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)

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

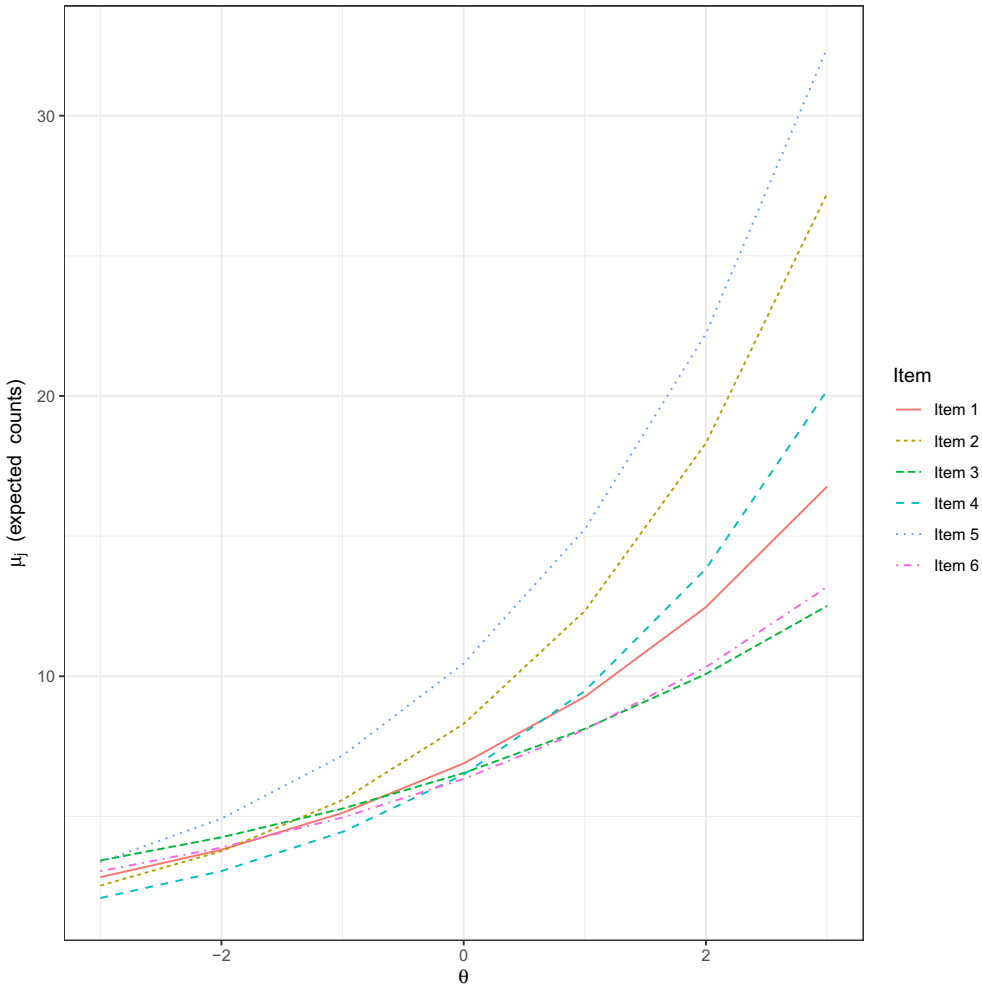

Figure 1. Item response functions (*i.e.*, plotting latent ability θ against the predicted counts μ_j for item *j*) of the 2PCMP model for six divergent thinking items (application example). Items are colour-coded and represented by different line types as indicated on the right-hand side.

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

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

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):

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

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

$\widehat{\text{Rel}} = 1 - \text{mean}\left(\widehat{\text{SE}}(\hat{\theta}_i)^2\right) / \widehat{\text{Var}}(\theta_i)$, where $\widehat{\text{SE}}(\hat{\theta}_i)$ 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 $\text{Rel} = \text{Cor}(\theta_i, \hat{\theta}_i)^2$ (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.

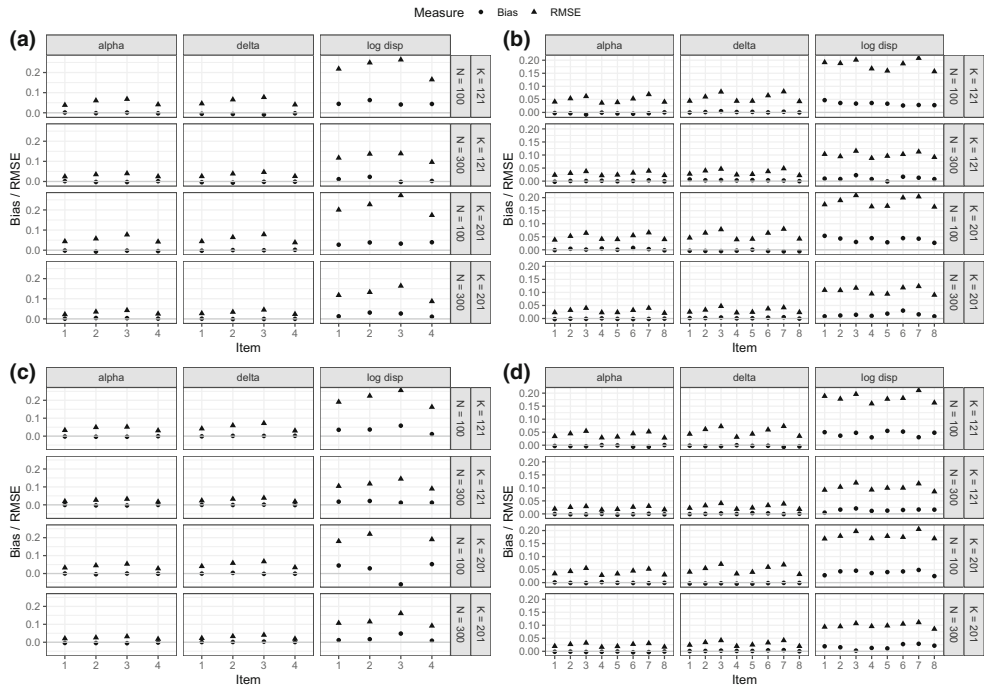

Figure 2. Bias (dots) and RMSE (triangles) for each parameter of each item for all conditions with equidispersed items ((a) four items, (b) eight items) and underdispersed items ((c) four items, (d) eight items). Each column within each plot shows the results for a different parameter (alpha = slope, 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

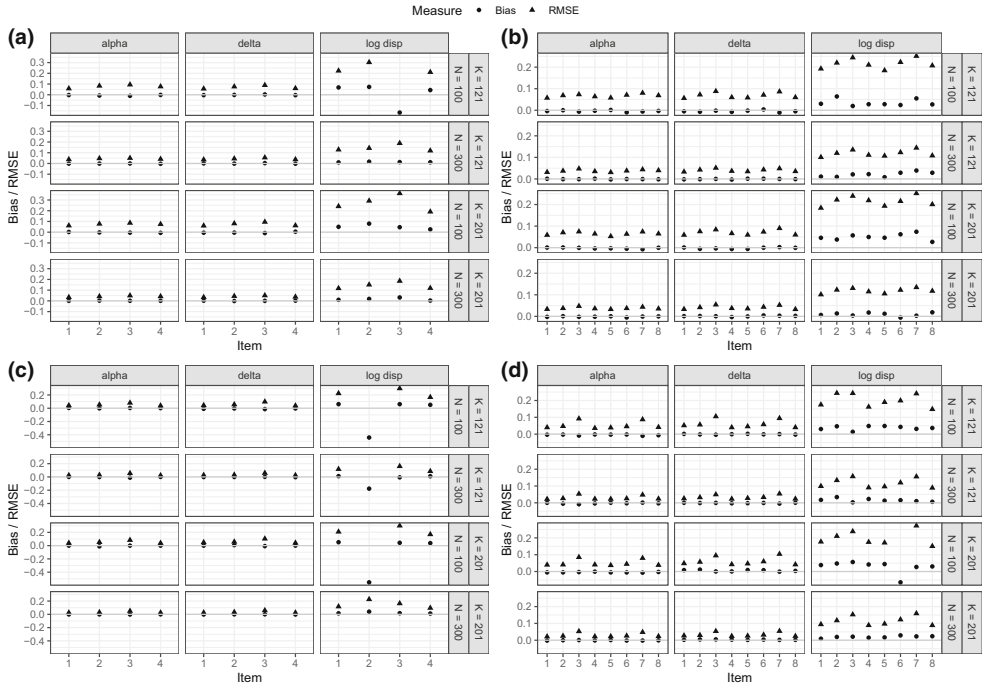

Figure 3. Bias (dots) and RMSE (triangles) for each parameter of each item for all conditions with overdispersed items ((a) four items, (b) eight items) and items with different types of dispersion ((c) four items, (d) eight items). Each column within each plot shows the results for a different parameter (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_{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	$\text{med}\left(\text{Cor}\left(\theta, \hat{\theta}\right)\right)$	med(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 (med(Rel)), and the (model-implied) empirical/estimated reliability $\left(\text{med}\left(\hat{\text{Rel}}\right)\right)$ 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 $\left(r^*_{11k}, \ldots, r^*_{f_k Mk}\right)^T$ 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 $b_{jk} := \sum_{i \in F_k} \log\left(r^*_{ijk}!\right) = \sum_{i=1}^{N} \mathbb{I}_{i \in F_k} \log\left(r^*_{ijk}!\right)$, and obtain

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

Denote the vector $\left(r^*_{111}, \ldots, r^*_{f_K MK}\right)^T$ 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

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

From the factorization of the likelihood, one can see that f_k, r_{jk}, and b_{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,

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

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)

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

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

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

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

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

for all $k2\{1, \ldots, K\}$. With analogous operations and using the definitions of $\mathrm{r_{jk}}$ and $\mathrm{h_{jk}}$ one obtains

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

and

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

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):

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

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

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

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.