PSYCHOMETRIKA
https://doi.org/10.1007/s11336-023-09921-w

RANDOM EFFECTS MULTINOMIAL PROCESSING TREE MODELS: A MAXIMUM LIKELIHOOD APPROACH

STEFFEN NESTLER
UNIVERSITÄT MÜNSTER

EDGAR ERDFELDER
UNIVERSITÄT MANNHEIM

The present article proposes and evaluates marginal maximum likelihood (ML) estimation methods for hierarchical multinomial processing tree (MPT) models with random and fixed effects. We assume that an identifiable MPT model with S parameters holds for each participant. Of these S parameters, R parameters are assumed to vary randomly between participants, and the remaining $S - R$ parameters are assumed to be fixed. We also propose an extended version of the model that includes effects of covariates on MPT model parameters. Because the likelihood functions of both versions of the model are too complex to be tractable, we propose three numerical methods to approximate the integrals that occur in the likelihood function, namely, the Laplace approximation (LA), adaptive Gauss–Hermite quadrature (AGHQ), and Quasi Monte Carlo (QMC) integration. We compare these three methods in a simulation study and show that AGHQ performs well in terms of both bias and coverage rate. QMC also performs well but the number of responses per participant must be sufficiently large. In contrast, LA fails quite often due to undefined standard errors. We also suggest ML-based methods to test the goodness of fit and to compare models taking model complexity into account. The article closes with an illustrative empirical application and an outlook on possible extensions and future applications of the proposed ML approach.

Key words: multinomial processing tree models, random effects models, hierarchical models, maximum likelihood estimation.

Multinomial processing tree (MPT) models are stochastic models for categorical data frequently used in different branches of behavioral science, primarily in cognitive psychology and social cognition research (for overviews and a recent tutorial see Batchelder & Riefer, 1999; Erdfelder et al., 2009, 2020; Hütter & Klauer, 2016; Schmidt et al., 2023). They have been applied to a wide range of phenomena, including, for example, recognition memory (e.g., Riefer & Batchelder, 1995; Xu & Bellezza, 2001), source monitoring (e.g., Meiser & Broder, 2002), recall memory (Batchelder & Riefer, 1986), and judgmental illusions, such as the hindsight bias (Erdfelder & Buchner, 1998; Nestler & Egloff, 2009; Nestler et al., 2012) . Specifically, MPT models are cognitive process models that refer to a particular experimental task or paradigm in which participants' judgments are categorized into a well-defined set of responses. It is assumed that the observed frequencies of responses who fall into these categories follow a multinomial distribution and that the probabilities underlying these frequencies are determined by latent cognitive processes that drive observed response behavior.

The primary goal of fitting MPT models to observed response frequencies is to estimate the cognitive process parameters, that is, the latent probabilities that certain cognitive processes were

The research reported in the present paper was supported in part by a grant from the Deutsche Forschungsgemeinschaft to the Research Training Group Statistical Modeling in Psychology (SMiP, GRK 2277).

Correspondence should be made to Steffen Nestler, Institut für Psychologie, Universität Münster, Fliednerstr. 21, 48149Münster, Germany. Email: steffen.nestler@uni-muenster.de

Correspondence should be made to Edgar Erdfelder, Universität Mannheim, Fakultät für Sozialwissenschaften A5, 68159Mannheim, Germany. Email: erdfelder@uni-mannheim.de

Published online: 29 May 2023

successful or not (e.g., memory processes, such as encoding, storage, or retrieval). Furthermore, by estimating models that impose psychologically motivated restrictions on these parameters (e.g., equality constraints or parameter fixations), model comparisons can be used to statistically test psychological assumptions and cognitive hypotheses that are linked to the model.

In most past applications, MPT models have been estimated by aggregating observed category frequencies across participants. This approach presumes that individual differences in cognitive process parameters can be neglected. However, when there are substantial individual differences, parameter estimates may be biased, and the results of inferential statistical procedures might not be optimal (e.g., the standard errors of the parameter estimates may be misleading, Batchelder & Riefer, 1999; Erdfelder et al., 2009; Klauer, 2006; 2010; Smith & Batchelder, 2010). In addition to these statistical problems, the assumption that there are no between-person differences in relevant cognitive processes seems rather implausible (Lee & Webb, 2005; Smith & Batchelder, 2010). Such an assumption also precludes the exploration of interesting research questions about the origin of individual differences in process parameters and about their relationships with other model parameters as well as covariates (e.g., Coolin et al., 2015, 2016). Hence, it is desirable to quantify potential interindividual differences and to investigate which person variables can explain them.

A number of extensions have therefore been proposed to model and predict the heterogeneity of parameters between individuals. First, Klauer (2006) suggested a latent class MPT framework in which a single person is assumed to be a member of one latent class, and model parameters may differ between latent classes. Second, Smith and Batchelder (2010) proposed a hierarchical MPT model in which participants' model parameters are assumed to be sampled from *independent* beta distributions (i.e., a beta distribution for each MPT model parameter; hence the name "beta-MPT model"). More recently, Klauer (2010) introduced another hierarchical extension of the standard MPT model that allows researchers to capture not only the variability in the (probit-transformed) parameters across individuals but also the covariances or correlations between these parameters. This latent-trait MPT model is based on the assumption that the person parameters stem from a multivariate normal distribution, and their expectations and covariance matrix are estimated from the observed frequency data.

Klauer (2010) suggested a Bayesian approach for parameter estimation and inferences in which the posterior distribution is approximated using Monte Carlo-Markov chain methods (see Heck et al., 2018a, for an implementation in `R`). In the present article, we show how the parameters of the latent-trait MPT model can be obtained through marginal maximum likelihood (ML) estimation. Our implementation of the ML method introduces a frequentist approach for hierarchical MPT models that is perhaps more familiar to researchers because standard errors, confidence intervals, and goodness-of-fit tests can be computed on the basis of well-known asymptotic properties of ML estimates. Moreover, in addition to the asymptotic optimality of ML estimates (Reed & Cressie, 1988) , the subtleties of specifying appropriate prior distributions for Bayesian estimation can be avoided when using the ML method. In particular, one does not have to worry about whether and how the obtained estimates and model comparisons are affected by the prior (for a recent discussion, see Sarafoglou et al., 2022) or whether de-facto equivalent models may become nonequivalent as a consequence of the choice of the prior (Kellen & Klauer, 2020). Importantly, the most efficient numerical ML algorithm is also typically faster than a Bayesian estimator, and the convergence of the ML algorithm is simpler to determine.

1. The Pair-Clustering Model

Before we introduce MPT models and the ML estimation methods in more detail, we briefly describe the pair-clustering model (e.g., Batchelder & Riefer, 1980, 1986) that we use throughout the article to illustrate the proposed methods. The pair-clustering model can be used to analyze data

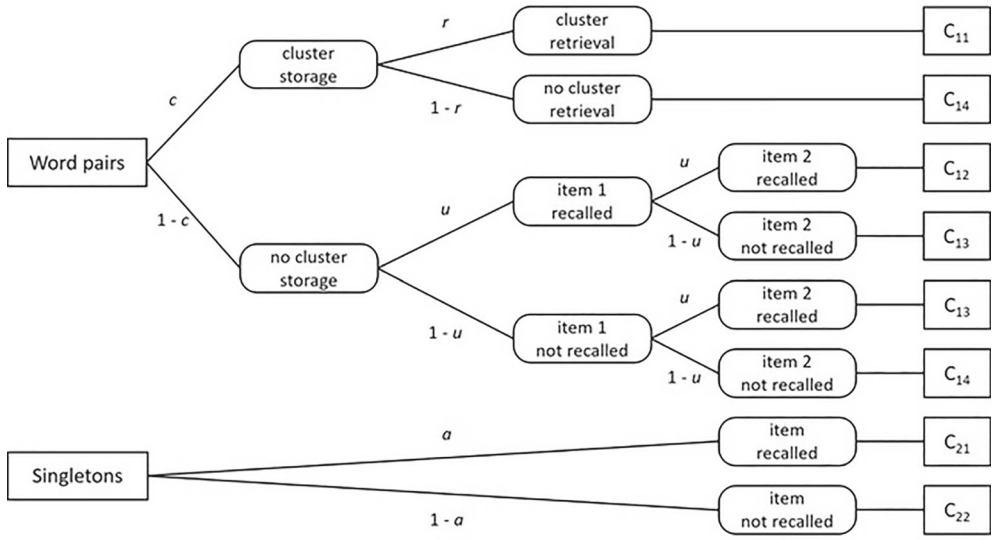

FIGURE 1.
The pair-clustering MPT model (adapted from Riefer & Batchelder, 1988, p. 330, Figure 2). Rectangles indicate stimulus classes (left) and observable responses (right). Rectangles with rounded corners represent latent cognitive states. Parameters attached to the branches indicate transition probabilities from left to right, specifically, storing a word pair as a cluster (c), retrieving a stored cluster in free recall (r), storing and retrieving a word from a non-clustered pair in free recall (u), and storing and retrieving a singleton in free recall (a).

from a free-recall experiment in which participants are presented with a list of word pairs plus a number of singletons. All words are preselected to be equally difficult in terms of memorizability. Each word pair consists of semantically related words (e.g., rose–tulip), whereas singletons are unrelated to other words in the list. These pairs and the singletons are presented one word at a time, and participants are later asked to recall the items from the list in any order. On the basis of participants' free recall performance, the studied word pairs and singletons are then assigned to one of six response categories. It is assumed that the probability of each response category can be modeled by two processing trees that include a total of four parameters (see Fig. 1 for a graphical illustration of the model).

Specifically, the model defines four response categories for the word pairs: Category C_{11} includes all cases where both words in a pair are recalled adjacently, C_{12} represents cases in which both words are recalled nonadjacently, C_{13} corresponds to cases where exactly one word is recalled, and finally C_{14} represents cases in which neither word from the pair is recalled. In addition, the singletons are scored in two categories: singletons recalled successfully (C_{21}) versus not successfully (C_{22}).

The pair-clustering model proposes four cognitive process parameters that jointly determine the response probabilities: (1) the probability c that a word pair is stored as a cluster in memory, (2) the probability r that a stored cluster is successfully retrieved in free recall, (3) the probability u that one word from a pair that is not stored as part of a cluster is successfully retrieved, and finally, (4) the probability a that a singleton is successfully stored and retrieved. These parameters can be used to derive response probabilities for each of the six categories by calculating the product of all parameters along a branch and then summing up all branch probabilities that refer to the same category. For example, because there is only one branch that terminates in category C_{11} (see Fig. 1), the probability of this category (i.e., both words recalled adjacently) is just the product $c \cdot r$ (i.e., successful storage as a cluster followed by successful retrieval of this cluster). Applying

the same logic to all branches in Fig. 1 results in the following six model equations:

$$\begin{aligned}
P(C_{11}) &= c \cdot r \\
P(C_{12}) &= (1-c) \cdot u^2 \\
P(C_{13}) &= (1-c) \cdot u \cdot (1-u) + (1-c) \cdot (1-u) \cdot u \\
P(C_{14}) &= c \cdot (1-r) + (1-c) \cdot (1-u)^2 \\
P(C_{21}) &= a \\
P(C_{22}) &= 1-a
\end{aligned} \tag{1}$$

The goal of MPT modeling, applied to the pair-clustering model, is to estimate the four cognitive process parameters and to test reasonable hypotheses about these parameters. One example of such a hypothesis would be that unclustered words from a pair behave like singletons in memory, that is, they are stored and retrieved with the same probability (i.e., $u = a$).

In the next section, we begin with a formal introduction of a typical MPT model without random effects. This is followed by an outline of the latent-trait MPT model with some extensions. In the third section, we present three methods of marginal ML estimation for the latent-trait model. Next, we describe the results of a small simulation study in which we compare the different methods with respect to accuracy and speed. In Sect. 3.2, we present an application of our method to an illustrative example using real data. In the final section, we discuss possible extensions and open questions for future research.

2. MPT Models Without Person-Level Random Effects

The pair-clustering model we just described is a specific instance of an MPT model. On a more general level, MPT models are statistical models of response frequencies of mutually exclusive, independent categories. To define the model, we assume that there are $k = 1, \ldots, K$ category systems and that each category system consists of $j = 1, \ldots, J_k$ categories C_{kj}. In the pair-clustering model, there are $K = 2$ category systems, one for the word pairs and one for the singletons. The first system consists of four categories: C_{11}, C_{12}, C_{13}, and C_{14}; and the singleton system contains two categories: C_{21} and C_{22}. We further assume that data from $t = 1, \ldots, T$ individuals is available. For each single individual t, the data for category system k are given as a vector of frequencies, $\boldsymbol{n}_{kt} = (n_{k1t}, \ldots, n_{kJ_kt})$. For instance, $\boldsymbol{n}_{1t}$ contains the four frequencies of person t for the four word pair categories (i.e., $k = 1$). These frequencies are assumed to stem from a multinomial distribution

$$f(\boldsymbol{n}_{kt}|\boldsymbol{\theta}) = \binom{N_{kt}}{n_{k1t}\ldots n_{kJ_kt}} p_{k1t}^{n_{k1t}} \cdot p_{k2t}^{n_{k2t}} \cdots p_{kJ_kt}^{n_{kJ_kt}} = \binom{N_{kt}}{n_{k1t}\ldots n_{kJ_kt}} \prod_{j=1}^{J_k} p_{kjt}^{n_{kjt}} \tag{2}$$

where $N_{kt} = n_{k1t} + \ldots + n_{kJ_kt}$ and $p_{k1t} + \cdots + p_{kJ_kt} = 1$. This definition requires the vector of all frequencies across all category systems K of person t, that is, $\boldsymbol{n}_t = (\boldsymbol{n}_{1t}, \ldots, \boldsymbol{n}_{Kt})$, to follow a product multinomial distribution:

$$f(\boldsymbol{n}_t|\boldsymbol{\theta}) = \prod_{k=1}^{K} \binom{N_{kt}}{n_{k1t}\ldots n_{kJ_kt}} \prod_{j=1}^{J_k} p_{kjt}^{n_{kjt}}. \tag{3}$$

The idea behind MPT models is that the probability of the occurrence of an event category can be modeled as a function of the cognitive process parameters. In the standard model (i.e., without person-level random effects), these process parameters $\boldsymbol{\theta}$ are unknown constants that do not vary between individuals. Thus, p_{kjt} is written as

$$p_{kjt} = P(C_{kj}|\boldsymbol{\theta}) \tag{4}$$

where $\boldsymbol{\theta}$ contains the S cognitive process parameters and is thus an element of $[0, 1]^S$, $s = 1, ..., S$. In the example, $\boldsymbol{\theta} = (c, r, u, a)$, and the probability of C_{14} (see above) is $p_{14t} = P(C_{14}|\boldsymbol{\theta}) = c(1 - r) + (1 - c)(1 - u)^2$. To write $P(C_{kj}|\boldsymbol{\theta})$ more generally, we define I_{kj} to be the number of branches of the tree that terminate in C_{kj}, with $i = 1, ..., I_{kj}$ indexing a specific branch B_{kji} that terminates in category j of category system k. As described by Hu and Batchelder (1994), we define:

$$P(B_{kji}|\boldsymbol{\theta}) = \prod_{s=1}^{S} \theta_s^{a_{skji}} (1 - \theta_s)^{b_{skji}} \tag{5}$$

where a_{skji} and b_{skji} indicate how often a process parameter θ_s and its complement $1 - \theta_s$, respectively, appear on the branch B_{kji}. For instance, there are two paths that terminate in C_{14}. To illustrate, we just consider the first of these branches, B_{141}:

$$\begin{aligned} P(B_{141}|\boldsymbol{\theta}) &= \prod_{s=1}^{4} \theta_s^{a_{s141}} (1 - \theta_s)^{b_{s141}} \\ &= \theta_1^{a_{1141}} (1 - \theta_1)^{b_{1141}} \cdot \theta_2^{a_{2141}} (1 - \theta_2)^{b_{2141}} \cdot \theta_3^{a_{3141}} (1 - \theta_3)^{b_{3141}} \cdot \theta_4^{a_{4141}} (1 - \theta_4)^{b_{1141}} \\ &= c^1 (1 - c)^0 \cdot r^0 (1 - r)^1 \cdot u^0 (1 - u)^0 \cdot a^0 (1 - a)^0 = c(1 - r) \end{aligned} \tag{6}$$

According to these definitions, the probability of C_{kj} is

$$P(C_{kj}|\boldsymbol{\theta}) = \sum_{i=1}^{I_k} P(B_{kji}|\boldsymbol{\theta}) = \sum_{i=1}^{I_k} \prod_{s=1}^{S} \theta_s^{a_{skji}} (1 - \theta_s)^{b_{skji}} \tag{7}$$

and the multinomial probability of the data for a single individual t is

$$f(\boldsymbol{n}_t|\boldsymbol{\theta}) = \prod_{k=1}^{K} \binom{N_{kt}}{n_{k1t} \ldots n_{kJ_kt}} \prod_{j=1}^{J_k} \left[\sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} \theta_s^{a_{skji}} (1 - \theta_s)^{b_{skji}} \right]^{n_{kjt}} . \tag{8}$$

Equation 8 can be used to obtain ML estimates of the process parameters $\boldsymbol{\theta}$. For instance, Hu and Batchelder (1994) showed how the parameters can be estimated with an EM algorithm. Alternatively, methods based on the analytical gradient or finite difference methods can be used for ML parameter estimation. Both approaches are implemented in the `R` package `MPTinR` (Singmann & Kellen, 2013; Singmann et al., 2020) or the package `mpt` (Wickelmaier & Zeileis, 2011, 2020) . Also, a Bayesian approach can be used to estimate the parameters (e.g., Lee & Wagenmakers, 2014).

3. MPT Models with Person-Level Random Effects

In the previous section, the process parameters contained in $\boldsymbol{\theta}$ were assumed not to differ between individuals. In the latent-trait MPT model, this assumption is relaxed because $R \leq S$ elements of $\boldsymbol{\theta}$ are written as a function of person-specific random effects that stem from a multivariate normal distribution

$$\boldsymbol{b}_t \sim MNV(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

where $\boldsymbol{\mu}$ is an $R \times 1$ vector of expectations and $\boldsymbol{\Sigma}$ is an $R \times R$ covariance matrix. Because each single θ_{st} is a probability parameter, we need to make sure that the resulting coefficient remains in the unit interval [0, 1]. On the basis of the literature on the Generalized Linear Model (GLM), we use the logit-link function (see Coolin et al., 2015)

$$\theta_{st} = \frac{1}{1 + exp\left[-(\beta_s + b_{st})\right]} \tag{10}$$

or the probit-link function (see Klauer, 2010)

$$\theta_{st} = \Phi(\beta_s + b_{st}) \tag{11}$$

to map a real-valued person parameter b_{st} on a probability parameter θ_{st}, with β_s serving as a parameter-specific intercept constant (in Eq. 11, Φ denotes the cumulative normal distribution). The two approaches typically result in very similar parameter estimates (at least in the GLM framework). To allow for comparisons in the framework of random effects MPT models, we have implemented both link functions in the `R` package that implements the statistical methods proposed in this article.

We can now define the *conditional* distribution of the category frequencies of person t given the random effects $\boldsymbol{b}_t$:

$$f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{b}_t) = \prod_{k=1}^{K} \binom{N_{kt}}{n_{k1t} \ldots n_{kJ_kt}} \prod_{j=1}^{J_k} \left[\sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} \theta_{st}^{a_{skji}} (1 - \theta_{st})^{b_{skji}} \right]^{n_{kjt}} \tag{12}$$

where θ_{st} is defined as in Eqs. 10 or 11. Along with the multivariate normal density of the random effects $\boldsymbol{b}_t$, Eq. 12 can then be used to obtain ML estimates of the model parameters. We refer to this model as the random effects MPT model.

Before we proceed, we would like to point out that the mean structure in our model is not identified because the expectation of the random effects is $\boldsymbol{\mu}$ and the conditional distribution contains the vector of intercept terms $\boldsymbol{\beta}$. To identify the mean structure, we define all elements of $\boldsymbol{\beta}$ to be zero when *all* process parameters are assumed to differ between individuals (i.e., $R = S$, where R denotes the number of random effects parameters). In this case, our random effects model corresponds to the model described by Klauer (2010). However, when only some (but not all) of the process parameters are assumed to be random, we estimate the respective entries in $\boldsymbol{\beta}$ for the $(S - R)$ fixed parameters while setting the remaining R entries (corresponding to random parameters) to zero and reducing the dimensions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly. Thus, by writing the model as proposed here, we achieve a high degree of flexibility to estimate MPT models with both random and fixed effects. This can be very useful, for example, when MPT models include guessing parameters that need to be estimated and are assumed to be constant across individuals.

Furthermore, we can also extend the model to include person-level covariates contained in the person-specific column vectors $\boldsymbol{X}_t$. This can be achieved in two ways: First, when the person-level covariates are used to predict R process parameters that are assumed to be random, $\boldsymbol{\mu}$ is person-specific, and we write it as function of the predictor values:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{X}_t, \tag{13}$$

where $\boldsymbol{\Gamma}$ is an $R \times p$ matrix of weights, and p is the number of person-level predictors used to predict the random process parameters. Note that when there is an assumption that a process parameter is *not* affected by a predictor in $\boldsymbol{X}_t$, the respective entry in $\boldsymbol{\Gamma}$ is simply set to zero. Furthermore, $\boldsymbol{\mu}$ represents the values of the person parameters when the covariates are zero. Second, when a fixed process parameter θ_s is assumed to vary as a function of person-level covariates so that the intercept β_{st} may differ between individuals t, we follow the procedure outlined by Coolin et al. (2015) and write

$$\beta_{st} = \beta_s + \boldsymbol{\gamma_s}\boldsymbol{X}_t \tag{14}$$

where β_s is the value of the parameter-specific intercept when all person-level covariates are zero and $\boldsymbol{\gamma_s}$ is a row vector containing the weights of the p covariates used to predict θ_s.

3.1. Marginal Maximum Likelihood Estimation

The goal is to estimate the free parameters contained in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the $(S - R)$-dimensional vector of intercepts and the $(S - R) \times p$ matrix of predictor weights, respectively, in the fixed-effects part of our model (corresponding to the model of Coolin et al., 2015) while $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ denote to the R-dimensional vector of parameter means, the $R \times p$ matrix of predictor weights, and the $R \times R$ parameter covariance matrix, respectively, in the random-effects part of our model (corresponding to the model of Klauer, 2010). We employ a marginal ML approach that is based on the marginal density of the response frequencies. For a single individual t, this density is

$$f(\boldsymbol{n}_t) = \int_{\boldsymbol{b}_t} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t) f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) d\boldsymbol{b}_t. \tag{15}$$

and for the entire sample, it is

$$f(\boldsymbol{n}) = \prod_{t=1}^{T} f(\boldsymbol{n}_t) \tag{16}$$

Eq. 16 can be used to define the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = \prod_{t=1}^{T} \int_{\boldsymbol{b}_t} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t) f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) d\boldsymbol{b}_t = \prod_{t=1}^{T} \int_{\boldsymbol{b}_t} L_t d\boldsymbol{b}_t \tag{17}$$

and the log-likelihood of the data

$$ll(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = \sum_{t=1}^{T} \log \left(\int_{\boldsymbol{b}_t} L_t d\boldsymbol{b}_t \right). \tag{18}$$

One problem with the marginal ML approach is that there is no analytical solution for the values of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ that maximize Eq. 18. Rather, numerical approximations have to be used to estimate the model parameters. For these approximations, we use the analytical gradient of the log-likelihood function given by

$$\frac{\partial ll}{\partial \tau_k} = \sum_{t=1}^{T} \frac{1}{f(\boldsymbol{n}_t)} \int_{\boldsymbol{b}_t} L_t \frac{\partial \log(L_t)}{\partial \tau_k} d\boldsymbol{b}_t \tag{19}$$

where τ denotes an element of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, or $\boldsymbol{\Sigma}$. To implement the gradient, one needs the first derivative of the log-likelihood function for a single person t with regard to a parameter. Note that $\log(L_t)$ is

$$\log f\left(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t\right) + \log f\left(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}\right) \tag{20}$$

Thus, for a single element σ_l contained in $\boldsymbol{\Sigma}$, the derivative is

$$\frac{\partial \log L_t}{\partial \sigma_l} = -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_l}\right) + \frac{1}{2}\left(\boldsymbol{b}_t - \boldsymbol{\mu}_t\right)' \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_l} \boldsymbol{\Sigma}^{-1} (\boldsymbol{b}_t - \boldsymbol{\mu}_t) \tag{21}$$

with

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_l} = \mathbf{1}_l + \mathbf{1}_l' - \mathbf{1}_l \cdot \mathbf{1}_l' \tag{22}$$

and $\mathbf{1}_l$ is an R x R matrix that contains a 1 in the position of σ_l and zeroes in all other positions. For elements in $\boldsymbol{\mu}$ or $\boldsymbol{\Gamma}$, the derivative is

$$\frac{\partial \log L_t}{\partial \mu_{tl}} = \left(\boldsymbol{b}_t - \boldsymbol{\mu}_t\right)' \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}_t}{\partial \mu_{tl}} \tag{23}$$

where

$$\frac{\boldsymbol{\mu}_t}{\partial \mu_l} = \mathbf{1}_l \text{ and } \frac{\boldsymbol{\mu}_t}{\partial \gamma_l} = \mathbf{1}_l \boldsymbol{X}_t \tag{24}$$

and $\mathbf{1}_l$ is an R x 1 vector or an R x p matrix that contains a 1 in the position of the parameter that is being estimated and zeroes in all other positions. Finally, the derivative of $\log(L_t)$ with regard to the elements in $\boldsymbol{\beta}$ is

$$\frac{\partial \log L_t}{\partial \beta_l} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J_k} \frac{n_{kjt}}{P(C_{kj}|\boldsymbol{\theta}_t)} \sum_{i=1}^{I_{kj}} \left[a_{skji} \cdot \left(\frac{\partial \theta_{st}}{\partial \beta_l}\right)^{-1} - b_{skji} \cdot \left(1 - \frac{\partial \theta_{st}}{\partial \beta_l}\right)^{-1} \right] \cdot P(B_{kji}|\boldsymbol{\theta}_t) \tag{25}$$

where $\frac{\partial \theta_{st}}{\partial \beta_l}$ is the derivative of the chosen link function with respect to the intercept parameter β_l. The same formula can be used for the parameters in $\boldsymbol{\gamma}$, but the derivative of the link function has to be computed for the element γ_{sl} rather than β_l (i.e., $\frac{\partial \theta_{st}}{\partial \gamma_{sl}}$).

A second problem with using the marginal ML approach is that the likelihood or log-likelihood function, respectively, and the gradient involve integrals that are not tractable. However, one can approximate the integrals using a number of numerical techniques (see Tuerlinckx et al., 2006; Nestler, 2020, 2021). In the `R` package that implements marginal ML estimation methods, we implemented three approaches: the Laplace approximation, the Adaptive Gauss–Hermite Quadrature (AGHQ), and Quasi Monte Carlo (QMC) Sampling.

Laplace Approximation The basic idea behind the Laplace approximation is that it can be used to replace the function within the integral with another function that has a closed-form expression. Imagine that the modes of the random effects $\boldsymbol{b}_t$ of

$$l(\boldsymbol{b}_t) = f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t) f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \tag{26}$$

are available for each of the T individuals. One can then show (see, e.g., Pinheiro & Bates, 1995) that the likelihood function given in Eq. 17 can be approximated by

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \approx \prod_{t=1}^{T} (2\pi)^{S/2} |\hat{\boldsymbol{\Omega}}|^{1/2} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \hat{\boldsymbol{b}}_t) f(\hat{\boldsymbol{b}}_t|\boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}) \tag{27}$$

where $\hat{\boldsymbol{b}}_t$ denotes the modes, and $\hat{\boldsymbol{\Omega}}$ is the asymptotic covariance matrix of these modes.

A problem with using Eq. 27 is that one has to estimate the modes $\hat{\boldsymbol{b}}_t$ for all T individuals given the (unknown) parameters contained in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$. In practical implementations, this problem is circumvented by first estimating the modes given the current parameter estimates. Then, the model parameters are estimated by maximizing Eq. 27 given the current modes $\hat{\boldsymbol{b}}_t$. This two-step procedure is repeated until the algorithm converges. The Laplace approximation is very fast (compared with the other methods), and, in the well-known `R` package `lme4` (Bates et al., 2015) , it is the default method for estimating the parameters of a Generalized Linear Mixed Model (GLMM).

Gauss–Hermite Quadrature The basic idea behind quadrature approaches is that they can be used to approximate the numerical value of the integral. When one assumes that the random effects are normally distributed (as we did), one first generates M vectors of size R x 1 of Gauss–Hermite (GH) nodes $\boldsymbol{x}$ and weights $\boldsymbol{w}$. For each of the node vectors (e.g., $\boldsymbol{x}_m$), one then computes $f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t = \boldsymbol{x}_m)$. The integral for person t can then be approximated by a weighted sum

$$L_t(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \approx \sum_{n_1=1}^{M} \cdots \sum_{n_S=1}^{M} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_{n_1 \cdots n_R}) \boldsymbol{w}_{n_1} \cdots \boldsymbol{w}_{n_R}. \tag{28}$$

The main problem with this GH quadrature is that the number of points M increases exponentially with the size of R. For example, when $M = 10$ and $R = 4$, there are $10^4 = 1000$ single node vectors. This is problematic because $f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t = \boldsymbol{x}_m)$ needs to be computed for each vector $\boldsymbol{x}_m$ *and* each person t. Thus, when M is large, the computational burden associated with approximating the integral is too large to be feasible. For instance, when $M = 10$ and $R = 10$, one would need to compute 10^{10} vector values for each person t. For this reason, we use Adaptive GH quadrature in our implementation (AGHQ, Rabe-Hesketh et al., 2002; Tuerlinckx et al., 2006). In AGHQ, the GH quadrature points $\boldsymbol{x}$ for each individual are first centered and scaled with the individual's modes $\hat{\boldsymbol{b}}_t$. These scaled nodes and weights are then used to compute a weighted sum that is similar to the one provided in Eq. 28. By scaling the node vectors with the modes, fewer nodes are required to achieve a precise approximation of the integrals.

Quasi Monte Carlo Integration Even when using AGHQ, the computational burden is too high for high-dimensional random effect distributions. An alternative one could use is Monte Carlo (MC) integration. MC integration rests on the observation that the integral in Eq. 15 can be seen

as an expectation of the function $f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t)$ with respect to the random effect distribution $f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$:

$$\int_{\boldsymbol{b}_t} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t) f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) d\boldsymbol{b}_t = \mathrm{E}\left[f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t)\right]. \tag{29}$$

One can thus draw M random samples from $f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ to approximate the integral (Robert & Casella, 2010) with

$$L_t(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \approx \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_{tm}). \tag{30}$$

where one inserts the m-th random draw for $\boldsymbol{b}_t$. The advantage of MC integration is that the number of draws M does not have to increase when another random effect is added. However, one disadvantage of the technique is that M must be sufficiently large to be precise. Furthermore, because samples are randomly drawn from the random effect distribution, a (Monte Carlo) sampling error is introduced into the estimation. For these two reasons, here, we use Quasi Monte Carlo (QMC) integration, which builds on deterministic sequences of points instead of randomly drawn points in MC integration (hence the name "Quasi"). There are different ways to generate such sequences (see González et al., 2006, for an introduction). In accordance with the GLMM literature (e.g., Crowther, 2017; González et al., 2006), we use Halton sequences in our implementation because smaller numbers of draws M are required to achieve a precise approximation. To further decrease the computational burden, we additionally scale the Halton numbers with $\hat{\boldsymbol{b}}_t$ and $\hat{\Omega}$.

3.2. Standard Errors, Goodness-of-Fit Tests, and Random Effects

Once the ML estimates have been determined, the estimates and their corresponding standard errors can be used to compute z statistics and confidence intervals. From standard ML theory, it follows that the parameters are asymptotically normally distributed with a covariance matrix that is obtained by calculating the inverse of the information matrix. This matrix is the negative of the matrix of second derivatives that is given by

$$\frac{\partial ll}{\partial \tau_k \tau_j} = \sum_{t=1}^{T} \frac{1}{f(\boldsymbol{n}_t)^2} \left[f(\boldsymbol{n}_t) \frac{\partial d_t}{\partial \tau_j} - d_t d_t^T \right] \tag{31}$$

where

$$d_t = \int_{\boldsymbol{b}_t} L_t \frac{\partial \log(L_t)}{\partial \tau_k} d\boldsymbol{b}_t \tag{32}$$

and

$$\frac{\partial d_t}{\partial \tau_j} = \int_{\boldsymbol{b}_t} \left[\frac{\partial^2 \log(L_t)}{\partial \tau_k \tau_j} + \frac{\partial \log(L_t)}{\partial \tau_k} \frac{\partial \log(L_t)}{\partial \tau_j} \right] L_t d\boldsymbol{b}_t \tag{33}$$

Again, due to the integrals involved in Eqs. 32 and 33, the matrix of second derivatives is hard to compute. In our implementation of the ML approach users can therefore choose whether standard errors are based on a numerical approximation of the Hessian with finite-difference methods using the analytical gradient (see Eq. 19) or on the exact hessian computed with Eq. 31.

Furthermore, before researchers interpret MPT parameter estimates, they need to show that their model actually fits the observed data. In addition, researchers very often want to compare the

fit of a current model with the fit of a more restricted model that imposes psychologically motivated constraints on the parameters (e.g., equality constraints or parameter fixations). Goodness-of-fit tests and model comparisons can both be conducted by means of a likelihood ratio test

$$LR = -2 \cdot (ll_r - ll_u) \tag{34}$$

where ll_r (ll_u) is the log-likelihood value of the restricted (unrestricted) model. Under certain regularity conditions (see Reed & Cressie, 1988), the LR test statistic asymptotically follows a central χ^2 distribution with degrees of freedom equal to the difference between the parameters of the unrestricted and restricted models if the more restricted model actually holds. The likelihood ratio test can be used to compare the fit of two models that differ in the mean structure (e.g., $u = a$ in the pair-clustering model) or in the covariance structure (e.g., $\sigma_{ua} = 0$).

The likelihood ratio test is based on the assumption that the models being compared are nested. For non-nested model comparisons, we suggest using the Akaike or the Bayesian Information Criterion (AIC or BIC, respectively):

$$\begin{aligned} AIC &= -2 \cdot ll + 2 \cdot df \\ BIC &= -2 \cdot ll + \log(n) \cdot df, \end{aligned} \tag{35}$$

where $n = \sum_{t=1}^{T} n_t$ and $df = R + (S - R) + p_{\boldsymbol{\Gamma}} + p_{\boldsymbol{\gamma}} + \dim(\boldsymbol{\Sigma})$. Here, R is the number of estimated cognitive process parameters set to be random (i.e., parameters in $\boldsymbol{\mu}$), $S - R$ is the number of fixed cognitive process parameters (i.e., parameters in $\boldsymbol{\beta}$), $p_{\boldsymbol{\Gamma}} + p_{\boldsymbol{\gamma}}$ gives the total number of to-be-estimated weights of the person-level covariates in the random and the fixed effects part of the model (i.e., parameters in $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$), respectively, and $\dim(\boldsymbol{\Sigma})$ represents the number of estimated covariance parameters (see Bates et al., 2015).

Finally, it can also be of interest to obtain estimates of the individual random effects for each participant t. As a random effects estimator, $\hat{\boldsymbol{b}}_t$, it is possible to use a participant's mode, which can be obtained by maximizing Eq. 26, while treating the final model parameter estimates as fixed. Fortunately, these values are already required when estimating the model so that they do not have to be estimated again. Another choice would be the empirical Bayes estimator (Bock & Aitken, 1981) :

$$\hat{b}_t = \prod_{t=1}^{T} \frac{1}{f(\boldsymbol{n}_t)} \int_{\boldsymbol{b}_t} \boldsymbol{b}_t \log(f(\boldsymbol{n}_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}_t)) f(\boldsymbol{b}_t|\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) d\boldsymbol{b}_t. \tag{36}$$

In this approach, we approximate the integrals using one of the integral approximation methods described above.

4. Simulation Study

We performed a simulation study to assess the frequentist properties of the suggested ML estimation approaches and also compare it with the performance of a Bayesian approach. Specifically, we examined the effect of the number of participants and the number of responses per participant on the bias of the parameter estimates and the coverage rate of the corresponding confidence or credibility intervals, respectively. In addition, we examined for the ML estimator how the number of quadrature points in the AGHQ approach and the size of the Halton sequence in the QMC method affect the adequacy of the model parameter estimates.

Population Model and Simulation Conditions. We used the pair-clustering model for this simulation. This MPT model comprises two category systems, including four and two categories, respectively, with response probabilities described by four process parameters c, r, u, and a. In our simulation study, the population covariance matrix of these four parameters (describing how they vary across participants) was set to

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.50 & 0.08 & 0.04 & 0.00 \\ 0.08 & 0.35 & 0.03 & 0.00 \\ 0.04 & 0.03 & 0.20 & 0.07 \\ 0.00 & 0.00 & 0.07 & 0.20 \end{pmatrix} \tag{37}$$

where the non-zero covariance terms reflect correlations of 0.30, 0.20, and 0.10. The mean values of the parameters were set to $\mu_c = 0.50$, $\mu_r = 0.40$, $\mu_u = 0.25$, and $\mu_a = 0.15$ (see Batchelder & Riefer, 1986).

We manipulated the number of simulated participants (75 vs. 125) and the number of simulated responses per participant (25 vs. 75 vs. 125). In accordance with Batchelder and Riefer (1986), about 80% of the responses were assigned to the first category system (i.e., 20 vs. 60 vs. 100), leaving 20% for the second system (i.e., 5 vs. 15 vs. 25). The `R` package `mvtnorm` and the function `rmultinorm` were used to generate the samples. We drew 500 samples from the population for each of the six simulation conditions.

Estimators All the functions that were required to estimate the parameters with ML were implemented in `R` (R Core Team, 2020; a working version of the package can be downloaded from https://osf.io/w97m5/). To examine the properties of the ML estimation procedures, the number of quadrature points in the AGHQ method was set to 3, 4, or 5, resulting in $4^3 = 64$, $4^4 = 256$ or $4^5 = 1,024$ node vectors per participant. For the QMC method, the size of the Halton sequence was set to 500, 1000, or 2000. To obtain Bayesian estimates, we employed the `TreeBUGS` package (Heck et al., 2018a) . `TreeBUGS` uses the JAGS-MCMC sampler (Plummer, 2003) to approximate the posterior distribution of the model's parameters. For each replication, we fitted the model with the `traitMPT` function using the default settings. That is, weakly informative priors were specified for all parameters (i.e., normal distributions for the means and a scaled Wishart distributions for the covariance matrix). Furthermore, three chains of 20,000 samples were generated from the posterior distributions, whereby the first 2000 samples were discarded for parameter estimation (i.e., burn-in period). Since `TreeBUGS` by default provides the means of the posterior distributions as parameter estimates, we decided to used them as the Bayes estimates.

Dependent Measures We used the relative bias (RB) of the parameter estimates and the coverage rate (CR) to investigate the statistical performances of the ML approaches and the Bayes estimator. For the relative bias, we first computed the average parameter estimate in a simulation condition. We then computed the difference between this average and the true parameter and thereafter divided the difference by the true parameter. We consider relative biases below 10% to be acceptable, biases of 10–20% to be substantial, and biases above 20% to be unacceptable (e.g., Forero et al., 2009; Morris et al., 2019). For ML, the confidence interval with the standard error[1] of an estimate was computed in each replication to determine the observed coverage of the 95% confidence intervals. The coverage was then coded 1 if the true parameter value was included in the interval and 0 if the true parameter was not. We used the same approach to determine the

[1] We used the numerically approximated Hessian matrix to compute the standard errors of the parameter estimates in a replication. Additional analyses with 100 replications from the simulation condition with 75 participants and 25 responses showed that differences between standard errors based on the numerically approximated Hessian and the analytically computed Hessian were very small. Therefore, we conclude that all unsatisfactory results regarding the coverage rate are not due to the numerical approximation of the Hessian matrix.

TABLE 1.
Relative frequencies of converged replications (CR) and average computation time (in s), depending on the estimator, the type of ML approximation method, the number of individuals T, and the number of responses N per individual.

Variable	T	N	Bayes	Laplace	ML-AGHQ			ML-QMC		
					3	4	5	500	1000	2000
CR	75	25	0.56	0.01	0.68	0.84	0.85	0.56	0.61	0.62
		75	0.90	0.38	0.75	0.89	0.84	0.83	0.86	0.89
		150	0.95	0.64	0.95	0.99	0.98	0.96	0.97	0.98
	125	25	0.66	0.01	0.71	0.88	0.89	0.61	0.62	0.64
		75	0.91	0.33	0.90	0.98	0.95	0.88	0.93	0.96
		150	0.94	0.65	1.00	1.00	1.00	1.00	1.00	1.00
Time	75	25	53.53	4.29	12.35	18.78	31.44	25.76	36.18	57.50
		75	53.65	6.73	15.64	22.46	31.46	32.52	42.65	69.05
		150	54.38	7.72	12.60	19.50	29.35	23.75	32.63	54.57
	125	25	85.46	8.52	17.63	27.46	49.13	56.91	83.99	106.4
		75	86.84	10.78	30.60	35.69	41.09	70.24	88.03	111.9
		150	88.18	12.65	26.98	32.92	40.13	46.21	61.35	87.32

AGHQ = Adaptive Gauss–Hermit quadrature with 3, 4, or 5 nodes; QMC = Quasi Monte Carlo integration with 500, 1000, or 2000 points. In case of Bayes, convergence rates were calculated using $\hat{R}$ values. Computation times were determined on an Intel Core i7-6700 with four cores and 16GB RAM.

observed coverage of the 95% credibility intervals, but employed the 95% credibility intervals as provided by `TreeBUGS` as the basis for the coding.

Results We first examined the percentage of samples in which the estimation algorithm converged for the two estimators, the different approximation methods, the different numbers of simulated participants, and the different numbers of simulated responses per participant. In case of ML, a sample was counted as converged when there were neither inadmissible estimates nor undefined standard errors in the final solution. For Bayes, judging the convergence is a more difficult issue (Hoff, 2009) . Here, we decided to use $\hat{R} > 1.05$ as the criterion, because it can be used well in simulation studies and it is suggested in the literature (e.g., Lynch, 2007). However, we acknowledge that a more or less stringent cutoff may lead to different results and that this should be taken into account in the interpretation of our findings.

As can be seen in Table 1, convergence rates for both estimators increased with the number of participants and the number of responses. In case of ML, the larger the number of points used by a method to approximate the integrals, the better the convergence of the ML estimator. When the number of responses was 75 or 125, convergence rates were acceptable for the AGHQ and QMC methods. In conditions with 25 responses, convergence rates were highest for conditions with the largest size of points. In this case, convergence rates were also similar to the convergence rate of the Bayes estimator. With regard to the Laplace approximation, we found that convergence rates were always below 70% (and even near 0% when number of responses was 25). Further analyses showed that all convergence failures occurred because some (or all) of the standard errors were undefined for the final estimates. These estimates were also very biased. We think that this bias can be explained by noting that the precision of the Laplace approximation depends on whether the shape of the function that is being integrated resembles a multivariate normal distribution. Hence, the method does not perform well for highly non-normal cases (e.g., when the responses are Bernoulli distributed; see Engel, 1998), and we suspect that similar performance problems occur for the MPT model. Finally, Table 1 also contains the average run times of the different methods. We note that comparing the computation times across ML and Bayes is difficult, because

TABLE 2.
Relative bias of parameter estimates in percent, depending on the approximation method, the number of individuals T, and the number of responses N per individual.

Method	No. Points	Parameter	T =75			T = 125		
			N = 25	N = 75	N = 125	N = 25	N = 75	N = 125
AGHQ	4^3	μ	2.72	1.35	0.36	− 1.07	− 0.34	0.74
		σ_b^2	− 23.6	− 1.62	− 2.94	− 12.4	− 3.86	− 3.53
		σ_{bb}	12.5	11.8	7.42	15.0	9.91	3.13
	4^4	μ	2.11	1.34	0.60	0.27	− 0.59	0.71
		σ_b^2	8.44	0.41	− 1.85	− 1.31	− 1.26	− 2.05
		σ_{bb}	− 13.8	1.76	3.76	4.53	2.04	− 1.20
	4^5	μ	3.99	1.98	0.70	1.71	− 0.79	0.69
		σ_b^2	7.38	2.29	− 1.11	1.28	0.02	− 1.45
		σ_{bb}	− 8.65	1.59	3.59	− 1.30	− 0.05	− 2.39
QMC	500	μ	− 5.64	− 2.65	− 1.59	− 6.71	− 3.17	− 1.93
		σ_b^2	11.4	− 5.07	− 4.98	9.01	− 7.39	− 5.58
		σ_{bb}	− 55.6	− 3.84	2.40	− 36.4	− 8.32	− 3.28
	1000	μ	− 2.19	− 0.91	− 0.77	− 3.95	− 2.68	− 0.75
		σ_b^2	2.96	− 0.95	− 2.01	− 1.73	− 2.95	− 2.53
		σ_{bb}	− 33.3	− 3.77	− 0.55	− 19.4	− 6.28	− 5.96
	2000	μ	0.94	− 0.14	− 0.32	− 0.25	− 1.99	− 0.09
		σ_b^2	4.52	− 1.55	− 3.05	− 3.67	− 4.03	− 3.44
		σ_{bb}	− 3.67	6.63	4.20	0.82	3.69	− 0.25
Bayes	–	μ	− 2.18	− 2.59	− 2.19	− 2.69	− 3.34	− 1.99
		σ_b^2	1.73	1.83	2.51	− 1.44	1.44	1.13
		σ_{bb}	− 55.7	− 16.5	− 9.31	− 33.4	− 11.3	− 8.91

AGHQ = Adaptive Gauss–Hermit quadrature with 3, 4, or 5 nodes; QMC = Quasi Monte Carlo integration with 500, 1000, or 2000 points.

they depend on how the approaches are implemented in R (i.e., using C++ in the background or parallelization), how many chains are generated etc. In case of Bayes, all methods were slower for larger numbers of participants. Similarly, ML methods became slower the larger numbers of points used to approximate the integrals and the larger the number of participants.

In the following, we drop the Laplace approximation from further consideration when we discuss the precision of the ML estimates (i.e., RB) and the confidence intervals (i.e., CR). Furthermore, to facilitate the interpretation of the results, we decided to average the results for the indices per parameter group (i.e., for the cognitive process parameter mean values c, r, u, and a, called μ; the variance parameters in $\mathbf{\Sigma}$, termed σ_b^2; and the covariance parameters in $\mathbf{\Sigma}$, termed σ_{bb}). The resulting values for the relative bias are displayed in Table 2. When we consider relative biases below 10% to be acceptable and biases of 10–20% to be substantial (Forero et al., 2009; Morris et al., 2019) , we found for both ML methods that relative biases decreased as the numbers of node vectors, sample sizes, and numbers of responses per participant increased. Importantly, relative biases were generally low and acceptable in all simulation conditions for the cognitive process parameters and the variance parameters. For the latter, however, biases were somewhat larger but still acceptable in case of AGHQ when the number of responses was 25 and the number of persons was 75. For the covariance parameters, the relative biases were larger and more substantial the smaller the number of points used and the smaller the number of responses. For the Bayes estimator, the relative bias was negligible for the cognitive process parameters and the variance parameters in all conditions. However, replicating the results of Klauer (2010), the

TABLE 3.
Coverage rate of parameter estimates, depending on the approximation method, the number of individuals T, and the number of responses N per individual.

Method	No. Points	Parameter	$T = 75$			$T = 125$		
			$N = 25$	$N = 75$	$N = 125$	$N = 25$	$N = 75$	$N = 125$
AGHQ	4^3	μ	84.5	92.3	92.9	87.8	93.9	94.3
		σ_b^2	79.7	89.8	91.1	75.3	90.3	91.4
		σ_{bb}	85.1	91.6	94.5	83.3	91.6	94.3
	4^4	μ	88.7	94.3	93.7	92.1	94.7	94.8
		σ_b^2	88.1	90.9	92.4	88.5	93.0	92.8
		σ_{bb}	86.9	94.0	95.1	89.6	93.8	95.2
	4^5	μ	94.6	94.6	93.9	93.9	95.0	94.8
		σ_b^2	96.7	93.4	92.9	93.1	94.0	93.0
		σ_{bb}	93.4	95.5	95.4	92.7	94.1	95.2
QMC	500	μ	77.9	89.0	92.4	80.1	89.3	94.0
		σ_b^2	71.6	83.2	88.1	78.1	83.7	88.7
		σ_{bb}	68.0	85.5	92.7	74.9	86.6	93.2
	1000	μ	81.8	91.9	93.4	78.9	93.4	94.6
		σ_b^2	74.9	86.3	90.5	79.6	88.4	91.1
		σ_{bb}	66.4	89.8	94.2	76.6	90.3	94.3
	2000	μ	86.7	92.2	93.5	84.5	93.6	94.5
		σ_b^2	77.3	86.7	90.5	77.4	90.1	91.1
		σ_{bb}	83.2	90.5	94.4	82.1	90.3	94.2
Bayes	–	μ	94.0	94.3	94.5	95.4	94.9	94.8
		σ_b^2	95.1	94.3	94.7	95.9	94.7	95.3
		σ_{bb}	98.8	96.6	97.3	98.4	96.0	96.0

AGHQ = Adaptive Gauss–Hermit quadrature with 3, 4, or 5 nodes, QMC = Quasi Monte Carlo integration with 500, 1000, or 2000 points.

covariance parameters were unacceptably and substantially biased when the number of responses was 25 or 75. When the number of responses was 125, the relative biases were still large but acceptable.

With regard to the CR (see Table 3), we found that for all three types of parameters, the CR moved closer to the nominal value as the sample size and the number of responses increased. For AGHQ, the CR was near its nominal value when the number of points was 4^4 and the number of responses at least 75. When the number of responses is 25, the CR was near the nominal value when 4^5 points were used. A similar pattern of results was observed for the QMC approximation, although the amount of undercoverage was greater than for the AGHQ approximation. Furthermore, when the number of responses was 25, undercoverage occurred irrespective of how many points were investigated in our simulation. In case of Bayes, we found that the coverage was nominal for the cognitive process parameters and the variance parameters. However, overcoverage occurred for the covariance parameters although this tendency disappeared the larger the number of responses.

To summarize, the simulation study reveals that the AGHQ method works better than the QMC method when the number of nodes is at least four. Relative biases were low even for small sample sizes and few responses per participant and confidence interval estimates were close to nominal. When the number of responses was 75, the QMC also provided acceptable results, at least when the number of points was 1000. Finally, when the number of responses was small (i.e.,

$R = 25$), AGHQ with five nodes yield estimates that are at least as good as the estimates of a Bayesian approach.

5. Illustrative Example

To illustrate the proposed ML approach, we analyzed neuropsychological data from Schilken (1998), who employed the pair-clustering model to analyze and compare the memory performance of 22 epileptic patients with a right-temporal focus of epileptic seizures, 21 epileptic patients with a left-temporal focus, and 20 healthy controls matched with respect to age and intelligence. Each participant learned word lists consisting of 10 semantically strongly related word pairs (e.g., armchair-sofa) and 5 singletons that were not related to other words in the list (e.g., sunflower). All words were comparable in terms of difficulty and memorizability when considered in isolation. In addition, three words were added to the beginning and another three words to the end of the word list to absorb primacy and recency effects in free recall. Because these primacy and recency buffer words were excluded from further analysis, $N = 15$ responses per participant and study-test cycle remained for analysis-10 responses for the pairs and 5 for the singletons. The studying of the word list and the subsequent free recall test were repeated 6 times to examine learning effects on the c, r, and u parameters of the pair-clustering model, resulting in a total of six study-test trials per participant.

The same data set was also analyzed by Klauer (2006, 2010), thus providing us with the opportunity to compare our results with Klauer's, which were based on the Bayesian Latent-Trait model (cf. Klauer, 2010, pp. 86/87). To maximize heterogeneity between participants, we followed the procedure outlined by Klauer (2006, 2010) and analyzed all $T = 22+21+20 = 63$ participants conjointly in our first model. Also following Klauer's guidelines, we restricted our attention to the first two study-test cycles for each participant. This left us with $T = 63$ participants, $K = 4$ category systems (i.e., those for word pairs and singletons in Trials 1 and 2) with $J_1 = J_3 = 4$ and $J_2 = J_4 = 2$ categories (for word pairs and singletons, respectively), and $N_1 = N_3 = 10$ as well as $N_2 = N_4 = 5$ responses per participant within the four category systems. Finally, again in line with Klauer (2010)'s suggestions, we imposed the restriction that unclustered words in a pair must match the singletons in terms of trial-specific storage and retrieval probabilities (i.e., $u = a$). Hence, there were three parameters to estimate per trial, $c^{(1)}$, $r^{(1)}$, $u^{(1)}$ for Trial 1 and $c^{(2)}$, $r^{(2)}$, $u^{(2)}$ for Trial 2.

To test the goodness of fit, we estimated the model that we just specified and a more general model with four parameters $c^{(d)}$, $r^{(d)}$, $u^{(d)}$, and $a^{(d)}$ per trial. For both models, we used the AGHQ approximation method with 4 nodes to fit the two models. The LR test statistic comparing the original and the more general model was $LR = 9.02$, which is not significantly different from zero, $\chi^2_{crit} = 30.1$, $p = .98$, $df = 19$. Table 4 shows the parameters of the mean structure (i.e., μ) in the restricted model. As can be seen, parameter values increased from Trial 1 to Trial 2, and the probability-transformed parameters were very similar to the probability-transformed parameters reported in Klauer (2010). Variance and correlation parameters were also quite similar across the two approaches (see Table 5). A notable exception was the variance of $r^{(1)}$, where the ML estimate was considerably larger than the Bayesian estimates. Also, the correlations involving this parameter were smaller for ML compared with Bayes.

We decided to go one step further than Klauer (2010) and also analyze effects of the clinical group on the parameter estimates. For this purpose, we added two dummy variables as covariates to our model, the first one coded "1" for the right-temporal epileptic patients (patients in the two other groups were coded 0) and the second dummy variable coded "1" for the left-temporal epileptic patients (again, all other patients were coded "0"). This model provided us with the opportunity to estimate the average (negative) effect of each clinical group relative to the control

TABLE 4.
Parameter estimates of the mean structure for the illustrative data example.

	Model 1				Model 2			
	Est	CI	TE	Bayes	Est	CI	$\Delta_{D.}$	TE
$\mu_{c^{(1)}}$	−0.58	[−0.87, −0.27]	0.28	0.31	−0.21	[−0.39, −0.03]		0.42
$\delta_{c^{(1)},D_1}$					−0.90	[−1.14, −0.64]	−1.11	0.13
$\delta_{c^{(1)},D_2}$					−0.47	[−0.71, −0.23]	−0.68	0.24
$\mu_{r^{(1)}}$	−0.33	[−0.68, 0.01]	0.37	0.35	−0.35	[−0.56, −0.14]		0.36
$\delta_{r^{(1)},D_1}$					0.16	[−0.01, 0.32]	−0.19	0.42
$\delta_{r^{(1)},D_2}$					0.10	[−0.06, 0.27]	−0.25	0.40
$\mu_{u^{(1)}}$	−0.93	[−1.06, −0.79]	0.18	0.18	−0.72	[−0.90, −0.53]		0.23
$\delta_{u^{(1)},D_1}$					−0.37	[−0.61, −0.12]	−1.09	0.14
$\delta_{u^{(1)},D_2}$					−0.30	[−0.56, −0.04]	−1.02	0.15
$\mu_{c^{(2)}}$	0.10	[−0.13, 0.34]	0.53	0.52	0.17	[0.02, 0.32]		0.57
$\delta_{c^{(2)},D_1}$					−0.12	[−0.40, 0.14]	−0.29	0.39
$\delta_{c^{(2)},D_2}$					−0.31	[−0.59, −0.04]	−0.48	0.32
$\mu_{r^{(2)}}$	−0.06	[−0.25, 0.12]	0.47	0.50	0.28	[−0.08, 0.64]		0.61
$\delta_{r^{(2)},D_1}$					−0.55	[−0.98, −0.13]	−0.83	0.20
$\delta_{r^{(2)},D_2}$					−0.43	[−0.91, 0.06]	−0.71	0.24
$\mu_{u^{(2)}}$	−0.36	[−0.51, −0.21]	0.36	0.34	−0.35	[−0.55, −0.15]		0.36
$\delta_{u^{(2)},D_1}$					−0.13	[−0.40, 0.12]	−0.48	0.32
$\delta_{u^{(2)},D_2}$					−0.07	[−0.32, 0.17]	−0.42	0.34

Parameters were obtained with AGHQ with 4 nodes. Est = Parameter estimates; CI = confidence interval; TE = probability-transformed estimates; Bayes = Bayesian Estimates as reported in Table 2 in Klauer (2010). D1 = First dummy variable, that is, 1 for right-temporal epileptic patients (all other participants are coded zero). D2 = Second dummy variable, that is, 1 for left-temporal epileptic patients (all other participants are coded zero). $\Delta_{D.}$ is the estimate of the parameter for the group coded in the respective dummy variable (i.e., $\delta_{.,D.} + \mu_{.}$). For model 2, probability transformed estimates (column TE) indicate parameter means in the three groups (control group, right-temporal epileptics, left-temporal epileptics, respectively).

TABLE 5.
Parameter estimates of the covariance structure for the illustrative example with real data.

	Variances		Correlations					
	Est	Bayes	1.	2.	3.	4.	5.	6.
1. $c^{(1)}$	0.37	0.35	–	0.54	0.73	0.78	0.72	0.71
2. $r^{(1)}$	0.31	0.10	0.25	–	0.55	0.69	0.60	0.64
3. $u^{(1)}$	0.12	0.06	0.73	0.22	–	0.72	0.70	0.78
4. $c^{(2)}$	0.32	0.34	0.76	0.23	0.66	–	0.69	0.75
5. $r^{(2)}$	0.13	0.09	0.71	0.22	0.62	0.65	–	0.69
6. $u^{(2)}$	0.14	0.15	0.95	0.29	0.83	0.87	0.83	–

The first two columns present variance estimates based on the AGHQ method with 4 nodes (Est) and the corresponding Bayesian estimates (Bayes) as reported by Klauer (2010, Table 2), respectively. The correlation matrix displays the Bayesian estimates of Klauer (2010) above the diagonal and the corresponding AGHQ estimates using 4 nodes below the diagonal.

group on each of the model parameters, that is, $c^{(1)}, r^{(1)}, u^{(1)}$ for Trial 1 and $c^{(2)}, r^{(2)}, u^{(2)}$ for Trial 2. The results are also shown in Table 4. The results indicate that estimates for the $c^{(\cdot)}$ parameters were lower for epileptic patients compared with control patients whereas this pattern is less clear for the remaining parameters.

6. Discussion

The aim of this article was to describe how the parameters in the latent-trait MPT model can be estimated with a marginal ML approach. Specifically, we introduced three methods to approximate the integrals that are involved when the goal is to maximize the marginal log-likelihood function, and we investigated the statistical properties of these methods in a simulation study. Finally, we presented an empirical example that illustrated the suggested approaches.

The results of the simulation study showed that AGHQ and QMC performed well with regard to the relative bias and the coverage rate. However, we also found that AGHQ performed somewhat better than QMC in most simulation conditions. We would therefore recommend that researchers use AGHQ as the default method for parameter estimation but switch to QMC when the number of random effects becomes too large to approximate the integral with AGHQ in a reasonable amount of time. Strictly speaking, however, our results refer to the range of simulation conditions examined here only. Hence, further simulation research is needed to examine the performance of our ML approaches for other MPT models or, for example, models with low variance components. In these simulation studies, one could then also examine some additional approximation methods we ignored in the current study, such as a Monte Carlo EM algorithm (Booth & Hobert, 1999) , variational approximation (Ormerod & Wand, 2010) , or Laplace importance sampling (Kuk, 1999) . The latter method is a modification to the Laplace approximation which we dropped from further consideration because of unsatisfactory performance in our simulation study. Laplace importance sampling is an interesting alternative that is definitely worth to investigate.

So far, the parameters of the latent-trait MPT model can be estimated with a Bayesian approach only (Heck et al., 2018a; Klauer, 2010) . Our simulation study suggests that the maximum likelihood approach introduced here-specifically, the AGHQ approximation method-provides estimates that are at least as good as Bayesian estimates (provided the number of nodes is sufficiently high). For covariance parameters, in particular, both relative estimation bias and coverage rates of confidence intervals appear to be clearly superior for AGHQ-based ML estimates compared to their Bayesian counterparts. These results are important for empirical applications, especially those focusing on parameter correlations. In addition, the maximum likelihood approach has some pragmatic advantages compared to the Bayesian approach, for example, because prior distributions are not required, the convergence of the estimation algorithm is easier to determine, and the asymptotic optimality properties of ML-estimated parameters have been well-known in the statistical literature for decades. This does not mean that we are rejecting a Bayesian approach; rather, we believe that the two approaches complement each other and that there likely are situations in which one approach is preferable to the other (Wasserman, 2004) .

We believe that further simulation research is needed to determine the best-performing method for a variety of situations. For example, on the one hand, estimation of covariances between MPT parameters may turn out to be a specific strength of ML methods. On the other hand, the Bayesian approach may outperform ML methods when the number of participants and/or the number of responses is small. In fact, the results of our simulation study suggest these tentative interpretations. However, our results require replication and definitely need to be extended to other MPT models before they can provide a basis for general recommendations. Furthermore, another interesting question for future research is whether there are circumstances under which the two methods produce discrepant results concerning model comparisons. Finally, we also think

that it is interesting to investigate whether a combination of the two estimators (e.g., using the ML estimates as starting values for Bayesian MCMC estimation) has better asymptotic properties compared to each single approach alone.

Additionally, there are a number of further research questions that we think would be worthwhile to study. First, it would be interesting to extend the ML estimator proposed here to handle both random participant and random item effects (Matzke et al., 2015) . Implementing such a crossed random effects MPT model would be a challenging task for future research. Another challenging issue concerns recent extensions of MPT models to include continuous variables such as response times (Heck et al., 2018b; Klauer & Kellen, 2018) . These generalized MPT models could also be embedded in a hierarchical random effects framework and analyzed using the marginal ML methods proposed here.

A major problem involves convergence problems of marginal ML methods, for example, when the number of responses per participant is very small or when the true variance components are small. It would be interesting to investigate whether a penalized maximum likelihood estimator (Chung et al., 2013) can solve the convergence issues of the ML estimator. Finally, both the Bayesian approach and the ML approach proposed here assume that the random effects are multivariate normally distributed. From a statistical point of view, however, this assumption does not need to be true, and it would be interesting to examine how robust the two approaches are with respect to such a misspecification when the true underlying distribution is actually, for example, a finite mixture distribution (a reasonable assumption for the illustrative example). We note that one can specify arbitrary distributions for the random effects in ML with QMC sampling and that the implementation of these "robust" ML estimators would also be interesting for future research.

In summary, the present article shows how marginal maximum likelihood estimation can be used to obtain the parameters of a random effects MPT model with or without covariates. Using the pair-clustering model as a running example, we found for both simulated and real data that the ML approach is a reasonable alternative to Bayesian hierarchical MPT analyses that are based on the Latent-Trait Model (Heck et al., 2018a; Klauer, 2010) . Future research should extend these results to other, more complex MPT models and perhaps also explore alternative numerical methods of marginal ML estimation.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review, 87*, 375–397. https://doi.org/10.1037/0033-295X.87.4.375

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology, 39*, 129–149. https://doi.org/10.1111/j.2044-8317.1986.tb00852.x

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86. https://doi.org/10.3758/BF03210812

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

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*. https://doi.org/10.1007/BF02293801

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*, 265–285. https://doi.org/10.1111/1467-9868.00176

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78*, 685–709. https://doi.org/10.1007/s11336-013-9328-2

Coolin, A., Erdfelder, E., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2015). Explaining individual differences in cognitive processes underlying hindsight bias. *Psychonomic Bulletin & Review, 22*, 328–348. https://doi.org/10.3758/s13423-014-0691-5

Coolin, A., Erdfelder, E., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2016). Inhibitory control underlies individual differences in older adults' hindsight bias. *Psychology and Aging, 31*, 224–238. https://doi.org/10.1037/pag0000088

Crowther, M. J. (2017). Extended multivariate generalised linear and non-linear mixed effects models. arXiv:1710.02223

Engel, B. (1998). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal, 40*, 141–154. https://doi.org/10.1002/(SICI)1521-4036(199806)

Erdfelder, E., Auer, T., Hilbig, B., Assfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie-Journal of Psychology, 217*, 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Erdfelder, E., & Buchner, A. (1998). A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 387–414. https://doi.org/10.1037/0278-7393.24.2.387

Erdfelder, E., Hu, X., Rouder, J., & Wagenmakers, E. (2020). Cognitive psychometrics: Recent contributions in honor of William H. Batchelder (1940–2018). *Journal of Mathematical Psychology, 99*, 1–7. https://doi.org/10.1016/j.jmp.2020.102468

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 625–641. https://doi.org/10.1080/10705510903203573

González, J., Tuerlinckx, F., De Boeck, P., & Cools, R. (2006). Numerical integration in logistic-normal models. *Computational Statistics & Data Analysis, 51*, 1535–1548. https://doi.org/10.1016/j.csda.2006.05.003

Heck, D., Arnold, N., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50*, 264–284. https://doi.org/10.3758/s13428-017-0869-7

Heck, D., Erdfelder, E., & Kieslich, P. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika, 83*, 893–918. https://doi.org/10.1007/s11336-018-9622-0

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Springer.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59*, 21–47. https://doi.org/10.1007/BF02294263

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology, 27*, 116–159. https://doi.org/10.1080/10463283.2016.1212966

Kellen, D., & Klauer, K. C. (2020). Selecting amongst multinomial models: An apologia for normalized maximum likelihood. *Journal of Mathematical Psychology, 97*, 102367. https://doi.org/10.1016/j.jmp.2020.102367

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika, 71*, 1–31. https://doi.org/10.1007/s11336-004-1188-3

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75*, 70–98. https://doi.org/10.1007/s11336-009-9141-0

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology, 82*, 111–130. https://doi.org/10.1016/j.jmp.2017.12.003

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation, 63*, 143–158. https://doi.org/10.1080/00949659908548521

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12*, 605–621. https://doi.org/10.3758/BF03196751

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Matzke, D., Dolan, C., Batchelder, W., & Wagenmakers, E. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80*, 205–235. https://doi.org/10.1007/s11336-013-9374-9

Meiser, T., & Broder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(116–137), 1. https://doi.org/10.1037/0278-7393.28.1.116

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nestler, S. (2020). Modeling interindividual differences in latent within-person variation: The confirmatory factor level variability model. *British Journal of Mathematical and Statistical Psychology, 73*, 452–473. https://doi.org/10.1111/

bmsp.12196
Nestler, S. (2021). Modeling intraindividual variability in growth with measurement burst designs. *Structural Equation Modeling: A Multidisciplinary Journal, 28*, 28–39. https://doi.org/10.1080/10705511.2020.1757455
Nestler, S., & Egloff, B. (2009). Increased or reveresed: The effect of surprise on the hindsight bias depends on the hindsight component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1539–1544. https://doi.org/10.1037/a0017006
Nestler, S., Egloff, B., Küfner, A., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: The case of hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology, 103*, 698–717. https://doi.org/10.1037/a0029461
Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician, 64*, 140–153. https://doi.org/10.1198/tast.2010.09058
Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational Graphics and Statistics, 4*, 12–35. https://doi.org/10.2307/1390625
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, 20–22 March 2003, p. 1–10.
R Core Team. (2020). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria. http://www.R-project.org/
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal, 2*, 1–21. https://doi.org/10.1177/1536867X0200200101
Reed, T. R. C., & Cressie, N. A. C. (1988). *Goodness of fit statistics for discrete multivariate data*. New York: Springer.
Riefer, D., & Batchelder, W. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition, 23*, 611–630. https://doi.org/10.3758/BF03197263
Riefer, D., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*, 318–339. https://doi.org/10.1037/0033-295X.95.3.318
Robert, C., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York: Springer.
Sarafoglou, A., Kuhlmann, B. G., Aust, F., & Haaf, J. M. (2022). Theory-informed refinement of Bayesian hierarchical MPT modeling. https://doi.org/10.31234/osf.io/kvyt5
Schilken, E. (1998). *Speicherung und Abruf von verbalem Material bei Patienten mit Temporallappenepilepsie*. Universität Bonn: Unpublished Diploma Thesis.
Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). Tutorial on multinomial processing tree modeling: How to develop, test, and extend MPT models. *Psychological Methods*. https://doi.org/10.1037/met0000561
Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods, 45*(2), 560–575. https://doi.org/10.3758/s13428-012-0259-0
Singmann, H., Kellen, D., Gronau, Q., Mueller, C., & Bhel, A. S. (2020). *MPTinR: Analyze multinomial processing tree models [Computer software manual]*. https://CRAN.R-project.org/package=MPTinR (R package version 1.13-0).
Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*, 167–183. https://doi.org/10.1016/j.jmp.2009.06.007
Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology, 59*, 225–255.
Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer. https://doi.org/10.1007/978-0-387-21736-9
Wickelmaier, F., & Zeileis, A. (2011). Multinomial processing tree models in R. https://www.R-project.org/conferences/useR-2011/TalkSlides/Contributed/17Aug_1705_FocusV_3-Psychometrics_1-Wickelmaier.pdf (Presented at the R User Conference 2011, August 16–18, Coventry, UK).
Wickelmaier, F., & Zeileis, A. (2020). MPT: Multinomial processing tree models [Computer software manual]. https://CRAN.R-project.org/package=mpt (R package version 0.6-2).
Xu, M., & Bellezza, F. (2001). A comparison of the multimemory and detection theories of know and remember recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1197–1210. https://doi.org/10.1037/0278-7393.27.5.1197

Manuscript Received: 4 FEB 2022