latent ability θ against the predicted counts μ_j for item j) of the 2PCMP model for six divergent thinking items (application example). Items are colour-coded and represented by different line types as indicated on the right-hand side.

Overdispersion (underdispersion) occurs if $\nu < 1$ ($\nu > 1$). For $\nu = 1$, the case of equidispersion is obtained and Equation (2) simplifies to the Poisson density. This makes it immediately clear that the 2PCMP model contains the 2PPCM as a special case. The dispersion parameter ν can be modelled either as equal across items or as item-specific. Here, I formulate the 2PCMP model in the most general form with item-specific dispersions $\nu_j, j = 1, \ldots, M$. A model with equal dispersion across items can be obtained by imposing the constraint that $\nu_1 = \ldots = \nu_M$.

Combining Equations (1) and (2), the probability of a person i responding with a count x_{ij} to item j, given a latent ability θ_i for person i and item parameters α_j and δ_j as well as an item-specific dispersion ν_j, is then given by

$$P(X_{ij} = x_{ij}|\theta_i, \alpha_j, \delta_j, \nu_j) = \text{CMP}_\mu\left(x_{ij}; \mu_{ij}, \nu_j\right), \tag{4}$$

with μ_{ij} as in Equation (1). Under the assumption of conditional independence, the probability of observing the response vector $\mathbf{x}_i$ for a person i to all M items is given by the product over items $j = 1, \ldots, M$, that is,

$$P(\mathbf{X}_i = \mathbf{x}_i|\theta_i, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\nu}) = \prod_{j=1}^{M} \text{CMP}_\mu\left(x_{ij}; \mu_{ij}, \nu_j\right), \tag{5}$$

with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M)^T$, $\boldsymbol{\delta} = (\delta_1, ..., \delta_M)^T$ and $\boldsymbol{\nu} = (\nu_1, ..., \nu_M)^T$. For ease of reading, the vector concatenating item and dispersion parameters for all items ($\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\nu}$) will be denoted by $\boldsymbol{\zeta}$. In terms of maximum likelihood estimation, the marginal maximum likelihood (MML) method represents the most viable approach for the 2PCMP model, as the joint maximum likelihood method could result in an inconsistent estimator (because with each additional observation, we would have to include an additional parameter for the person's ability) and the conditional maximum likelihood method is not an option for two-parameter IRT models (Baker & Kim, 2004).

For MML estimation, assume that the latent ability parameters $\theta_1, \ldots, \theta_N$ are independent and identically standard normally distributed as $\theta_i \sim N(0, 1)$, $i = 1, \ldots, N$. Note that in two-parameter IRT models, the latent ability variance needs to be fixed to 1 to ensure identification of the model (Baker & Kim, 2004). Denote the density function of the standard normal distribution by ϕ. The joint probability of observing a person i with a latent ability θ_i and a response vector $\mathbf{x}_i$ is given by $P(\mathbf{x}_i, \theta_i|\boldsymbol{\zeta}) = P(\mathbf{x}_i|\theta_i, \boldsymbol{\zeta})\phi(\theta_i)$. Consequently, the marginal likelihood of the item and dispersion parameters under the data $\mathbf{x}$ (across all N persons and all M items) is given by

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) = \prod_{i=1}^{N} \int P(\mathbf{x}_i|\theta_i, \boldsymbol{\zeta})\phi(\theta_i)\mathrm{d}\theta_i. \tag{6}$$

The goal is to obtain the parameter estimates for $\boldsymbol{\zeta}$ which maximize the marginal likelihood in Equation (6) (or rather, the logarithm of Equation (6)). Due to the integral in Equation (6) which does not exist in closed form, this is challenging to do directly. An elegant way to solve this issue is to employ the EM algorithm.

3. Marginal maximum likelihood estimation with the EM algorithm

The EM algorithm (Dempster, Laird, & Rubin, 1977; for a general introduction see, for example, McLachlan & Krishnan, 2007; for an IRT-specific introduction see Bock & Aitkin, 1981) is an algorithm for iterative maximum likelihood (ML) estimation. This section introduces an EM algorithm for the 2PCMP model in a compact and computationally advantageous representation. The corresponding derivation (which first derives a different representation and shows that it is mathematically equivalent to the more compact and computationally advantageous one) is shown in Appendix A.

The EM algorithm for the 2PCMP model uses fixed Gauss–Hermite quadrature to numerically approximate the integral in Equation (6) that does not exist in closed form. Gauss–Hermite quadrature tends to be a sensible choice in lower-dimensional IRT models for binary and ordinal data (Chalmers, 2012). The integral over a continuous variable (in

this case, θ_i) is approximated by a sum over a discretized version of the variable (which I denote by Q_i). The levels of the discretized variable are referred to as quadrature nodes, denoted by $q_1, \ldots, q_k$ for K nodes. Increasing the number of nodes yields better approximations, but increases the computational cost. The quadrature nodes are weighted according to their probability of occurrence with quadrature weights, denoted by w_k for nodes $k = 1, \ldots, K$. Rewriting the marginal likelihood in Equation (6) in quadrature notation yields

$$L_m(\boldsymbol{\zeta}; \mathbf{x}) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{x}_i | q_k, \boldsymbol{\zeta}) w_k, \tag{7}$$

where the expected counts implied by Equation (7) are $\mu_{jk} = \exp(\alpha_j q_k + \delta_j)$.

In MML estimation problems like in IRT, one can consider responses $\mathbf{x}$ as observed data and the latent abilities $\boldsymbol{\theta}$ $(= (\theta_1, \ldots, \theta_N)^T)$ as unobserved data, together forming the complete data $(\boldsymbol{x}, \boldsymbol{\theta})$. The EM algorithm, built for this type of incomplete-data problem, maximizes the complete-data (log) likelihood. It iterates between two steps: In each expectation (E) step, the parameters ($\boldsymbol{\zeta}$) sought are assumed to be known and the expected complete-data (log) likelihood is determined. In each maximization (M) step, the expected complete-data (log) likelihood from the previous E-step is maximized in terms of $\boldsymbol{\zeta}$ (under the parameter estimates from the previous M-step $\boldsymbol{\zeta}'$). The EM algorithm oscillates between E- and M-steps until a convergence criterion is met. Each EM cycle increases the marginal likelihood until the fixed point of the algorithm is reached (McLachlan & Krishnan, 2007).

To be able to take the expectation in each E-step, one needs to calculate the probability distribution over $\boldsymbol{\theta}$ given $\boldsymbol{\zeta}'$ from the previous M-step and the observed data $\mathbf{x}$. One employs Bayes' theorem to this end and approximates the posterior distribution of θ_i by the posterior probabilities of the quadrature nodes $q_1, \ldots, q_k$. The posterior probability for node k and item j given a response vector $\mathbf{x}_i$ is

$$P(q_k | \mathbf{x}_i, \boldsymbol{\zeta}') = \frac{\prod_{j=1}^{M} \mathrm{CMP}_{\mu}\left(x_{ij} | q_k \boldsymbol{\zeta}_j'\right) w_k}{\sum_{k'=1}^{K} \prod_{j=1}^{M} \mathrm{CMP}_{\mu}\left(x_{ij} | q_{k'} \boldsymbol{\zeta}_j'\right) w_{k'}}, \tag{8}$$

where $\boldsymbol{\zeta}_j'$ denotes the set of item and dispersion parameters for item j from the previous M-step. The quadrature weights w_k constitute the prior probabilities for the quadrature nodes, approximating the prior distribution for θ_i, which is assumed to be $N(0, 1)$ for $i = 1, \ldots, N$ under the 2PCMP model.

For the 2PCMP, the expected complete-data log likelihood, $\mathbb{E}(LL_c)$, is proportional to the following expression (see Appendix A for the derivation):

$$\mathbb{E}(LL_c) \propto \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[\left(x_{ij} \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \nu_j \log(x_{ij}!) - \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right) P(q_k | \mathbf{x}_i, \boldsymbol{\zeta}') \right) \right]. \tag{9}$$

Equation (9) can then be maximized in terms of the item parameters for each item $j = 1, \ldots, M$ during the following M-step, where one assumes the $P(q_k | \mathbf{x}_i, \boldsymbol{\zeta}')$ to be given.

Any omitted terms in Equation (9) are constant with respect to $\boldsymbol{\zeta}$, so that they can be disregarded when optimizing for $\boldsymbol{\zeta}$. The maximization is carried out by iteratively finding the roots of the first derivatives with respect to the item parameters. For each α_j, the gradient is given by

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \alpha_j} = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\mu_{jk} q_k}{V\left(\mu_{jk}, \nu_j\right)} \left(x_{ij} - \mu_{jk}\right) P\left(q_k | x_{ij}, \boldsymbol{\zeta}_j'\right), \tag{10}$$

where

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

denotes the variance of the CMP_μ distribution (Huang, 2017), and for each δ_j,

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \delta_j} = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\mu_{jk}}{V\left(\mu_{jk}, \nu_j\right)} \left(x_{ij} - \mu_{jk}\right) P\left(q_k | x_{ij}, \boldsymbol{\zeta}_j'\right). \tag{12}$$

For the dispersion parameters ν_j, it is advantageous in terms of both estimation and interpretation (Forthmann, Gühne, et al., 2020) to optimize for the log dispersions $\log\nu_j$. The estimation-related advantage is an unconstrained parameter space. For each $\log\nu_j$, the gradient is

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \log \nu_j} = \sum_{k=1}^{K} \sum_{i=1}^{N} \nu_j \left(A\left(\mu_{jk}, \nu_j\right) \frac{x_{ij} - \mu_{jk}}{V\left(\mu_{jk}, \nu_j\right)} - \left(\log\left(x_{ij}!\right) - B\left(\mu_{jk}, \nu_j\right)\right) \right) P\left(q_k | x_{ij}, \boldsymbol{\zeta}_j'\right), \tag{13}$$

where one can utilize the results by Huang (2017) that $A = \mathbb{E}_X(\log(X!)(X - \mu))$ and $B = \mathbb{E}_X(\log(X!))$. From the gradients of all three types of parameters, it is easy to see that gradients for the 2PCMP model with equality constraints (*i.e.*, $\alpha_1 = \ldots = \alpha_M$ or $\nu_1 = \ldots = \nu_M$) are simply obtained by taking the derivative in terms of a constant (across items) α or $\log\nu$ which merely adds a sum over M to the gradients shown above.

As explained in more detail in Appendix A, the expression in Equation (9) for the expected complete-data log likelihood and the resulting gradients for the M-step (Equations (10–13)) offer computational advantages. They allow one to express the EM equations, in particular the derivatives for the dispersion parameters, in efficient terms with regard to computational costs and numerical stability.

3.1. Standard errors for model parameters

MML estimation with the EM algorithm has the disadvantage that standard errors are not as immediately available as they are from Newton–Raphson type estimation procedures (McLachlan & Krishnan, 2007), as the observed-data log likelihood $LL_m = LL_m(\boldsymbol{\zeta}; \mathbf{x})$ is not maximized directly. Instead, the expected complete-data log likelihood $\mathbb{E}(LL_c)$ is maximized. The observed information matrix (from which one can obtain the asymptotic covariance matrix of the model parameters) can be expressed in terms of the expected

complete-data log likelihood (Oakes, 1999). To express the fact that the $\mathbb{E}(LL_c)$, which is maximized with respect to $\boldsymbol{\zeta}$, depends on the parameter estimate $\boldsymbol{\zeta}'$ from the previous M-step, write $\mathbb{E}(LL_c(\boldsymbol{\zeta}|\boldsymbol{\zeta}'))$. Then, Oakes's identity (Oakes, 1999) states that at the fixed point (when $\boldsymbol{\zeta} = \boldsymbol{\zeta}'$),

$$\frac{\partial^2 LL_m(\boldsymbol{\zeta};\mathbf{x})}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T} = \left|\frac{\partial^2\mathbb{E}(LL_c(\boldsymbol{\zeta}|\boldsymbol{\zeta}'))}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T} + \frac{\partial^2\mathbb{E}(LL_c(\boldsymbol{\zeta}|\boldsymbol{\zeta}'))}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}'^T}\right|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}'}. \tag{14}$$

Chalmers (2018) provided a finite-differences based numerical approximation technique to Oakes's identity. With this method, one numerically approximates the two summands in Equation (14). This method does not require any additional results to those in Equations (10–13).

3.2. Estimation of ability parameters

Once item parameter estimates have been obtained, one may also use the 2PCMP model to estimate person parameters. To this end, one assumes the item parameters as known. An ML ability estimation technique is given in Appendix B. Under the assumptions of this method, ability parameters are estimated separately for each person. For the CMP_μ distribution this can quickly become computationally expensive for larger samples. A Bayes EAP ability estimation method based on the last E-step is computationally much cheaper in this case and will be used both for the simulation studies and the empirical example below.

The EM algorithm for the 2PCMP model estimates an approximation to the posterior distribution of θ, given the data and the item parameters, in each E-step (Equation (8)). From the (approximative) posterior distribution of the last E-step at the point of convergence, one can estimate the ability of a person i, $i \in \{1, \ldots, N\}$, as the posterior mean (known as the EAP estimator; Baker & Kim, 2004),

$$\hat{\theta}_{i,\text{EAP}} = \sum_{k=1}^{K} q_k P(q_k|\mathbf{x}_i, \boldsymbol{\zeta}), \tag{15}$$

where $\boldsymbol{\zeta}$ are assumed as known (in actuality, one uses the model parameter estimates at convergence). As the (final) E-step yields an approximation of the full posterior, one can just as easily estimate a corresponding standard error,

$$\widehat{\text{SE}}\left(\hat{\theta}_{i,\text{EAP}}\right) = \sqrt{\sum_{k=1}^{K}\left(q_k - \hat{\theta}_{i,\text{EAP}}\right)^2 P(q_k|\mathbf{x}_i, \boldsymbol{\zeta})}, \tag{16}$$

and determine the .025 and .975 quantiles to obtain a 95% credible interval. As the posterior probabilities can be saved from the last E-step, this estimation requires only negligible additional computation time.

3.3. Computational aspects and implementation

The algorithm for the estimation of the 2PCMP model as well as the methods for obtaining standard errors and ability estimates outlined above have been implemented in R and

C++, integrated into the R code with the help of the Rcpp package (Eddelbuettel et al., 2011). The code is available in the R package countirt available on GitHub (https://github.com/mbsmn/countirt). Details of the computational implementation are given in Appendix C. The two main challenges in the numerical implementation of the EM algorithm for the 2PCMP model are numerical stability and computational efficiency. The algorithm repeatedly requires a number of approximations of several infinite series and the solving of Equation (3) for each item and quadrature node combination. For extreme quadrature node, slope, and dispersion values, this may result in numerical instability and/or noticeably increased computation time. To circumvent this, I tabled the most important statistics ($\lambda(\mu, \nu)$, $Z(\lambda(\mu, \nu), \nu)$ and $V(\lambda(\mu, \nu), \nu)$) for a fine grid of μ and ν values. Values for these statistics are interpolated from the grid using two-dimensional bicubic interpolation. Computation time can also be reduced by cutting down the number of iterations until convergence with the choice of starting values. Starting values for slope and intercept parameters of the 2PCMP model are determined by fitting a 2PPCM using a comparatively fast Poisson density based EM algorithm (also implemented in countirt; see the Online Supplementary Materials for details on the algorithm). With this method of choosing starting values, the EM algorithm for the 2PCMP model requires only relatively few EM iterations, as illustrated in the following two sections.

4. Simulation studies

For the first simulation study, the aim was to examine the 2PCMP model's statistical properties, primarily in terms of parameter recovery, in different data settings. For the second simulation study, I wanted to compare the 2PCMP model's performance in a realistic data setting to the performance of established methods which are generalized by the 2PCMP model. Both simulation studies were conducted in R (R Core Team, 2021). Details of the implementation of the simulation studies are given in Appendix C. This work is accompanied by an OSF repository with supplementary materials (https://osf.io/hx5js/). All scripts used to run the simulations and to prepare the results, the simulation results (rds files) as well as additional tables and figures (in the Online Supplementary Materials) are available on the OSF repository.

4.1. Simulation study I

4.1.1. Design and data generation

The design of the first simulation study was inspired by Forthmann, Gühne, & Doebler, (2020). In alignment with their simulations, the number of items simulated in this study was either $M = 4$ or $M = 8$ and the sample sizes (number of persons) simulated were either $N = 100$ or $N = 300$. I set the number of quadrature nodes to either $K = 121$ or $K = 201$ so that I could assess the speed–accuracy trade-off due to the number of quadrature nodes used. I simulated four different kinds of item sets: all items equidispersed, all items overdispersed, all items underdispersed, or a combination of all three types of dispersion among the items (referred to as mixed items). The levels of these design factors were fully crossed to yield 32 different simulation conditions. The true parameter values were inspired by Myszkowski and Storme (2021) as well as my reanalysis of the same data set (see Section 5); they are shown for all conditions with four items in Table 1 (see the Online Supplementary Materials for details). For conditions with

Table 1. True parameter values for simulation study I

j	α_j	δ_j	Equidispersion ν_j (log(ν_j))	Overdispersion ν_j (log(ν_j))	Underdispersion ν_j (log(ν_j))	Mixed dispersion ν_j (log(ν_j))
1	0.33	2.40	1.00 (0.000)	0.40 (−0.916)	1.60 (0.470)	1.00 (0.000)
2	0.47	1.80	1.00 (0.000)	0.50 (−0.693)	1.87 (0.626)	2.40 (0.875)
3	0.60	1.50	1.00 (0.000)	0.60 (−0.511)	2.40 (0.875)	0.30 (−1.204)
4	0.20	2.10	1.00 (0.000)	0.30 (−1.204)	2.13 (0.756)	1.00 (0.000)

eight items, I duplicated the four items with the parameter combinations as shown in Table 1.

I set the number of simulation trials per condition to 250. Note that due to the numerical complexity of the CMP density, estimation of the 2PCMP model as well as standard error computation are computationally expensive, thus limiting the number of simulation trials feasible. For each simulation trial in each condition, I randomly drew N person ability parameters from a standard normal distribution. Using code from Forthmann, Gühne, et al., (2020), I then simulated a data set from a CMP_μ distribution under the respective parameter constellations for the condition (see Forthmann, Gühne, et al., 2020 for details). I fitted a 2PCMP model to the data set and computed standard errors for the item parameters as well as Bayes EAP ability parameter estimates. I recorded all computation times.

4.1.2. Performance criteria

To assess the 2PCMP model's performance in the different simulation conditions, I used the following criteria. Denote a simulation trial by t and the number of simulation trials by T.

Bias. For each model parameter p, I estimated the bias as $\text{Bias}_p = \text{mean}(\hat{p}_t) - p$, that is, the difference between the mean of estimates $\hat{p}_t$ across trials $t = 1, \ldots, T$ and the true parameter p.

Root mean squared error (RMSE). For each model parameter p, I estimated the RMSE as $\text{RMSE}_p = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{p}_t - p)^2}$, that is, the average squared difference between the estimates $\hat{p}_t$ (for $t = 1, \ldots, T$) and the true parameter p. In comparison to the bias, the RMSE additionally takes the variance of the estimator into account, with smaller values indicating that the estimator showed little bias and had small variance.

Coverage of the 95% confidence intervals (CIs). This is the percentage of simulation trials for which the 95% CI for parameter p covered the true value of p. If the nominal α-level of .05 is retained, the coverage should be .95. Using a Wald approximation, the lower boundary of the CI for parameter p in simulation trial t is given by $\text{CI}_{\text{lower}} = \hat{p}_t - 1.96\ \text{SE}(\hat{p}_t)$ and the upper boundary by $\text{CI}_{\text{upper}} = \hat{p}_t + 1.96\ \text{SE}(\hat{p}_t)$, where $\text{SE}(\hat{p}_t)$ denotes the respective estimator's standard error.

Ability parameters. For each simulation trial, I computed the correlation between the true ability parameters in that trial and the ability parameter estimates. To compare performance across conditions and to account for the potential lack of interval scaling of correlations, I computed the median correlation for each condition. Furthermore, I computed (model-implied) empirical reliability estimates of the 2PCMP model as described in Forthmann, Gühne, et al., (2020), that is, as

$\widehat{\text{Rel}} = 1 - \text{mean}\left(\widehat{\text{SE}}\left(\hat{\theta}_i\right)^2\right) / \widehat{\text{Var}}(\theta_i)$, where $\widehat{\text{SE}}\left(\hat{\theta}_i\right)$ denotes the estimate of the standard error for the latent ability estimator for person i and $\widehat{\text{Var}}(\theta_i)$ denotes the estimate of the latent ability variance. In each condition, I calculated the median across trials for the empirical reliability estimates. To be able to evaluate the results, I also calculated the (model-implied) true reliability as $\text{Rel} = \text{Cor}\left(\theta_i, \hat{\theta}_i\right)^2$ (Embretson & Reise, 2013), that is, the variance of the estimated abilities ($\hat{\theta}_i$) explained by the true abilities (θ_i), in each trial. Again, I calculated the median across trials.[1]

Additionally, I examined the numerical stability and convergence, average computation time across trials and average number of EM iterations required to reach convergence. I recorded the computation times, including the computation of the initial values.

4.1.3. Results

All models in all trials converged once their estimation started properly. However, in certain conditions, the situation arose in a very small number of trials (depending on the condition, between 0.4% and 6.8%) that the model estimation fell victim to numerical instability. That is, certain parameter value combinations did not allow for the gradient to be computed numerically stably. This occurred early on in the estimation process, mostly in the first iteration. The conditions concerned were mostly those with underdispersed or mixed items (see the Online Supplementary Materials on OSF for more detailed reporting). In all other trials across conditions, the model estimation started and converged properly.

Computation times and number of EM iterations. In terms of computation times and number of iterations until convergence (shown in detail in the Online Supplementary Materials on OSF), as expected, settings with equidispersed items exhibited faster computation times and required fewer iterations than settings with the other item types (equidispersed items, $M_{\text{ct}} = 418.110-1656.076$ s and $M_{\text{iter}} \approx 17-20$ iterations; overdispersed items, $M_{\text{ct}} = 637.754-3324.056$ s and $M_{\text{iter}} \approx 22-28$ iterations; underdispersed items, $M_{\text{ct}} = 682.159-4287.334$ s and $M_{\text{iter}} \approx 40-69$ iterations; mixed items, $M_{\text{ct}} = 1042.459-3673.292$ s and $M_{\text{iter}} \approx 29-54$ iterations). An increase in the number of items tended to lead to a decrease in the number of iterations (especially for settings with mixed items), but to an increase in computation time. This means that each iteration was computationally a lot more expensive for $M = 8$ due to the greater number of gradients for which roots need to be found. The number of quadrature nodes tended not (or only slightly) to affect the average number of iterations, but, as expected, it made each iteration more expensive, leading in part to considerable increases in computation times. Note that computation times depend on and will differ between machines.

Bias and RMSE for item parameters. Bias and RMSE estimates are shown for conditions with equidispersed (top row) and underdispersed (bottom row) items in Figure 2 and for conditions with overdispersed items (top row) and mixed items (bottom row) in Figure 3. Only values smaller than 1 in absolute value are shown; all exact values are shown in the Online Supplementary Materials on OSF. The results showed that across conditions, bias was very small for the slope and intercepts parameters. RMSE estimates

[1] Note that a comparison with reliability estimators such as Cronbach's coefficient α is not useful here as one of the main assumptions of Cronbach's coefficient α, equal slope parameters, is violated by the 2PCMP model, from which data are simulated.

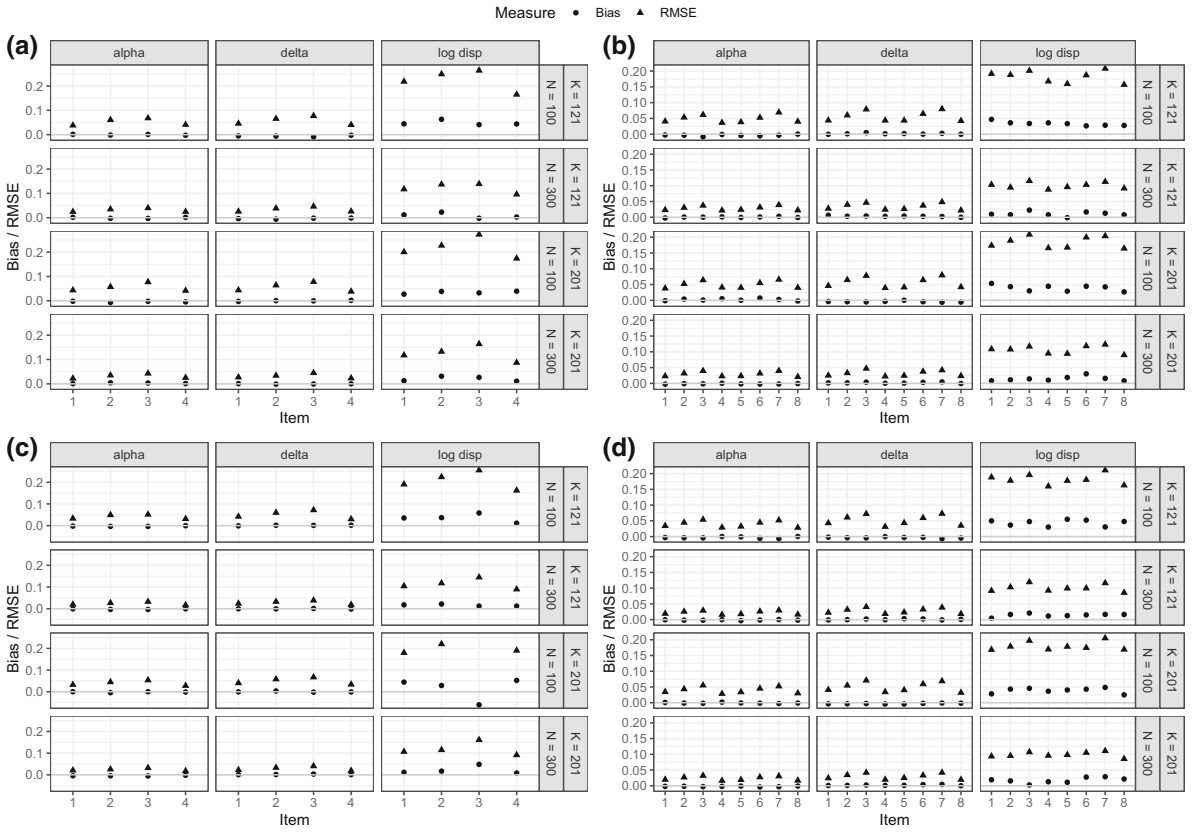


Figure 2. Bias (dots) and RMSE (triangles) for each parameter of each item for all conditions with equidispersed items ((a) four items, (b) eight items) and underdispersed items ((c) four items, (d) eight items). Each column within each plot shows the results for a different parameter (alpha = slope, delta = intercept, log disp = log dispersion). The rows within each plot indicate the sample size (N) and the number of nodes (K). The item number is shown on the x-axis. The horizontal lines indicate 0. Only values less than | 1 | are shown; see the Online Supplementary Materials for all values.

for these parameters tended to be smaller than 0.1 across conditions. The RMSE estimates for slopes and intercepts tended to decrease for conditions with $N = 300$. This effect of the sample size was even more evident for the dispersion parameters. For these, the results showed more noticeable bias for $N = 100$, which was visibly reduced for conditions with $N = 300$. The same pattern emerged for the RMSE. The RMSE estimates for dispersions even exceeded values in absolute magnitude larger than 1 (compare the Online Supplementary Materials). This only occurred for conditions with four items for under- and overdispersed items, and happened for more conditions with four than with eight items for mixed items. These large RMSE estimates predominantly occurred for $N = 100$, and at the very least stabilized for larger N and more quadrature nodes. It is also interesting that these are the only cases where increasing K had any noticeable effect. Otherwise, $K = 121$ seemed to suffice. This is clearly advantageous in terms of computation time.

Coverage of 95% CI for item parameters. Results for the coverage of the 95% CI are shown in the Online Supplementary Materials on OSF. The exact values are also listed in the Online Supplementary Materials. Overall, the results in terms of coverage were promising. Across all conditions, coverage estimates tended to be very close to the nominal level, but note that they were still sometimes slightly liberal (see the Online

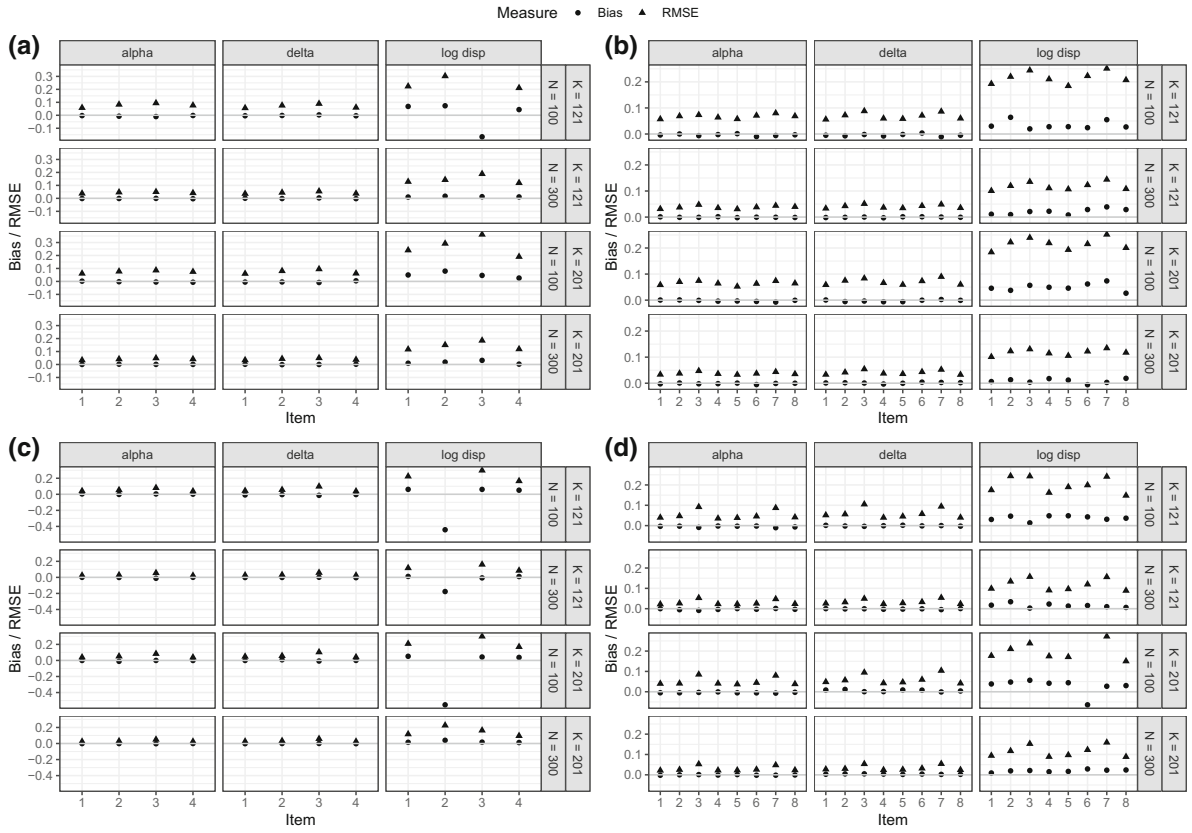


Figure 3. Bias (dots) and RMSE (triangles) for each parameter of each item for all conditions with overdispersed items ((a) four items, (b) eight items) and items with different types of dispersion ((c) four items, (d) eight items). Each column within each plot shows the results for a different parameter (alpha = slope, delta = intercept, log disp = log dispersion). The rows within each plot indicate the sample size (N) and the number of nodes (K). The item number is shown on the x-axis. The horizontal lines indicate 0. Only values less than $|\ 1\ |$ are shown; see the Online Supplementary Materials for all exact values.

Supplementary Materials). Coverage tended to improve with larger N, but did not generally tend to benefit from more quadrature nodes.

Person parameter estimates. Person parameter estimates were assessed using median correlations between the true and the estimated abilities (shown in detail in the Online Supplementary Materials on OSF). These were higher for settings with underdispersed (median r values from .940 to .969) and mixed items (median r values from .910 to .953) and reached the lowest values for overdispersed items (median r values from .831 to .908). Equidispersed items showed median correlations between .897 and .946. Otherwise, only the number of items had a clearly noticeable effect (*e.g.*, for mixed items, $N = 100$, $K = 121$: .910 for $M = 4$ and .952 for $M = 8$). As the (median) model-implied true reliabilities are closely related to the (median) correlations between true and estimated abilities, they showed a very similar pattern of results (see the Online Supplementary Materials on OSF). In terms of median (model-implied) empirical reliabilities, those tended to more noticeably underestimate the true reliabilities in settings with only four items. But there were differences between item groups in this regard, with better results for the underdispersed items (*e.g.*, for $N = 300$, $K = 121$: .887 for the true and .877 for the estimated reliability) and less favourable results for the overdispersed items (*e.g.*, for $N = 300, K = 121$: .703 for the true and .597 for the estimated reliability). For eight items,

model-implied reliabilities were estimated quite well (at least in the median) for all item groups except overdispersed items (*e.g.*, for $M = 8$, $N = 300$, $K = 121$: .823 for the true and .795 for the estimated reliability).

A summary of the main conclusions from simulation study I is provided in Discussion.

4.2. Simulation study II

The aim of the second simulation study was the comparison of the 2PCMP model to established methods in a realistic data setting where the complexity of the 2PCMP is warranted (*i.e.*, a setting with varying slopes and varying dispersions). The models I included for comparison were the 2PPCM (Myszkowski & Storme, 2021) and the CMPCM (Forthmann, Gühne, et al., 2020). I estimated them both once as described by the respective authors and once as constrained 2PCMP models to examine any potential differences in estimation algorithms. By design, all models in this study but the full 2PCMP model are misspecified. The aim of the study was to examine how impaired performance of the established models is by realistic misspecification and thus what advantage the 2PCMP model can offer.

4.2.1. Design

For the realistic data setting, I used parameter estimates obtained by reanalysing divergent thinking tasks data (Silvia, 2008a, 2008b; Silvia et al., 2008) made available by the author with permission to reanalyse (Silvia, 2013) (for the parameter estimates, see Table 5). Mimicking the real data, I simulated $M = 6$ items and $N = 242$ participants in each simulation trial. As in simulation study I, I drew the underlying abilities of the participants (θ_i, $i = 1, \ldots, N$) from a standard normal distribution and then simulated data from a CMP_μ distribution based on code by Forthmann, Gühne, et al., (2020) with $\mu_{ij} = \exp\left(\tilde{\alpha}_j\theta_i + \tilde{\delta}_j\right)$ and $\nu_j = \exp\left(\tilde{\nu}_{\log,j}\right)$, where $\tilde{\alpha}_j$, $\tilde{\delta}_j$, and $\tilde{\nu}_{\log,j}$ are the parameter estimates for the slopes, intercepts, and log dispersions, respectively, obtained through the reanalysis (Table 5). I ran 500 simulation trials.

4.2.2. Models for comparison and performance criteria

I fitted the 2PCMP model using the EM algorithm presented above with 121 quadrature nodes (as the first simulation study indicated that these would suffice in most cases). I further included the CMPCM (Forthmann, Gühne, et al., 2020) which constitutes a special case of the 2PCMP, with slope parameters constrained so that $\alpha_1 = \ldots = \alpha_M$. I fitted the CMPCM using two different implementations: (1) with the EM algorithm for the 2PCMP presented above, and (2) as described in Forthmann, Gühne, et al., 2020 using the glmmTMB package (Brooks et al., 2017). These implementations differ not only with regard to the algorithm used for model estimation, but also slightly in the model formulation. Yet they both constitute a one-parameter CMP model. For the first implementation, the latent ability variance is fixed at 1 and I estimate one slope parameter (constrained to be the same across items). With the second implementation, the slope parameters of all items are fixed at 1 and I estimate the latent ability variance freely (see Forthmann, Gühne, et al., 2020, for details). In order to compare dispersion estimates from these two implementations, I inverted the estimates provided by glmmTMB.

The third model included is the 2PPCM (Myszkowski & Storme, 2021). This model is contained as a special case within the 2PCMP with the constraint that $\nu_1 = \ldots = \nu_M = 1$. There are existing estimation algorithms and corresponding software implementations for the 2PPCM – for example with the software MPlus (Muthén & Muthén, 1998, see Myszkowski & Storme, 2021, for an overview) – but for convenience I also implemented an EM algorithm for the 2PPCM based on the Poisson density in the countirt package (see the Online Supplementary Materials on OSF for details). I fitted the 2PPCM once with that Poisson-density-based EM algorithm and once using the EM algorithm for the 2PCMP based on the CMP density under the constraint that $\nu_1 = \ldots = \nu_M = 1$. For an explanation regarding the relation between the EM algorithms based on the Poisson and CMP density, see the Online Supplementary Materials.

I used the same performance criteria as in simulation study I. Additionally, I computed the median (across trials) correlations between the ability scores as produced by the five models.

4.2.3. Results

None of the models experienced any numerical instability in any of the 500 trials. They all converged in each trial.

Computation times and number of EM iterations. On average across trials, the computation time was longest for the CMPCM fitted with glmmTMB ($M_{ct} = 1372.287$ s). The (full) 2PCMP took on average the second longest time ($M_{ct} = 714.221$ s) and on average required $M_{\text{iter}} \approx 20$ iterations until convergence. This was followed closely by the 2PCMP with equal slopes (*i.e.*, CMPCM with alternative formulation; $M_{ct} = 711.903$ s and $M_{\text{iter}} \approx 26$ iterations), the 2PCMP with dispersions fixed at 1 (*i.e.*, a 2PPCM; $M_{ct} = 403.988$ s and $M_{\text{iter}} \approx 47$ iterations), and the 2PPCM ($M_{ct} = 10.542$ s and $M_{\text{iter}} = 46$ iterations). These results reflect that the Poisson density and gradients are much easier and less computationally expensive to evaluate than the CMP density and gradients. The starting value determination approach for the full and the equal slopes 2PCMP model led to considerably smaller numbers of iterations (as compared to the two 2PPCMs which use a different approach). Note that computation times depend on and will differ between machines. Standard deviations for computation times and number of iterations are presented in the Online Supplementary Materials on OSF together with additional considerations.

Bias, RMSE, and coverage of 95% CIs for item parameters. Table 2 displays the estimates for the bias, the RMSE and the coverage of the 95% CIs. As in simulation study I, the bias for the (full) 2PCMP was small to negligible across parameters (with comparatively larger biases for the dispersion parameters). As expected, bias tended to be greater for the four misspecified models. In particular, at least for some parameters and models, the bias tended to be larger than the average standard error for the respective parameter, while for the parameters in the full 2PCMP model, the bias was always smaller (in absolute magnitude) than the average standard error (see the Online Supplementary Materials on OSF for more details and standard error ranges). This pattern was more pronounced for slope and dispersion parameters than for intercepts which were overall the least inflicted parameters in regard to impaired performance (*i.e.*, the biases on the intercepts were mostly smaller than the respective average standard errors). RMSE estimates also tended to be larger for the misspecified models. The coverage of the 95% CI was overall quite good for the 2PCMP model, with coverage estimates for the intercepts and slopes very close to the nominal level for the majority of items. For the dispersion

Table 2. Bias, RMSE and coverage of the 95% CIs for all five models of simulation study II

	Bias					RMSE					Coverage				
Param.	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
α_1	−0.001	0.014	−0.027	−0.027	–	0.025	0.022	0.037	0.037	–	.940	.896	.794	.864	–
α_2	0.003	−0.086	0.040	0.040	–	0.035	0.088	0.054	0.054	–	.978	.000	.518	.742	–
α_3	0.002	0.094	−0.008	−0.008	–	0.026	0.096	0.027	0.027	–	.966	.000	.930	.958	–
α_4	0.000	−0.068	−0.041	−0.041	–	0.026	0.070	0.049	0.049	–	.962	.036	.602	.744	–
α_5	0.001	−0.067	0.047	0.047	–	0.032	0.069	0.056	0.056	–	.950	.036	.386	.634	–
α_6	0.001	0.066	−0.021	−0.021	–	0.024	0.068	0.031	0.031	–	.952	.022	.906	.930	–
δ_1	−0.001	−0.005	0.007	0.007	−0.005	0.027	0.027	0.028	0.028	0.027	.928	.948	.910	.956	.948
δ_2	−0.003	0.028	−0.017	−0.017	0.028	0.038	0.046	0.043	0.043	0.046	.976	.874	.716	.920	.874
δ_3	−0.002	−0.027	−0.000	−0.000	−0.027	0.028	0.038	0.028	0.028	0.038	.952	.882	.932	.950	.882
δ_4	−0.002	0.023	0.013	0.013	0.022	0.030	0.037	0.032	0.032	0.037	.956	.848	.884	.968	.848
δ_5	−0.003	0.019	−0.021	−0.021	0.019	0.035	0.039	0.041	0.041	0.039	.954	.912	.688	.894	.912
δ_6	−0.002	−0.020	0.004	0.004	−0.020	0.025	0.032	0.025	0.025	0.032	.954	.934	.952	.972	.930
$\log\nu_1$	0.018	0.048	–	–	0.046	0.115	0.123	–	–	0.121	.944	0940	–	–	.944
$\log\nu_2$	0.024	−0.089	–	–	−0.089	0.133	0.152	–	–	0.152	.934	.858	–	–	.858
$\log\nu_3$	0.015	0.021	–	–	0.019	0.106	0.116	–	–	0.116	.942	.944	–	–	.946
$\log\nu_4$	0.014	−0.229	–	–	−0.232	0.149	0.258	–	–	0.260	.942	.486	–	–	.472
$\log\nu_5$	0.012	−0.077	–	–	−0.077	0.119	0.135	–	–	0.135	.950	.908	–	–	.906
$\log\nu_6$	0.013	0.052	–	–	0.050	0.106	0.124	–	–	0.123	.940	.930	–	–	.932

Notes. If a model did not estimate a parameter, the respective cell is left empty.
1 = 2PCMP; 2 = 2PCMP with $\alpha_1 = \ldots = \alpha_6$; 3 = 2PCMP with $\nu_1 = \ldots = \nu_6 = 1$; 4 = 2PPCM; 5 = CMPCM; Param. = parameter.

Table 3. Evaluation of person parameter and reliability estimates in simulation study II

Model	$\text{med}\left(\text{Cor}\left(\theta, \hat{\theta}\right)\right)$	$\text{med}(\text{Rel})$	$\text{med}\left(\hat{\text{Rel}}\right)$
2PCMP	.921	.848	.826
2PCMP, $\alpha_1 = \ldots = \alpha_6$	.915	.838	.799
2PCMP, $\nu_1 = \ldots = \nu_6 = 1$	.891	.794	.805
PPCM	.891	.794	.805
CMPCM	.915	.838	.827

Note. Median correlations between the true and the estimated person parameters $\left(\text{med}\left(\text{Cor}\left(\theta, \hat{\theta}\right)\right)\right)$, the (model-implied) true reliability ($\text{med}(\text{Rel})$), and the (model-implied) empirical/estimated reliability ($\text{med}\left(\hat{\text{Rel}}\right)$) for all models in simulation study II.

Table 4. Median correlations between models' ability estimates in simulation study II

	1	2	3	4	5
2PCMP (1)	1.000	.993	.966	.966	.993
2PCMP, $\alpha_1 = \ldots = \alpha_6$ (2)		1.000	.954	.954	1.000
2PCMP, $\nu_1 = \ldots = \nu_6 = 1$ (3)			1.000	1.000	.954
2PPCM (4)				1.000	.954
CMPCM (5)					1.000

parameters, the nominal level was exactly met for the fifth item and slightly undercut for the other items, but still above 0.93 for all items. For the misspecified models, coverage overall tended to be lower, but was generally less impaired for the intercepts and log dispersions (but see more detailed descriptions and considerations in the Online Supplementary Materials). Coverage on the slope parameters in the misspecified models (if estimated) was in part very poor (in particular, for the 2PCMP with $\nu_1 = \ldots = \nu_6 = 1$). The pattern with regard to bias and standard errors offers a possible explanation for these results, as they occur in particular for item and model combinations where the bias is substantially larger than the average respective standard error.

Person parameter estimates. Table 3 shows that the highest median correlation between the true and the estimated person parameters was achieved by the (full) 2PCMP model, followed by the two versions of the CMPCM and the two versions of the 2PPCM, respectively. Note that due to the simulation design, the (model-implied) true reliability of the 2PCMP model constitutes the ground truth in this simulation study. The (median) model-implied true reliability is therefore already negatively biased for the misspecified models, more so for the 2PPCMs than for the CMPCMs. Further, the (full) 2PCMP and the two versions of the CMPCM slightly underestimated their respective model-implied true reliabilities in the median across trials. Different results for the two CMPCMs are likely due to the different estimation procedures. As expected, the two versions of the 2PPCM showed the same result for the median model-implied estimated reliability. They slightly overestimated their model-implied true reliability in the median across trials, but still underestimated the median reliability implied by the true underlying model.

Table 4 shows the median correlations (across trials) between the ability scores as produced by the five models. The pattern of results aligns with that seen in Table 3. Those models which are equivalent exhibited perfect correlations as one would expect. The correlations of the 2PCMP model ability estimates with those of the other models were

very high, especially between the one-parameter (*i.e.*, the two versions of the CMPCM) and the two-parameter 2PCMP models. This pattern is also found for other comparison of one- and two-parameter models (see, for example, Bürkner, 2020; Loken & Rulison, 2010) and will be discussed in Section 6.

5. Application example

For an empirical application example of the 2PCMP model, I reanalysed divergent thinking fluency tasks data as published in Silvia (2008a, 2008b) and Silvia et al. (2008) and made available by Silvia *via* OSF (https://osf.io/8vrck/) together with permission for the reanalysis (Silvia, 2013). Myszkowski and Storme (2021) recently reanalysed the same data using, among other models, the 2PPCM. They also assessed whether the equidispersion assumption was justified and found evidence to the contrary for the 2PPCM. This makes this data set particularly interesting for reanalysis with the 2PCMP model which loosens the equidispersion assumption of the 2PPCM.

For a detailed description of the data set, see Silvia et al. (2008). In short, the data set contains response data from $N = 242$ college students on $M = 6$ items. The items were divergent thinking fluency tasks which instruct participants to provide as many creative responses as possible to a prompt. Three different types of tasks were employed. They were alternate use uses tasks (AUT), where participants name alternate uses for everyday objects (a brick in item 1 and a knife in item 4), instances tasks, where participants are asked to name instances of a more general class (round things in item 2 and things that make noise in item 5), and consequences tasks, where participants list consequences of an event (no more sleep in item 3 and 12 inches height in item 6). The items were administered with a time limit of 3 min per item. Tasks like this can be scored in different ways to assess different underlying abilities (Silvia et al., 2008). For the 2PCMP model, I simply computed the number of responses given by each participant to each item. This is in line with the data preparation performed by Myszkowski and Storme (2021) and is considered to measure fluency.

I fitted the 2PCMP model to the data using 121 quadrature nodes (see the OSF repository for the R code). The model converged after 15 iterations. The parameter estimates are presented in Table 5. The model estimated the reliability at .821 (see Section 4.2 for how the reliability is estimated from the 2PCMP model). The slope parameters α_j (which are equal to item discriminations a_j) represent how well differences in latent ability (*i.e.*, divergent thinking fluency) are depicted by differences in responses.[2] Item 2 (an instances task) displayed the highest discrimination, indicating the best ability to differentiate between participants in terms of their divergent thinking fluency. Items 5 (also an instances task) and 4 (AUT) followed in terms of their discriminatory ability. The other AUT (item 1) was slightly less discriminatory. The two consequences tasks (items 3 and 6) were least well able to differentiate between participants in terms of their divergent thinking fluency. This pattern is visualized in Figure 1 which depicts the item response functions. The better the discrimination of an item, the steeper the item response curve – implying larger differences in expected responses (y-axis) for different latent ability (x-axis). Difficulties (d_j) can be obtained from slopes (α_j) and intercepts (δ_j) as $d_j = -\delta_j/\alpha_j$.

[2] Note that with a latent variance fixed at 1 (as is the case here for identification purposes), due to the exponential response function in the 2PCMP model, one would not necessarily expect discrimination values close to or even larger than 1. This would imply quite large expected counts for higher latent abilities quite quickly. Of course, whether this is sensible depends on the type of data at hand.

Table 5. Parameter estimates of the 2PCMP model for six divergent thinking items (application example)

Item	Parameter	Estimate	*SE*	95% CI
1	Slope	0.296	0.024	[0.249, 0.344]
	Intercept	1.930	0.027	[1.877, 1.984]
	Log dispersion	0.548	0.114	[0.324, 0.772]
2	Slope	0.396	0.035	[0.327, 0.466]
	Intercept	2.116	0.039	[2.040, 2.193]
	Log dispersion	−0.531	0.121	[−0.768, −0.295]
3	Slope	0.216	0.027	[0.163, 0.269]
	Intercept	1.879	0.028	[1.825, 1.933]
	Log dispersion	0.148	0.102	[−0.052, 0.347]
4	Slope	0.378	0.026	[0.327, 0.429]
	Intercept	1.871	0.030	[1.812, 1.930]
	Log dispersion	0.863	0.152	[0.564, 1.162]
5	Slope	0.377	0.033	[0.312, 0.442]
	Intercept	2.347	0.037	[2.276, 2.419]
	Log dispersion	−0.596	0.120	[−0.830, −0.361]
6	Slope	0.244	0.024	[0.197, 0.292]
	Intercept	1.846	0.026	[1.796, 1.897]
	Log dispersion	0.515	0.106	[0.308, 0.722]

The item with the largest difficulty in absolute value is the most difficult, which in this case are the consequences items (item 3 with $d_3 = -8.713$ and item 6 with $d_6 = -7.560$). They are followed by item 1 (AUT, $d_1 = -6.513$) and item 5 (instances, $d_5 = -6.227$), and then item 2 (instances, $d_2 = -5.345$). Item 4 (AUT) was the easiest, with $d_4 = -4.947$. The log dispersion parameters indicate how much responses are expected to vary, given a certain latent ability (*i.e.*, due to randomness). Looking at Figure 1, that would mean how much one expects responses for one given person (with one value on the *x*-axis) to vary from the expected response based on item difficulty and discrimination as shown by the item response curves. Here, items 2 and 5 (instances tasks) were the most dispersed (they were the only two items with overdispersion). The least dispersed (implying responses conditional on latent ability varied least around the expected response) were items 1 and 4 (AUT) which exhibited underdispersion. Items 3 and 6 (consequences tasks) fell in the middle in terms of dispersion (for item 3, equidispersion cannot be rejected). These results can inform researchers' item selection. It is not uncommon to only use one type of task to measure divergent thinking (*e.g.*, only AUT) in a study (*e.g.*, Beisemann, Forthmann, Bürkner, & Holling, 2020). Analyses of different divergent thinking items with the 2PCMP model can indicate which items are best at discriminating between divergent thinking abilities. They can also help to further psychometric understanding of these different items which were not constructed in an IRT framework.

Within the 2PCMP model, it is easy to test the assumptions of the established models contained within the 2PCMP model as special cases – the 2PPCM and the CMPCM. Starting with the 2PPCM, I fitted a 2PCMP model with the constraint that $\nu_1 = \ldots = \nu_M = 1$. Comparing the two models with a likelihood ratio test (*i.e.*, testing the equidispersion assumption of the 2PPCM), I found evidence of a significantly better fit of the (full) 2PCMP model, $\chi^2(6) = 87.903$, $p < .001$. This result is also reflected by the 95% CI for the log

dispersions in Table 5. Based on the marginal log likelihood which is evaluated in each iteration of the EM algorithm, the test statistic for the likelihood ratio test is $-2(LL_{m0}-LL_{m1})$ (with LL_{m0} as the marginal likelihood of the constrained model and LL_{m1} as the marginal likelihood of the unconstrained model, both at convergence). This test statistic is approximately χ^2 distributed with as many degrees of freedom as we have constrained parameters. Testing the assumptions of the CMPCM, I also fitted a 2PCMP model with the constraints that $\alpha_1 = \ldots = \alpha_M$. The comparison *via* the likelihood ratio test (*i.e.*, testing the assumption of equal slopes of the CMPCM) indicated significantly better fit of the 2PCMP model, $\chi^2(5) = 43.550$, $p < .001$. Note that we here have five constrained parameters, as one slope parameter is estimated for all six items. For both the 2PPCM and the CMPCM, the respective assumptions were violated for this data set, requiring the model complexity offered by the 2PCMP model.

6. Discussion

The present work introduces the 2PCMP model, a two-parameter count IRT model. The model allows item discriminations to be varied, which can help researchers with item selection. With the use of the mean parameterized CMP distribution (Huang, 2017), the model can account and test for over-, under- and equidispersion at an item-specific level. The model constitutes a generalization of the recently introduced CMPCM (Forthmann, Gühne, et al., 2020) as well as the 2PPCM (Myszkowski & Storme, 2021), both of which extend the RPCM (Rasch, 1960). All three of these models are contained within the 2PCMP model as special cases, so that the 2PCMP model offers an easy approach of testing (and if necessary loosening) their respective assumptions. Since, to the best of my knowledge, no estimation methods for the 2PCMP model were previously available (Forthmann, Gühne, et al., 2020), I derived an MML estimation method based on the EM algorithm (Dempster et al., 1977) for the 2PCMP model. Simulation studies showed promising performance of the 2PCMP model. The empirical example illustrated how easily the assumptions of the CMPCM and the 2PPCM can be tested within the 2PCMP model, and that this constitutes a realistic concern.

6.1. Evaluation of the 2PCMP model and recommendations

The simulation study results revealed overall satisfactory performance in terms of parameter recovery and reliability in a number of different settings varying with regard to the number of items, the type of underlying item-specific dispersion, the sample size, the number of quadrature nodes, and under realistic parameter values. Based on the results, I would recommend larger sample sizes than $N = 100$ for the 2PCMP model and administration of more than four items, especially if one is interested in very accurate estimates of the dispersion parameters. Not surprisingly, a greater number of items also results in better, and in fact quite good, estimates of model-implied reliability. These recommendations should minimize the risk of encountering numerical instabilities, which were overall relatively rare and in practice might be addressed by varying the starting values slightly. Numerical instabilities may likely be caused by certain parameter constellations, especially in terms of slopes and log dispersions, when both tend to larger (absolute) values. A second simulation study comparing the 2PCMP in a realistic data setting to the CMPCM and the 2PPCM showed that the use of the 2PCMP model is beneficial in a setting where the assumptions of established methods are violated. This is

true for parameter estimation accuracy, but in particular in terms of coverage of the 95% confidence intervals which in some cases falls drastically below the nominal level in the misspecified models, especially for the slope parameters.

In terms of the ability parameter and reliability estimation, one could also see an (albeit only slight) advantage of the 2PCMP model. Ability point estimates for the compared models were strongly correlated, in particular between the one- and two-parameter version (CMPCM and 2PCMP). This is a common pattern also found in comparisons of the one-parameter logistic (1PL) and two-parameter logistic (2PL) models for binary data (see, for example, Bürkner, 2020; Loken & Rulison, 2010). While point estimates tend to be very similar even if the 2PL model holds and the 1PL is violated, the differences between one- and two-parameter models are still reflected elsewhere, for example in the standard errors and the reliability estimates. As the 2PL model can be considered a border case of the 2PCMP model (the binomial distribution is a border case of the CMP distribution), it is unsurprising to observe similar results for the 2PCMP model. For the setting in the second simulation study, no numerical instabilities were observed. The comparison of computation times showed that the EM algorithm for the 2PCMP model is not only competitive compared to other software, but even showed faster computation time on average for the CMPCM than glmmTMB (Eddelbuettel et al., 2011) (which, however, is much more general software). The method employed for choosing starting values for the 2PCMP model proved advantageous in terms of average number of iterations.

6.2. Limitations

Notwithstanding promising results in terms of statistical properties from the simulation studies and in terms of numerical stability and relative computational efficiency of the proposed EM algorithm, the present work is also subject to certain limitations. The number of trials in the simulation studies was limited by the computation costs of fitting the 2PCMP model, so that only 250 or 500 simulation trials were run per scenario. For the item parameters' standard errors, only one method was used (based on a numerical approximation to Oakes's identity; Chalmers, 2018). Corresponding 95% confidence intervals were constructed using a Wald approximation. This may leave results for the coverage of the 95% confidence intervals confounded with the methods used for standard error and CI computation and does not allow any specific weaknesses of the methods to be deduced. Thus, this work cannot offer specific recommendations as to which methods to use for standard errors and CIs. Due to computation costs, only one method for person parameter estimation was evaluated, a Bayes EAP estimator (see Appendix B for an alternative ML method). The comparison of the 2PCMP model with established methods was focused on models which are special cases of the 2PCMP model and on a setting in which the assumptions of the established methods were violated. Thus, the comparison is unable to offer insights about comparative performance of other count IRT models (*e.g.*, for overdispersion, the negative binomial model; Hung, 2012) or about the compared models' performance in different types of settings. As only one set of parameter values was used in the second simulation study, the strength of the violation of assumptions of the established methods was not systematically varied.

6.3. Avenues for future research

With the 2PCMP model, future research can analyse count-data-generating psychometric tasks and self-report items with regard to their discriminatory power, difficulty, and

measurement precision. Such investigations can help inform item selection. Future research could also address some of the limitations of the present work. The 2PCMP model could be compared to other existing models – for example, for overdispersed count data, the IRT-ZIP (Wang, 2010) or the NBRM (Hung, 2012) – under different conditions. A model comparison *via* information criteria such as Akaike's might be helpful to this end; best fit for the 2PCMP model among models examined would provide strong validation for the 2PCMP model. In the future, different methods for standard error as well as confidence interval computation could be compared to allow for recommendations of the best methods for the 2PCMP model. The performance of other person parameter methods (such as ML; see Appendix B, but note computational cost) could be examined and compared to the Bayes EAP method used in this work. Computation time efficiency for the 2PCMP model EM algorithm could be further improved with the use of EM accelerators (for a recent review of available state-of-the-art methods, see Beisemann, Wartlick, & Doebler, 2020, who also compared the methods for binary IRT models). This could help to make more simulation trials feasible in future simulation studies to reduce Monte Carlo standard errors. The derived MML estimation technique for the 2PCMP model is based on a fixed Gauss–Hermite quadrature EM algorithm. Other EM variants such as adaptive Gauss–Hermite quadrature EM (see Schilling & Bock, 2005, for the binary case) could be explored. In general, other estimation techniques might be investigated, such as a Bayesian estimation approach which might be particularly helpful for smaller sample sizes. An extension of the 2PCMP model to include an offset would allow for modelling time limits imposed for the items which is not unusual for psychometric tests generating count data (*e.g.*, in Silvia et al., 2008, a time limit of 3 min per item was used). The 2PCMP model itself might be extended, for example to a multidimensional 2PCMP model or to allow for the inclusion of covariates. For instance, by including item covariates on dispersion parameters, researchers could investigate sources of under- and overdispersion. More complex extensions could include options to model multilevel count data or more complex factorial designs.

Acknowledgements

This work was supported by funding of the DFG (DO 1789/7-1) granted to Prof. Dr. Philipp Doebler. I would like to thank Prof. Dr. Philipp Doebler for his valuable suggestions and feedback on earlier versions of this manuscript. I would further like to thank Paul Silvia and his co-authors for making their data available for researchers to re-analyze. Open Access funding enabled and organized by Projekt DEAL.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Marie Beisemann: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Data availability statement

Results from the simulation studies conducted in this work are openly available in an OSF repository at http://doi.org/10.17605/OSF.IO/HX5JS. An anonymous link for peer review has been created at https://osf.io/hx5js/?viewonly=7a53dd7cb1fa4bb593e3c49504c6a10a. Data used for the application example in this work are available in an OSF repository at https://osf.io/8vrck/. These data were collected for and published in Silvia (2008a, 2008b), Silvia et al. (2008). They were made publicly available for re-analysis by the authors via OSF.

References

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. https://doi.org/10.1177/0033294118797577

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. https://doi.org/10.1177/0031512518812183

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

Beisemann, M., Forthmann, B., Bürkner, P.-C., & Holling, H. (2020). Psychometric evaluation of an alternate scoring for the remote associates test. *The Journal of Creative Behavior*, *54*(4), 751–766. https://doi.org/10.1002/jocb.394

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

Blocker, A. W. (2018). *fastghquad: Fast 'rcpp' implementation of gauss-hermite quadrature* [Computer software manual] (R package version 1.0). Retrieved from https://CRAN.R-project.org/package=fastGHQuad

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

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., . . . Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Bürkner, P.-C. (2020). Analysing standard progressive matrices (spm-ls) with Bayesian item response models. *Journal of Intelligence*, *8*(1), 5. https://doi.org/10.3390/jintelligence8010005

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

Chalmers, R. P. (2018). Numerical approximation of the observed information matrix with oakes' identity. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 415–436. https://doi.org/10.1111/bmsp.12127

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *xtable: Export tables to latex or html* [Computer software manual] (R package version 1.8–4). Retrieved from https://CRAN.R-project.org/package=xtable

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.lindif.2015.01.013

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. https://doi.org/10.18637/jss.v040.i08

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Hove: Psychology Press.

Faddy, M., & Bosch, R. (2001). Likelihood-based modeling and analysis of data underdispersed relative to the poisson distribution. *Biometrics*, *57*(2), 620–624. https://doi.org/10.1111/j.0006-341X.2001.00620.x

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., & Doebler, P. (2021). Reliability of researcher capacity estimates and count data dispersion: A comparison of poisson, negative binomial, and Conway-Maxwell-poisson models. *Scientometrics*, *126*(4), 3337–3354. https://doi.org/10.1007/s11192-021-03864-8

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-tests. *Journal of Psychoeducational Assessment*, *38*(6), 692–705. https://doi.org/10.1177/0734282919889262

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*, 32–50. https://doi.org/10.1111/bmsp.12184

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*(3), 257–269. https://doi.org/10.1080/10400419.2017.1360059

Francois, R., Eddelbuettel, D., & Eddelbuettel, M. D. (2010). *Package rcppgsl*. R [Computer software manual] (R package version 0.3.8). Retrieved from https://CRAN.R-project.org/package=RcppGSL

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., . . . Rossi, F. (2010). *GNU Scientific Library Reference Manual* (3rd ed., pp. 103–180).

Gaujoux, R. (2020). *doRNG: Generic reproducible parallel backend for 'foreach' loops* [Computer software manual] (R package version 1.8.2). Retrieved from https://CRAN.R-project.org/package=doRNG

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Hasselman, B. (2018). *nleqslv: Solve systems of nonlinear equations* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=nleqslv (R package version 3.3.2)

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, *17*(6), 359–380. https://doi.org/10.1177/1471082X17697749

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

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. https://doi.org/10.1177/014662169501900307

Jansen, M. G. (1997). Rasch's model for reading speed with manifest explanatory variables. *Psychometrika*, *62*(3), 393–409. https://doi.org/10.1007/BF02294558

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

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). Tmb: Automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*. Retrived from https://doi.org/10.48550/arXiv.1509.00660

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. https://doi.org/10.1348/000711009X474502

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. https://doi.org/10.3102/1076998617694878

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

Microsoft Corporation. & Weston, S. (2020). *doParallel: Foreach parallel adaptor for the 'parallel' package* [Computer software manual] (R package version 1.0.16). Retrieved from https://CRAN.R-project.org/package=doParallel

Muthén, L., & Muthén, B. (1998, 2010). *Mplus user?s guide*. Los Angeles, CA: Muthén & Muthén.

Mutz, R., & Daniel, H.-D. (2018). The bibliometric quotient (bq), or how to measure a researcher?S performance capacity: A Bayesian Poisson Rasch model. *Journal of Informetrics*, *12*(4), 1282–1295. https://doi.org/10.1016/j.joi.2018.10.006

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. https://doi.org/10.1002/jocb.490

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 479–482. https://doi.org/10.1111/1467-9868.00188

Ogasawara, H. (1996). Rasch's multiplicative Poisson model with covariates. *Psychometrika*, *61*(1), 73–92. https://doi.org/10.1007/BF02296959

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. https://doi.org/10.1080/09644000903055799

R Core Team. (2021). *R: A language and environment for statistical computing [computer software manual]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from. https://www.R-project.org/

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

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*(3), 533–555. https://doi.org/10.1007/s11336-003-1141-x

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. https://doi.org/10.1111/j.1467-9876.2005.00474.x

Silvia, P. J. (2008a). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual Differences*, *44*(4), 1012–1021. https://doi.org/10.1016/j.paid.2007.10.027

Silvia, P. J. (2008b). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139. https://doi.org/10.1037/1931-3896.2.3.139

Silvia, P. J. (2013, Nov). *Assessing creativity with divergent thinking tasks (silvia et al., 2008, study 2, psychology of aesthetics, creativity, and the arts)*. OSF. Retrieved from https://osf.io/8vrck/

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., . . . Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68. https://doi.org/10.1037/1931-3896.2.2.68

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York, NY: Chapman and Hall/CRC.

Verhelst, N., & Kamphuis, F. (2009). A Poisson-gamma model for speed tests. *Measurement and Research Department Reports*, *2*, 2010–2011.

Wang, L. (2010). IRT–ZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, *35*(6), 671–692. https://doi.org/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. https://doi.org/10.1016/S0047-259X(03)00020-4

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

Wickham, H. (2021). *tidyr: Tidy messy data* [Computer software manual] (R package version 1.1.3). Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Francois, R., Henry, L., & Müller, K. (2021). *dplyr:A grammar of data manipulation* [Computer software manual] (R package version 1.0.5). Retrieved from https://CRAN.R-project.org/package=dplyr

Received 29 September 2021; revised version received 5 April 2022

Supporting Information

The following supporting information may be found in the online edition of the article:

Supplementary Materials

Appendix A:

Derivation of the EM algorithm for the 2PCMP model

To derive the complete-data log likelihood for the 2PCMP model, the complete data are chosen to be $(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{x}$ are the responses and $\boldsymbol{\theta}$ are the latent abilities. To find the corresponding likelihood, assume that each latent ability θ_i can be divided up into a finite set of K discrete categories, denoted by $q_k, k = 1, \ldots K$, yielding the discrete variable $\boldsymbol{Q} = (Q_1, \ldots Q_N)^T$. With $\mathbb{I}_{\{.\}}$ as the indicator function, let $f_k = \sum_{i=1}^{N} \mathbb{I}_{\{Q_i = q_k\}}$ $(k = 1, \ldots K)$ denote the number of participants with discrete latent ability of level q_k in our sample of N participants. Note that $\sum_{k=1}^{K} f_k = N$. Denote by $\mathbf{f} = (f_1, \ldots, f_K)^T$ the vector containing the number of participants in each of the K latent ability categories. Under the assumption that the N discrete latent abilities (*i.e.*, the N participants) are sampled pairwise independently, one can assume a multinomial distribution for the discrete latent abilities, with probabilities $w_1, \ldots, w_k$ for each of the K categories, as given in the following. Thus, the probability of $\mathbf{f}$ is given by

$$P(\mathbf{f}; w_1, \ldots, w_K) = \left[\frac{N!}{f_1! \ldots f_K!}\right] \prod_{k=1}^{K} w_k^{f_k}. \tag{17}$$

This same assumption is also made in the derivation of the EM algorithm for other IRT models, for example for binary data (Baker & Kim, 2004). The notation in this section is deliberately similar to that used in Baker and Kim (2004) to highlight similarities and differences. For better readability, define $F_k = \{\forall i : Q_i = q_k\}$ as the set of person indices where the persons have latent ability level q_k. Note that each set F_k has f_k elements. Let r^*_{ijk} $(j = 1, \ldots, M, k = 1, \ldots, K)$ denote the response given by a person i of discrete latent ability q_k to item j, that is, $r^*_{ijk} = \mathbb{I}_{i \in F_k} x_{ij}$. For an arbitrary but fixed ability level q_k $(k \in \{1, \ldots, K\})$, write $\mathbf{r}^*_\mathbf{k}$ to denote the response vector $\left(r^*_{11k}, \ldots, r^*_{f_k M k}\right)^T$ of all persons $i \in F_k$ answering M items. Then the probability of observing $\boldsymbol{r}^*_\boldsymbol{k}$ under the 2PCMP model is given by

$$P\left(\boldsymbol{r}^*_\mathbf{k}; \boldsymbol{\zeta}, q_k\right) = \prod_{j=1}^{M} \prod_{i \in F_k} \left(\frac{\lambda\left(\mu_{jk}, \nu_j\right)^{r^*_{ijk}}}{\left(r^*_{ijk}!\right)^{\nu_j}} \frac{1}{Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)} \right) \tag{18}$$

$$= \prod_{j=1}^{M} \frac{\lambda\left(\mu_{jk}, \nu_j\right)^{\sum_{i \in F_k} r^*_{ijk}}}{\exp\left(\nu_j \sum_{i \in F_k} \log\left(r^*_{ijk}!\right)\right)} \frac{1}{Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)^{f_k}}. \tag{19}$$

Define $r_{jk} := \sum_{i \in F_k} r^*_{ijk} = \sum_{i=1}^{N} \mathbb{I}_{i \in F_k} r^*_{ijk}$ (*i.e.*, the sum of the responses of all f_k participants with ability level q_k on item j) and $h_{jk} := \sum_{i \in F_k} \log\left(r^*_{ijk}!\right) = \sum_{i=1}^{N} \mathbb{I}_{i \in F_k} \log\left(r^*_{ijk}!\right)$, and obtain

$$P\left(\boldsymbol{r}^*_\mathbf{k}; \boldsymbol{\zeta}, q_k\right) = \prod_{j=1}^{M} \frac{\lambda\left(\mu_{jk}, \nu_j\right)^{r_{jk}}}{\exp\left(\nu_j h_{jk}\right)} \frac{1}{Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)^{f_k}}. \tag{20}$$

Denote the vector $\left(r^*_{111}, \ldots, r^*_{f_K M K}\right)^T$ of all responses by $\mathbf{r}^*$. The probability of observing $\mathbf{r}^*$ is given by $\prod_{k=1}^{K} P\left(\mathbf{r}^*_\mathbf{k}; \boldsymbol{\zeta}, q_k\right)$. Consequently, the joint probability of $\mathbf{f}$ and $\mathbf{r}^*$, that is, the complete-data likelihood L_c, is given by

$$L_c = P(\boldsymbol{f}, \boldsymbol{r}^*; \boldsymbol{\zeta}) = \left(\prod_{k=1}^{K} \prod_{j=1}^{M} \frac{\lambda\left(\mu_{jk}, \nu_j\right)^{r_{jk}}}{\exp\left(\nu_j h_{jk}\right)} \frac{1}{Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)^{f_k}} \right) \left(\left[\frac{N!}{f_1! \ldots f_K!} \right] \prod_{k=1}^{K} w_k^{f_k} \right). \tag{21}$$

From the factorization of the likelihood, one can see that f_k, r_{jk}, and h_{jk}, for all $j \in \{1, \ldots, M\}$, for all $k \in \{1, \ldots, K\}$, constitute sufficient statistics for the complete data under the 2PCMP model. Taking the logarithm and omitting constants,

$$\begin{aligned}
&\log P(\boldsymbol{f}, \boldsymbol{r}^*; \boldsymbol{\zeta}) = LL_c \\
&\propto \left(\sum_{k=1}^{K} \sum_{j=1}^{M} r_{jk} \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \nu_j h_{jk} - f_k \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right)\right) \\
&\quad + \left(\sum_{k=1}^{K} f_k \log(w_k) \right) \\
&\propto \sum_{k=1}^{K} \sum_{j=1}^{M} r_{jk} \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \nu_j h_{jk} - f_k \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right).
\end{aligned}$$

The right summand which is omitted above from the second to the third line does not depend on $\boldsymbol{\zeta}$ and thus will not influence the optimization in terms of $\boldsymbol{\zeta}$. As this is what the log likelihood is used for here, any terms not dependent on $\boldsymbol{\zeta}$ (*i.e.*, which do not have an index j) can be ignored. Take the expectation over $\mathbf{Q}$ given the observed data $\mathbf{x}$ and $\boldsymbol{\zeta}'$. The expected complete-data log likelihood is proportional to (and equal to save for constant terms)

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}(LL_c) = \mathbb{E}(LL_c) \\
&\propto \sum_{k=1}^{K} \sum_{j=1}^{M} \mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}\left(r_{jk}\right) \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \nu_j \mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}\left(h_{jk}\right) \\
&\quad - \mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}(f_k) \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right) =: ELL_{\mathbf{C}}.
\end{aligned} \tag{22}$$

With the posterior probability of node q_k, $P(q_k|\mathbf{x}_i, \boldsymbol{\zeta}')$, as defined in Equation (8), we have

$$\mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}(f_k) = \mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}\left(\sum_{i=1}^{N} \mathbb{I}_{\{Q_i = q_k\}}\right) \tag{23}$$

$$= \sum_{i=1}^{N} \mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}\left(\mathbb{I}_{\{Q_i = q_k\}}\right) \tag{24}$$

$$= \sum_{i=1}^{N} P(q_k \mid \mathbf{x}_i, \boldsymbol{\zeta}') =: \overline{\mathrm{f}}_{\mathrm{k}}, \tag{25}$$

for all $k \in \{1, \ldots, K\}$. With analogous operations and using the definitions of $\mathrm{r_{jk}}$ and $\mathrm{h_{jk}}$ one obtains

$$\mathbb{E}_{\mathbf{Q}|\mathbf{x},\boldsymbol{\zeta}'}\left(r_{jk}\right) = \sum_{i=1}^{N} x_{ij} P(q_k|\mathbf{x}_i, \boldsymbol{\zeta}') =: \overline{\mathrm{r}}_{jk} \tag{26}$$

and

$$\mathbb{E}_{\mathbf{Q}|\mathbf{x},\zeta'}(h_{jk}) = \sum_{i=1}^{N} \log(x_{ij}!) P(q_k \mid \mathbf{x}_i, \zeta') =: \overline{\mathsf{h}}_{jk} \tag{27}$$

for all $k \in \{1, \ldots, K\}$, for all $j \in \{1, \ldots, M\}$. Using these and Equation (22) for the E-step (as one can and as EM algorithms for, for example, logistic IRT models typically do with analogous equations; see Baker & Kim, 2004), results in gradients with numerically challenging terms for the M-step (compare Equations (29–31), with, in particular, challenging terms in the gradients for the dispersion parameters, see Equations (32–34)). To alleviate this problem, I substitute the definitions of $\overline{f}_k$, $\overline{r}_{jk}$, and $\overline{h}_{jk}$ into Equation (22) and rearrange in the search for a more compact formulation of the expected complete-data log likelihood (and especially more compact expressions of the resulting gradients for the M-step). It is easy to show that this expression can be rearranged into Equation (9):

$$\begin{aligned} ELL_c &= \sum_{k=1}^{K}\sum_{j=1}^{M} \overline{r}_{jk} \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \nu_j \overline{h}_{jk} - \overline{f}_k \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{M} \left(\sum_{i=1}^{N} x_{ij} P(Q_k \mid \mathbf{x}_i, \zeta')\right) \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) \\ &\quad - \nu_j \left(\sum_{i=1}^{N} \log(x_{ij}) P(Q_k \mid \mathbf{x}_i, \zeta')\right) \\ &\quad - \left(\sum_{i=1}^{N} P(Q_k \mid \mathbf{x}_i, \zeta')\right) \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M} \left[\left(x_{ij} \log\left(\lambda\left(\mu_{jk}, \nu_j\right)\right) - \log(x_{ij})\nu_j - \log\left(Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)\right)\right) P(Q_k \mid \mathbf{x}_i, \zeta')\right], \end{aligned} \tag{28}$$

thereby showing that the EM algorithms based on Equations (22) and (9) are equivalent representations of the same algorithm which maximizes the same expected complete-data log likelihood in each M-step. In fact, Equation (9) is a simplification of Equation (2), giving the justification for Equation (9). The advantage of the substitution of $\overline{f}_k$, $\overline{r}_{jk}$, and $\overline{h}_{jk}$ in Equation (22) and subsequent rearrangement is – as mentioned above – that the resulting term yields much more compact representations of the derivatives (note that if one were to first take the derivatives of Equation (22) and then substitute the respective definitions for $\overline{f}_k$, $\overline{r}_{jk}$, and $\overline{h}_{jk}$, one should arrive at the same representations as Equations (22) and (9) are equivalent). To illustrate this point, I provide the derivatives of Equation (22) in terms of the item parameter without substituting $\overline{f}_k$, $\overline{r}_{jk}$, and $\overline{h}_{jk}$. They are

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \alpha_j} = \sum_{k=1}^{K} \frac{q_k \mu_{jk}}{V\left(\mu_{jk}, \nu_j\right)} \left(\overline{r}_{jk} - \mu_{jk} \overline{f}_{jk}\right), \tag{29}$$

for the α_j, for all $j \in \{1, \ldots, M\}$,

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \delta_j} = \sum_{k=1}^{K} \frac{\mu_{kj}}{V\left(\mu_{jk}, \nu_j\right)} \left(\overline{r}_{jk} - \mu_{jk} \overline{f}_{jk}\right) \tag{30}$$

for the δ_j, for all $j \in \{1, \ldots, M\}$, and

$$\frac{\partial \mathbb{E}(LL_c)}{\partial \log \nu_j} = \sum_{k=1}^{K} \nu_j \left(\frac{\overline{r}_{jk}}{W_{jk}} - \overline{h}_{jk} + \overline{f}_{jk} R_{jk} \right) \tag{31}$$

for the $\log \nu_j$, for all $j \in \{1, \ldots, M\}$, with

$$R_{jk} = \sum_{x=0}^{\infty} \frac{\lambda\left(\mu_{jk}, \nu_j\right)^x}{(x!)^{\nu_j} Z\left(\lambda\left(\mu_{jk}, \nu_j\right), \nu_j\right)} \left(\frac{x}{W_{jk}} - \ln(x!) \right) \tag{32}$$

and

$$W_{jk} = \sum_{x=0}^{\infty} \frac{\left(x - \mu_{jk}\right)^2 \lambda\left(\mu_{jk}, \nu_j\right)^x}{(x!)^{\nu_j} S_{jk}}, \tag{33}$$

where

$$S_{jk} = \sum_{x=0}^{\infty} \ln(x!) \frac{\left(x - \mu_{jk}\right) \lambda\left(\mu_{jk}, \nu_j\right)^x}{(x!)^{\nu_j}}. \tag{34}$$

One can immediately see, in particular, that the derivatives for the log dispersions contain more complicated terms than in the previous section. In any implementation, the series R_{jk}, S_{jk}, and W_{jk} need to be numerically approximated, adding potential sources of numerical instability.

Appendix B:

Maximum likelihood ability estimation

Assume the item parameters $\boldsymbol{\zeta}$ as known, and that the responses of N participants are pairwise independent and conditionally independent between items given the participant's latent ability. The probability of the response vector for a participant i, $i \in \{1, \ldots, N\}$ arbitrary but fixed, given their latent ability θ_i under the 2PCMP model is

$$P(\mathbf{x}_i | \theta_i, \boldsymbol{\zeta}) = \prod_{j=1}^{M} \mathrm{CMP}_{\mu}\left(x_{ij}; \mu_{ij}, \nu_j\right). \tag{35}$$

As one assumes one participant's responses independent of other participants' responses, ML estimates of their ability may be found for one person at a time. To obtain the ML estimate of person i ($i \in \{1, \ldots, N\}$), one takes the logarithm of Equation (35) and iteratively optimizes the result with respect to the participant's ability θ_i. To this end, the first derivative of the logarithm of Equation (35), which is given by

$$\frac{\partial \log P(\mathbf{x}_i | \theta_i, \boldsymbol{\zeta})}{\partial \theta_i} = \sum_{j=1}^{M} \frac{\partial \log \mathrm{CMP}_{\mu}\left(x_{ij}; \mu_{ij}, \nu_j\right)}{\partial \theta_i} = \sum_{j=1}^{M} \frac{x_{ij} \alpha_j \mu_{ij}}{V\left(\mu_{ij}, \nu_j\right)} - \frac{\alpha_j \mu_{ij}^2}{\lambda\left(\mu_{ij}, \nu_j\right)}, \tag{36}$$

is set equal to 0 and then one iteratively solves for θ_i, for arbitrary but constant $i \in \{1, \ldots, N\}$. To this end, Newton–Raphson type methods or similar alternatives can be employed. These methods usually require second derivatives, which, if not provided analytically, are approximated numerically. In either case, the estimation is carried out separately for each person, leading to a large number of evaluations of the gradient in Equation (36) which may quickly lead to long computation times.

Appendix C:

Details of the computational implementation

2PCMP model EM algorithm

I generated grids for using $\lambda(\mu, \nu)$, $Z(\lambda(\mu, \nu), \nu)$, and $V(\lambda(\mu, \nu), \nu)$ using TMB Kristensen, Nielsen, Berg, Skaug, & Bell, 2015) *via* code I modified from glmmTMB (Brooks et al., 2017). I used the GSL library (Galassi et al., 2014) from C++ to interpolate values from the grid using two-dimensional bicubic interpolation, tied into the R code with the help of RcppGSL (Francois, Eddelbuettel, & Eddelbuettel, 2010). I still numerically approximate other infinite series (A and B from Equation (13)) in C++ using the same method as Kristensen et al. (2015), where I start evaluating the series at its mode and add increments in either direction of the mode until the absolute increments fall below a very small value $\varepsilon \in \mathbb{R}, \varepsilon > 0$.

I chose starting values for the α and δ parameters in the 2PCMP model by fitting a 2PPCM to the data. For the 2PPCM, I used part-whole corrected correlations to determine starting values for the slope parameters and logarithms of the item means for the intercepts. For the starting values for the log dispersions of the 2PCMP model, I use the starting values of the slopes and intercepts to generate a number of observations under the 2PPCM (with 1,000 as the default). The logarithms of item-specific ratios of the variance of the simulated responses to the variance of the observed responses are used as starting values for the log dispersions.

The fixed Gauss–Hermite quadrature was in part implemented with the help of the R package fastGHQuad (Blocker, 2018), that is, fastGHQuad was used to generate the quadrature nodes and weights. Weights were then adjusted to be appropriate for the standard normal distribution, and sums over the quadrature nodes were implemented in C++. In simulation study I, I investigated what number of nodes would be a good recommendation. Prior trial simulations had already shown that it is strongly recommended to use at least 100 quadrature nodes to achieve satisfactory accuracy in parameter estimation. The iterative root finding of the gradients in each M-step is carried out with the Broyden method as implemented in the R package nleqslv (Hasselman, 2018).

Simulation studies

In both simulation studies, the 2PCMP model and constrained versions of it as well as the 2PPCM in simulation study II were fitted using the countirt package. In both simulation studies, ability parameters for the 2PCMP model were estimated with the Bayes EAP estimator for better computational efficiency for the simulations. Further R packages used were the glmmTMB package Brooks et al., 2017) to fit the CMPCM (Forthmann, Gühne, et al., 2020), the doParallel package (Microsoft Corporation & Weston, 2020) and the doRNG package (Gaujoux, 2020) to implement parallel computation of simulation trials,

the tidyr (Wickham, 2021) and the dplyr (Wickham, Francois, Henry, & Müller, 2021) packages to prepare the simulation results, and the ggplot2 (Wickham, 2016) as well as the xtable (Dahl, Scott, Roosen, Magnusson, & Swinton, 2019) packages to create the tables and figures.

Please note that a corrigendum to this article has been published since the article's initial publication, correcting a typographical error in Equation (36) in Appendix B:

(2024). Corrigendum. *British Journal of Mathematical and Statistical Psychology, (77)*, 237–237. https://doi.org/10.1111/bmsp.12312

Article 2

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. https://doi.org/10.1080/00273171.2023.2288577

Article 3

Beisemann, M., Holling, H., & Doebler, P. (2024). Every trait counts: Marginal maximum likelihood estimation for novel multidimensional count data item response models with rotation or ℓ_1–regularization for simple structure. *PsyArXiv pre-print, version 1*. https://doi.org/10.31234/osf.io/fqyjs

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};

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

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

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

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

$$P(q_{k_1}, \dots, q_{k_L} | \mathbf{x}_i, \boldsymbol{\zeta}') = \frac{\prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_{k_1}, \dots, q_{k_L}, \boldsymbol{\zeta}'_j) w_{k_1} \dots w_{k_L}}{\sum_{k'_1=1}^{K} \dots \sum_{k'_L=1}^{K} \prod_{j=1}^{M} \text{CMP}_\mu(x_{ij} | q_{k'_1}, \dots, q_{k'_L}, \boldsymbol{\zeta}'_j) w_{k'_1} \dots w_{k'_L}}, \tag{7}$$

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

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

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_μ 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

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

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,

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

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

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

[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, $\boldsymbol{\Sigma}_\theta$, fixed. The diagonal of $\boldsymbol{\Sigma}_\theta$ 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 $\boldsymbol{\Sigma}_\theta$ 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 $\boldsymbol{\Sigma}_\theta$ 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

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

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		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)
Method	Parameter														
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.)*

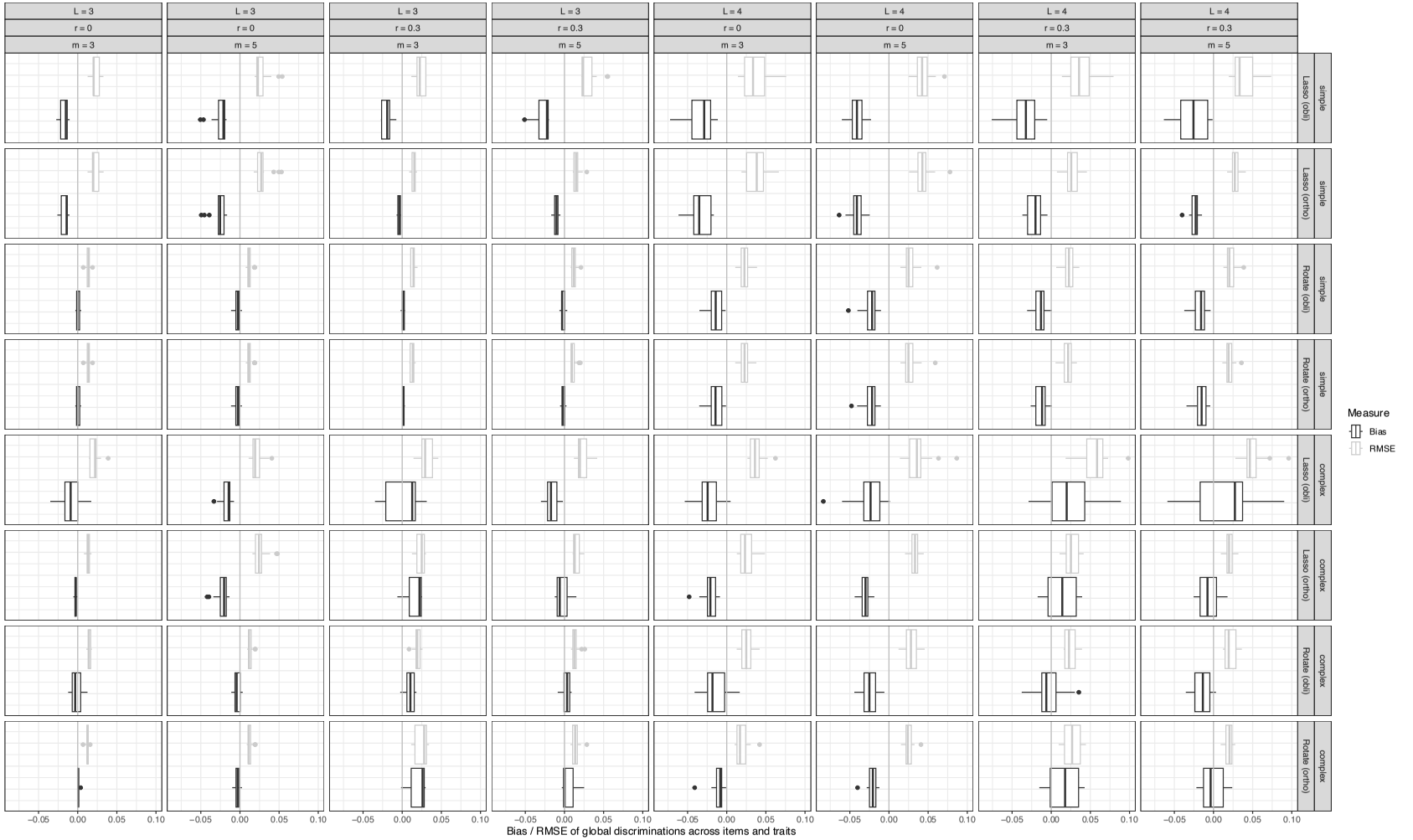
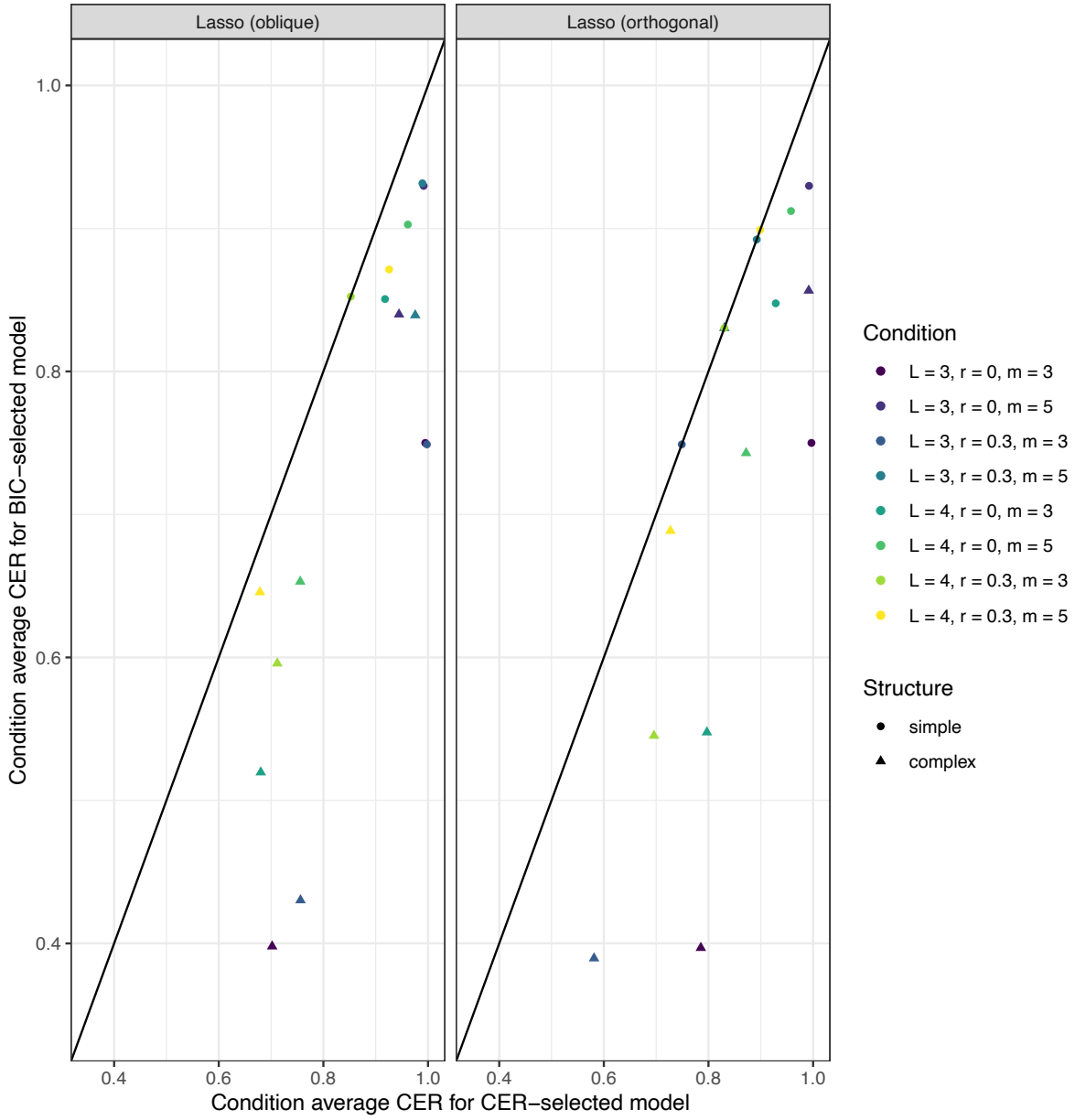


Figure 2

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,

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

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

and

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

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

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 $\mu^2_{jk_1,\ldots,k_L}$ 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.

Part III

Appendix