

**Tutorial on Multinomial Processing Tree Modeling: How to Develop, Test,
and Extend MPT Models**

Oliver Schmidt¹, Edgar Erdfelder², and Daniel W. Heck¹

¹University of Marburg

²University of Mannheim

Author Note

OS and DWH were supported by the research training group Breaking Expectations (GRK 2271, project number 290878970-GRK 2271, project 4), funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). EE and DWH were supported by the research training group Statistical Modeling in Psychology (GRK 2277), also funded by the DFG, and the William K. and Katherine W. Estes Fund.

Data and R code for all analyses are available at the Open Science Framework: <https://osf.io/24pbm/>. This tutorial paper is based on pre-conference workshops at the 51st congress of the German Psychological Society (DGPs) in Frankfurt, Germany, (September 2018) and the 64th conference of experimental psychologists (64. Tagung experimentell arbeitender Psychog:innen, TeaP) in Cologne, Germany, (March 2022), supported by the Estes fund. The corresponding materials including video lectures are available at <https://github.com/danheck/MPT-workshop>.

Correspondence concerning this article should be addressed to Oliver Schmidt, Department of Psychology, University of Marburg, Gutenbergstraße 18 Room 01057, D-35037 Marburg, Germany. E-mail: oliver.schmidt@uni-marburg.de.

We are very thankful to Franziska Meissner for all our joint MPT workshops.

Abstract

Many psychological theories assume that observable responses are determined by multiple latent processes. Multinomial processing tree (MPT) models are a class of cognitive models for discrete responses that allow researchers to disentangle and measure such processes. Before applying MPT models to specific psychological theories, it is necessary to tailor a model to specific experimental designs. In this tutorial, we explain how to develop, fit, and test MPT models using the classical pair-clustering model as a running example. The first part covers the required data structures, model equations, identifiability, model validation, maximum-likelihood estimation, hypothesis tests, and power analyses using the software multiTree. The second part introduces hierarchical MPT modeling which allows researchers to account for individual differences and to estimate the correlations of latent processes among each other and with additional covariates using the TreeBUGS package in R. All examples including data and annotated analysis scripts are provided at the Open Science Framework (<https://osf.io/24pbm/>).

Keywords: Cognitive modeling, hypothesis testing, parametric order constraints, Bayesian hierarchical models

Tutorial on Multinomial Processing Tree Modeling: How to Develop, Test, and Extend MPT Models

One primary use of MPT models is as data-analysis tools, capable of disentangling and measuring the separate contribution of different cognitive processes underlying observed data.

Batchelder and Riefer ([1999](#))

Multinomial processing tree (MPT) models are statistical models for discrete data that allow researchers to disentangle and measure contributions of various latent cognitive processes to observable behavior (Batchelder & Riefer, [1990](#)). A main feature of MPT models is their conceptual and mathematical simplicity. Given an appropriate empirical paradigm, MPT models are thus ideally suited to develop simple formal measurement models of the psychological processes involved in this paradigm. In many areas of psychology, observable behavior is often complex and not determined by a single process but the outcome of multiple latent processes. For a better understanding of psychological phenomena and for testing theories about these phenomena properly, disentangling underlying latent processes is necessary.

MPT models are suitable for discrete data with a finite number of observable categories. Given that many studies in psychology collect discrete judgments or choices with a small number of response categories, MPT models are well suited for answering many substantive research questions. MPT models define one or more processing trees that encompass the possible process sequences linking a specific stimulus to different behavioral outcomes. The conditional probabilities of moving from one process to the next serve as free parameters that measure the contribution of unique latent processes to observable responses. Since their development (e.g., Batchelder & Riefer, [1980](#), [1986](#)), MPT models have continuously gained attention because of the many advantages they offer for psychological research. In the past three decades, MPT models have been applied in many sub-disciplines of psychology including areas such as attention and perception, learning, memory, judgment and decision making, social cognition, and many more (for reviews, see Batchelder & Riefer, [1999](#); Erdfelder et al., [2009](#); Hütter & Klauer, [2016](#)).

In this tutorial, we show how to develop an MPT model to measure latent psychological processes and how to use these measures to test psychological hypotheses. By explaining the common steps of MPT model applications in practice, we cover the process of model development, model application, and statistical analysis. Going beyond the traditional analysis of MPT models, this tutorial also introduces recent, advanced methods of MPT modeling for hierarchical data structures as commonly obtained in repeated-measures designs. Throughout the tutorial, we use the pair-clustering model developed by Batchelder and Riefer (1980, 1986) as a running example and illustrate step by step how to use suitable software for MPT modeling. In each section, we first introduce the relevant concepts (highlighted in bold) before illustrating their application with suitable software. For modeling aggregated data, we introduce the freely-available program **multiTree** (Moshagen, 2010) which provides a user-friendly interface and many functionalities useful for model development. Moreover, we explain how to use the **TreeBUGS** package in R for modeling hierarchically nested data (Heck, Arnold, et al., 2018). All data, model files, and R code are available at the Open Science Framework (<https://osf.io/24pbm/>).

1 Model Development and Application

1.1 Experimental Paradigm and Psychological Processes

To our knowledge, Batchelder and Riefer (1980) proposed the first MPT model within psychology, namely, the pair-clustering model.¹ In their pioneering publication, the authors were primarily interested in effects of semantic clustering on episodic memory, that is, how clustering of semantically related information during learning improves later memory for this information. They thus investigated free recall of previously studied word lists composed of 16 semantically related word pairs (e.g., apple and banana, rose and tulip) and 6 single words unrelated to all other words in the list (see Figure 1). Only one word was presented at a time during learning (4 sec/word), with words arranged in

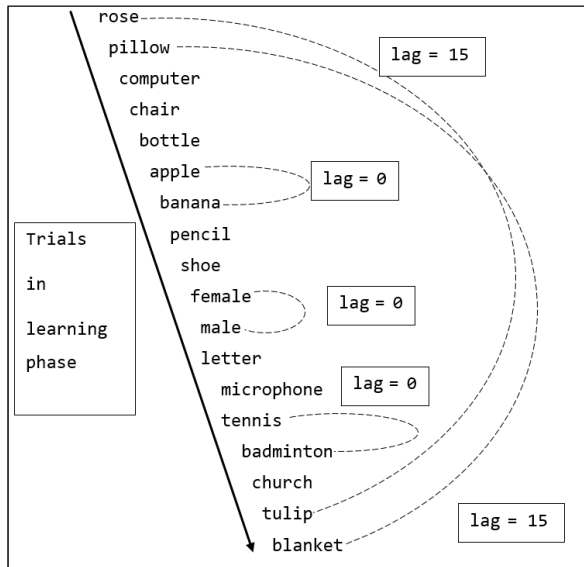
¹ Other disciplines such as statistical genetics previously used parameterized multinomial models employing the MPT model equation structure. Bernstein's (1925) famous AB0 blood-group model is a classical example.

random order except that the presentation lag of words from pairs was kept constant. Specifically, there were four word pairs assigned to each of four lag conditions: both words presented adjacently (lag 0), four words intervening (lag 4), 12 words intervening (lag 12), and 24 words intervening (lag 24). Somewhat surprisingly, they observed that recall memory did not benefit from short lags. Rather, both the proportion of recalled words and the proportion of recalled word pairs were unaffected by presentation lag (Batchelder & Riefer, 1980, p. 379).

Does this imply that semantic clustering has no impact on memory? This is clearly not the case, as revealed by a re-analysis of the same data using the pair-clustering model (Batchelder & Riefer, 1980, p. 382). Applying this model, the authors detected that the probability of cluster storage in fact strongly decreased with increasing presentation lag of word pairs, as predicted. In contrast, the probability of cluster retrieval strongly increased with presentation lag. As both effects are in opposite directions, they cancel out at the level of behavioral measures like recall rates that confound different underlying cognitive process. Behavioral measures may thus hide theoretically important effects that MPT modeling makes visible: Whereas short lags benefit cluster storage, long lags foster encoding variability (Bower, 1970; Melton, 1970) which in turn benefits cluster retrieval.

What are the key assumptions of the pair-clustering model? Batchelder and Riefer (1980) assumed that free-recall performance depends on three **latent cognitive processes**: clustering of semantically related information, retrieval of stored clusters, and retrieval of information that is not stored as a cluster. Clustering means that a word pair is stored as a cluster based on the semantic relation between the paired words.² The model assumes that clustering can only be applied to semantically related pairs but not to singletons. If two semantically related words are stored as a cluster, the retrieval process can either result in retrieving the cluster as a whole (i.e., recalling both words) or not retrieving the cluster at all (i.e., recalling none of the two words). If a word pair is not stored as a cluster, it is still possible to store and retrieve the two words independently (i.e., they are processed as if they were single words).

² Although not relevant for the research question addressed by Batchelder and Riefer (1980), clusters based on other types of relations could also be considered (e.g., rhyme relations).

Figure 1*Experimental Paradigm and Data Structure for the Pair-Clustering MPT Model.***(A) Experimental study list****(B) Observed frequencies in multiTree**

File Model Analysis Help	
Parameters Data Equations	
Data Tree Category Frequencies	
Title:	
old_lag0E1	42
old_lag0E2	5
old_lag0E3	63
old_lag0E4	290
old_F1	64
old_F2	336
young_lag0E1	90
young_lag0E2	14
young_lag0E3	84
young_lag0E4	212
young_F1	102
young_F2	298

Note. Panel A shows a study list presented to the participants. The pairs are separated by other words with lag 0 or lag 15. Panel B shows the observed frequencies in *multiTree* for the condition lag 0 separately for old and young participants (data of Bayen, 1990).

1.2 Data Structure and the Multinomial Distribution

In the free-recall paradigm, each participant provides a sequence of recalled words which includes both singletons and pairs in a mixed order. Since the free-recall paradigm does not supply a natural category structure, it is necessary to specify a **category system** suited for MPT modeling. The category system should encode all information relevant for modeling the psychological processes of interest. One can often define a category system in many different ways. Hence, this step usually requires researchers to try out different possibilities in order to select an approach that achieves a trade-off between informativeness (i.e., defining sufficiently fine-grained categories to disentangle the hypothesized latent processes) and simplicity (i.e., defining as few categories as possible).

For singletons, task performance can simply be represented by the frequency of recalled words (F_1) and non-recalled words (F_2). For word pairs, it would be possible to use three categories to distinguish whether both, one, or none of the two words were recalled in the test phase. However, to disentangle clustering and retrieval processes, it is

necessary to use a more fine-grained category system that encodes specific features of the exact sequence of recalled words.

Batchelder and Riefer (1986) proposed to define two separate categories for recalled pairs depending on the specific sequence of responses. By distinguishing between the adjacent and the non-adjacent recall of a word pair, one can disentangle successful storage and retrieval of paired clusters from the recall of unclustered individual words, respectively. Hence, we use four categories: whether both words of a pair are recalled adjacently (E_1), whether both words are recalled but with other words in between (E_2), whether only one of the two words is recalled (E_3), or whether none of the words is recalled (E_4). This category system requires a long list of words to ensure a sufficiently large number of observed frequencies for each response category. Moreover, with an increasing number of words, it becomes less likely that the independent recall of both words of a pair occurs adjacently merely by coincidence. In the data by Bayen (1990), which we analyze below, the category system of four word pair categories and two singleton categories is extended to a **factorial design** with the between-subjects condition *age* of the participants (young versus older) and the within-subjects factor *lag* between paired words (lag 0 versus lag 15; see Figure 1).

In general, the specification of an **informative category system** is an important step in developing new MPT models that offers researchers a lot of flexibility. If participants provide multiple discrete responses within each trial of an experiment, one can define a category system by all combinations of the possible responses (i.e., a contingency table). For instance, in source memory, participants have to learn words from two sources and then have to classify studied words not only as “old” or “new,” but also as stemming from “Source A” or “Source B.” MPT models for this paradigm focus on three of the four possible combinations of these two responses: “old – Source A,” “old – Source B,” and “new” (Batchelder & Riefer, 1990, if a word has been classified as new, responses about the source are usually omitted). In a similar vein, MPT models may also incorporate information about continuous variables such as response times by using discrete bins (e.g., “fast” versus “slow” responses, Heck and Erdfelder, 2016; Hilbig et al.,

2011)).³ Even multivariate quantitative judgments can be mapped on discrete response categories by looking at the set of possible rank orders of these judgments (e.g. Dehn & Erdfelder, 1998; Erdfelder & Buchner, 1998).

Given a specific category system, MPT models assume that responses follow a multinomial distribution for each condition. The multinomial distribution extends the binomial distribution to random variables with more than two response categories. For instance, in the free-recall paradigm, we assume that all singletons are stored and recalled independently. Thus, the frequencies n_1 of correctly recalled singletons and n_2 of non-recalled singletons are modeled by a binomial distribution with success probability p_1 . Similarly, the four observable categories for pairs (i.e., E_1 to E_4) result in the observed frequencies n_1, n_2, n_3, n_4 which follow a multinomial distribution with probabilities p_1, p_2, p_3, p_4 . When assuming that observations for singletons and pairs are independent, the six categories follow a joint multinomial distribution, meaning that the probability mass functions of the binomial distribution for the categories F_1 and F_2 and the multinomial distribution for E_1 to E_4 are simply multiplied (see Hu & Batchelder, 1994, for details).

1.3 Parameters and Tree Structure

As outlined in the introduction, the main goal of MPT modeling is to disentangle latent processes based on observable responses. For this purpose, the probabilities of hypothesized processes are represented as **parameters** that jointly determine the probability of observing a specific observation. If a model fits well and has been validated empirically (see Section 1.6), estimates for these parameters can be interpreted as measures of the postulated cognitive processes (Batchelder & Riefer, 1999). More precisely, an MPT model assumes $s = 1, \dots, S$ free parameters θ_s that represent the probabilities of a participant entering a specific latent state (e.g., whether clustering or retrieval of a word pair succeeds). Since the free parameters θ_s are defined as

³ In some paradigms such as the implicit association test (IAT), it may be necessary to implement a deadline for responding to obtain a sufficient amount of variance in the response frequencies. Without a deadline, error rates may be close to zero due to ceiling performance, thus rendering the data non-informative (Meissner & Rothermund, 2013; Nadarevic & Erdfelder, 2011)

probabilities, they must be in the interval $[0, 1]$. In turn, this implies that the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_S)$ must be in the S -dimensional parameter space $\Omega = [0, 1]^S$.

The parameters θ_s are combined in a **binary probability tree** that specifies the model-predicted probabilities for the observed response categories. Such a processing tree always contains two or more terminal nodes (i.e., the observed response categories) and, depending on the underlying paradigm and theory, a various number of intermediate nodes (i.e., latent states). Each non-terminal node leads to one of two consecutive states that refer to a latent process being successful or not with probabilities θ_s and $(1 - \theta_s)$, respectively. In MPT models, a category may be reached by different combinations of latent processes meaning that different processes can result in the same responses (for details, see Appendix A).

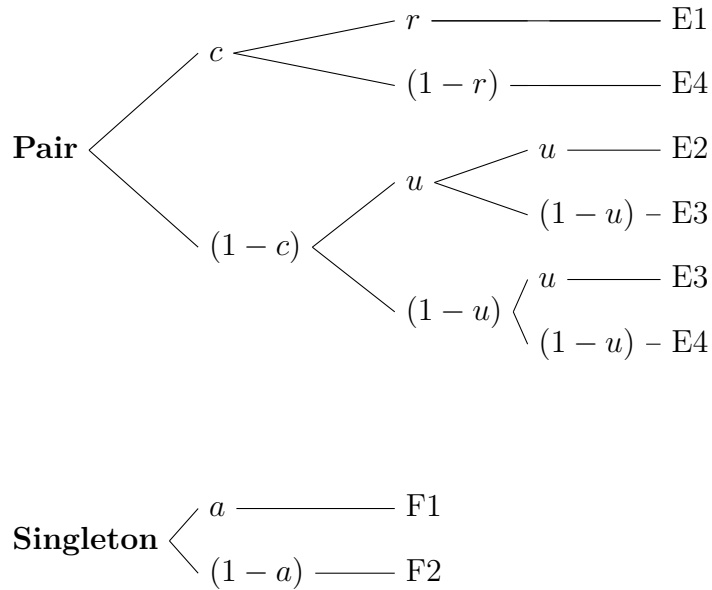
In the present example, we define separate processing trees for word pairs and singletons as shown in Figure 2. For singletons, it is assumed that words cannot be clustered because the study list does not contain semantically related words. Hence, the parameter a is defined as the probability of storing and recalling singletons independently. This process succeeds with probability a , resulting in a F_1 response, and fails with probability $1 - a$, resulting in a F_2 response.

For pairs, the MPT model assumes that multiple processes determine observable responses jointly. These processes are represented by the probability c of storing a pair as a cluster, the conditional probability r of successfully retrieving a stored cluster, and the probability u of storing and retrieving non-clustered words of a pair independently. First, with probabilities c and $1 - c$, a pair is stored as a cluster or not, respectively. In turn, pairs that have been stored as a cluster can either both be retrieved or not, resulting in recalling both words adjacently (E_1) with probability r or recalling none of both words (E_4) with probability $1 - r$. If a pair is not stored as a cluster with probability $1 - c$, the two items of the pair can be stored and retrieved independently. Hence, the first word is recalled or not with probabilities u and $1 - u$, respectively. Independently, the second item is recalled or not with probability u and $1 - u$, respectively (see Figure 2).

When developing a new MPT model, researchers should always start with a

Figure 2

Pair-Clustering Model by Batchelder and Riefer (1986).



Note. The parameters c and r are the probabilities for storing and retrieving a word pair as a cluster, respectively. Parameters u and a refer to the probabilities of recalling an unclustered single word for pairs and singletons, respectively.

simple model that considers as few latent processes as possible. This facilitates model development, often helps to ensure identifiability, and simplifies model validation. During model development, any auxiliary assumptions need to be kept in mind and should be explicitly written down. Doing so facilitates the communication of underlying assumptions and the generation of new ideas for further model adjustments. For instance, the pair-clustering model assumes that different pairs and singletons are recalled independently. Moreover, if clustering does not succeed, it is assumed that the two words of a pair are stored and retrieved independently with the same probability u . Obviously, for these assumptions to make sense, it is necessary to preselect stimuli carefully, so that pairs and singletons are unrelated and the two words of a pair are equally difficult. When developing a new model, it might be necessary to make additional auxiliary assumptions to render a model identifiable (e.g., by means of equality constraints; see Section 1.5) or to extend the model to additional experimental conditions using separate parameters (see Section 1.6).

1.4 Model Equations

Based on a processing tree, the predictions of an MPT model can be formalized mathematically. The **model equations** specify the model-implied probabilities of the observable response categories given the parameters θ_s . The pair-clustering model predicts that a singleton can either be recalled or not as represented by the categories F_1 and F_2 , respectively. Figure 2 shows that the model-implied probabilities for these categories are simply

$$P(F_1|\text{Singleton}) = a$$

$$P(F_2|\text{Singleton}) = (1 - a).$$

Moreover, the processing tree for pairs implies that the corresponding model equations are

$$P(E_1|\text{Pair}) = c \cdot r$$

$$P(E_2|\text{Pair}) = (1 - c) \cdot u \cdot u$$

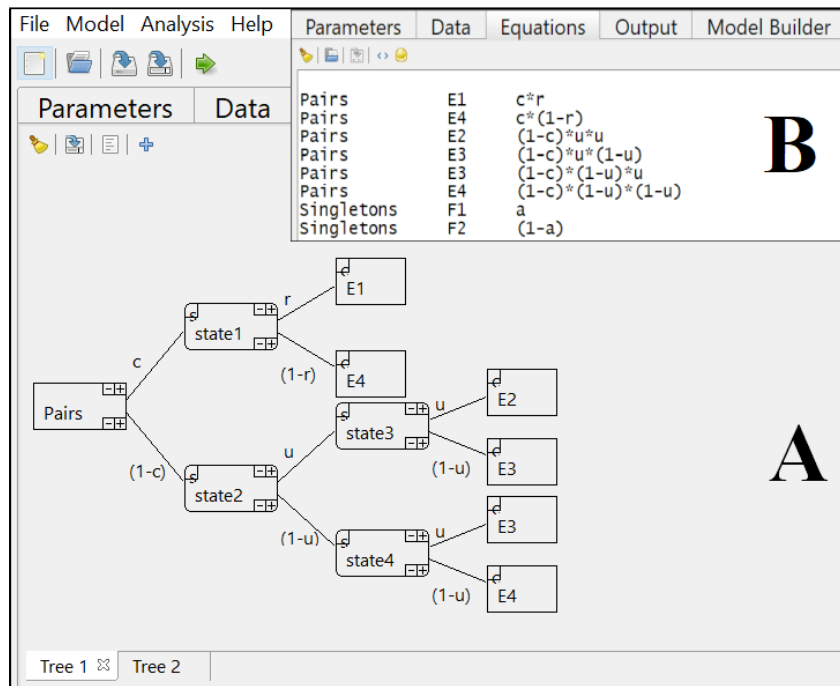
$$P(E_3|\text{Pair}) = (1 - c) \cdot u \cdot (1 - u) + (1 - c) \cdot (1 - u) \cdot u$$

$$P(E_4|\text{Pair}) = c \cdot (1 - r) + (1 - c) \cdot (1 - u) \cdot (1 - u).$$

The model predicts that a pair is stored as a cluster and retrieved successfully with joint probability $c \cdot r$ which gives the probability of adjacently recalling both words (E_1). Whereas there is only a single branch leading to category E_1 , there are two possible branches of recalling only one word of a pair (E_3). In this case, the model assumes a joint probability of $(1 - c) \cdot u \cdot (1 - u)$ that a word pair was not clustered and that the first but not the second word was recalled and a probability of $(1 - c) \cdot (1 - u) \cdot u$ for the reverse recall pattern. Since these two possibilities are mutually exclusive, the total probability of recalling only one word of a pair (E_3) is the sum of the joint probabilities of these two branches. Similarly, the last category of not recalling any word of a pair (E_4) is reached by two branches in the processing tree. Either the pair is stored as a cluster but not retrieved, $c \cdot (1 - r)$, or the pair is not stored and the independent recall of both word fails,

Figure 3

Implementation of the Pair-Clustering MPT Model in *multiTree*



Note. Model implementation is possible by either using the multiTree graphical model builder (A) or by clicking on the „Equations“ tab to specify the EQN file directly (B).

$(1 - c) \cdot (1 - u) \cdot (1 - u)$. Again, the total probability is given by the sum of these two joint probabilities. Figure 3 shows how to implement the pair-clustering model in *multiTree*.

When interpreting MPT models, it is important to keep in mind that only the model equations are relevant for fitting and testing the model. The visualization as a processing tree merely serves as a convenient illustration of the model. In fact, this representation is not always unique, meaning that for some MPT models, identical probabilities can be represented visually by different probability trees. For instance, in the pair-clustering model, the independent storage and retrieval of non-clustered words does not distinguish which of the two words is recalled with probability u and which is not. The model only tests whether the probabilities of recalling both, one, or none of the two words conditional on clustering failure are u^2 , $2u(1 - u)$, and $(1 - u)^2$, respectively. In particular, it is not adequate to interpret the order of the processes in an MPT tree as a temporal sequence of underlying latent processes. As a remedy, one may rely on extensions of MPT model for response times (see Section 2.7).

1.5 Model Identifiability

Model identifiability deals with the question whether it is possible to obtain unique estimates for the parameters of an MPT model. This topic is of special relevance for model development and for the extension of existing MPT models to new paradigms. A model is identifiable if it predicts different data for different values of the parameters. More precisely, an MPT model is identifiable if each parameter vector $\boldsymbol{\theta}$ generates unique predictions for the observed response probabilities \boldsymbol{p} (Bamber & van Santen, 2000). Technically, an identifiable MPT model thus defines a one-to-one mapping $\boldsymbol{\theta} \mapsto \boldsymbol{p}$ of the parameters to the category probabilities \boldsymbol{p} , meaning that every point in the parameter space Ω maps to a different point in the data space P (i.e., the space of all possible category probabilities). Since one-to-one mappings can be inverted, it is possible to obtain unique parameter estimates $\hat{\boldsymbol{\theta}}$ given model-consistent relative category frequencies $\hat{\boldsymbol{p}}$ by means of parameter estimation (i.e., $\hat{\boldsymbol{p}} \mapsto \hat{\boldsymbol{\theta}}$). If the MPT model is not identifiable, such an inversion does not provide unique parameter estimates, meaning that, for the same observed frequencies, different estimates can be obtained that account for the observed responses equally well.

In MPT modeling, two types of identifiability are distinguished. **Global identifiability** means that a model provides a one-to-one mapping for *all* possible parameter vectors $\boldsymbol{\theta}$ in Ω . In contrast, **local identifiability** means that the mapping is one-to-one only in the neighborhood of a specific parameter vector $\boldsymbol{\theta}_0$ in Ω . Practically, this means that a locally but not globally identified model is applicable only for a restricted set of parameter values close to $\boldsymbol{\theta}_0$. If a model is globally identifiable, it is also locally identified at all points in the parameter space. However, if local identifiability is achieved, this does not logically imply that a model is also identified globally. Usually, it is highly complex and time-consuming to prove global identifiability analytically (for an example, see Meiser & Bröder, 2002), but several methods are available for checking whether an MPT model is at least locally identifiable.

Checking the identifiability of an MPT model.

First, the number of free parameters (S) must not exceed the number of free categories (J) provided by the category system, that is, $S \leq J$. For discrete data, J is defined as the maximum number of logically independent frequencies which could vary freely in a given data set. Hence, an MPT model can only be identifiable if the **degrees of freedom** $df = J - S$ are non-negative (i.e., $df \geq 0$). For the pair-clustering model, the number of free categories is $J_1 = 2 - 1 = 1$ for singletons and $J_2 = 4 - 1 = 3$ for pairs. In total, the category system has $J = 1 + 3 = 4$ and thus enables the estimation of a maximum of $S = 4$ parameters. The pair-clustering model in Figure 2 has the parameters c , r , u , and a and satisfies this requirement. The software multiTree (Moshagen, 2010) computes the number of free categories, the number of parameters, and the degrees of freedom of an MPT model automatically. If a model features too many parameters ($S > J$), it is overparameterized and cannot be identifiable (see Figure 4). Note, however, that the requirement $S \leq J$ is only a necessary but not a sufficient condition for identifiability.

A powerful technique for checking local identifiability uses the **Jacobian matrix** which is defined as the matrix of first partial derivatives of the model equations with respect to the parameters θ_s . The maximum rank R of the Jacobian matrix provides information about the identifiability of an MPT model. If $R < S$, then the model can neither be locally nor globally identifiable; if $R = S$ the model is at least locally identifiable but not necessarily globally identifiable. The software multiTree computes the maximum rank of the Jacobian across random locations of the parameter vector θ via the option “Identifiability” in the “Analysis” menu. Applying this method to the pair-clustering model, we see that the maximum rank of the Jacobian is $R = 4$, meaning that the model is locally identifiable.

Constructing an identifiable model.

Since the pair-clustering model is identifiable, we use a modified model version to illustrate how to deal with non-identifiability in practice. For this purpose, we assume

that the process of recalling a single word of a pair differs between the first and the second word. This means that we distinguish between u_1 and u_2 thus increasing the number of parameters to $S = 5$. This assumption merely serves as an illustration and is neither theoretically motivated nor practically useful. Figure 4A shows that the modified model includes more model parameters than free categories and thus cannot be identifiable. In addition, Figure 4B illustrates the consequences of this issue: in a repeated analysis of the same data, we obtain different parameter estimates which all lead to an equally good fit.




Two approaches are usually applied to render an unidentifiable MPT model identifiable. First, if a model has too many parameters (e.g., if $J < S$) one may implement **equality constraints** based on theoretically justified assumptions. An equality constraint restricts two parameters to be equal based on theoretical assumptions, thus reducing the number of free parameters S . For the modified model, we may assume that the probability of recalling a non-clustered word of a pair does not depend on the order in which the words have been presented. Hence, we restrict the two model parameters to be equal, $u_1 = u_2$. The resulting model has only $S = 5 - 1 = 4$ free parameters resulting in one additional degree of freedom and in turn making the model identifiable (see Figure 4C). Alternatively, one can obtain an identifiable model by fixing a parameter to a **constant value** (e.g. $u_1 = .50$; see Figure 4D). However, often it is difficult to justify such a decision theoretically.

To obtain an identifiable MPT model, researchers can also **extend the category system** by adding new experimental conditions while including only few additional free parameters. Often, factorial designs are used to create more information (i.e., a larger number of free categories J). If one assumes that only some but not all of the parameters vary across the new experimental conditions, the increase in J may outweigh the increase in the number of parameters S . For instance, in the free-recall paradigm, manipulating the lag between word pairs (e.g., 0 versus 15) would double the number of categories for pairs (e.g., we obtain $E_{1,0}$ and $E_{1,15}$, with the second index indicating the lag). When fitting the pair-clustering model to these data, we could estimate separate sets of four parameters for each condition. However, it is plausible to assume that this manipulation

Figure 4*Identifiability Checks in multiTree (based on Data of Bayen, 1990)***(A) Unidentifiable MPT model**

Parameters	Data	Equations	Output	Model Builder
Hierarchical Model Families				
<input type="checkbox"/> Define current model as new baseline model (needs to be estimated before it can serve as a baseline).				
<input type="checkbox"/> Compare current model against baseline model				
a	free	0.16	Specification	
c	free	0.333	Number of trees	2
r	free	0.314	Number of categories	6
u1	free	0.137	Number of free categories	4
u2	free	0.136	Number of parameters	5
			Number of constrained parameters	0
			Degrees of freedom	-1

(B) Repeated analysis of the same data

Parameters	Data	Equations	Output	Model Builder		
<div>  </div>						
Run	a	c	r	u1	u2	Fit
1	0.16000	0.55309	0.18984	0.08705	0.32131	0.00000
2	0.16000	0.54344	0.19321	0.08776	0.31197	0.00000
3	0.16000	0.60964	0.17223	0.38416	0.08335	0.00000
4	0.16000	0.46371	0.22643	0.09503	0.24527	0.00000
5	0.16000	0.61429	0.17093	0.39007	0.08308	0.00000
6	0.16000	0.71930	0.14598	0.07780	0.57236	0.00000
7	0.16000	0.44323	0.23690	0.23030	0.09749	0.00000
8	0.16000	0.33386	0.31451	0.13657	0.13740	0.00000
9	0.16000	0.62890	0.16696	0.40953	0.08225	0.00000
10	0.16000	0.56599	0.18552	0.33436	0.08614	0.00000
Dev.	0.00000	0.38544	0.16853	0.33172	0.49011	

(C) Equality constraints of parameters

Parameters	Data	Equations	Output	Model Builder
Hierarchical Model Families				
<input type="checkbox"/> Define current model as new baseline model (needs to be estimated before it can serve as a baseline).				
<input type="checkbox"/> Compare current model against baseline model				
a	free	0.16	Specification	
c	free	0.333	Number of trees	2
r	free	0.314	Number of categories	6
u1	= u2	0.137	Number of free categories	4
u2	free	0.136	Number of parameters	5
			Number of constrained parameters	1
			Degrees of freedom	0

(D) Fixing parameters to constants

Parameters	Data	Equations	Output	Model Builder
Hierarchical Model Families				
<input type="checkbox"/> Define current model as new baseline model (needs to be estimated before it can serve as a baseline).				
<input type="checkbox"/> Compare current model against baseline model				
a	free	0.16	Specification	
c	free	0.333	Number of trees	2
r	free	0.314	Number of categories	6
u1	constant	0.5	Number of free categories	4
u2	free	0.136	Number of parameters	5
			Number of constrained parameters	1
			Degrees of freedom	0

Note. The software **multiTree** shows the model specification with the number of parameters and free categories. Panel A shows an unidentifiable MPT model with more parameters than free categories (resulting in $df < 0$). Panel B shows that the repeated analysis of an unidentifiable model leads to different parameter estimates for the same data. Panel C depicts a model version with the equality constraint $u_1 = u_2$ which renders the model identifiable. Panel D shows another solution for constructing an identifiable model by restricting a parameter to a constant value ($u_1 = .50$).

does not affect memory for single items (i.e., we can make the equality constraint $u_0 = u_{15}$). Reducing the number of free parameters in such a way may render a non-identifiable model identifiable. However, due to the extended experimental design, this approach may require additional data collection with a sufficient number of responses in each of the categories. Moreover, extending the experimental design can facilitate identifiability only if the number of additional parameters is strictly smaller than the degrees of freedom gained by the new categories. Hence, this approach does not necessarily ensure model identifiability.

1.6 Model Validation

MPT models are often used to formalize and test predictions of established psychological theories. Nevertheless, new MPT models need to be validated empirically.

Model validation usually focuses on testing the **construct validity** of the hypothesized latent processes and the corresponding parameters. Since each model parameter is supposed to represent a specific latent process, it is vital to examine whether parameters are selectively linked to the corresponding latent processes as assumed by the model. Once the construct validity of an MPT model and its parameters has been established, the model can be applied to answer substantive research questions such as the effect of manipulations or covariates on the postulated processes as measured by the parameters.

The substantive interpretation of MPT parameters is usually established by testing the **selective influence** on the model parameters (e.g. Bayen et al., 1996; Erdfelder & Buchner, 1998). For this purpose, it is necessary to apply an experimental manipulation which is assumed to affect only a specific underlying process but not the remaining processes. If a model parameter provides a valid measurement, it should be influenced by the corresponding experimental manipulation (i.e., convergent validity). Furthermore, no other parameters of the MPT model should be influenced by the specific experimental manipulation since this would indicate that the parameters do not disentangle the different hypothesized processes adequately (i.e., discriminant validity).

For instance, the clustering parameter c of the pair-clustering model should be selectively influenced by manipulating the instructions. Based on the idea of priming specific memory processes, participants are instructed either to learn as many words as possible or to explicitly cluster semantically related words. To establish the validity of the clustering parameter c , it should be significantly larger in the latter than in the former condition. Ideally, manipulating the instructions in such a way should not affect the remaining parameters. Since it is generally difficult to establish such theoretically strong manipulations which affect only a single parameter (especially if a model comprises many parameters), a weaker validation strategy focuses on testing whether theoretically targeted manipulations affect small subsets of parameters selectively, usually considered as core parameters of a model (e.g. Erdfelder et al., 2007; Erdfelder & Buchner, 1998).

Technically, the selective effect of a manipulation (i.e., convergent validity) is tested by restricting the corresponding MPT parameter to be equal across manipulation

conditions (see Section 1.8). Convergent validity is established when this test turns out to be significant. Moreover, discriminant validity is tested by restricting all remaining parameters to be equal while excluding the parameter that is assumed to be selectively influenced. A model has discriminant validity if its goodness of fit is not affected substantially by the latter constraint (see Section 1.7).

An alternative approach of assessing the construct validity of an MPT model focuses on the **correlation of individual parameters** with other psychological constructs or external criteria. This approach resembles the classic multi-trait multi-method approach of Campbell and Fiske (1959). On the one hand, the correlation between MPT parameters and theoretically related, external criteria should be high to indicate convergent validity (Bott et al., 2020). On the other hand, the correlation between MPT parameters and theoretically unrelated or irrelevant external criteria should be close to zero to indicate discriminant validity. However, this correlational approach is weaker than testing selective influence in a randomized experiment because the latter allows stronger causal conclusions regarding the effect of manipulations on MPT parameters. Moreover, the correlational approach requires individual parameter estimates on the person level, and thus, the application of hierarchical MPT models (see Section 2).

1.7 Parameter Estimation and Goodness of Fit

Statistical theory. Besides assessing construct validity, it is important that an MPT model provides a good description of the data. For this purpose, statistical goodness-of-fit tests assess whether there is a significant deviation between model predictions and the data. Multinomial model fitting relies on the **power-divergence statistic** PD^λ which quantifies the deviance between the model-implied and the observed category frequencies (Hu & Batchelder, 1994; Moshagen, 2010; Read & Cressie, 1988). The statistic PD^λ uses a calibration parameter λ that can be freely chosen by the researcher to specify how exactly the discrepancy between observed and model-implied frequencies is quantified. Common special cases can be derived by choosing specific values for λ . First, if $\lambda = 1$, the statistic PD^λ is identical to Pearson's X^2 statistic.

Second, if the limiting case $\lambda \rightarrow 0$ is used, the statistic PD^λ is identical to the **likelihood-ratio statistic** G^2 which is the commonly used default for MPT modeling (for further statistical details see Appendix B.1).

Model fitting proceeds by searching for a parameter vector θ that minimizes the statistic PD^λ for a given data set, given a specific fixed value of λ . As it turns out, minimizing $\text{PD}^{\lambda=0}$ (i.e., G^2) is equivalent to maximizing the likelihood of the parameters given the data (i.e., maximum-likelihood estimation). As closed-form equations for maximum likelihood parameter estimation are typically difficult to obtain and may also result in estimates outside $[0, 1]$, parameter estimation usually relies on an iterative estimation algorithm. For multinomial models, PD^λ can be minimized by using the expectation-maximization (EM) algorithm (Dempster et al., 1977; Hu & Batchelder, 1994). The EM algorithm starts with a random vector of parameter values. Then, the algorithm proceeds by alternating between two steps: The expectation step (E) computes the expected frequencies of the MPT branches given the current value of the parameters, whereas the subsequent maximization step (M) provides a new set of estimates for the parameters conditional on the branch frequencies estimated in the E-step.

Given a set of parameter estimates, we can test whether the model fits the data. The null hypothesis of the **goodness-of-fit test** states that the true category probabilities in the population are equal to the model-implied probabilities. Under the null hypothesis, the power-divergence statistic PD^λ asymptotically follows a χ^2 distribution with degrees of freedom equal to the difference between the number of free categories (J) and the number of free parameters (S), $\text{df} = J - S$ (this is the case for any choice of λ).⁴ A small value of PD^λ results in a large p -value indicating a good model fit, whereas a large value of PD^λ results in a small p -value indicating a bad model fit, respectively. Note that goodness of fit cannot be tested for saturated MPT models with $\text{df} = 0$.

Implementation in multiTree. We use the software `multiTree` (Moshagen,

⁴ Strictly speaking, this result only holds when certain regularity conditions are met (Read & Cressie, 1988). Most importantly, none of the parameters must lie on the boundary of the parameter space (i.e., be equal to 0 or 1).

2010) to estimate the parameters and to test the model's goodness of fit. Using the data of Bayen(1990) as our running example, we first fit a basic model version of the pair-clustering model using the responses of old participants in the condition with lag 0 only. This basic model has four parameters (c, r, u, a) to account for six observed categories $(E_1, E_2, E_3, E_4, F_1, F_2)$. This results in a saturated model with zero degrees of freedom, $df = 0$, meaning that model fit cannot be tested by a χ^2 test. However, it is still possible to obtain unique estimates of the model parameters.

The software `multiTree` allows users to specify several options for the analysis. Here, we use maximum-likelihood estimation (i.e., $\lambda = 0$) with the default settings of the EM algorithm. This means that the maximum number of iterations of the E- and M-steps is restricted to 5,000 and that parameter estimation is replicated two times using random starting values and a convergence criterion of $1.0E - 10$.⁵ The output of `multiTree` includes a section on model fit showing the overall goodness-of-fit statistic G^2 (labeled as `PD^lambda=0`), the degrees of freedom, and the corresponding p -value (see Figure 4). Furthermore, we obtain several model indices such as the log-likelihood value `ln(likelihood)` and the model-selection criteria AIC and BIC.⁶ The output also shows parameter estimates, standard errors, and confidence intervals. In our example, the probability of storing a word pair as a cluster is estimated to be $\hat{c} = .334$ with a 95% confidence interval of $[0; .840]$. The estimated probability of retrieving a stored pair is $\hat{r} = .314$ (CI: $[0; .795]$). The estimates of independently recalling a single word of a pair and of a singleton are very similar, $\hat{a} = .160$ (CI: $[\.124; .192]$) and $\hat{u} = .137$ (CI: $[\.029; .245]$), respectively.

1.8 Testing Equality Constraints

Statistical theory. Often, we are interested in testing whether two MPT parameters are equal, $\theta_1 = \theta_2$. To test such a hypothesis statistically, we can use a

⁵ For more complex MPT models, `multiTree` may indicate that the EM algorithm does not converge properly, in which case the user should opt for more iterations.

⁶ By default, `multiTree` also shows indices for AIC and BIC that compare the current model to the saturated unconstrained model (i.e., ΔAIC and ΔBIC).

Figure 5

Output for a Fitted MPT Model in *multiTree* (Based on Data of Bayen, 1990).

(A) Model fit and parameter estimates

File Model Analysis Help			
Parameters Data Equations Output Model Builder			
Estimation proceeded normally.			
Current model defined as baseline model.			
Model Fit			
PD λ lambda=0.0 (df=0)	=	0.00000	
ln(likelihood)	=	-502.14150	
AIC	=	1012.28300	
BIC	=	1031.02145	
Delta AIC	=	0.00000	
Delta BIC	=	0.00000	
Parameter Estimates, Standard Errors, and Confidence Intervals			
a	=	0.16000 (0.01833)	[0.12407 - 0.19593]
c	=	0.33290 (0.25815)	[-0.17206 - 0.83986]
r	=	0.31447 (0.24542)	[-0.16654 - 0.79548]
u	=	0.13699 (0.05493)	[0.02934 - 0.24465]

(B) Fitting and testing a nested model

File Model Analysis Help			
Parameters Data Equations Output Model Builder			
Difference to Baseline Model (Difference = Current - Baseline)			
PD λ lambda=0.0 (df=1)	=	0.14808	p = 0.70038
AIC difference	=	-1.85192	
BIC difference	=	-6.53653	
Ratio of AIC weights	=	0.71626	
Ratio of BIC weights	=	0.96332	
Baseline Current			
a	free	free	
c	free	free	
r	free	free	
u	free	= a	
Parameter Estimates, Standard Errors, and Confidence Intervals			
a	=	0.15793 (0.01738)	[0.12386 - 0.19200]
c	=	0.41563 (0.08733)	[0.24447 - 0.58680]
r	=	0.25263 (0.06107)	[0.13294 - 0.37231]
u	=	a	

Note. Since the model in Panel A is saturated ($df = 0$), it fits perfectly ($PD^{\lambda=0} = G^2 = 0$) but cannot be tested using the χ^2 distribution. Panel B shows the difference test ΔG^2 testing the restricted model against the baseline model. The restricted model does not fit the data significantly worse than the baseline model, $\Delta G^2(1) = 0.148$, $p = .700$.

likelihood-ratio test for nested models (more generally, we could again use the PD^{λ} statistic). Model A is nested if it can be obtained as a special case from a more general model B via equality constraints.⁷ The likelihood-ratio test assesses whether the more restrictive model fits the data significantly worse than the baseline model. Hence, we need to fit both models to the same data to obtain the test statistic $\Delta G^2 = G_A^2 - G_B^2$. If the more restrictive model holds, this difference test statistic asymptotically follows a χ^2 distribution with degrees of freedom corresponding to the difference in the degrees of freedom of the nested and the baseline model, that is, $df = df_A - df_B$. If ΔG^2 shows a significant deviation between both models, the nested model is rejected in favor of the more general model. Whenever possible, it is best practice to conduct ΔG^2 difference tests instead of checking the absolute model fit of a nested model using G^2 since the difference test usually has a higher statistical power (due to having smaller df).

The pair-clustering model in the most general version introduced above is not

⁷ More precisely, model A is nested in model B if the parameters of model A are a subset of the parameters of model B (Bamber & van Santen, 2000).

testable ($df = 0$). This is due to the fact that the probability of recalling a single word of a pair (u) is allowed to differ from that of recalling a singleton (a). However, the model was constructed with the idea in mind that these two probabilities should be identical if the model provides a valid psychological account of memory processes. The corresponding statistical hypothesis $H_0 : a = u$ leads to a nested model with one parameter less and is thus testable with $df = 1$.

Implementation in multiTree. In multiTree, equality constraints of the form $a = u$ can easily be added using the “Parameters” tab (see Figure 4). To test this constraint, we first have to specify the unrestricted model as a baseline model and fit it to the data. Second, we set the equality constraint $a = u$ and fit the restricted model. For our illustrative example, this analysis indeed shows that the restricted model does not fit significantly worse than the unrestricted model ($\Delta G^2(1) = 0.148, p = .700$). Hence, in line with the psychological assumptions underlying the model, we cannot reject the null hypothesis that the retrieval probability for single words in pairs and singletons differs. The restricted model yields almost the same parameter estimates as the saturated model ($\hat{a} = \hat{u} = .158$, CI: [.124 – .192]). This is in line with our observation that the two estimates were very similar in the baseline model. Moreover, the parameter estimates for c and r are also very similar to those of the unrestricted model but with smaller standard errors ($\hat{c} = .416$, CI: [.244; .587]; $\hat{r} = .253$, CI: [.133; .372]). As the nested model is more parsimonious than the unrestricted model, it is selected as a comparison standard for further tests.

Testing MPT parameters across groups

Statistical theory. Equality constraints on the parameters of an MPT model are often used to test differences in parameters between groups or experimental conditions. We illustrate such an application using an extended data set which includes the factor *age* (young vs. old). However, as before, we limit the analysis to singletons and word pairs presented with lag 0. First, we need to construct an extended MPT model for both age groups in which the two trees (pairs and singletons), the six categories ($E1$ to $F2$), and the four model parameters are doubled. Subsequently, the categories and

parameters must be renamed with labels indicating the age groups (e.g., a_y and a_o for the a parameter of young and older adults, respectively).

In evaluating sets of hypotheses concerning different MPT parameters (e.g., regarding c or regarding r), restrictions should always be tested against an established baseline model. Since parameter estimates and the corresponding test outcomes could differ depending on the sequence of implemented constraints, one should usually use a single baseline model. Using different baseline models is only suitable based on strong theoretical assumptions.

Implementation in *multiTree*. We first test the baseline model which again assumes that the retrieval and storage of singletons and non-clustered words within a pair does not differ. Implementing the assumptions $a_y = u_y$ and $a_o = u_o$ for both age groups leads to a good model fit, $G^2(2) = 0.155$, $p = .925$. The parameter estimates for young and older participants show only a small difference for the probability of recalling a pair as a cluster ($\hat{c}_y = .448$ [.328, .568]; $\hat{c}_o = .416$ [.244, .587]), and larger differences for the retrieval of clusters ($\hat{r}_y = .502$ [.359, .645]; $\hat{r}_o = .253$ [.133, .372]) and the independent retrieval of single items ($\hat{u}_y = .254$ [.214, .294]; $\hat{u}_o = .158$ [.124, .192]). This pattern of estimates suggests that age decrements in recall memory are primarily due to retrieval problems rather than to storage problems in older adults (see also Riefer & Batchelder, 1991). How can this be assessed statistically?

Starting with the baseline model ($df = 2$), we test whether storing a pair as a cluster differs between age groups, $H_0: c_y = c_o$. This requires us to fit both the baseline model and the restricted model ($df = 3$) and apply the ΔG^2 difference test for equality constraints. The nested model does not fit the data significantly worse than the baseline model ($\Delta G^2(1) = 0.095$, $p = .758$). As expected, the estimated probability of storing a pair as a cluster, $\hat{c}_y = \hat{c}_o = .437$ [.339, .535], is very similar to the separate estimates in the baseline model. We can thus maintain the assumption that the probability of clustering does not differ between age groups.

We also test whether the retrieval probability r differs between young and older participant by adding the equality constraint $H_0: r_y = r_o$ to the established baseline

model (while removing the constraints on the parameters c_y and c_o). Somewhat surprisingly given the strong discrepancy between \hat{r}_y and \hat{r}_o , the restricted model does not fit the data significantly worse than the baseline model, $\Delta G^2(1) = 3.766$, $p = .052$. Given this test outcome, we may argue that the probability of retrieving a word does not differ significantly between age groups, $\hat{r}_y = \hat{r}_o = .466$ [.343; .590]. However, the low statistical power of this test provides an explanation for this unexpected result (see Section 1.10).

Testing multiple MPT parameters at once

Building up on our previous example, we now test a set of equality constraints on multiple MPT parameters at once. We illustrate this in a 2×2 factorial design with the additional within-subjects factor *lag* (lag 0 versus lag 15). Again, this requires to use different labels for all trees, categories, and parameters such as $a_{y,0}$ and $a_{y,15}$. The resulting MPT model consists of six processing trees for pairs with lag 0, pairs with lag 15, and singletons for the young and the old age group (see `multiTree` file on OSF).

To implement the baseline model, we adopt the equality constraint $a = u$ for both lag conditions, $a_y = u_{y,0} = u_{y,15}$ and $a_o = u_{o,0} = u_{o,15}$. The resulting model has 10 free parameters and shows a satisfactory fit, $G^2(4) = 1.754$, $p = .781$ (see Table 1 provides an overview of parameter estimates of the baseline model). Using this baseline model, we test whether there is any effect of age on the storage parameter c . We implement this hypothesis by two additional equality constraints, $c_{o,0} = c_{y,0}$ and $c_{o,15} = c_{y,15}$. The restricted model does not fit significantly worse than the baseline model, $\Delta G^2(2) = 0.515$, $p = .773$. Hence, the probability c of storing a pair as a cluster does not differ significantly between age groups when tested in both lag conditions simultaneously.

1.9 Order Constraints and Interactions

Statistical theory. Order constraints express inequalities on the model parameters such as $\theta_1 \leq \theta_2$. For instance, in the pair-clustering model, retrieval of pairs could be predicted to be better for young than for old individuals so that $r_o \leq r_y$. For MPT models, we can add order constraints by reformulating the model structure and parameters, an approach called **reparameterization**. This section illustrates such an

Table 1

Comparison of Aggregated and Hierarchical Parameter Estimates (Based on Data of Bayen, 1990).

Age group	Parameter	Lag	Aggregated ML		Hierarchical Bayes	
			Estimate	95% CI	Estimate	95% ETI
Young	c	0	.440	[.321, .558]	.387	[.244, .520]
		15	.194	[.047, .340]	.208	[.148, .299]
	r	0	.512	[.366, .658]	.581	[.378, .864]
		15	.865	[.230, 1.00]	.841	[.569, .996]
	u		.250	[.214, .286]	.247	[.214, .282]
Old	c	0	.440	[.282, .597]	.379	[.189, .540]
		15	.110	[.000, .324]	.115	[.062, .221]
	r	0	.239	[.136, .342]	.299	[.149, .663]
		15	.682	[.000, 1.00]	.730	[.325, .992]
	u		.165	[.133, .198]	.155	[.125, .188]

Note. “Aggregated ML” refers to fitting the pair-clustering model to the aggregated response frequencies using maximum likelihood (with 95% CI being the confidence interval). “Hierarchical Bayes” refers to fitting the individual response frequencies using a hierarchical Bayesian, latent-trait MPT model (with 95% ETI being the equally-tailed credibility interval, see Section 2.3).

advanced application of MPT models. It may be skipped on first reading.

Knapp and Batchelder (2004) proposed two methods for implementing parametric order constraints in MPT models. We will only focus on the first approach which is the default used in practice and readily implemented in `multiTree`. Reparameterization of the order constraint $\theta_1 \leq \theta_2$ is based on the idea that the smaller parameter θ_1 is reduced by a multiplicative factor relative to the larger parameter θ_2 . If we refer to the factor by which the first parameter becomes smaller as the **auxiliary parameter** s , we obtain the reparameterization $\theta_1 = \theta_2 \cdot s$, with $0 \leq s \leq 1$. We can thus simply add s as a new parameter to the model and replace the parameter θ_1 by the product $\theta_2 \cdot s$ whenever it occurs in the model equations.

Importantly, the reparameterization does not change the total number of parameters. Although the model with order constraints is more restrictive than the model without order constraint, their goodness of fit tests thus have the same *dfs*. Note

that in extreme cases, s may approach the boundaries of the parameter space (i.e., 0 or 1) so that the regularity conditions for the asymptotic χ^2 distribution of the PD^λ goodness-of-fit statistic would no longer hold. In such cases, the parametric bootstrap can be used to approximate the distribution of PD^λ under H_0 (see Appendix B.4). This notwithstanding, since s is just another, standard MPT parameter in the model, the order constraint results in a new MPT model that can be analyzed like any other MPT model (Knapp & Batchelder, 2004).

In MPT modeling, order constraints can be used to test **interactions of multiple factors** (Kuhlmann et al., 2019). In the pair-clustering model, we might want to test whether the probability r of retrieval varies as a function of both lag and age. To disentangle the two main effects and the interaction, it is necessary to reparameterize the four r parameters of the baseline model. Like any MPT parameter, the probability parameters r are restricted to the range $[0, 1]$. This implies that the standard additive model underlying a 2-factorial ANOVA does make little sense in the MPT context because strong main effects would inevitably induce an interaction effect. Instead, we therefore rely on a **log-linear model** of the probabilities, which decomposes the *logarithm* of the parameters by an additive structure:

$$\log r_{y,0} = \log r_{y,0}$$

$$\log r_{o,0} = \log r_{y,0} + \log s_0$$

$$\log r_{y,15} = \log r_{y,15}$$

$$\log r_{o,15} = \log r_{y,15} + \log s_{15}.$$

The interaction is absent if the two auxiliary parameters s_0 and s_{15} are equal, $\log s_0 = \log s_{15}$, meaning that the difference in $\log r$ between old and young is identical for lag 0 and 15. Hence, the four r parameters can be represented only by the main effects of age and lag.

Instead of testing the interaction in a log-linear model directly, we will use the equivalent approach of implementing order constraints. For this purpose, in line with the

literature on cognitive aging (e.g., Riefer & Batchelder, 1991), we assume that the retrieval probability of clustered pairs (r) is higher for young than for old participants. We implement these two order constraints by using the two auxiliary parameters s_0 and s_{15} as follows:

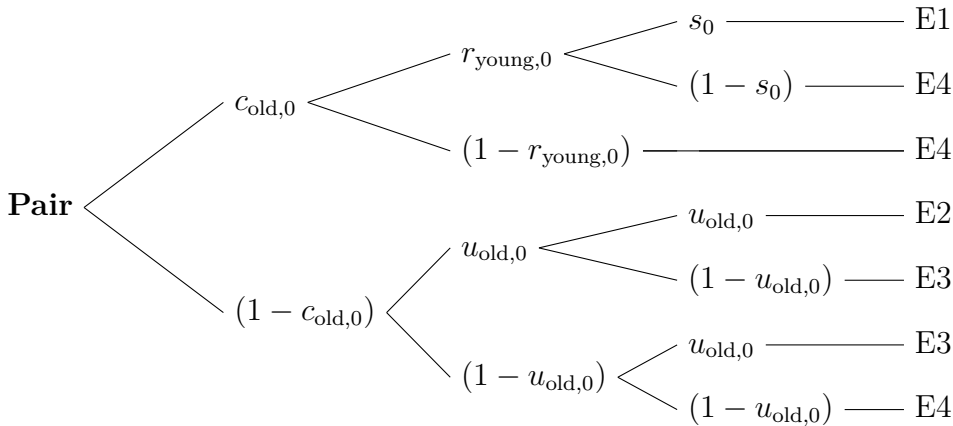
$$r_{o,0} = r_{y,0} \cdot s_0$$

$$r_{o,15} = r_{y,15} \cdot s_{15}.$$

These two equations are obtained from the log-linear model by exponentiation. Since the auxiliary parameters s_0 and s_{15} are probabilities, the parameters $r_{o,0}$ and $r_{o,15}$ cannot exceed $r_{y,0}$ and $r_{y,15}$, respectively. Moreover, Figure 6 shows that the resulting, reparameterized model is again an MPT model. Given an interaction between lag and age, the relative decrease of the r parameter for old compared to young participants will differ between the conditions lag 0 and lag 15. We thus need to test whether the two auxiliary parameters differ, $H_0 : s_0 = s_{15}$. This shows that the reparameterization by order constraints is equivalent to using a log-linear model on the logarithm of the parameters.

Figure 6

Reparameterization of the Pair-Clustering Model for Older Participants and Lag 0.



Note. The MPT model implements the order constraint $r_{o,0} \leq r_{y,0}$ via the reparameterization $r_{o,0} = r_{y,0} \cdot s_0$. Implementation of the same order constraint for the lag 15 condition proceeds analogously. The processing trees for young participants are not affected by these reparameterizations and thus not shown.

Implementation in multiTree. First, we test whether the two factors lag and

age have any effect on the parameter r , that is, $H_0 : r_{y,0} = r_{y,15} = r_{o,0} = r_{o,15}$. The ΔG^2 difference test for our illustrative data shows that the restricted model fits the data significantly worse than the baseline model ($\Delta G^2(3) = 13.556, p = .004$). Hence, the factors lag or age both affect the retrieval of clustered items. We implement the order constraints on the retrieval parameters r in **multiTree** by clicking on the button labeled “< >” in the “Equations” tab. Thereby, researchers can implement multiple order constraints on the parameters by adding auxiliary parameters as explained above. To test the interaction between lag and age, we need to compare the order-constrained model which assumes that the relative decrease of the parameter r in the old compared to the young group is identical in both lag conditions (i.e., $s_0 = s_{15}$, resulting in $df = 5$) against a baseline model without this assumption ($df = 4$). The nested likelihood ratio test shows that the restricted model does not fit the data significantly worse than the baseline model, $\Delta G^2(1) = 0.413, p = .521$. As the two s parameters do not differ significantly, there is no evidence in favor of an interaction effect of lag and age on the retrieval parameter r .

1.10 Power Analysis

Statistical theory. The statistical power $1 - \beta$ of a hypothesis test is defined as the probability of correctly rejecting the null hypothesis. For MPT models, power analysis is conducted with respect to the PD^λ statistic.⁸ The relevant quantities for determining power are the significance level α , the effect size w for χ^2 tests (Cohen, 1992), and the sample size N . An **a priori power analysis** is commonly used to gain information for designing a study before data collection. For this purpose, one computes the sample size N required for correctly rejecting the null hypothesis H_0 with a desired statistical power $1 - \beta$. Moreover, one needs to specify a significance level α and the expected effect size w under H_1 based on previous studies or theoretical considerations. Alternatively, a **post-hoc power analysis** can be conducted if the sample size is already given. Thereby, one computes the achieved statistical power to detect the expected effect size w under H_1 on the significance level α given a specific sample size N .

⁸ As mentioned above, λ is usually set to zero, implying that PD^λ is identical to the likelihood-ratio statistic G^2 .

Often, power analysis is based on **default effect sizes** that are considered to be “small,” “medium,” or “large.” In MPT modeling and other types of parameterized multinomial models, this common approach is of limited use because it is often not possible to provide general guidelines how the effect size w translates into the parameter differences under H_1 that are of primary interest (Erdfelder et al., 2005). Moreover, it is important to consider the proportion of observations per condition which is ignored when relying on default values for the expected effect size w .

MPT models offer an alternative strategy for specifying the **expected effect size** w in a power analysis based on theoretical considerations. First, plausible values need to be specified for all MPT parameters under the alternative and the null hypothesis H_1 and H_0 , respectively. This is usually quite easy because the parameters have an intuitive interpretation as probabilities of entering different cognitive states. Second, the relative number of observations N_k for each tree k needs to be defined. Based on this input, it is possible to derive the corresponding effect size w which can then be used to perform a power analysis. Technically, the effect size is calculated by fitting the nested model corresponding to H_0 to the expected frequencies under the alternative hypothesis H_1 by minimizing the G^2 statistic (Erdfelder et al., 2005; Moshagen, 2010). If the calculated discrepancy between H_0 and H_1 is small, this corresponds to a small value of the effect size w and vice versa. Once the effect size w has been obtained, it can be used for a priori or post-hoc power analyses as usual.

Importantly, statistical power in MPT modeling does not only depend on sample size but also on the experimental design and the location of the test-relevant MPT parameters in the probability tree (Heck & Erdfelder, 2019). Power is generally higher for MPT parameters that occur (a) in trees with large numbers of observations, (b) in multiple branches of the model, and (c) near to the root of the probability tree (e.g., parameters referring to unconditional probabilities). This is due to the fact that the conditional parameter structure of MPT models naturally reduces the number of observations available for estimating parameters on lower levels of the tree structure. Hence, the power to detect even large effects may be rather low for parameters which

only occur at the end of branches. For instance, the probability r of retrieving both words of a stored cluster can only be estimated precisely when the number of stored clusters is sufficiently large (i.e., when the parameter c is large). Correspondingly, in Section 1.8, the power to detect an effect of age on retrieval may thus have been rather low, meaning that the outcome $p = .052$ may be uninformative. Often, one can increase the power of a hypothesis test by optimizing the experimental design, for instance, by ensuring a high probability of clustering (e.g., $c \geq .80$; Heck & Erdfelder, 2019).

Implementation in multiTree. We illustrate how to perform an a priori power analysis in `multiTree` for testing the retrieval parameter r between age groups (Section 1.8). For the parameters of the baseline model under H_1 , we assume a relatively large difference in the retrieval probabilities for young and old participants in the underlying population, $r_y = .50$ and $r_o = .25$. All other parameters were assumed to resemble the parameter estimates under H_0 , that is, $c_0 = .236$, $c_y = .471$, $u_o = .135$, $u_y = .260$. When using a significance level of $\alpha = .05$, the power analysis shows that detecting the specified medium effect with a power of $1 - \beta = .80$ requires 3,568 observations per tree which corresponds to a total number of observations of $N = 14,270$. Hence, the number of observations in our data set ($N = 1,600$) was clearly too small to detect a difference of .25 in r between age groups.

As a second example, we perform a post-hoc power analysis in `multiTree` to detect an interaction of lag and age with respect to the retrieval parameter r . We again use the order-constrained, reparameterized model from Section 1.9 ($df = 6$) to test the null hypothesis $H_0 : s_0 = s_{15}$ ($df = 7$) which corresponds to a test of the lag \times age interaction. The χ^2 difference test of the null hypothesis thus has $df = 7 - 6 = 1$. Concerning the test-relevant parameters, we assume a large interaction effect of $s_0 = .70$ versus $s_{15} = .30$ which corresponds to retrieval probabilities of $r_{y,0} = .512$ and $r_{o,0} = .512 \cdot .70 = .358$ versus $r_{y,15} = .999$ and $r_{o,15} = .999 \cdot .30 = .300$, respectively. Concerning the remaining MPT parameters, we adopt the parameter estimates from the baseline model. Given $n = 400$ observations per tree as in our data set and using $\alpha = .05$, the power is $1 - \beta = .922$ to detect a large effect of the interaction of lag and age.

However, if we assume only a medium interaction effect of $s_0 = .60$ versus $s_{15} = .40$, the power reduces to $1 - \beta = .443$. Hence, medium or small effects may remain undetected given the relatively small sample size, meaning that the non-significant test in the previous section provides only limited evidence for the null hypothesis.

2 Hierarchical Modeling

MPT modeling often relies on repeated-measures designs in which each individual provides repeated responses across one or more within-subject conditions. Figure 7 shows the corresponding nested data structure in which each individual provides a vector of response frequencies in separate rows. There are generally three approaches to deal with such nested data structures. First, one may analyze aggregated frequencies on the group level as illustrated in the examples above (**complete pooling**). Second, one may conduct separate analyses per individual (**no pooling**), thus independently fitting an MPT model at the individual level. Third, hierarchical modeling (**partial pooling**) offers an elegant trade-off between the first two approaches by specifying how the data are distributed both on the individual and the group level.

2.1 Complete Pooling: Consequences of Aggregation

Traditionally, MPT modeling uses category frequencies aggregated across both participants and items which is equivalent to computing the column sums of the matrix

Figure 7

Nested Data Structure of a Within-Subjects Design (Data of Bayen, 1990).

1	id,	group,	age,	sex,	IST70,	lag0E1,	lag0E2,	lag0E3,	lag0E4,	F1,	F2
2	1,	old,	76,	f,	3,	1,	0,	1,	8,	2,	8
3	2,	old,	75,	f,	16,	0,	0,	1,	9,	0,	10
4	3,	old,	80,	f,	12,	2,	0,	7,	1,	6,	4
5	...										
6	41,	young,	23,	m,	16,	4,	0,	1,	5,	5,	5
7	42,	young,	24,	f,	16,	2,	1,	1,	6,	1,	9
8	43,	young,	26,	m,	10,	3,	0,	2,	5,	5,	5

Note. Hierarchical modeling with **TreeBUGS** requires that the column names are identical to the labels of the MPT response categories. Response frequencies for the pair-clustering model (without the lag 15 word pairs) are provided in the columns **lag0E1** to **F2**. The column **IST70** is an intelligence score (Intelligenz-Struktur-Test; Amthauer, 1970) which is used as an external covariate. For analyses of lag effects, four additional columns corresponding to **lag15E01**, **lag15E02**, **lag15E03**, and **lag15E04** frequencies per participant are required.

in Figure 7. These aggregated frequencies were also used in the examples above. Pooling the data results in precise parameter estimates and ensures a relatively high statistical power for hypothesis tests. Sometimes, there is no alternative to MPT analyses of aggregated data, for example, when there are many participants but each participant provides only a single data point (e.g., in the MPT analyses of the Wason Selection Task by Klauer et al., 2007).

However, aggregation is based on the assumption that observed responses and latent processes are independent and identically distributed (i.i.d.) meaning that the parameters do not vary between individuals. Assuming an absence of inter-individual differences seems to be questionable in many contexts. For instance, in the pair-clustering model, it is unlikely that all individuals have the same parameter values for clustering c , retrieval r , and the recall of singletons u . Empirically, the i.i.d. assumption is violated if response patterns systematically differ across participants (Smith & Batchelder, 2008).

If the i.i.d. assumption is violated, aggregation can result in incorrect statistical inferences such as biased point estimates of parameters, incorrect confidence intervals due to ignorance of interindividual variability, and inflated model-fit indices (Klauer, 2006; Smith & Batchelder, 2008). This is due to the fact that, even if an MPT model holds for each individual with different parameters, it is not necessarily the case that the MPT model is also a valid description of the aggregate data due to its nonlinear structure. Hence, results may in general differ between (a) averaging response frequencies and fitting the model compared to (b) fitting the MPT model separately per individual and averaging the parameter estimates. However, if we focus only on MPT models in which none of the parameters or their complements occurs repeatedly in any single branch of the model and if, in addition, all parameters are uncorrelated across individuals, a model is necessarily aggregation invariant meaning that the MPT model with averaged individual parameters also describes the aggregated frequencies at the group level (Erdfelder, 2000). Note that the pair-clustering model is not aggregation invariant because the u parameter and its complement $(1 - u)$ occur repeatedly in the branches for non-clustered word-pairs. This may induce biases in parameter estimates and

goodness-of-fit tests based on aggregated data. The magnitude of this bias largely depends on the variance of u across individuals and also on its correlation with c .

2.2 No Pooling: Fitting an MPT Model for Each Individual

Less commonly, researchers apply MPT model fitting and parameter estimation separately per individual (e.g., Bröder et al., 2013; Hilbig & Moshagen, 2014). This corresponds to using one vector of response frequencies per individual as input (i.e., one row of the matrix in Figure 7). In a second step, individual parameter estimates are then averaged at the group level. Similarly, it is possible to obtain an overall goodness-of-fit statistic by summing the individual G^2 statistics and the corresponding degrees of freedom. More recently, a sequential testing approach has been proposed that can be applied to MPT data of single participants to minimize the expected number responses required to reach a statistical decision with controlled error probabilities (Schnuerch et al., 2020). Just like the complete pooling approach, the no pooling approach is sometimes unavoidable. This is the case whenever there are data of a single participant only (or very few participants), but each participant provides a large sample of responses (e.g., the recognition memory analyses presented in Heck & Erdfelder, 2016).

Whereas the no-pooling approach accounts for individual differences, it has several drawbacks. First, it is only applicable if a sufficiently large number of responses per participant is available. Often, this number is limited by the experimental design, resulting in a low precision of parameter estimates and a low statistical power. Second, parameter estimates may have a bias (i.e., over- or underestimate the true parameter systematically) as a consequence of the small sample size per individual, because maximum likelihood estimates in general are unbiased only asymptotically (i.e., for large N). Third, the issue of a small number of responses is especially problematic for MPT modeling because we aim at estimating *conditional* probabilities. For instance, in the pair-clustering model, it is not possible to precisely estimate the probability of recalling a cluster (parameter r) if there are only a few cases of clustering (parameter c) since the probability of r is conditional on c . In general, this problem is larger for parameters at

the second or even third level of a tree (such as r) than for parameters at the first level (such as c).

2.3 Partial Pooling: Latent-Trait MPT

Hierarchical MPT models offer an elegant solution for model fitting and parameter estimation while addressing the limitations of the complete- and no-pooling approaches (Klauer, 2010; Smith & Batchelder, 2010). Essentially, hierarchical models assume that the same processing-tree structure is valid for all individuals, but with a different vector of parameters θ_i for each person i (level 1). On the group-level (level 2), a certain distribution of the individual parameters θ_i is assumed depending on the specific modeling approach (e.g., latent-trait or beta-MPT). Thereby, information is partially pooled across individuals, meaning that individual parameter estimates are informed by each other. This provides more statistical power than individual fitting and comes without the drawbacks of analyzing aggregated data. However, to exploit its full potential, the partial pooling approach requires data of more than just a few participants with more than just a few responses per participant.

Latent-trait MPT models are the most commonly used hierarchical approach and assume that individual parameters differ on a latent continuum (Klauer, 2010). Similar as in many psychometric models (e.g., item response theory; Plieninger & Heck, 2018), differences between individuals are assumed to follow a multivariate normal distribution at the group level. However, since the normal distribution assumes parameter values θ_{si}^* ranging from $[-\infty, +\infty]$, these parameter values need to be transformed into probabilities on the interval $[0, 1]$ to obtain valid MPT parameters. The latent-trait MPT model thus uses a probit-link function to obtain individual MPT parameters θ_{si} on the probability scale. This means that the latent parameters θ_{si}^* are the z -values corresponding to the cumulative density function of the standard-normal distribution, $\theta_{si} = \Phi(\theta_{si}^*)$. The resulting parameters θ_{si} are valid probabilities which are then plugged into the MPT model equations separately for each individual i .

By assuming a multivariate normal distribution, latent-trait MPT models allow

researchers to estimate the correlation between model parameters (Klauer, 2010). For instance, with respect to the pair-clustering model, it seems plausible that there is a correlation between the model parameters c , r , and u reflecting a common underlying memory capacity which varies across individuals. However, large numbers of participants and responses per participant are usually required to estimate correlations between MPT model parameters with high precision (Jobst et al., 2020).

Latent-trait MPT models also offer the benefit of resulting in shrinkage of the individual parameter estimates. This means that the individual estimates of a hierarchical model (partial pooling) have a smaller variance and show less extreme values compared to independently estimating the MPT parameters separately for each participant (no pooling). Thereby, the overall error of estimation is minimized meaning that on average, individual parameter estimates are closer to the true value (Efron & Morris, 1977).

As briefly outlined in Section 2.6, fitting hierarchical MPT models can in principle be achieved by using marginal-maximum-likelihood estimation. However, in contrast to traditional MPT models which are also usually fitted with maximum-likelihood estimation, hierarchical MPT models so far have almost exclusively relied on Bayesian inference. The Bayesian approach requires two components (for an introduction, see Lee & Wagenmakers, 2014). First, we need the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ of the data \mathbf{y} given all of the model parameters $\boldsymbol{\theta}$. The likelihood for the data matrix in Figure 7 is a specific, parameterized joint multinomial distribution as defined by the latent-trait MPT model. Second, we need a prior distribution $p(\boldsymbol{\theta})$ to specify which values of the parameters $\boldsymbol{\theta}$ are more or less plausible before seeing the data. The density of the posterior distribution of the parameters $\boldsymbol{\theta}$ given the data \mathbf{y} is then obtained by applying Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \quad (1)$$

In practice, it is usually not possible to obtain parameter estimates based on the probability density function $p(\boldsymbol{\theta}|\mathbf{y})$ of the posterior distribution in Eq. (1). As a remedy, Bayesian parameter estimation relies on Markov Chain Monte Carlo (MCMC) sampling (for an introduction, see van Ravenzwaaij et al., 2018). Essentially, parameter estimation

is simplified by drawing random samples for the parameters from the posterior distribution. MCMC sampling is a method that returns such random samples from the posterior distribution for all parameters both at the individual and the group level. Descriptive statistics of the MCMC samples (e.g., the sample average or variance) can then be used to describe relevant features of the posterior distribution (e.g., the posterior mean or variance, respectively).

Fitting a latent-trait MPT Model.

We illustrate the application of latent-trait MPT modeling by revisiting some of the research questions already considered above using traditional MPT modeling. Hierarchical modeling with the R package **TreeBUGS** (Heck, Arnold, et al., 2018) requires a data matrix consisting of the response frequencies of each participant (per row) for all possible response categories (columns) as shown in Figure 7. Figure 8 shows how to fit the pair-clustering model separately for the two groups of young and old participants (lines 5 – 11) and inspect the convergence of MCMC sampling (line 14). Convergence of MCMC sampling is a prerequisite for interpreting the results of parameter estimation. When plotting the MCMC chains as shown in Figure 9A, it is important that the posterior samples for each parameter move randomly like “hairy caterpillars” without showing any systematic patterns (e.g., upwards or downwards trends). If this is not the case, it is necessary to adapt model fitting by obtaining more MCMC samples or using more chains (see R script for details). Table 1 shows that the parameter estimates of the latent-trait model are similar to those obtained with the analysis of the aggregated data. Slightly larger differences between the point estimates only occur for parameters that are estimated with low precision (i.e., a large confidence or credible interval). Overall, this indicates that the results are robust with respect to the analysis method despite the fact that, strictly speaking, the pair-clustering model is not aggregation invariant.

The assessment of model fit focuses on the comparison of the observed frequencies with those predicted by the posterior distribution of the fitted MPT model (lines 17 – 18 in Figure 8). Conceptually, this is similar to assessing the discrepancy of observed versus predicted frequencies via the PD^λ or G^2 statistics used in traditional MPT modeling. In

Figure 8*Fitting a Latent-Trait MPT Model Using the TreeBUGS Package in R.*

```

1 # load package
2 library(TreeBUGS)
3
4 # fit models separately per group
5 trait_young <- traitMPT(eqnfile = "models/pc_lag.eqn",
6                         data = bayen_young,
7                         restrictions = list("a=u_0=u_15"))
8
9 trait_old <- traitMPT(eqnfile = "models/pc_lag.eqn",
10                      data = bayen_old,
11                      restrictions = list("a=u_0=u_15"))
12
13 # visual check of MCMC convergence
14 plot(trait_old)
15
16 # assess model fit
17 plotFit(trait_old)
18 PPP(trait_old)
19
20 # get posterior summary statistics
21 summary(trait_old)
22
23 # between-subjects: difference of "c" between young & old
24 betweenSubjectMPT(model1 = trait_young,
25                   model2 = trait_old,
26                   par1 = "c_0")
27
28 # within-subjects: difference of "c" between lag=0 & lag=15
29 test_c <- transformedParameters(
30   fittedModel = trait_young,
31   transformedParameters = list("delta_c = c_0 - c_15"))
32 summarizeMCMC(test_c)

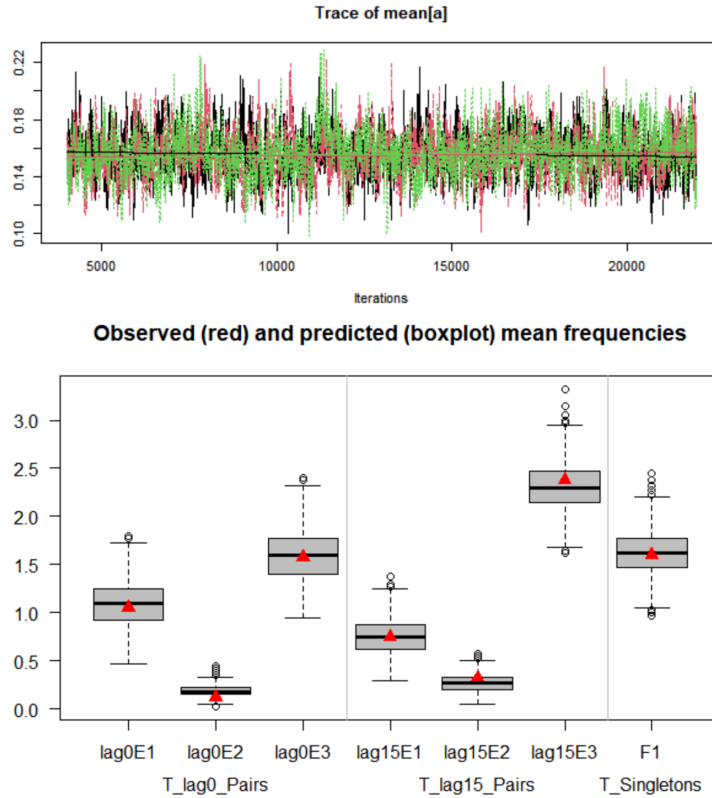
```

Note. The data include the within-subjects factor lag and the between-subjects condition age. After loading the TreeBUGS package (line 2), the model is fitted separately for each age group (young and older participants; lines 5-11). Next, MCMC convergence is checked visually (line 14), model fit is assessed (lines 17 – 18), and posterior summary statistics are returned (line 21). The difference of the c_0 parameter between age groups is assessed (lines 24 – 26). Finally, the difference in the parameter c between lag 0 and lag 15 is estimated for young participants (lines 29 – 32).

hierarchical models, however, we can focus on two different aspects of the data: the mean frequencies averaged across individuals and the covariance of individual frequencies. Both aspects can be assessed graphically. Figure 9B shows that the observed mean frequencies (red triangles) are perfectly in line with the posterior-predicted frequencies (gray boxplots). Model fit can also be quantified by computing the fit statistics T_1 and T_2 which focus on the mean and covariance of the individual frequencies, respectively (Klauer, 2010). Each statistic is computed once for the observed and once for the posterior-predicted data. By comparing these two cases, one obtains a

Figure 9

Output for a Fitted Latent-Trait MPT Model (Based on Data of Bayen, 1990).



Note. Panel A shows MCMC samples for one MPT parameter with each color referring to one of the three chains. Panel B shows the posterior-predictive check for comparing observed and predicted mean frequencies across individuals (indicated by red triangles and gray box plots, respectively).

posterior-predictive p -value (PPP) which indicates the amount of model misfit. There is a general agreement that PPP values larger than .05 reflect satisfactory model fit whereas small values close to zero indicate misfit. For the present data, model fit is satisfactory for both age groups with respect to the mean frequencies ($p_{T_1}^y = .633$, $p_{T_1}^o = .572$) and the covariance ($p_{T_2}^y = .698$, $p_{T_2}^o = .525$).

Between-subjects comparison.

We now examine the hypothesis whether the clustering process c differs between young and old participants in the lag 0 condition. For conducting between-subjects comparisons, we use the `betweenSubjects()` function as shown in lines 24 – 26 of Figure 8. Based on the MCMC samples for each group, TreeBUGS computes the difference of the specified parameters at the group level, $\Delta_{c_{age}} = c_{y,0} - c_{o,0}$. The resulting

transformed MCMC samples are then used to obtain the posterior mean as well as the 95% credibility interval. The results indicate only a small difference in the clustering of word pairs ($\Delta c_{\text{age}} = .008[-.212, .232]$) with the credibility interval overlapping zero.⁹

Within-subjects comparison.

Next, we examine the hypothesis whether the clustering probability c differs within young participants depending on the lag condition (lag 0 versus lag 15). For within-subject comparisons, we use the `transformedParameters` function as shown in lines 29 – 32 in Figure 8. Thereby, TreeBUGS computes the difference in the specified parameter, $\Delta c = c_0 - c_{15}$ at the group level while using all MCMC samples to obtain estimates for the posterior mean and the 95% credibility interval. The results show that there is a substantial difference in clustering depending on the lag between word pairs, $\Delta c = .183 [.032, .322]$, with c being larger for short lags compared to long lags, as already found by Batchelder and Riefer (1980).

2.4 Other Hierarchical Models: Beta-MPT and Latent-Class MPT

There are two alternative approaches for hierarchical MPT modeling which are less commonly used than latent-trait MPT and come with different possibilities and drawbacks. First, **beta-MPT models** assume that the person parameters are uncorrelated across participants (Smith & Batchelder, 2010). Hence, the parameters θ_{si} (e.g., the probability of recall r for individual i) follow independent beta distributions at the group level. The beta distribution is ideally suited for modeling (conditional) probability parameters since it only allows values in the interval $[0, 1]$. At the group level, the beta distribution is described by the two shape parameters α_{si} and β_{si} which allow to estimate the group-level mean of the MPT parameters via the transformation $\alpha/(\alpha + \beta)$. The supplementary material shows how to fit a beta-MPT model to the pair-clustering data using TreeBUGS. Whereas the beta-MPT model is conceptually simple, its

⁹ Note that based on this evidence alone the inference that clustering does not differ between age groups would be problematic. We did neither perform a sensitivity analysis for the present sample size nor did we compute a Bayes factor in favor of the null vs. alternative hypothesis (Heck et al., 2022).

assumptions are violated if the model parameters are correlated, possibly leading to biased parameter estimates. Compared to the latent-trait approach, it is also more difficult to extend the beta-MPT model to include external covariates.

In contrast to latent-trait and beta-MPT models, **latent-class MPT models** do not assume that the person parameters differ on a latent continuum. Instead, it is assumed that parameter heterogeneity emerges because each participant belongs to one of a small number of latent classes (Klauer, 2006). Latent-class MPT models thus assume that participants can be clustered with respect to latent characteristics. Within each of these latent classes, the parameter values are assumed to be identical according to the i.i.d. assumption. Maximum-likelihood estimation of the parameters and the group membership are obtained based on the observable response frequencies using an expectation-maximization algorithm (Klauer, 2006). The number of latent groups is determined by using model-selection criteria (i.e., AIC, BIC, R^2) of competing models assuming a different number of latent classes. Latent-class MPT models can be fitted using the software **HMMTree** (Stahl & Klauer, 2007).

2.5 External Covariates

The latent-trait approach allows researchers to regress individual MPT parameters on external covariates (Heck, Arnold, et al., 2018; Michalkiewicz et al., 2018). For instance, we could predict memory performance as measured by the parameter u by general intelligence. Essentially, this amounts to a level-2 regression of the person parameter θ_{si} on a design matrix \mathbf{X} of covariates,

$$\theta_{si} = \Phi(\mathbf{X}\boldsymbol{\beta} + \delta_{si}), \quad (2)$$

where Φ does again refer to a probit link function, $\boldsymbol{\beta}$ is a vector of regression coefficients, and δ_{si} are normally distributed person-random effects for the parameter θ_{si} . A Monte Carlo simulation study showed that this approach works well in practice if a sufficient number of responses and individuals are available (Jobst et al., 2020).

When fitting a latent-trait MPT model with **TreeBUGS**, data for external

covariates can be provided via the argument `covData`. Corresponding regression structures for the parameters are defined via the argument `predStructure`. `TreeBUGS` also offers the option to define separate regression models for each parameter. For instance, one could predict the parameter u by general intelligence while predicting the parameter c both by intelligence and working-memory capacity. We illustrate the regression approach by predicting individual differences in u by general intelligence measured via the Intelligenz-Strukturtest (IST-70) by using the argument `predStructure=list("u ; IST70")`. This results in estimated probit-regression coefficients of $\beta = 0.012$ $[-0.017, 0.039]$ and $\beta = 0.037$ $[0.008, 0.067]$ for young and old participants, respectively.

`TreeBUGS` also provides the function `BayesFactorSlope()` to compute a Bayes factor for testing the hypothesis $H_0 : \beta = 0$ assuming a modified Jeffreys-Zellner-Siow (JZS) prior (see Heck, Arnold, et al., 2018, for details).¹⁰ Applying this function to the fitted model shows that the evidence in favor of a positive correlation of intelligence with u is $BF_{10} = 0.12$ and $BF_{10} = 2.76$ for young and old individuals, respectively. This means that there is some evidence for a null correlation in young participants and only weak evidence for a positive correlation for old participants.

2.6 Marginal Maximum Likelihood Estimation of Hierarchical MPT Models

Nestler and Erdfelder (2022) recently proposed marginal-maximum-likelihood (ML) estimation and evaluation of Klauer's (2010) hierarchical latent-trait MPT model with or without covariates as an alternative to the Bayesian methods implemented in `TreeBUGS`. The key problem of ML estimation for hierarchical MPT models results from the fact that (a) analytical solutions that maximize the marginal likelihood of the data do not exist, and (b) numerical approximations based on the gradient of the log-likelihood function involve complex integrals that are not tractable. Nestler and Erdfelder (2022) therefore explored several numerical integral approximation methods and found the Adaptive Gauss-Hermite Quadrature (AGHQ) to work best (Tuerlinckx et al., 2006). In

¹⁰ When adding more than one predictor per MPT parameter, Bayes factors can only be computed if a normally distributed prior on β is assumed (Heck, 2019).

fact, their ML re-analysis of the pair-clustering data used by Klauer (2010) to illustrate Bayesian MCMC estimation resulted in estimates very similar to those reported in Table 2 of Klauer (2010). Given the well-known asymptotic optimality of ML and PD^λ estimators in general (Read & Cressie, 1988), we would expect that ML and Bayesian estimation methods will generally converge in their results when both the number of individuals and the number of responses per individual are large. For small samples, Nestler and Erdfelder (2022) speculated that the Bayesian approach may outperform ML estimation in terms of unbiasedness (p. 24). Whether this is actually the case needs to be studied in future comparative Monte Carlo experiments for a representative sample of MPT models and must remain an open question so far.

In any case, the ML framework for hierarchical MPT models based on the latent-trait model has now been worked out in detail, including options for regressing model parameters on external covariates with random or fixed intercepts, thus including the MPT regression method proposed by (Coolin et al., 2015) as a special case. The ML modeling framework is flexible as it allows to treat some of the parameters as random (as assumed in the latent-trait model) and the remaining parameters as fixed effects; it also allows to replace the probit link function used in the latent-trait model by a logit link function. Conceptually, the ML approach may have some advantages compared to Bayesian methods because researchers do not need to consider the choice of prior distributions and their possible influence on the estimation results. In practical applications, convergence of the ML estimation algorithm may be easier to determine and faster to obtain than with Bayesian MCMC methods. Yet, these assessments must be considered preliminary and need a rigorous evaluation before they can guide applications of hierarchical MPT models in future research.

2.7 Extensions of MPT Models for Continuous Variables

MPT models are inherently limited to discrete data. In the following, we provide a short overview of recent methodological advances for extending MPT models to continuous data. More precisely, we are concerned with **data structures** in which we

observe one or more continuous variables (e.g., response time) alongside the discrete response within each trial. Typically, the focus is on modeling both choices and response times. Note that response times are not available in free recall, and hence, such extensions are not relevant for the pair-clustering model discussed as a running example in this tutorial. Hence, we do not report examples of how to apply these modeling approaches.

MPT models are based on the core assumption that the observed data can be modeled by a **mixture distribution** over a finite set of latent (cognitive) states. With respect to response times, this implies that each branch of an MPT tree is associated with a component distribution that describes the corresponding speed of responding. Across the different branches, multiple branches may lead to the same type of response, thus leading to a mixture of the component distributions with the mixture weights determined by the corresponding MPT branch probabilities. The following approaches differ only with respect to the specific assumptions about how to model the component distributions.

First, one may **categorize the continuous variable into discrete bins** (Heck & Erdfelder, 2016). For instance, we can classify responses as “fast” or “slow” relative to the median RT per participant (determined independently from the to-be-analyzed data). Thereby, we can compare the relative speed of different processing branches. This is achieved by extending the MPT model to the contingency table obtained by splitting all response categories into “fast” and “slow.” Thereby, we obtain an extended, larger MPT model with more categories which can again be fitted using standard MPT software (e.g., Brainerd et al., 2019; Heck & Erdfelder, 2017, 2020).

Second, **generalized processing tree models** may be used for extending MPT models to continuous data (Heck, Erdfelder, & Kieslich, 2018). These models assume specific parametric component distributions such as a normal or ex-Gaussian distribution. Then, the standard MPT parameters are estimated alongside the parameters describing the latent component distributions (e.g., the mean of response times associated with a specific memory process). Generalized processing tree models offer the benefit of being applicable not only to response times but also to other types of continuous variables (e.g., process-tracing measures such as mouse-tracking; Heck, Erdfelder, & Kieslich, 2018). The

R package `gpt` facilitates model fitting but is limited to independent and identically distributed data (i.e., complete or no pooling).

Third, Klauer and Kellen (2018) developed **RT-MPT models**, a class of models for discrete choices and response times. In contrast to the other two approaches, RT-MPT models assume that the transitions between different latent states along a processing branch occur in a strictly serial fashion (Hu, 2001). Hence, observed response times are assumed to be consistent with the sum of several independent processing times associated with the different processes. Moreover, the model assumes an additive non-decision time that varies for different response options. Bayesian hierarchical RT-MPT models can be fitted using the `rtmpt` package in R (Hartmann et al., 2020).

3 Discussion

Multinomial processing tree (MPT) models assume that the distribution of observed responses can be described by a mixture of latent (cognitive) processes (Batchelder & Riefer, 1990). By estimating the probabilities of being in different latent states (i.e., the outcomes of the cognitive processes involved), MPT models can explain patterns of results that are difficult to interpret when limiting the analysis to directly observable responses (e.g., error rates). Once the latent processes are disentangled and measured with MPT modeling, the pattern of parameter estimates may explain findings that are otherwise puzzling.

An intriguing example is the classical finding of Batchelder and Riefer (1986) that memory performance appears to be unrelated to the lag of semantically related words in a studied word list. It seems very plausible that short lags should invite semantic processing and clustering, and semantic clustering should in turn boost recall memory subsequently—yet it did not in Batchelder and Riefer’s free recall data. How to explain this apparent contradiction? By analyzing their data with the pair clustering MPT model, Batchelder and Riefer (1986) were able to provide a straightforward answer: Short lags indeed enhance semantic clustering in memory (as measured by the cluster storage probability c) but at the same time short lags are detrimental for memory retrieval (as

measured by the cluster retrieval probability r), presumably because retrieval in free recall benefits from variability in encoding contexts that increases with lags. Notably, the cognitive aging data from Bayen (1990) we used as a running example in the present tutorial conceptually replicated the results of Batchelder and Riefer (1986): c decreased and r increased significantly from lag 0 to lag 15 for both younger and older adults.

Another important use of MPT models is to provide converging (or discrepant) evidence for cognitive theories previously tested with other approaches (e.g., using different behavioral or psychophysiological measures). For example, it has long been argued that the episodic memory deterioration often observed with increasing age is not due to memory storage deficits but rather due to memory retrieval deficits in elderly people (retrieval-deficit hypothesis of cognitive aging). Assuming that free recall is more sensitive to retrieval deficits than recognition performance, the retrieval-deficit theory of cognitive aging is supported by the finding that the age decline is small or even absent in recognition performance but clearly visible in free recall data (see, e.g., Schonfield & Robertson, 1966). Given such findings, it is highly informative to see that young and older adults do not differ in the storage parameter c but in cluster retrieval r , with younger adults outperforming older adults in the latter parameter (Riefer & Batchelder, 1991). Again, this finding was conceptually replicated in the Bayen data used here, although in our case the age difference in r did not pass the significance threshold due to lack of power (as we have seen in Section 1.10).

In addition, MPT models have played an important role in determining and measuring the multitude of cognitive and affective determinants of performance in supposedly unidimensional measurement instruments. Think, for example, about the influential implicit association test (IAT, Greenwald et al., 1998) and similar procedures. Originally conceived as unidimensional measures of implicit attitudes, MPT approaches to these paradigms have clearly revealed that each paradigm is multidimensional from a social-cognition perspective. Interestingly, however, the MPT models developed so far disagree on whether four (Conrey et al., 2005) or three (Meissner et al., 2019; Nadarevic & Erdfelder, 2011) parameters are required to fully explain performance in the IAT and

how these parameters should be defined. Ultimately, systematic validation research programs are required to determine how many and which parameters are required for the IAT and to identify the paradigm variant that captures these parameters best and with sufficient precision (Calanchini et al., 2021).

This is but a small subset of the advantages gained so far in testing cognitive theories and developing appropriate measurement devices using MPT modeling. Readers are encouraged to study the reviews available so far to gain an impression of the scope and the flexibility of the many MPT models developed and used in different branches of behavioral research (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016)

As we have seen, MPT models are highly relevant for many fields of psychology, cognitive and social psychology in particular. The scope of MPT models is large because most paradigms in psychological research rely on discrete data. Moreover, it is often questionable whether and how observed responses reflect latent (cognitive) capacities, meaning that MPT models are a suitable method for obtaining more valid measurements of latent processes. This idea of linking MPT modeling and psychometric measurement has been coined **cognitive psychometrics** (Riefer et al., 2002) and was the topic of a recent special issue in the *Journal of Mathematical Psychology* (Erdfelder et al., 2020). The present tutorial provided a gentle introduction to the basics of MPT modeling while also highlighting recent developments for modeling individual differences and response times. We hope that this toolbox for MPT modeling will allow researchers to test novel and more precise hypotheses and to compare competing verbal theories using formal modeling.

References

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.
- Amthauer, R. (1970). *Intelligenz-Struktur-Test: IST 70*. Hogrefe.
- Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, 44, 20–40.
<https://doi.org/10.1006/jmps.1999.1275>
- 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. (1990). Multinomial processing models of source monitoring. *Psychological Review*, 97, 548–564.
<https://doi.org/10.1037/0033-295X.97.4.548>
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86.
<https://doi.org/10.3758/BF03210812>
- Bayen, U. J. (1990). Zur Lokalisation von Altersdifferenzen im episodischen Gedächtnis Erwachsener: Eine Querschnittsuntersuchung auf der Basis eines mathematischen Modells. *Berichte aus dem psychologischen Institut der Universität Bonn*, 16, 1–125.
- Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 197–215.
<https://doi.org/10.1037/0278-7393.22.1.197>

- Bernstein, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Zeitschrift für Abstammungs-und Vererbungslehre*, 37, 237–270.
- Bott, F. M., Heck, D. W., & Meiser, T. (2020). Parameter validation in hierarchical MPT models by functional dissociation with continuous covariates: An application to contingency inference. *Journal of Mathematical Psychology*, 98, 102388.
<https://doi.org/10.1016/j.jmp.2020.102388>
- Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, 9, 529–533.
[https://doi.org/https://doi.org/10.1016/S0022-5371\(70\)80096-2](https://doi.org/https://doi.org/10.1016/S0022-5371(70)80096-2)
- Brainerd, C. J., Nakamura, K., & Lee, W.-F. A. (2019). Recollection is fast and slow. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45, 302–319. <https://doi.org/10.1037/xlm0000588>
- Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 21, 916–944. <https://doi.org/10.1080/09658211.2013.767348>
- Calanchini, J., Meissner, F., & Klauer, K. C. (2021). The role of recoding in implicit social cognition: Investigating the scope and interpretation of the ReAL model for the implicit association test. *PLOS ONE*, 16, e0250068.
<https://doi.org/10.1371/journal.pone.0250068>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
<https://doi.org/10.1037/0033-2909.112.1.155>
- Conrey, R., Sherman, J., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89, 469–87.
<https://doi.org/10.1037/0022-3514.89.4.469>
- 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>

Dehn, D. M., & Erdfelder, E. (1998). What kind of bias is hindsight bias? *Psychological Research*, 61, 135–146.

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, 1–38.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236, 119–127.

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie*. Unpublished habilitation thesis.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., 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., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition*, 25, 114–131.

<https://doi.org/10.1521/soco.2007.25.1.114>

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: 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., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3., pp. 1565–1570). London: Wiley.

Erdfelder, E., Hu, X., Rouder, J. N., & Wagenmakers, E.-J. (2020). Cognitive psychometrics: The scientific legacy of William H. Batchelder (1940–2018). *Journal of Mathematical Psychology*, 99, 102468.

<https://doi.org/10.1016/j.jmp.2020.102468>

- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.
<https://doi.org/10.1037//0022-3514.74.6.1464>
- Hartmann, R., Johannsen, L., & Klauer, K. C. (2020). rtmpt: An R package for fitting response-time extended multinomial processing tree models. *Behavior Research Methods*, *52*, 1313–1338. <https://doi.org/10.3758/s13428-019-01318-x>
- Heck, D. W. (2019). A caveat on the Savage-Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, *72*, 316–333. <https://doi.org/10.1111/bmsp.12150>
- Heck, D. W., Arnold, N. R., & 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. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., . . . Hoijsink, H. (2022). A review of applications of the Bayes factor in psychological research. *Psychological Methods*. <https://doi.org/10.1037/met0000454>
- Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87.
<https://doi.org/10.1016/j.jmp.2019.03.004>
- Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. <https://doi.org/10.3758/s13423-016-1025-6>
- Heck, D. W., & Erdfelder, E. (2017). Linking process and measurement models of recognition-based decisions. *Psychological Review*, *124*, 442–471.
<https://doi.org/10.1037/rev0000063>
- Heck, D. W., & Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Computational Brain & Behavior*, *2*, 202–209. <https://doi.org/10.1007/s42113-019-00035-0>

- Heck, D. W., & Erdfelder, E. (2020). Benefits of response time-extended multinomial processing tree models: A reply to Starns (2018). *Psychonomic Bulletin & Review*, 27, 571–580. <https://doi.org/10.3758/s13423-019-01663-0>
- Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, 83, 893–918. <https://doi.org/10.1007/s11336-018-9622-0>
- Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, 60, 29–34. <https://doi.org/10.1016/j.jmp.2014.06.002>
- Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 827.
- Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, 21, 1431–1443. <https://doi.org/10.3758/s13423-014-0643-0>
- Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, 45, 603–634. <https://doi.org/10.1006/jmps.2000.1340>
- 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>
- Jobst, L. J., Heck, D. W., & Moshagen, M. (2020). A comparison of correlation and regression approaches for multinomial processing tree models. *Journal of Mathematical Psychology*, 98, 102400. <https://doi.org/10.1016/j.jmp.2020.102400>

- Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, 71, 7–31. <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>
- Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 680–703. <https://doi.org/10.1037/0278-7393.33.4.680>
- Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, 48, 215–229. <https://doi.org/10.1016/j.jmp.2004.03.002>
- Kuhlmann, B. G., Erdfelder, E., & Moshagen, M. (2019). Testing interactions in multinomial processing tree models. *Frontiers in Psychology*, 10, 2364. <https://doi.org/10.3389/fpsyg.2019.02364>
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 116.
- Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. *Frontiers in Psychology*, 10. <https://www.frontiersin.org/article/10.3389/fpsyg.2019.02483>
- Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, 104, 45–69. <https://doi.org/10.1037/a0030734>

- Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 9, 596–606.
[https://doi.org/https://doi.org/10.1016/S0022-5371\(70\)80107-4](https://doi.org/https://doi.org/10.1016/S0022-5371(70)80107-4)
- Michalkiewicz, M., Arden, K., & Erdfelder, E. (2018). Do smarter people employ better decision strategies? The influence of intelligence on adaptive use of the recognition heuristic. *Journal of Behavioral Decision Making*, 31, 3–11.
<https://doi.org/10.1002/bdm.2040>
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42, 42–54.
<https://doi.org/10.3758/BRM.42.1.42>
- Nadarevic, L., & Erdfelder, E. (2011). Cognitive processes in implicit attitude tasks: An experimental validation of the trip model. *European Journal of Social Psychology*, 41, 254–268. <https://doi.org/10.1002/ejsp.776>
- Nestler, S., & Erdfelder, E. (2022). Random effects multinomial processing tree models: A maximum likelihood approach. *Manuscript submitted for publication*.
- Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, 53, 633–654. <https://doi.org/10.1080/00273171.2018.1469966>
- Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer.
- Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, 14, 184–201. <https://doi.org/10.1037/1040-3590.14.2.184>
- Riefer, D. M., & Batchelder, W. H. (1991). Age differences in storage and retrieval: A multinomial modeling analysis. *Bulletin of the Psychonomic Society*, 29, 415–418.
<https://doi.org/10.3758/BF03333957>
- Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42, 40–47.

- Schnuerch, M., Erdfelder, E., & Heck, D. W. (2020). Sequential hypothesis tests for multinomial processing tree models. *Journal of Mathematical Psychology*, 95, 102326. <https://doi.org/10.1016/j.jmp.2020.102326>
- Schnuerch, M., Heck, D. W., & Erdfelder, E. (in press). Waldian t tests: Sequential Bayesian t tests with controlled error probabilities. *Psychological Methods*.
- Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 20, 228–236. <https://doi.org/10.1037/h0082941>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.
- Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, 15, 713–731. <https://doi.org/10.3758/PBR.15.4.713>
- 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>
- Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, 39, 267–273. <https://doi.org/10.3758/BF03193157>
- 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. <https://doi.org/10.1348/000711005X79857>
- van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, 25, 143–154. <https://doi.org/10.3758/s13423-016-1015-8>
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196. <https://doi.org/10.3758/BF03206482>

Appendix A

Likelihood Function of MPT Models

In the following, we derive the likelihood function of MPT models (Hu & Batchelder, 1994), which is required both for frequentist and Bayesian inference. For readability, we drop the index k for the different trees of an MPT model. We start with a simple example of only two observed categories with known category probabilities p_1 and $p_2 = 1 - p_1$. If observations are independent and identically distributed, the observed frequencies n_1 and n_2 follow a binomial distribution with probability mass function

$$P(n_1, n_2) = \frac{N!}{n_1! n_2!} p_1^{n_1} p_2^{n_2} \quad (\text{A1})$$

where $N = n_1 + n_2$ refers to the total number of observations. Multinomial modeling generalizes this idea to variables with $J + 1$ categories. Note that we use $J + 1$ as an upper index to highlight that the data provide only J degrees of freedom. Given a vector of category probabilities $\mathbf{p} = (p_1, p_2, \dots, p_{J+1})$ and assuming independent sampling, the vector of observed response frequencies $\mathbf{n} = (n_1, n_2, \dots, n_{J+1})$ follows a **multinomial distribution**,

$$P(n_1, n_2, \dots, n_{J+1}) = \frac{N!}{n_1! n_2! \dots n_{J+1}!} p_1^{n_1} p_2^{n_2} \dots p_{J+1}^{n_{J+1}}. \quad (\text{A2})$$

MPT models account for the category probabilities $\mathbf{p} = (p_1, \dots, p_{J+1})$ by assuming a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_S)$ of S parameters (e.g., the probabilities of clustering or retrieval). Essentially, the tree structure of an MPT model defines a multivariate function f which assigns certain category probabilities to each set of possible parameter values,

$$\begin{aligned} p_1(\boldsymbol{\theta}) &= f_1(\theta_1, \theta_2, \dots, \theta_S) \\ p_2(\boldsymbol{\theta}) &= f_2(\theta_1, \theta_2, \dots, \theta_S) \\ &\dots \\ p_{J+1}(\boldsymbol{\theta}) &= f_{J+1}(\theta_1, \theta_2, \dots, \theta_S). \end{aligned} \quad (\text{A3})$$

Note that the parameterized category probabilities $p_j(\boldsymbol{\theta})$ are simply the model equations discussed in Section 1.4.

To define the general form of the function f for any MPT model, we first derive the **branch probability** $p_{ij}(\boldsymbol{\theta})$ that a certain branch i results in category j . This probability equals the product \prod of all parameter values occurring along the branch,

$$p_{ij}(\boldsymbol{\theta}) = c_{ij} \prod_{s=1}^S \theta_s^{a_{ijs}} \cdot (1 - \theta_s)^{b_{ijs}}. \quad (\text{A4})$$

Here, c_{ij} denotes a positive constant that accounts for parameters in a branch that are fixed to specific numbers (e.g., a guessing probability of .50). Moreover, the indices a_{ijs} and b_{ijs} denote non-negative integers (often 0 or 1) that simply count how often the parameter θ_s or its complement $1 - \theta_s$ occur in the branch i leading to category j , respectively.

In an MPT model, different branches $i = 1, \dots, I_j$ of the probability tree can lead to the same response category j . Because the different branches are mutually exclusive, the **category probabilities** $p_j(\boldsymbol{\theta})$ of an MPT model are obtained as the sum of all branch probabilities leading to this category,

$$p_j(\boldsymbol{\theta}) = \sum_{i=1}^{I_j} p_{ij}(\boldsymbol{\theta}) = \sum_{i=1}^{I_j} c_{ij} \prod_{s=1}^S \theta_s^{a_{ijs}} \cdot (1 - \theta_s)^{b_{ijs}}. \quad (\text{A5})$$

The probabilities $p_j(\boldsymbol{\theta})$ of all categories necessarily sum up to one as long as the parameters θ_s are in the interval $[0,1]$ (Hu & Batchelder, 1994). This corresponds to the requirement that the vector of parameter values is in the parameter space Ω ,

$$\boldsymbol{\theta} \in \Omega = [0, 1]^S. \quad (\text{A6})$$

The **likelihood function** is defined as the probability of the data given the parameters. For MPT models, we obtain the likelihood by plugging in the category probabilities equations from Equation (A5) into the multinomial probability distribution

in Equation (A2),

$$\begin{aligned}
 L(\boldsymbol{\theta} \mid \mathbf{n}) &= \frac{N!}{n_1! n_2! \cdots n_{J+1}!} \prod_{j=1}^{J+1} p_j(\boldsymbol{\theta}) \\
 &= \frac{N!}{n_1! n_2! \cdots n_{J+1}!} \prod_{j=1}^{J+1} \sum_{i=1}^{I_j} c_{ij} \prod_{s=1}^s \theta_s^{a_{ijs}} \cdot (1 - \theta_s)^{b_{ijs}}
 \end{aligned} \tag{A7}$$

Usually, the full likelihood of an MPT model requires the multiplication across another index to account for the $k = 1, \dots, K$ conditions (or trees) of the model (which are essentially independent multinomial distributions; Hu & Batchelder, 1994).

Appendix B

Model-Fit Statistics

B.1 The Power-Divergence Statistic PD^λ

Goodness-of-fit statistics such as the power-divergence statistic PD^λ measure the distance between observed and expected response frequencies (Read & Cressie, 1988).

Given real-valued, fixed value of λ , the statistic is defined as

$$PD^\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{j=1}^{J+1} n_j \left[\left(\frac{n_j}{n_j(\hat{\theta})} \right)^\lambda - 1 \right], \quad (B1)$$

with $PD^\lambda = 0$ is defined as the limit when λ approaches zero. Here, n_j denote the observed frequencies where $n_j(\hat{\theta}) = N \cdot p_j(\hat{\theta})$ denote the expected frequencies for category j . If observed and expected frequencies are identical for all categories, the ratio in the inner brackets equals one. In this case, $PD^\lambda = 0$ indicates that the model fits perfectly. Under the null hypothesis that the true category probabilities are in line with the model equations, PD^λ is χ^2 -distributed with

$$df = J - S, \quad (B2)$$

provided that none of the parameter estimates lies at the boundary of the parameter space (see Section B.4 if an estimate equals 0 or 1). This means that the degrees of freedom are simply calculated by subtracting the number of parameters S from the degrees of freedom J of the data (i.e., the number of free categories). If an MPT model contains $k = 1, \dots, K$ trees, $J = \sum_{k=1}^K J_k$.

B.2 The Likelihood-Ratio Statistic G^2

The commonly used default setting for the PD^λ statistic in MPT modeling is to set $\lambda = 0$. In this case, the power divergence statistic PD^λ is identical to the likelihood-ratio statistic G^2 (Read & Cressie, 1988),

$$\lim_{\lambda \rightarrow 0} PD^\lambda = 2 \sum_{j=1}^{J+1} n_j \log \left(\frac{n_j}{n_j(\theta)} \right) = G^2. \quad (B3)$$

If the observed frequencies n_j are identical to the expected frequencies $n_j(\boldsymbol{\theta})$, it follows that $G^2 = 0$ since $\log(1) = 0$. Note that maximizing the likelihood is equivalent to minimizing G^2 .

B.3 Pearson's X^2 Statistic

Sometimes, the PD^λ statistic is used with $\lambda = 1$. In this case, the statistic is identical to Pearson's X^2 (Read & Cressie, 1988),

$$\text{PD}^{\lambda=1} = \sum_{j=1}^{J+1} \frac{(n_j - n_j(\boldsymbol{\theta}))^2}{n_j(\boldsymbol{\theta})} = X^2. \quad (\text{B4})$$

Here, model fit is measured using the squared differences between the observed frequencies n_j and the expected frequencies $n_j(\boldsymbol{\theta})$, divided by the expected frequencies. Again, if $n_j = n_j(\boldsymbol{\theta})$ for all categories, $X^2 = 0$.

B.4 Parametric Bootstrap

If one of the true parameters in a hypothesis tests is at the boundary of the parameter space (i.e., $\theta_s = 0$ or $\theta_s = 1$), the PD^λ statistic does not follow an asymptotic χ^2 distribution (Read & Cressie, 1988). This may often happen when testing order constraints as in Section 1.9. In such a case, the distribution of the test statistic PD^λ can be approximated with a parametric bootstrap. This means that the estimated parameters (which may include cases at the boundary such as $\theta_s = 0$) are used as true values to simulate a large number of data sets (e.g., 500 replications) with the same sample sizes per tree of the model as in the observed data set. Next, the specified model is fitted to each of these simulated data sets, thus resulting an empirically simulated distribution of the test statistic PD^λ for the given experimental design, sample size, and data-generating parameters. A p -value can easily be computed as the proportion of simulated test statics exceeding the empirically observed value. `multiTree` offers a parametric bootstrap via the option “Bootstrapped Model Fit” in the panel on the right.

Appendix C

Model Comparison

C.1 Testing Nested Models Using ΔG^2

To allow direct comparisons of nested models, the goodness-of-fit statistic ΔG^2 is applied. To infer whether a nested Model A fits the data significantly worse than a more complex Model B, the difference in fit between both models is quantified by

$$\Delta G_{A-B}^2 = G_A^2 - G_B^2. \quad (\text{C1})$$

Under certain regularity conditions (Read & Cressie, 1988), the statistic ΔG_{A-B}^2 is χ^2 -distributed with $df_{A-B} = df_A - df_B$.

C.2 Model Selection with AIC and BIC

To compare non-nested MPT models, we need to rely on information criteria which take both model fit as well as model complexity into account. The **Akaike information criterion** (AIC; Akaike, 1998) assesses model fit via the maximized log-likelihood (i.e., the G^2 statistic) and model complexity via the number S of free parameters,

$$\text{AIC} = G^2 + 2 \cdot S. \quad (\text{C2})$$

The model with the smallest AIC value should be selected (Wagenmakers & Farrell, 2004). The AIC-best-fitting model from a set of candidate models can be shown to be the one that best approximates the true model in terms of minimizing the Kullback-Leibler distance between model predictions and true data. In comparison to other information criteria, the AIC tends to prefer more complex models.

Alternatively, model selection of non-nested models can be performed with the **Bayesian information criterion** (BIC; Schwarz, 1978). In contrast to the AIC, model complexity is quantified by the model parameters S multiplied with the logarithm of the sample size,

$$\text{BIC} = G^2 + \log(N) \cdot S. \quad (\text{C3})$$

Again, the model with the smallest BIC value should be selected. The BIC generally tends to prefer simpler models.

By default, `multiTree` compares the AIC and BIC values of the substantive MPT model M_0 against the saturated model M_1 (i.e., the model with as many parameters as degrees of freedom). The difference in AIC values is referred to as

$$\Delta\text{AIC} = \text{AIC}(M_0) - \text{AIC}(M_1). \quad (\text{C4})$$

Similarly, `multiTree` computes the difference in BIC values. If the model fits the data well, ΔAIC and ΔBIC should be smaller than 0.

C.3 Model Selection with the Fisher Information Approximation (FIA)

The Fisher information approximation (FIA) is a model-selection criterion based on the minimum-description-length framework (Rissanen, 1996). Instead of merely counting the number of parameters as in AIC and BIC, FIA considers the functional relationship between the parameters to assess model complexity. The importance of the model structure especially applies to reparameterized models for testing order constraints. As discussed in Section 1.9, such models are less complex than their unconstrained counterparts but still use the same number of free parameters. FIA is defined as

$$\text{FIA} = 0.5 \cdot G^2 + C_{\text{FIA}}(N), \quad (\text{C5})$$

where $C_{\text{FIA}}(N)$ is a complexity term which considers the number of free parameters, the number of observations, and the functional shape of the model via the Fisher information matrix (for details, see Rissanen, 1996). Note that FIA might result in erroneous results with small samples, and thus, it is important to check that N is sufficiently large for its application (Heck et al., 2014).

C.4 Model Selection with the Bayes Factor (BF)

Bayes Factors are the natural Bayesian solution for quantifying the relative evidence for a set of competing models (Heck et al., 2022). Essentially, the Bayes factor is the knowledge-updating factor required for going from the prior odds for two competing models M_1 and M_2 to the posterior odds, thereby quantifying which model is more plausible given the data,

$$\underbrace{\frac{p(M_1 | \mathbf{n})}{p(M_2 | \mathbf{n})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{n} | M_1)}{p(\mathbf{n} | M_2)}}_{\text{BF}_{12}} \underbrace{\frac{p(M_1)}{p(M_2)}}_{\text{prior odds}}. \quad (\text{C6})$$

Technically, the Bayes factor is thus defined as the ratio of two marginal likelihoods, that is, the probabilities of the observed data \mathbf{n} given the specific models M_1 and M_2 . In contrast to the maximum likelihood approach, the marginal likelihood does not focus on the best fitting parameter values $\hat{\boldsymbol{\theta}}$ for model selection, but rather on the full range of possible parameter values weighted by the corresponding prior distribution. As the Bayes factor also takes the functional shape of MPT models into account, it is well suited to evaluate order constraints (Heck & Davis-Stober, 2019). Regarding its interpretation, large Bayes Factors ($\text{BF}_{12} > 10$) correspond to strong evidence for M_1 relative to M_2 , whereas small values ($\text{BF}_{12} < 0.1$) indicate strong evidence for M_2 relative to M_1 . Note, however, that stricter interpretational guidelines may be necessary to match standard frequentist error probabilities for tests between models (for an analogous argument relating to Bayesian t tests, see Schnuerch et al., in press). The **TreeBUGS** package allows users to compute the Bayes factors for standard (non-hierarchical) MPT models (for a worked example, see Heck et al., 2022).