Title

Evaluating the Robustness of Parameter Estimates in Cognitive Models: A Meta-Analytic Review of Multinomial Processing Tree Models Across the Multiverse of Estimation Methods

Journal

Psychological Bulletin, 150(8)

ISSN

0033-2909

Authors

Singmann, Henrik

Heck, Daniel W

Barth, Marius

et al.

Publication Date

2024-08-01

DOI

10.1037/bul0000434

Psychological Bulletin

Evaluating the Robustness of Parameter Estimates in Cognitive Models: A Meta-Analytic Review of Multinomial Processing Tree Models Across the Multiverse of Estimation Methods

Henrik Singmann, Daniel W. Heck, Marius Barth, Edgar Erdfelder, Nina R. Arnold, Frederik Aust, Jimmy Calanchini, Fabian E. Gümüsdagli, Sebastian S. Horn, David Kellen, Karl C. Klauer, Dora Matzke, Franziska Meissner, Martha Michalkiewicz, Marie Luisa Schaper, Christoph Stahl, Beatrice G. Kuhlmann, and Julia Groß

Online First Publication, June 27, 2024. https://dx.doi.org/10.1037/bul0000434

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Bulletin

ISSN: 0033-2909

https://doi.org/10.1037/bul0000434

Evaluating the Robustness of Parameter Estimates in Cognitive Models: A Meta-Analytic Review of Multinomial Processing Tree Models Across the Multiverse of Estimation Methods

Henrik Singmann[1, 2], Daniel W. Heck[3], Marius Barth[4], Edgar Erdfelder[5], Nina R. Arnold[6], Frederik Aust[4, 7], Jimmy Calanchini[8], Fabian E. Gümüsdagli[9], Sebastian S. Horn[10], David Kellen[11], Karl C. Klauer[12], Dora Matzke[7], Franziska Meissner[13, 14], Martha Michalkiewicz[9], Marie Luisa Schaper[9], Christoph Stahl[4], Beatrice G. Kuhlmann[5], and Julia Groß[5]

[1] Department of Experimental Psychology, University College London
[2] Department of Psychology, University of Warwick
[3] Department of Psychology, University of Marburg
[4] Department of Psychology, University of Cologne
[5] Department of Psychology, School of Social Sciences, University of Mannheim
[6] Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg
[7] Department of Psychology, University of Amsterdam
[8] Department of Psychology, University of California, Riverside
[9] Faculty of Mathematics and Natural Sciences, Institute for Experimental Psychology, Heinrich Heine University Düsseldorf
[10] Psychologisches Institut, Universität Zürich
[11] Department of Psychology, Syracuse University,
[12] Institut für Psychologie, Albert-Ludwigs-Universität Freiburg
[13] Department of General Psychology II, Institute of Psychology, Friedrich Schiller University Jena
[14] Institute of General Practice and Family Medicine, Jena University Hospital, Friedrich Schiller University Jena

Researchers have become increasingly aware that data-analysis decisions affect results. Here, we examine this issue systematically for multinomial processing tree (MPT) models, a popular class of cognitive models for categorical data. Specifically, we examine the robustness of MPT model parameter estimates that arise from two important decisions: the level of data aggregation (complete-pooling, no-pooling, or partial-pooling) and the statistical framework (frequentist or Bayesian). These decisions span a *multiverse* of estimation methods. We synthesized the data from 13,956 participants (164 published data sets) with a meta-analytic strategy and analyzed the *magnitude of divergence* between estimation methods for the parameters of nine popular MPT models in psychology (e.g., process-dissociation, source monitoring). We further examined moderators as potential *sources of divergence*. We found that the absolute divergence between estimation methods was small on average (<.04; with MPT parameters ranging between 0 and 1); in some cases, however, divergence amounted to nearly the maximum possible range (.97). Divergence was partly explained by few moderators (e.g., the specific MPT model parameter, uncertainty in parameter estimation), but not by other plausible candidate moderators (e.g., parameter trade-offs, parameter correlations) or their interactions. Partial-pooling methods showed the smallest divergence within and across levels of pooling and thus seem to be an appropriate default method. Using MPT models as an example, we show how transparency and robustness can be increased in the field of cognitive modeling.

Henrik Singmann https://orcid.org/0000-0002-4842-3657
Daniel W. Heck https://orcid.org/0000-0002-6302-9252
Beatrice G. Kuhlmann https://orcid.org/0000-0002-3235-5717
Julia Groß https://orcid.org/0000-0002-1555-1070

Julia Groß and Beatrice G. Kuhlmann share senior authorship. The authors have no conflicts of interest to disclose. The meta-analytic data, all analysis scripts, as well as the supplemental results are available on Open Science Framework at https://osf.io/waen6/.

The authors thank Joachim Vandekerckhove, Eric-Jan Wagenmakers, Richard Chechile, and Michael D. Lee for discussions at earlier stages of the project. This work was funded by Grant GR-4649/2-1 (Scientific network) from the German Research Foundation (DFG) awarded to Julia Groß and Beatrice G. Kuhlmann. Henrik Singmann was additionally supported by Grant 100014_179121 from the Swiss National Science Foundation (SNSF).

Henrik Singmann played a lead role in formal analysis, methodology, software, visualization, writing–original draft, and writing–review and editing, a supporting role in conceptualization, and an equal role in investigation. Daniel W. Heck played a lead role in software, a supporting role in methodology, writing–original draft, and writing–review and editing, and an equal role in investigation. Marius Barth played a supporting role in methodology, software, writing–original draft, and writing–review and editing and an equal role in investigation. Edgar Erdfelder played a

continued

Public Significance Statement
Cognitive models are formal instantiations of psychological theories that are becoming increasingly popular in psychology. Using multinomial processing tree (MPT) models as an example, we show how transparency and robustness can be increased in the field of cognitive modeling. Specifically, we conducted a large-scale meta-analysis to investigate how the choice of parameter estimation method affects results. Overall, the choice of method resulted in small differences, with few exceptions, lending support for the robustness of the results of MPT modeling in psychological research. Furthermore, the most advanced method (partial-pooling, or hierarchical modeling) provided the best compromise between methods, making it an appropriate default estimation method.

Keywords: multiverse analysis, parameter estimation, transparency, cognitive modeling, multinomial processing tree models

Psychological research involves many decisions that may impact the results. For example, consider a researcher who wants to conduct a study on automatic and controlled processes in a memory task (e.g., familiarity and recollection). Before data collection, they need to decide on an experimental setup (e.g., how many items to present how many times) and choose appropriate measures to capture the relevant concepts (e.g., number of recalled items from the study list). After data collection, they need to decide on the processing of data (e.g., the handling of outliers and missing data). For data analysis, they need to choose an appropriate model (e.g., a purely statistical model, like analysis of variance [ANOVA], or a cognitive model, like a multinomial processing tree [MPT] model). In addition, they have to decide whether (and how) to take into account variability in participants and items and choose a method of inference or for parameter estimation (e.g., frequentist or Bayesian approach).

Each decision in the research process is a choice on one dimension in a multidimensional decision matrix. In most situations, different choices yield results that differ at least to some degree. However, in a published article, typically only a single cell of this decision matrix (or path through a "garden of forking paths," Gelman & Loken, 2014) is reported along with a single result, whereas alternative (yet reasonable) cells remain undisclosed (Steegen et al., 2016). As a consequence, the uncertainty/robustness of the conclusions depending on these decisions is not communicated (Wagenmakers et al., 2022). In addition, in recent years, the research community has become increasingly aware that "researcher's degrees of freedom" (Simmons et al., 2011) can invite questionable research practices such as selective reporting of desirable results. These practices are likely linked to the low reproducibility of research findings (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012).

There is now increasing effort to show how these decisions affect conclusions on the presence and size of effects, and how the decisions can best be made transparent (e.g., Baribault et al., 2018; Dutilh et al., 2019; Landy et al., 2020; Starns et al., 2019; Steegen et al., 2016). For example, Steegen et al. (2016) set up a "multiverse analysis" in which they tested the presence of an effect across different plausible data-processing choices (e.g., data exclusion, data transformation). They found that conclusions drawn from the same data set varied substantially depending on data-processing choices. Similarly, Baribault et al. (2018) set up a "meta-study" to systematically explore effects of choices in study design (e.g., mask duration in a priming experiment within a reasonable range). Again, the observed effect size varied substantially across the conducted microexperiments. Other approaches that are less systematic but more representative of empirical practice let different teams of researchers design a study ("crowdsourcing hypothesis tests," Landy et al., 2020) or analyze a data set ("many analysts," Hoogeveen et al., 2023; Silberzahn et al., 2018). In addition, many-

supporting role in formal analysis, methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Nina R. Arnold played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Frederik Aust played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Jimmy Calanchini played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Fabian E. Gümüsdagli played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Sebastian S. Horn played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. David Kellen played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Karl C. Klauer played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Dora Matzke played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Franziska Meissner played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Martha Michalkiewicz played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Marie Luisa Schaper played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Christoph Stahl played a supporting role in methodology, writing–original draft, and writing–review and editing and an equal role in investigation. Beatrice G. Kuhlmann played a lead role in conceptualization, funding acquisition, methodology, project administration, and writing–original draft, a supporting role in formal analysis and writing–review and editing, and an equal role in investigation. Julia Groß played a lead role in conceptualization, funding acquisition, methodology, project administration, writing–original draft, and writing–review and editing, a supporting role in formal analysis, and an equal role in investigation.

Correspondence concerning this article should be addressed to Henrik Singmann, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. Email: singmann@gmail.com

analysts projects have explored the impact of choice of the statistical or cognitive model or model variant ("modeler's degrees of freedom") on the robustness of results (see also Boehm, Annis, et al., 2018; Dutilh et al., 2019; Starns et al., 2019). Again, these projects revealed large heterogeneity in the teams' approaches to design a study or analyze the data.

These and similar studies promote transparency in the research process, but they vary substantially with regard to the dimension of the process they examine (e.g., data processing, hypothesis testing, data analysis) and with regard to how systematically they examine the respective degrees of freedom (e.g., if they explore all possible or only selected decisions). Typically, in these studies, the resulting divergence between results is documented. But what are the factors that lead to divergence? Such further examination or test is usually missing. Additionally, given the increasing popularity of cognitive modeling in psychological research, the impact of cognitive-modeling decisions on divergence in results is still underrepresented in the current literature (but see, e.g., Boehm, Annis, et al., 2018; Dutilh et al., 2019).

The aim of this article is to systematically explore two important cognitive-modeling decisions—the level of pooling and the statistical framework—and make transparent how these decisions affect the results. In addition, this is the first study to go a step further and examine potential sources of divergence to better understand in what situations divergence does or does not occur, and how it can be minimized.

The remainder of the introduction is structured as follows. In the next section, we describe the two modeling decisions investigated here. We will then introduce the general principle of the model class we examined in this article: MPT models (Riefer & Batchelder, 1988). We will proceed with defining the *multiverse* for our analysis—a two-dimensional matrix resulting from the two modeling decisions, implying a set of nine different estimation methods. We will continue with considerations on the conditions under which we expect divergence to (not) occur, based on basic statistical theory. We will then summarize the goals and results of our meta-analysis.

Cognitive Modeling Decisions

As a running example that we will return to throughout this article, let us assume that a researcher uses a word-stem completion task to study familiarity and recollection in memory. They have decided on all relevant aspects of the experimental design, finished collecting data, and now want to analyze the data. An increasingly popular approach in experimental psychology is to use a cognitive model to describe and explain the observed data. Cognitive models establish a link between theory and data by formally instantiating the mechanisms that are assumed to generate the data. Most cognitive models have free parameters whose values are estimated from the data. These parameters characterize the latent psychological processes (e.g., familiarity and recollection) that are assumed to underlie the observed behavior (e.g., responding with an item from the study list; e.g., Jacoby, 1991). Thus, in contrast to "purely" statistical models (e.g., ANOVA), whose parameter estimates reflect information about manifest variables (e.g., response times, choices, or number of recalled items), cognitive models aim to directly address the latent psychological processes involved in the specific paradigm under consideration (e.g., evidence accumulation in fast two-choice tasks, Brown & Heathcote, 2008; Ratcliff, 1978; reward sensitivity in decision making under risk, Tversky & Kahneman, 1992; Wallsten et al., 2005; encoding and retrieval in multitrial free recall; Alexander et al., 2016; Batchelder & Riefer, 1986). Therefore, cognitive models are useful tools for theoretical progress as they allow the researcher to test whether theoretical assumptions are consistent with the data and to determine to what degree an assumed cognitive process contributes to the observed behavior.

Here, we consider two crucial dimensions on which researchers have to make a decision when applying a cognitive model: They need to decide how they take into account similarities and differences between units of observation (i.e., pooling of data), and whether they want to estimate model parameters in a frequentist or in a Bayesian statistical framework. As with any result in psychological research, parameter estimates from cognitive models should ideally be reproducible and robust across these modeling decisions (Lee et al., 2019; Vandekerckhove et al., 2019). We focus on the level of pooling and the statistical framework, because a researcher can hardly avoid deliberating on these two aspects, whereas for other aspects (e.g., prior distributions), the researcher can resort to standards. We do not consider decisions that are not specific to cognitive modeling and have been addressed elsewhere (e.g., data processing, Steegen et al., 2016).

Pooling of Data

One of the key assumptions of commonly used statistical and cognitive models is that of *independent and identically distributed* (i.i.d.) observations or residuals. Here, we will mainly focus on the independence assumption, which states that once the structure of the model is taken into account, all observations should be unrelated. Researchers often assume independence, for example, if each observation (e.g., a response in a task) comes from a different participant. However, in most experimental psychological research, it is common to collect multiple observations from the same participant. As these observations are likely to be more similar to each other than observations from different participants, the independence assumption is likely violated in this case, if it is not properly accounted for in the model.[1]

What we describe below as the different levels of pooling are different modeling approaches that can be used when the independence assumption is likely to be violated. Consider a situation in which a researcher has collected responses from multiple participants, and the data from each participant are sufficient to estimate all model parameters. Let us further assume this model has a parameter vector $\boldsymbol{\theta}$ of fixed length. For example, the researcher might have run an experiment on familiarity and recollection in a word-stem completion task with multiple participants. Each participant was first asked to learn a list of words and then—under *inclusion* instructions—was asked to complete word stems with studied words or—under *exclusion* instructions—was asked to complete word stems with another word *not* from the studied list. If the researcher wanted to

[1] The same argument can also be made for items as a grouping factor: Responses to the same item are likely to be more similar to each other than responses to different items, which can also introduce nonindependence (Matzke et al., 2015). Here, we only focus on participants as a grouping factor, as most of the data analyzed in our meta-analysis were available only on a by-participant level.

predict the number of word stems completed with studied words in both conditions using a regression or ANOVA model, $\boldsymbol{\theta}$ would be of length two (i.e., the intercept, representing, for instance, the mean number of word stems completed with studied words under inclusion instructions, and the slope, representing the difference between the two conditions).

Alternatively, the researcher could apply the process-dissociation model (Jacoby, 1991), which yields two parameters representing different latent processes: conscious recollection, r, and automatic activation/familiarity, a (described in more detail below). In this case, $\boldsymbol{\theta}$ would also be of length two, now representing the latent processes that can be interpreted as probabilities of familiarity and recollection.

The different levels of pooling describe how many $\boldsymbol{\theta}$ are being estimated in total (e.g., Gelman & Hill, 2007). The simplest approach is *complete-pooling*. Complete pooling ignores the dependency in the data and applies the model with shared parameters to the aggregated data.[2] That is, a total of one $\boldsymbol{\theta}$ is estimated. In cognitive modeling, complete-pooling is historically the most common approach (e.g., Ratcliff, 1978; Riefer & Batchelder, 1988), but it has at least two major problems. First, in the case of substantial individual variation among parameters (i.e., a violation of the identically distributed assumption), estimates from the complete-pooling approach can be biased resulting in *aggregation artifacts* (Estes, 1956; Estes & Maddox, 2005). Second, ignoring the violations of the independence assumption often leads to overconfident results, as uncertainty estimates (e.g., standard errors) are too narrow due to *pseudoreplication* (e.g., Hurlbert, 1984; Kenny & Judd, 1986). However, complete-pooling is still used today in situations in which it is not possible to collect enough data at the individual level (e.g., in research on children or clinical patients), or when some responses occur only rarely.

An alternative to complete-pooling that avoids violations of the independence assumptions when there is only a single source of nonindependence is *no-pooling*. No pooling means estimating separate models for each participant, thus obtaining one $\boldsymbol{\theta}$ for each participant (i.e., the total number of parameters is the length of $\boldsymbol{\theta}$ times the number of participants). Many no-pooling applications also include a second analysis step in which the individual-level parameter estimates are themselves analyzed statistically (e.g., by using a t test). No pooling can only be applied if there are ample data on the individual level. With little data, individual-level parameter estimates can be biased (MLE estimates are only asymptotically unbiased; e.g., Riefer & Batchelder, 1991) or the precision can be too low for them to be of any practical use. Even with sufficient data, the no-pooling approach has the drawback of ignoring information about the similarity between participants, which can lead to an overestimation of the group variance (e.g., Boehm, Marsman, et al., 2018). There are two more problems with no-pooling when combined with a second analysis step: First, uncertainty surrounding each of the individual parameter estimates (i.e., their standard errors) is not properly propagated to the subsequent analysis step; when taking the role of data, the parameter point estimates are all treated alike, regardless of their precision (but see Jobst et al., 2020). Second, if the hypothesis of interest involves two or more MPT parameters jointly, the subsequent analysis can become statistically nontrivial and is only straightforward if the hypothesis pertains only to a single MPT parameter (in that case a t test or similar can be used).

Finally, *partial-pooling* represents the most recently developed approach, which accounts for both differences and similarities between participants by simultaneously estimating parameters at the individual level and at the group level (also known as a *hierarchical*, or *multilevel approach*; e.g., Boehm, Marsman, et al., 2018; Gelman & Hill, 2007; Lee, 2011). Partial-pooling estimates one $\boldsymbol{\theta}$ for each participant plus one group-level $\boldsymbol{\theta}$ (plus parameters of the group-level distribution, such as variances or covariances). The individual-level parameters are constrained by a specified group-level distribution, leading to hierarchical shrinkage (e.g., extreme values are "shrunk" toward the overall mean). On average, partial-pooling improves parameter accuracy by exploiting the information available from the other participants (Efron & Morris, 1977). Partial-pooling has become increasingly popular in psychology in the last two decades, either in the form of linear mixed-effects models (e.g., Singmann & Kellen, 2019) or hierarchical Bayesian models (e.g., Lee & Wagenmakers, 2013) and is often considered the "gold standard," because it avoids the problems of the two other (nonhierarchical) approaches. However, there are also drawbacks to partial-pooling. First, partial-pooling requires an assumption about the distribution of individual-level parameters, which can be false (e.g., Bartlema et al., 2014). In many applications, individual-level parameters are assumed to be unimodally distributed (e.g., Klauer, 2010). Violations of this assumption might lead to imprecision and bias in parameter estimation (e.g., Schielzeth et al., 2020), and "true" outliers, multimodal distributions, latent classes, or groups of participants with different mixtures of cognitive processes are difficult to detect. Second, the benefits of partial-pooling fail to realize, if there are strong interdependencies among parameters (Scheibehenne & Pachur, 2015). In such situations, a no-pooling approach would be more flexible, as individual-level estimates are not constrained by (or shrunk toward) the group-level estimates. Third, estimating all parameters in a fully parameterized partial-pooling model (i.e., with a maximal random effects structure; e.g., Matzke et al., 2015) can be computationally challenging.

Statistical Framework

The second important data-analysis decision we consider here concerns the statistical framework. The frequentist and Bayesian frameworks provide different bases for statistical inference and parameter estimation (for overviews, see Dienes, 2008; Kruschke & Liddell, 2018). In the frequentist framework, probabilities correspond to expected outcomes of repeated sampling (i.e., a relative frequency). Parameter estimation is typically based on maximum likelihood (ML) methods or—more generally—on minimum power divergence methods (e.g., Read & Cressie, 1988). ML methods aim at finding the set of parameter values $\boldsymbol{\theta}$ that maximizes the probability of observing the data $\boldsymbol{D}$ given those parameters, $P(\boldsymbol{D}|\boldsymbol{\theta})$. The result of maximum likelihood estimation (MLE) is a point value, the ML estimate, with an associated measure of uncertainty, the standard error (e.g., Pawitan, 2014).

[2] In the case of categorical (i.e., multinomial) data, which are the focus of the present article, complete pooling means summing all observations across participants and items. In the case of other models (e.g., regression models), multiple approaches are possible, such as treating all observations as independent, or aggregating the observations within participants first (the latter approach does not violate the independence assumption).

In the Bayesian framework, uncertainty for $\boldsymbol{\theta}$ is conveyed by probability distributions (e.g., Gelman et al., 2013). The researcher can express relative uncertainty before seeing the data in the form of the prior distribution $P(\boldsymbol{\theta})$. The prior is then updated in light of the observed data $\boldsymbol{D}$ via Bayes' theorem, resulting in a posterior probability distribution, $P(\boldsymbol{\theta}|\boldsymbol{D})$. In a Bayesian framework, the posterior distribution of a parameter provides all information available about the parameter after seeing the data. For example, the mean, median, or maximum of the posterior distribution can be used as point estimates, whereas specific intervals (such as the 2.5% and the 97.5% quantiles) can be used as measures of uncertainty. In contrast to the frequentist framework, parameter estimation in the Bayesian framework requires a commitment about the prior belief regarding which parameter values are plausible. However, in practice, the choice of prior is often inconsequential for parameter estimation as long the prior is sufficiently noninformative and there is a sufficient amount of data.

The frequentist statistical framework has long dominated research in psychological science (including cognitive psychology), but Bayesian methods are becoming increasingly popular (e.g., Lee & Wagenmakers, 2013; van de Schoot et al., 2021; Vandekerckhove et al., 2018). In part, the growing prominence of Bayesian methods is due to advances in computer-driven sampling methods such as Markov Chain Monte Carlo (MCMC; e.g., Gilks et al., 1996; van Ravenzwaaij et al., 2018). These methods allow researchers to sample from the posterior distribution, which is often difficult to obtain analytically.

For both dimensions—level of pooling and statistical framework—researchers might have good reasons to make a specific choice. Regarding the level of pooling, when the amount of data per individual is limited, a no-pooling approach is often impossible. Conversely, when wishing to investigate individual differences, the choice of a no-pooling or partial-pooling approach is necessary. Regarding the statistical framework, frequentist estimation is often computationally cheaper (i.e., faster) than Bayesian estimation, and therefore often preferred during iterative model building, when analyzing many data sets, or for teaching. On the other hand, to date, implementations of partial-pooling methods for cognitive models only exist in a Bayesian statistical framework; choosing partial-pooling therefore currently implies choosing a Bayesian estimation approach (but see Nestler & Erdfelder, 2023). In many situations, however, different choices are similarly appropriate.

Before we elaborate our approach for systematically investigating divergence in modeling results that arise as a consequence of deciding on the level of pooling and statistical framework, we will first introduce the general principle of the MPT model class, which we will focus on in this meta-analysis.

General Principle of MPT Models

MPT models are stochastic measurement models for categorical data (i.e., data that follow a multinomial distribution). They allow researchers to explain participants' behavioral responses in experimental tasks—specifically, the observed frequencies across different response categories—by estimating the contribution of different latent psychological processes that underlie these responses (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016). MPT models are tailored to specific experimental paradigms, so that each model implies a different set of unique latent processes leading to different response categories. MPT models can be flexibly used to formulate and test psychological theories of cognition, for example, in the areas of memory (e.g., Batchelder & Riefer, 1986; Bayen et al., 1996; R. E. Smith & Bayen, 2004), judgment and decision making (e.g., Erdfelder & Buchner, 1998; Gawronski et al., 2017; Hilbig, Erdfelder, & Pohl, 2010), reasoning (e.g., Klauer et al., 2000), and implicit-attitude measurement (e.g., Conrey et al., 2005; Meissner & Rothermund, 2013). For a recent tutorial on MPT-modeling methods and guidelines, see Schmidt et al. (2023).

MPT models can be depicted as processing trees in which the root node signals the start of the processing sequence for each item type and the leaf nodes represent the observable response categories. The intermediate nodes represent latent cognitive processes or mental states. The edges beginning in each node are associated with a set of model parameters that represent the probabilities with which the subsequent node (or mental state) is reached. Each path from the root node to a leaf node describes one possible sequence of latent cognitive processes leading to a specific observable response in the task. For some MPT models, a distinction can be made between parameters that represent relevant theoretical processes (henceforth referred to as *core parameters*) and parameters that are important for the model's architecture, but are theoretically less important (referred to as auxiliary parameters).

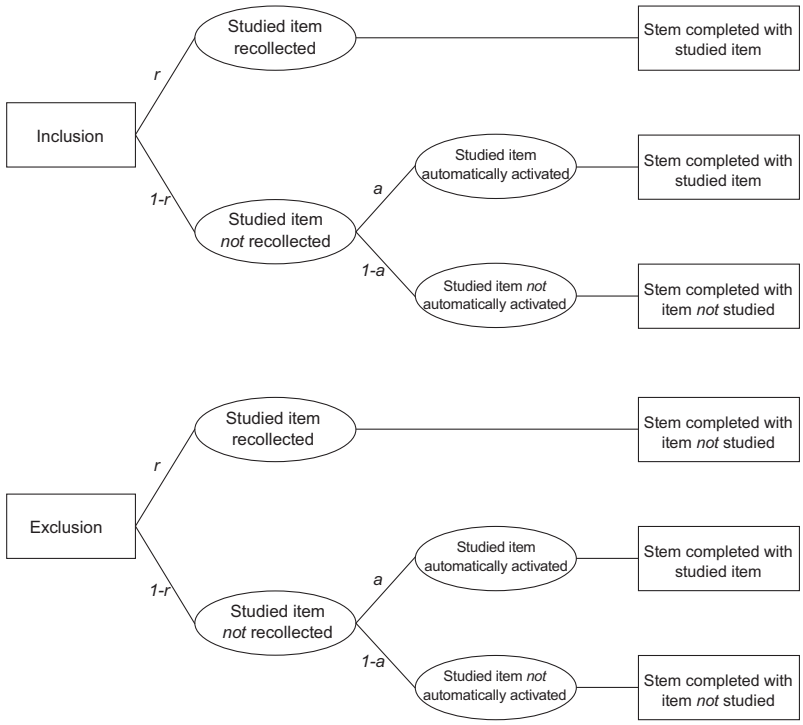
To remain with our introductory example from episodic memory, let us consider the process-dissociation model (Jacoby, 1991; Figure 1) as an example.[3] The process-dissociation model disentangles automatic from controlled processes in various paradigms. Here, we consider conscious recollection and automatic activation in an indirect memory task (word-stem completion). Participants study a list of words and, at test, are given to-be-completed stems (e.g., "ap_____" may be completed to form the word "apartment"). In the process-dissociation procedure, there are two testing conditions (see also Pooling of Data section). In the inclusion condition, participants are asked to complete the stem with a word from the study list; in the exclusion condition, participants are asked to complete a stem with a word that was not in the study list. In both the inclusion and the exclusion condition, responses can belong to one of two response categories: The stem is completed with a studied item, or the stem is completed with an item *not* studied.

Figure 1 shows how the model assumes that the latent processes combine to produce observable behavior in the response categories. There are two processing trees, one for the inclusion condition and one for the exclusion condition. Stem completion with a studied item in the inclusion condition results either from conscious recollection, with probability r, or from automatic activation, with probability a, given a recollection failure, $(1 - r)$. Stem completion with a studied item in the exclusion condition results only from automatic activation with probability a, given a recollection failure, $(1 - r)$. The model thus assumes two latent states: With probability r, a studied item is consciously recollected. With probability a, a studied item is automatically activated (given that it is not consciously recollected). By contrasting the inclusion and

[3] Note that the process-dissociation procedure and model originated independently of the MPT framework, but subsequent extensions (e.g., Buchner et al., 1995) extensively relied on the MPT framework to estimate parameters and test hypotheses.

Figure 1

Multinomial Processing Model Tree Representing the Process-Dissociation Model for Recollection and Automatic Activation in Word-Stem Completion



Note. r = probability that a studied item is consciously recollected; a = probability that a studied item that is not recollected is automatically activated.

exclusion conditions, estimates of these latent processes can be obtained. A formal introduction to the MPT model class is provided in Appendix A.

Defining the Multiverse of MPT Estimation Methods

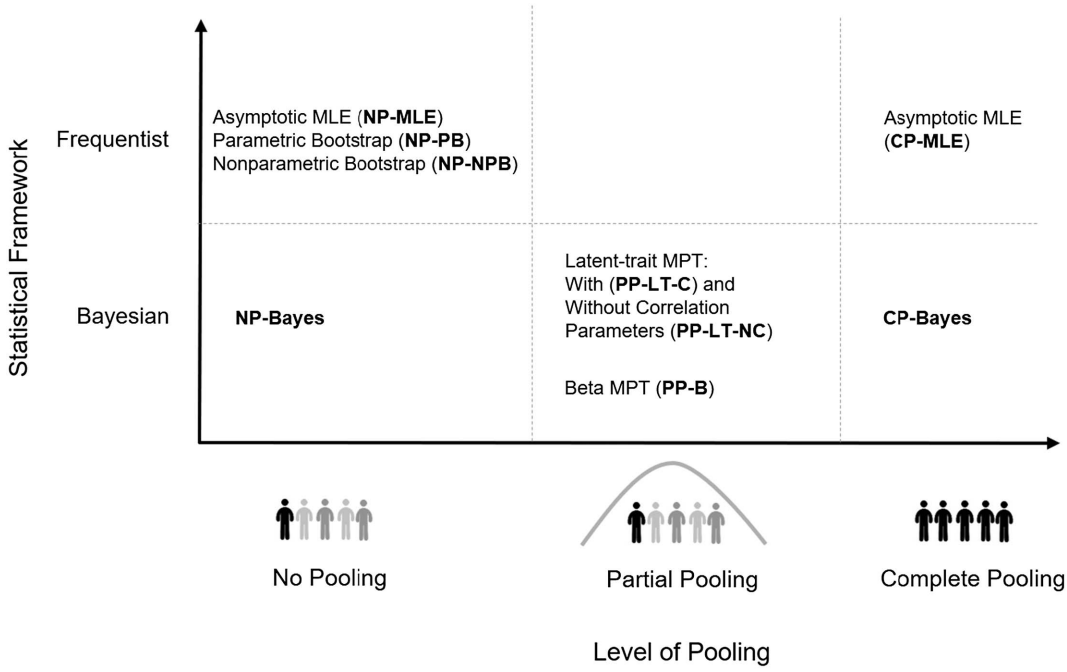
In our meta-analysis, we will explore the magnitude of divergence between MPT parameter estimates across the multiverse of estimation methods. The multiverse results from crossing the three levels of pooling (complete-pooling, no-pooling, partial-pooling) with the two statistical frameworks (frequentist, Bayesian) and is shown in Figure 2. The majority of the methods in this multiverse are established methods for MPT modeling that have evolved over the past decades in response to methodological discussions (e.g., problems associated with violations of assumptions) and technical advancement (e.g., MCMC sampling). Striving to systematically examine the complete MPT estimation multiverse, we also included any feasible and plausible combination of estimation method and level of pooling, regardless of whether they had been discussed or used previously.

To facilitate a multiverse analysis of MPT data, we have developed a freely available software tool, `MPTmultiverse` (Singmann et al., 2020) that simultaneously applies up to nine different methods to a given data set. Here, we present an application of this tool to more than 160 data sets and nine different MPT models.

One of the challenges of such a multiverse analysis is to ensure that the results from the different methods can be appropriately compared with one another. For example, a frequentist analysis provides an MLE point estimate with a standard error, whereas a Bayesian analysis provides an entire posterior distribution from which point estimates and uncertainty information need to be derived. Further, whereas a complete-pooling result refers to the group level, a no-pooling result refers to the individual level, and a partial-pooling result yields estimates on both the individual level and the group level. Thus, in order to be able to evaluate the divergence between results across the multiverse, it is necessary to derive quantities that can be compared across estimation approaches. For example, one might compare the mean of the group-level posterior distribution to a complete-pooling ML point estimate. Our multiverse analysis shows how such commensurability can be achieved (for details, see Summary Measures and Methods of Synthesis section).

On the frequentist side, we implemented complete-pooling and three different variants of no-pooling. MPT modeling has long relied on MLE complete-pooling (CP-MLE) as the method of choice (e.g., Batchelder & Riefer, 1986; Hu & Batchelder, 1994; Riefer & Batchelder, 1988). In some cases, researchers have also used MLE

Figure 2
The MPT Multiverse of Estimation Methods



Note. MPT = multinomial processing tree; NP-MLE = no-pooling, maximum likelihood estimation; NP-PB = no-pooling, parametric bootstrap; NP-NPB = no-pooling, nonparametric bootstrap; NP-Bayes = no-pooling, Bayesian parameter estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; PP-LT-NC = partial-pooling, latent-trait without correlation parameters; CP-MLE = complete-pooling, maximum likelihood estimation; CP-Bayes = complete-pooling, Bayesian parameter estimation; PP-B = partial-pooling beta. Names in bold are used throughout this article to refer to a given estimation method.

no-pooling (NP-MLE; Figure 2, top left; e.g., Meissner & Rothermund, 2013; Simons et al., 2002). To address the problem that no-pooling can lead to unstable estimates and standard errors in studies with low numbers of observations, we additionally implemented no-pooling with parametric bootstrap (NP-PB) and nonparametric bootstrap (NP-NPB). Both methods are readily available in standard MPT software (e.g., Moshagen, 2010; Singmann & Kellen, 2013).[4] All frequentist methods used in the current multiverse analysis were implemented via `MPTinR` (Singmann & Kellen, 2013).

On the Bayesian side, we implemented methods for each level of pooling. In response to some of the problems associated with complete-pooling and no-pooling, several (Bayesian) partial-pooling methods for MPT models have been developed (e.g., Klauer, 2010; Matzke et al., 2015; J. B. Smith & Batchelder, 2010). These methods differ with respect to the assumption of the group-level distribution. The *beta-MPT* approach (PP-B; J. B. Smith & Batchelder, 2010) assumes that the different types of individual-level parameters are independent and follow beta distributions on the group level. The *latent-trait* approach (PP-LT-C; Klauer, 2010) represents individual-level parameters as displacements from a group-level mean, drawn from a zero-centered multivariate normal distribution (in probit space). While parameter correlations are not part of PP-B (they can only be estimated in a subsequent step), they are specified as free parameters in PP-LT-C and can therefore be estimated directly. As an example for parameter correlations, consider that the probability of recollection and the probability of automatic process activation could be positively correlated across individuals in the process-dissociation model.

In addition to the established methods PP-B and PP-LT-C, we implemented a latent-trait variant without parameter correlations, PP-LT-NC. This method can inform whether divergence between the two partial-pooling methods, PP-B and PP-LT-C, is associated with distributional assumptions or covariance parameters. We furthermore included Bayesian variants of complete-pooling (CP-Bayes) and no-pooling (NP-Bayes). Although these latter two methods are not commonly used, they complement the existing multiverse. All Bayesian methods were implemented via `TreeBUGS` (Heck et al., 2018), either using `JAGS` (Plummer, 2003) or a custom MCMC sampler.

[4] When performing our multiverse analyses, the latent-class approach by Klauer (2006) was, to the best of our knowledge, the only readily available frequentist implementation of a partial-pooling approach (assuming that each participant belongs to a subgroup of participants with a fixed set of parameters). We initially implemented analyses with this approach but were not able to obtain reliable parameter estimates for a majority of data sets because of numerical overflow/underflow problems. We therefore did not include results from this approach into our analyses. Only after finalizing all our analyses, we became aware that Nestler and Erdfelder (2023) recently proposed an alternative partial-pooling approach based on marginal ML methods.

Under What Conditions Can We Expect Divergence Versus Perfect Agreement Between MPT Estimation Methods?

If different methods can reasonably be applied to a given data set for a given model, to what extent are cognitive-modeling results robust across, or differ between, methods? An important first step toward answering that question is to identify properties of MPT models that we expect to affect the degree of divergence between different estimation methods based on theoretical considerations. We have identified one such property, the *structural-aggregation invariance* of MPT models. Structural aggregation invariance holds for an MPT model if none of the model parameters (or its complement) occurs more than once in any of the branches of the model (Erdfelder et al., 2023). By contrast, structural-aggregation invariance does not hold for an MPT model if at least one model parameter occurs repeatedly on a branch. We provide an example, the pair-clustering model, in Appendix A. In this model, parameter u repeats within multiple branches for word pairs. As shown in Appendix B, structural-aggregation invariance is a necessary condition for predicting perfect agreement between estimation methods under conditions we will outline below. Thus, we expect larger divergences for MPT models for which structural-aggregation invariance does not hold compared to models for which this property holds.

If structural-aggregation invariance holds for a given MPT model, we can derive the following empirically testable *prediction of perfect agreement* from basic statistical theory (see Appendix B, for the formal proof): Assuming that a structurally aggregation invariant MPT model holds for each participant (with parameter values that may vary between individuals), the agreement between any pair of consistent estimation methods considered in our multiverse analysis should be perfect (i.e., there should be no divergence between group-level parameter estimates), if all correlations between parameters that occur on the same branch are approximately zero and if both the number of participants and the number of responses per participant is sufficiently large so that standard errors of the estimates approach zero. Note that whereas structural-aggregation invariance is a property of an MPT model-independent of the actual data, the prediction of perfect agreement is specific to the interplay of MPT model and data (i.e., the correlation as well as the standard error are empirical statistics).

With the large data set obtained in our meta-analysis, we are in the position to indirectly test one assumption that is shared among all common MPT estimation methods, but generally cannot be tested in a specific application of an MPT model: That the model holds for each participant.[5] This indirect test relies on the fact that the prediction of perfect agreement only holds if all of the four assumptions mentioned above (i.e., the MPT model holds for each participant, structural-aggregation invariance, correlations and standard errors of approximately zero) hold. If any of the four assumptions does not hold, then perfect agreement is not expected. Importantly, only with a large data set there is the possibility to come across conditions where three of the four assumptions hold (i.e., all but the assumption that the MPT model holds for each participant). Thus, the large data set obtained in this meta-analysis allows us to test the prediction of perfect agreement. If the three assumptions hold, but the prediction of perfect agreement does not hold, this would indicate that the MPT does not hold for each participant.[6] Such an outcome would threaten the validity of many results previously obtained with MPT models, as all estimation methods assume the same model holds for each participant. By contrast, if the remaining three assumptions hold and the prediction of perfect agreement also holds, this would provide corroborating evidence for the validity of this central assumption.

Can We Identify Moderators That Explain Divergence Between Methods?

Even though the prediction of perfect agreement provides an important test bed for the MPT model class, this prediction only applies in a limited set of circumstances. Whereas most MPT models are structurally aggregation invariant, in the vast majority of MPT applications in psychology, there is a nonzero correlation among parameters that occur on the same branch and/or a nonzero standard error. An important additional goal of this article is therefore to identify empirical conditions that lead to larger or smaller divergences among different estimation methods in situations that fall outside the scope of the prediction of perfect agreement.

Our empirical approach for identifying such conditions can be seen as complementary to parameter recovery simulation studies in which the "ground truth" is known. For example, Chechile (2009) found that across different generic MPT models, complete-pooling (based on MLE) produced more accurate results than no-pooling (partial-pooling was not considered; see also Batchelder & Riefer, 1986). Groß and Pachur (2020) found that partial-pooling methods yielded more accurate estimates than complete-pooling or no-pooling methods for two exemplary MPT models (see also Jobst et al., 2020; Rouder & Lu, 2005; Shiffrin et al., 2008); however, within the partial-pooling methods, the latent-trait approach with explicit modeling of correlations (PP-LT-C; Klauer, 2010) seemed to be overparameterized for a 13-parameter MPT model (HB13; Erdfelder & Buchner, 1998), yielding less accurate results than the simpler beta-MPT method (PP-B).

Whereas simulation studies can shed light on the conditions under which we should expect divergences in empirical (i.e., real) data, the extent to which they sufficiently capture the conditions that appear in empirical data remains unclear. In our analysis, we will therefore consider a number of potential moderators and examine the degree to which they can explain diverging results between different methods.

First, in line with the theoretical predictions described above, any condition related to uncertainty in estimation (i.e., estimation error) could explain divergence between any of the methods considered. For example, the amount of data that is available to estimate a specific parameter might affect estimation uncertainty. Likewise, the presence of parameter trade-offs (i.e., structural relationships between parameters) might make it difficult to estimate a parameter independently of other parameters in the model, thereby affecting

[5] By "holds for each participant," we mean that each individual is associated with a multinomial probability distribution over the responses in line with the respective MPT model.

[6] Strictly speaking, for the reasons outlined in Appendix B, this conclusion is generally true only for MPT models with no more than two parameters per branch. Theoretically, for MPT models with three or more parameters per branch, complex patterns of stochastic dependence within branches may occur despite all correlations within branches being zero. This would violate aggregation invariance and could thus counteract perfect agreement between estimates. However, such a scenario would be in conflict with all hierarchical models considered in our multiverse analysis and seems unlikely in practice.

uncertainty in estimation (e.g., Krefeld-Schwalb et al., 2022; Spektor & Kellen, 2018). Second, for small to moderate sample sizes, parameter heterogeneity could explain divergence between methods that do (vs. do not) take heterogeneity into account (i.e., complete-pooling vs. other methods). Third, parameter correlations across individuals between parameters that occur within the same branch of an MPT model might explain divergence between methods that do (vs. do not) explicitly model these correlations (i.e., latent-trait vs. other methods), because not considering those correlations in a model can lead to biased parameter estimates (Erdfelder, 2000; Erdfelder et al., 2023; Klauer, 2010). Fourth, the actual parameter values might explain divergence between methods. For example, in MPT models, the parameters are probabilities bound between 0 and 1; parameter values near these boundaries are often estimated less precisely and therefore might show greater divergences across estimation methods. Finally, model fit (or misfit) might signal a divergence between methods. In the Method section, we provide a detailed description of all moderators considered.

The Present Study

The goal of our multiverse meta-analysis is to document how researchers' decisions regarding different estimation methods in cognitive-modeling affect results and to identify empirical conditions that help explain the divergence between methods that we observe. To do so, in the present study, we performed a systematic meta-analysis and gathered the available data from 164 published data sets that applied one of nine predefined MPT models. We selected these nine models because of their popularity and widespread use within the fields of memory, judgment and decision making, and social cognition. The models, along with their core parameters, are described in detail in Table 1. We then applied the multiverse of estimation methods, using the `MPTmultiverse` software package, to these data sets and examined divergence in core model parameters between nine implemented estimation methods that adopt different levels of pooling within different statistical frameworks. This large-scale meta-analysis allowed us to answer three main research questions:

1. What is the actual *magnitude of divergence* between results coming from the application of different estimation methods to the same empirical data?

By examining divergence between parameters estimated from empirical data, our findings can inform whether results and conclusions can be considered robust with regard to the estimation method used. By capitalizing on a large set of published data across a broad selection of MPT models and research questions, our findings provide information about a wide range of typical MPT modeling conditions (e.g., populations and paradigms).

2. Does the *prediction of perfect agreement* hold for the MPT model class?

The different estimation methods considered in our multiverse can accommodate different degrees of individual differences and are derived from different statistical frameworks. Despite these differences, we have derived specific conditions, based on basic statistical theory, under which we expect the different estimation methods to agree perfectly: For structurally aggregation invariant MPT models for which both the correlations among parameters on the same branch as well as the standard error approach zero. These conditions occur only rarely in regular-sized empirical data sets, but are likely to be present in our large meta-analytic data set. If this prediction does not hold, the validity of MPT modeling in its entirety would be threatened. However, if this prediction holds, the statistical assumptions underlying MPT modeling would be to some degree justified.

Table 1
MPT Models Included in the Meta-Analysis

Paradigm	Model	Model separately measure	Core parameter
1. Recognition memory	Two high-threshold model for confidence ratings (Bröder et al., 2013), for 6-point and 8-point scales.	Item recognition and guessing	D_N, D_O, g
2. Source monitoring	Two high-threshold model of source monitoring for two sources (Bayen et al., 1996).	Item recognition, source memory, and various forms of guessing	b, D_1 ($= D_2 = D_N$), g ($= a$), ($d_1 = d_2$)
3. Free recall	Pair-clustering model without singletons (Batchelder & Riefer, 1980, 1986).	Storage and retrieval processes in free recall	c, r, u
4. Prospective memory	Prospective memory model (R. E. Smith & Bayen, 2004).	Prospective and retrospective components of prospective memory, ongoing-task ability	P, M, C_1, C_2
5. Hindsight bias	Hindsight bias model (Erdfelder & Buchner, 1998).	Reconstruction and recollection processes in hindsight judgments	b, c, r_C, r_E
6. Recognition-based inference	r-model (Hilbig, Erdfelder, & Pohl, 2010).	Recognition and further knowledge as bases of inference	r, a, b
7. Implicit-attitude tasks	Quad model (Conrey et al., 2005; with the model specification used in Calanchini et al., 2014).	Automatic association activation, discriminability, and overcoming bias	AC, D, G, OB
8. Implicit-attitude tasks	ReAL model (Meissner & Rothermund, 2013).	Evaluative association, recoding, and label-based identification	Re, A_1, A_2, L_1, L_2, L_3, L_4
9. Various paradigms	Process-dissociation model (without guessing parameter; Jacoby, 1991).	Controlled and automatic processes (e.g., recollection and automatic activation)	r_I ($= r_E$), a

Note. The definition of the core parameters for each model can be found in Tables 2 to 9.

3. What are the *sources of divergence*?

As the conditions under which we expect the different methods to agree perfectly only occur rarely in empirical data, an important question for MPT practitioners is to understand the conditions under which they can predict the degree of divergence between the different estimation methods. Our study is the first to examine these conditions empirically in a large-scale meta-analysis. We consider several candidate moderators that capture aspects of estimation uncertainty, heterogeneity, parameter correlations, parameter trade-offs, and model fit (details are provided in the Method section). Note that some of the moderators can be controlled by the researcher prior to data collection (e.g., the number of observations, or population under investigation), whereas others cannot be controlled (e.g., parameter trade-offs), or only partly be controlled (e.g., relative information to estimate a parameter). Our analysis is thus not only informative with regard to the robustness (or instability) of previously published results, but will also be of high practical relevance for future MPT-modeling applications.

Method

Transparency and Openness

This meta-analysis followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines (Moher et al., 2009) and was initially preregistered. Our preregistration protocol can be found at https://osf.io/bpuwj/ (published in July, 2018). Once we obtained all data, we decided to improve the preregistered analyses plan in several core aspects (e.g., the dependent variable, definition of some key independent variables, and the theoretical framing). As a consequence, only the data-collection part should be considered formally preregistered.

All meta-analytic data, analysis code, as well as supplemental results are available on the Open Science Framework (OSF) at https://osf.io/waen6/ (Singmann et al., 2024). Data were analyzed using R, and the final analysis is based on Version 4.3.2 (R Core Team, 2022).

Inclusion and Exclusion Criteria

Eligible for inclusion in the reanalysis were all empirical data sets generated by human participants (i.e., no simulated data, no animals) for which an MPT analysis based on one of the nine MPT models (Table 1) had been published in English or German in a book chapter or peer-reviewed journal (including footnotes, appendices, and additional online material is available at OSF: https://osf.io/waen6/).[7] For practical reasons, the search was limited to publications prior to June 1, 2018.

Information Sources

We developed literature search strategies separately for each model using text strings identical or related to the respective MPT model name (e.g., "process dissociation model," "process-dissociation model"; search strings for each model are documented in the additional online material is available at OSF: https://osf.io/waen6/). We searched APA PsycInfo, PubMed, Scopus, and Web of Science.[8] We also scanned the literature for references that cite the original model publication.

Study Selection

The study selection process is depicted in Figure 3. The steps were carried out separately for each of the nine selected MPT models by experts of the respective model. In an initial step, records were screened for eligibility and excluded if they did not meet the criteria based on screening of the abstract and/or the full text. In a second step, the remaining full-text articles were assessed for eligibility by detailed inspection, and reasons for exclusions were listed. The excluded articles belonged to one (or more) of four categories: No original data were reported; no MPT modeling was conducted; a paradigm was used that precluded use of the MPT model; a different MPT model variant than listed in Table 1 was used. In a third step, for the remaining eligible studies, the availability of individual-level data was checked. In a fourth step, we prepared the data for reanalysis which required splitting some of the experiments into independent data sets for technical and/or conceptual reasons. The results reported below are based on these 164 independent data sets from which we derived a total of 1,779 group-level core parameter estimates.

Data Collection

We extracted individual-level category frequencies for each data set in each selected study. These were required in order to apply all nine estimation methods. If individual-level data were not available in the study, we contacted the authors of the study. We used data that were made available to us by October 15, 2018. We extracted the class of MPT model, study population (i.e., college students, older adults, children, clinical patients, other), as well as any grouping variable that specifies an experimental or quasi-experimental condition of the study.

Summary Measures and Methods of Synthesis

We re-analyzed the category frequencies for all available data sets with the R package `MPTmultiverse`, which was developed specifically for that purpose (Singmann et al., 2020). `MPTmultiverse` combines the packages `MPTinR` (Singmann & Kellen, 2013) and `TreeBUGS` (Heck et al., 2018). The estimation methods were applied to complete data sets (i.e., across between-subjects or within-subjects conditions). Next, we provide a description of how we obtained group-level parameter estimates and corresponding standard errors for each method per data set. When the data set consisted of multiple conditions, we estimated separate group-level parameters per condition.

Frequentist Methods

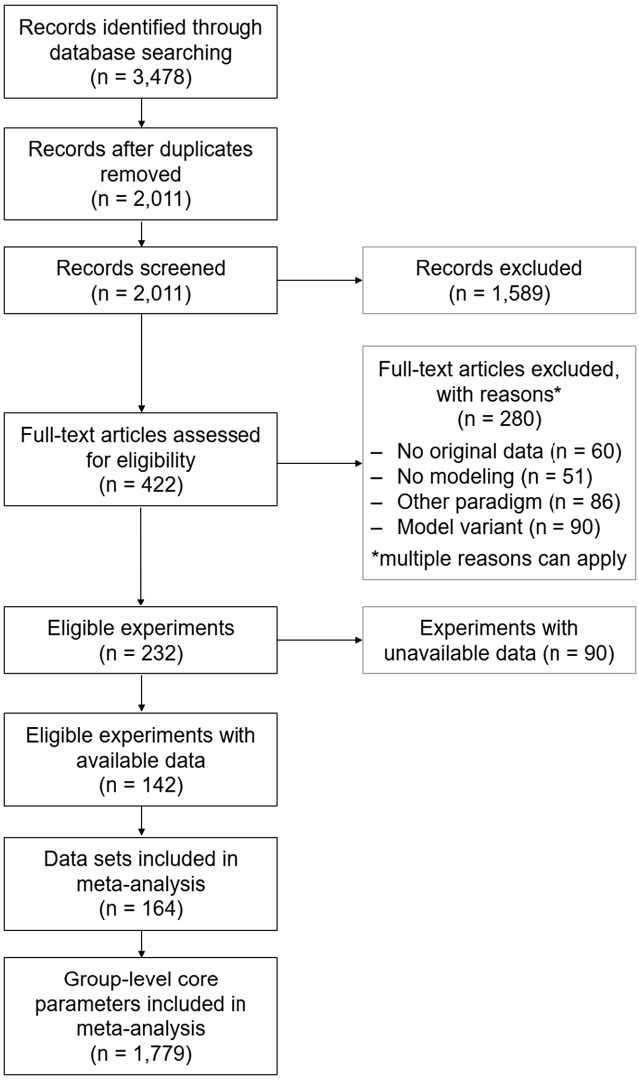
Estimates for all frequentist methods were obtained through gradient-based numerical optimization using 10 fitting runs with different random starting values for each data set. For the

[7] We initially wanted to include only those studies that used one of the nine estimation methods from our multiverse analysis (Figure 2). We discarded this criterion after thorough discussion, as our focus was not on the specific results of the initially published analyses, but rather on the divergence between the estimation methods included in our multiverse analysis.

[8] Because each search was performed by a different group of authors at different institutions, we cannot provide further details regarding which individual databases were accessed for each Web of Science search.

Figure 3

Meta-Analysis Flowchart for Study Selection



Note. No original data = no original data were reported; no modeling = no MPT modeling was conducted; other paradigm = a paradigm was used that precluded use of the MPT model; model variant = a variant of the MPT model was applied. MPT = multinomial processing tree.

complete-pooling method, the group-level estimates were the estimates from the fitting run that produced the largest likelihood, and the standard error was based on either the analytical Hessian matrix or—when calculating the analytical Hessian matrix failed—the numerical Hessian matrix (i.e., standard errors are computed as the square root of the diagonal of the inverse of the analytical/numerical Hessian matrix; e.g., Pawitan, 2014). For the no-pooling estimates, the group-level estimates were the mean of the individual-level estimates, and the standard errors were the standard errors of the mean (i.e., standard deviation of individual-level estimates divided by the square root of the number of estimates), excluding all individual-level estimates (on a case-wise basis) that were empirically not identified. For the asymptotic no-pooling method, empirical identifiability was assessed by comparing results of independent fitting runs that produced approximately the same likelihood (absolute difference on log-likelihood scale < .01) and we retained those estimates that were also approximately equal (i.e., maximum absolute difference < .01).

For the bootstrap no-pooling methods, we created a bootstrap distribution of estimates for each individual participant using 1,000 bootstrap samples. That is, for each participant, we created 1,000 sets of synthetic data of the same size as the original data and then fit the model to each of these synthetic data sets and recorded all parameter estimates. The distribution of the estimates from fits to the synthetic data is the bootstrap distribution of the parameter estimates. The nonparametric samples are random samples from a multinomial distribution in which the probabilities are given by the data. The parametric samples are random samples from a multinomial distribution in which the probabilities are given by the predictions of the model fit to the data (i.e., these samples assume that the model fits the data). We discarded estimates as not being empirically identifiable if the width of the 95% bootstrap CI (i.e., the 97.5% quantile minus the 2.5% quantile of the bootstrap distribution) spanned almost the full parameter range (i.e., was larger than .99).

Bayesian Methods

All Bayesian estimates were based on MCMC sampling using either custom-written samplers (complete-pooling and no-pooling) or using `JAGS` (partial-pooling methods; Plummer, 2003). For all methods, results were based on three independent chains. We only retained a result if all parameters achieved an $\hat{R} < 1.05$ and the number of effective samples was >2,000. The minimum number of post burn-in samples was 50,000 per chain (with 20,000 burn-in and 10,000 adaptation samples) with a thinning rate of 10. Due to large differences in model complexity and data sizes, the number of samples and thinning rate had to be adapted by the researchers in order to reach the convergence criteria (for some data sets, repeated attempts to reach convergence criteria were without success). We used the default `TreeBUGS` priors for all Bayesian methods. Specifically, for the beta-MPT method (PP-B), we defined the priors for the shape parameters α and β of the group-level distributions as gamma distributions with shape = 1 and rate = 0.1, with the lower bound truncated at 1. For the latent-trait method (PP-LT), the priors for the group-level parameters were standard normal distributions ($\mu = 0$, $\sigma^2 = 1$), implying uniform distributions in probability space.

For the complete-pooling and partial-pooling methods, the group-level estimates were the posterior means of the corresponding parameters, and the standard errors are the posterior standard deviations. For the no-pooling method, the individual-level estimates were averaged for each sample and then group-level estimates and standard errors were obtained as for the other methods.

Quantifying Divergence

To quantify divergence, we used two measures. First, to quantify divergence across estimates, we used the root-mean-squared error (RMSE). It represents the expected prediction error from predicting the value of the estimate of one method from the value of the estimate of the other method. Importantly, the RMSE is on the same scale as the MPT parameter estimates and can therefore be readily interpreted. For example, an RMSE of .1 indicates that the expected prediction error is .1 parameter value (i.e., 10% of the possible parameter range).

As for any squared deviation measure, larger divergences affect the RMSE more strongly than smaller divergences.

Second, to quantify divergence for specific estimates, we used the absolute deviation. Like the RMSE, absolute deviation is on the same scale as the MPT parameter value and thus is easy to interpret.

Overview of Analysis

Our results are split into three parts along the three research questions listed above. In Part 1, we provide an overview of the *magnitude of divergence* across estimation methods. In this part, we also provide a test whether parameter estimates for models for which structural-aggregation invariance holds show smaller divergences than estimates for models for which it does not hold. In Part 2, we test whether the theoretically derived *prediction of perfect agreement* holds under the specified conditions. In Part 3, we check whether we can identify *sources of divergence*; that is, moderators that allow us to predict the degree of divergence.

In Part 1 of the results, we look at the magnitude of divergence. In the first subsection, we focus on the divergence across data sets, MPT models, and parameters. In the second subsection, we focus on the divergence across all $(9 \times 8)/2 = 36$ pairs of estimation methods. To avoid an overwhelming amount of details and to provide a compact presentation of the results, Parts 2 and 3 focus on a selected set of five method pairs. These methods reflect the most commonly used ones within the method multiverse. Specifically, we designated two *reference methods*, complete-pooling MLE (CP-MLE) and the latent-trait method with correlations (PP-LT-C), and then calculated the absolute deviation between the estimates of the reference method and those of selected *comparison methods*. As comparison methods, we designated no-pooling MLE (NP-MLE), the partial-pooling beta method (PP-B), and the latent-trait method without correlations (PP-LT-NC). The complete results for all comparison methods are presented in the additional online material available at OSF: https://osf.io/waen6/.

Moderators

As moderators that could potentially explain the magnitude of divergence, we considered any variable obtained in our analysis that could affect parameter estimation. For some of the moderators, we had a priori predictions of how they might affect the magnitude of divergences as noted below; for others, we did not. The moderators can be classified as either model-dependent or model-independent. We call a moderator *model-dependent* if the value of the moderator depends on the specific MPT model. For example, the model parameter is specific to each MPT model (e.g., parameter *a* in the process-dissociation model may exhibit larger divergence than parameter *r*) and is thus model-dependent. All model-dependent moderators are categorical. By contrast, we call a moderator *model-independent* if the value of the moderator does not depend on the specific MPT model under consideration. For example, the standard error of any estimate from any MPT model can in principle take on any value between 0 and 1. Model-independent moderators can be continuous (e.g., standard error) or categorical (e.g., population of participants).

From a theoretical perspective, the model-independent moderators are more interesting, because if we were to find that the largest sources of divergence were model-dependent moderators, our results were only applicable to the set of MPT models considered here. However, if we were to find that model-independent moderators can help explain the observed divergences, then it is reasonable to assume that our results generalize beyond the MPT models considered here. In the following, we list all moderators considered in our analysis, beginning with the model-dependent moderators.

Model-Dependent Moderators. We consider four model-dependent moderators. These are the *model* (a factor with nine levels shown in Table 1); the *submodel* (an alternative version of the model factor with 13 levels, which also considers subtypes of some of the models; e.g., for the source monitoring model 2HTSM, we consider submodels 4 and 5d from Bayen et al., 1996, and a six-parameter variant from Bell & Buchner, 2010 which we henceforth refer to as submodel 6e); model *parameter* with 53 levels (here we consider parameter nested in submodel, e.g., parameter *b* from the 2HTSM is represented in three different levels, one for each submodel); and *data set* (with 164 levels).

Model-Independent Categorical Moderators. We consider two categorical moderators that are model-independent: the *population of participants*, which has five different levels (college students, older adults, children, clinical patients, and other), and the *scientific goal*, which has two levels representing two distinct reasons why the researchers of the original articles used an MPT model—either for parameter estimation (e.g., to compare model parameters across conditions) or for model selection (e.g., compare different models with each other). In addition, we considered a number of model-independent continuous moderators.

Standard Error. Appendix B shows that one potential source of divergence is the estimation uncertainty expressed in the standard errors (*SE*s) of the estimates. The way the *SE*s were estimated are specific for each method and described in Summary Measures and Methods of Synthesis section. To streamline presentation, we present results of the combined (i.e., average) *SE* of both methods in a pair, after winsorizing each individual *SE* at a maximum of 0.25 (to reduce the influence of a few large outliers). In line with the theoretical predictions, we expected larger *SE*s to be associated with larger divergence. Results of individual *SE*s (with and without winsorizing) are qualitatively the same and presented in the additional online material is available at OSF: https://osf.io/waen6/.

Parameter Correlations. The other moderator identified in Appendix B is the correlation between parameters on the level of the participants, which are modeled explicitly only in the fully parameterized latent-trait model (PP-LT-C). To assess the impact of parameter correlations, we recorded the posterior means of the individual-level correlation parameters from the latent-trait model, ρ. Following Appendix B, of particular importance are the correlations among parameters that appear on the same tree branch. In the analysis, we used the maximum correlation for a specific parameter and data set with all parameters that appear on the same branch.[9] In line with the theoretical predictions, we expected larger parameter correlations to be associated with larger divergence.

[9] Results using a summary statistic other than the maximum correlation with all parameters that appear on the same branch (maximum correlation, median correlation, mean correlation, and proportion of correlations larger than .5 for correlations with all parameters that appear on the same branch as well as correlations with all other parameters, not only those on the same branch) are qualitatively the same and presented in the additional online material available at OSF: https://osf.io/waen6/ (Singmann et al., 2024). The same holds for the results regarding parameter trade-offs discussed below.

Values of Parameter Estimates. Another potential source of divergence between methods is the value of the estimate itself. For example, values near the boundaries of the parameter space (i.e., 0 or 1) may show larger divergences than values near the mid of the parameter space (i.e., .5). Therefore, we will test for a u-shaped (or other quadratic) relationship between the parameter value with the absolute deviation.

Interindividual Differences: Standard Deviation. To assess the effect of interindividual differences on divergence, we used two measures. First, we assessed interindividual differences by calculating the standard deviation (*SD*) of the individual-level posterior means from the latent-trait partial-pooling model (after a retransformation to the probit scale).[10] We expected larger *SD*s to be associated with larger divergence.

Interindividual Differences: Heterogeneity. Second, we assessed the heterogeneity of the observed response frequencies across participants in a nonparametric manner. Specifically, we applied the asymptotic χ^2 test proposed by J. B. Smith and Batchelder (2008) to each data set. As a measure of heterogeneity, we used Cohen's *w*, an effect size measure for χ^2 tests. We expected larger effect sizes to be associated with larger divergence.[11]

Parameter Trade-Offs. One potential reason for estimation uncertainty is the presenece of structural parameter trade-offs, or parameter fungibility (e.g., Krefeld-Schwalb et al., 2022; Spektor & Kellen, 2018). Such structural relationships can produce distortions in parameter estimation and can make it difficult to interpret the model parameters independently from each other. We estimated parameter trade-offs via the MCMC chains of the latent-trait model. Specifically, we defined the parameter trade-off for a specific parameter and data set as the maximum across-chain correlation of the group-level parameter with other group-level parameters that appear on the same branch. We chose the maximum (instead of mean, or median) absolute group-level correlation because a high trade-off with one parameter cannot be compensated by a low trade-off with other parameters. We expected larger parameter trade-offs to be associated with larger divergence.

Relative Information. One property of the MPT model class that could also affect parameter estimation is the tree structure. Specifically, estimates of parameters near the root of the tree govern how much information is available for estimating parameters appearing in the subsequent branches. For example, if the estimate for *r* in the process-dissociation model (Figure 1) is high, then there is only little information available for estimating *a*. In the extreme case of $r = 1$, the estimate of *a* is not identifiable (i.e., *a* can take on any value with the model making the same prediction).

To assess the effect of the proportion of information that is available for estimating each parameter, we first calculated the probability for each branch (i.e., the product of the parameters in the branch). The relative information for a parameter is operationalized as the sum of all estimated branch probabilities containing this parameter and varies between 0 and 1. As this measure was highly correlated across methods ($r \approx 1$), we only report results based on the relative information of the reference method.[12] Thus, we obtained one relative-information estimate for each parameter and data set. We expected smaller relative information to be associated with larger divergence.

Relative *N*. Relative information does not account for the sample size of a data set. Therefore, we also considered relative *N*, which we defined as the relative information multiplied by the total number of observations for a data set. We expected smaller relative *N* to be associated with larger divergence.

Model Fit. We also considered the fit of the model. As a measure, we used the *p* value of the fit statistic. The null hypothesis is that the model fits the data. Consequently, we expected that small *p* values (i.e., indicating a model does not fit the observed data well) are associated with larger divergence. The method for obtaining the *p* values differed across estimation methods. For the frequentist methods, we used G^2 tests, either asymptotic or based on the bootstrap distribution. For the no-pooling method the G^2 statistics were summed across participants before calculating the *p* value. For the Bayesian methods, we used posterior predictive tests based on the T_1 statistic (Klauer, 2010).

Potential Biases Due to Study Selection

Our method of identifying suitable studies is potentially biased in several ways: We only included published studies from English or German sources that were published prior to June 1, 2018 and could be found using our search strategy. Studies in other languages, more recent studies, and studies only discoverable through other databases not included in our search are not part of this meta-analysis. While this might have affected the selection of studies, we are confident that this did not introduce any bias in our conclusions. The reason for this is that our aim differed from that of a typical meta-analysis. We were not interested in estimating the size or direction of a specific effect for a specific model parameter, but in assessing the divergence of parameter estimates across estimation methods for all core model parameters. Hence, a selection bias for a specific effect—such as a possible publication bias toward studies showing a significant difference for a specific model parameter—is unrelated to and inconsequential for our actual research question. This also implies that tests for publication bias, such as funnel plots, cannot be applied to our meta-analysis. The only requirement for our conclusions to be valid is that the selection of studies is representative of studies using MPT models at large. Given our focus on nine of the most popular MPT models, and that our researcher team has extensively worked with these models, we are confident that our study selection meets this requirement.

Another potential bias that is specific to our meta-analysis is that the original estimation method is disproportionately likely to be a frequentist method (which used to be common for decades) as compared to a Bayesian method (which is commonly used only recently). However, because we re-analyzed the data with each approach of the MPT multiverse, the uneven distribution of original estimation methods also does not pose a threat to the validity of our analyses. Further, studies with an application of MPT models other than the ones used here (Table 1) are missing. Different results might emerge for MPT models we did not consider. However, due to our

[10] We did not record the posterior estimate of the *SD*, which is a model parameter of the latent-trait model.

[11] We used additional measures based on the same statistical test as alternative measures of heterogeneity. These were the χ^2-value of the test, the *p* value of the test, and the effect size measure Cramer's *V*. All of these showed qualitatively the same pattern as Cohen's *w* (see additional online material available at OSF: https://osf.io/waen6/; Singmann et al., 2024).

[12] The reason these differ slightly across methods is that the relative information is calculated from the estimated parameter values which differ across methods.

Table 2
Overview of Recognition Memory Data Sets for the Confidence-Rating Two High-Threshold (c2HT) Model

c2HT: 6-point scale				Mean absolute deviation		
Data set	Population	N	K	D_N	D_O	g
Dube and Rotello (2012), pictures	Students	27	400	.07	.03	.05
Dube and Rotello (2012), words	Students	22	400	.06	.02	.02
Heathcote et al. (2006), Experiment 1	Students	16	558	.03	.01	.01
Heathcote et al. (2006), Experiment 2	Students	23	560	.02	.01	.01
Jaeger et al. (2012)	Students	63	120	.14	.06	.09
Jang et al. (2009)	Students	33	140	.10	.03	.04
Koen et al. (2013), Experiment 2 (full attention)	Students	48	198	.06	.02	.03
Koen et al. (2013), Experiment 4 (immediate)	Students	48	300	.04	.02	.02
Koen and Yonelinas (2010), pure list	Students	32	320	.10	.02	.04
Koen and Yonelinas (2011)	Students	20	600	.03	.02	.02
Pratte et al. (2010)	Students	97	480	.03	.01	.01
D. G. Smith and Duncan (2004)	Students	30	140	.06	.02	.03
c2HT: 8-point scale				Mean absolute deviation		
Data set	Population	N	K	D_N	D_O	g
Benjamin et al. (2013)	Students	124	120	.08	.02	.03
Onyper et al. (2010), pictures	Students	136	768	.03	.00	.01
Onyper et al. (2010), words	Students	131	768	.04	.01	.01

Note. N = number of participants, K average number of responses per participant. The rightmost columns show the mean absolute deviations the core model parameters across method pairs and (if present) within-subject conditions. D_N = detect new; D_O = detect old; g = guess old.

inclusion of a large number of model-independent moderators, as well as the careful selection of MPT models from various research fields, this seems unlikely.

Results

Tables 2–9 list all 164 data sets included in our meta-analysis, one table for each MPT model (the Quad and ReAL model are presented in one table; see also Table 1), including the study population, number of participants, and mean number of trials per participant. We applied the multiverse of all nine estimation methods (Figure 2) to all 164 data sets. For 87% of the data sets, we successfully obtained estimates from all nine methods; for 10%, we obtained estimates from eight methods (i.e., one method failed for those data sets), and for 3%, we obtained estimates from seven methods. That is, if a specific method failed for a data set, this was not a strong indication that another method would fail as well. The Bayesian methods were more likely to fail than the frequentist methods. Failure rates were 5% each for PP-LT-C and PP-LT-NC, 4% for NP-Bayes, 2% for NP-NPB, 1% for PP-B, and 0% for all other estimation methods.

Part 1: Magnitude of Divergence

Divergence Across Data Sets and Models

Tables 2–9 also list the mean absolute deviation for each data set and core model parameter across all method pairs, except for the NP-Bayes method for reasons discussed below. Because the mean absolute deviation is on the same scale as the MPT parameter, it can range from 0 (i.e., no divergence) to 1 (i.e., maximum divergence). Inspection of the mean absolute deviations suggests that the largest systematic differences occurred across models and model parameters, but not so much across the data sets within an MPT model.

The model with the largest mean absolute deviations is the pair-clustering model (Table 4, Batchelder & Riefer, 1986, also introduced in Appendix A)—the only model which does not possess the structural-aggregation invariance (SAI) property.[13] For this model, u shows the smallest mean absolute deviation (≈.04) and r the largest (≈.1 or more). These results provide a first corroboration of the prediction that models for which SAI holds show smaller divergences than models for which it does not hold.

However, also for the remaining models for which SAI holds, there are considerable differences on the level of the parameter. Some parameters—such as OB (Table 8) and r_I (Table 9)—have comparatively large mean absolute deviations of ≈.1 or larger, whereas others—such as D_N (Table 2), d (Table 3), b (Table 6), and Re (Table 8)—have medium mean absolute deviations of ≈.05. Still others have small mean absolute deviations near 0—such as D, (Table 3), a (Table 7), and a (Table 9). In addition, single data sets showed markedly larger divergences when compared to other data sets for a given model (e.g., R. E. Smith et al., 2014, Experiment 1, Table 5 and Van Dessel et al., 2017, Table 8). Neither the study population,

[13] Notably, to maximize effects of SAI violation, we deliberately excluded singletons from our multiverse analyses of the pair-clustering model and focused on recall patterns for word pairs only. As shown by Erdfelder et al. (2023), systematic estimation bias in the pair-clustering model can be prevented to some degree by including many singletons into the analysis, because singletons—in contrast to unclustered words from pairs—yield unbiased estimates of parameter u that counteract the systematic biases induced by SAI violation. Because we were mainly interested in examining effects of SAI violation on RMSE, we focused on pair-clustering data for word pairs exclusively.

Table 3
Overview of Source Monitoring Data Sets for the Two High-Threshold Model of Source Monitoring (2HTSM)

2HTSM submodel 4				Mean absolute deviation			
Data set	Population	*N*	*K*	*b*	*D*	*g*	*d*
Arnold et al. (2013)	Students	48	64	.01	.01	.02	.08
Besken and Gülgöz (2008)	Older adults and students	80	54	.07	.02	.05	.06
Bayen and Kuhlmann (2011) Experiment 1	Students	48	96	.04	.02	.02	.05
Bayen and Kuhlmann (2011) Experiment 2	Students	72	96	.01	.01	.03	.09
Giang et al. (2012)	Students	58	80	.01	.01	.01	.05
Kuhlmann et al. (2012) Experiment 1	Students	72	96	.00	.01	.01	.07
Kuhlmann et al. (2016) Experiment 1	Older adults and students	144	72	.02	.01	.01	.02
Kuhlmann et al. (2016) Experiment 2	Students	72	72	.02	.00	.02	.02
Kuhlmann and Touron (2017)	Older adults and students	159	100	.03	.01	.02	.02
Meiser (2003)	Students	66	72	.03	.00	.01	.04
Meiser and Hewstone (2001)	Students	40	108	.02	.00	.02	.04
Simons et al. (2002), Experiment 1	Older adults and patients	22	240	.03	.01	.00	.01
Simons et al. (2002), Experiment 2	Older adults and patients	16	120	.05	.00	.00	.00
Simons et al. (2002), Experiment 3	Patients	5	120	.03	.01	.00	.05
Schütz and Bröder (2011), Experiment 1	Students	50	150	.01	.00	.01	.01
Schütz and Bröder (2011), Experiment 2	Students	48	150	.00	.00	.00	.01
Schütz and Bröder (2011), Experiment 3	Students	50	90	.00	.00	.00	.03
Schütz and Bröder (2011), Experiment 4	Students	48	90	.01	.00	.00	.02
Schütz and Bröder (2011), Experiment 5	Students	50	90	.00	.00	.00	.01

2HTSM submodel 5d				Mean absolute deviation				
Data set	Population	*N*	*K*	*b*	*D*	*g*	d_1	d_2
Bell et al. (2015), Experiment 3	Students	100	80	.01	.01	.03	.09	.10
Bell et al. (2015), Experiment 4	Students	135	80	.03	.01	.02	.07	.08
Dodson and Shimamura (2000), Experiment 1	Students	45	90	.04	.01	.03	.08	.06
Dodson and Shimamura (2000), Experiment 3	Students	36	120	.03	.01	.03	.08	.07
Dodson and Shimamura (2000), Experiment 4	Students	75	120	.05	.01	.02	.06	.05
Klauer and Meiser (2000), Experiment 3	Students	40	98	.02	.01	.03	.08	.11
Klauer and Meiser (2000), Experiment 4	Students	120	108	.02	.01	.01	.05	.09
Küppers and Bayen (2014), Experiment 1	Students	48	96	.02	.00	.03	.08	.08
Mieth et al. (2016a), Experiment 1	Students	112	80	.02	.01	.02	.07	.08
Mieth et al. (2016a), Experiment 2	Students	96	80	.02	.01	.01	.08	.08
Mieth et al. (2016a), Experiment 3	Students	101	80	.02	.01	.03	.07	.09
Süssenbach et al. (2016)	Students	104	80	.02	.01	.04	.04	.08

2HTSM submodel 6e				Mean absolute deviation					
Data set	Population	*N*	*K*	*b*	D_1	D_2	*g*	d_1	d_2
Bell et al. (2015) Experiment 1	Students	138	80	.04	.01	.01	.03	.08	.08
Bell et al. (2015) Experiment 2	Students	114	80	.02	.01	.02	.03	.09	.09
Kroneisen and Bell (2018)	Students	40	160	.01	.01	.00	.01	.11	.10
Mieth et al. (2016b)	Students	216	80	.03	.01	.01	.04	.10	.08

Note. See Table 2. *b* = old/new guessing; *D*, D_1, D_2 = item memory; *g* = source guessing; *d*, d_1, d_2 = source memory.

nor number of participants, nor mean number of observations per participant appear to show systematic effects on deviations.

Divergence Across Method Pairs

Next, we zoomed in on the divergence separately for each method pair. Figure 4 provides an overview of all pairwise deviations. If all data points were exactly on the main diagonal (for which $y = x$), this pattern of results would indicate perfect agreement between two methods. As can be seen, no two methods agreed perfectly. However, most of the data points are distributed along the main diagonal. This is a reassuring result, as it indicates that methods generally produce similar estimates.

Figure 4 provides a further visualization of the predicted effect of SAI on divergence. Estimates for the pair-clustering model (for which SAI does not hold) are shown as blue triangles, whereas estimates for all other models are shown as grey circles. Figure 4 also provides two different RMSEs for each method pair: one based solely on estimates from models for which SAI holds (top left corner); the other based on all estimates (lower right corner). By comparing the two RMSEs in each panel we gain further support for the prediction that, across almost all method pairs, models for which SAI holds show smaller divergences than models for which it does not hold. In addition, the most egregious and systematic divergences—observed mainly for the cases in which the CP-MLE method produced estimates of ≈1 while the other method produced estimates between 0 and 1—also come exclusively from the pair-

Table 4
Overview of Free Recall Data Sets for the Pair-Clustering (PC) Model

PC model				Mean absolute deviation		
Data set	Population	*N*	*K*	*c*	*r*	*u*
Bröder et al. (2008)	Patients, older adults, and other	68	120	.04	.09	.06
Francis et al. (2018), English	Students	126	40	.07	.18	.03
Francis et al. (2018), English dominant	Students	62	40	.06	.15	.03
Francis et al. (2018), Spanish dominant	Students	63	40	.07	.17	.03
Golz and Erdfelder (2004), afternoon	Patients	36	60	.08	.14	.04
Golz and Erdfelder (2004), forenoon	Patients	35	60	.08	.16	.04
Matzke et al. (2015)	Students	65	60	.07	.18	.04
Riefer et al. (2002), alcoholics	Patients	42	120	.06	.12	.03
Riefer et al. (2002), schizophrenics	Patients	54	121	.05	.09	.02

Note. See Table 2. c = probability that an item pair is clustered and stored in memory; r = probability that an item pair is retrieved from memory, given that it was clustered; u = probability that a member of an item pair is stored and retrieved from memory, given that the item pair was not stored as a cluster.

Table 5
Overview of Prospective Memory Data Sets for the Prospective Memory (PM) Model

PM model				Mean absolute deviation			
Data set	Population	*N*	*K*	C_1	C_2	*M*	*P*
Arnold, Bayen, and Böhm (2015)	Students	125	327	.02	.01	.02	.02
Schnitzspahn et al. (2012)	Older adults and students	86	306	.01	.01	.05	.04
Horn et al. (2011), Experiment 2A	Students	27	128	.01	.01	.03	.05
Horn et al. (2011), Experiment 2B	Students	29	128	.01	.01	.04	.05
Pavawalla et al. (2012)	Patients and other	34	62	.01	.02	.06	.03
Rummel et al. (2011)	Students	61	62	.03	.03	.03	.04
R. E. Smith and Hunt (2014)	Older adults and students	138	78	.02	.01	.07	.04
R. E. Smith et al. (2014), Experiment 1	Students	81	62	.20	.22	.24	.10
R. E. Smith and Bayen (2005), Experiment 1	Students	20	136	.01	.00	.01	.02
R. E. Smith and Bayen (2005), Experiment 2	Students	21	136	.00	.01	.02	.02
R. E. Smith et al. (2010)	Children and students	118	66	.02	.01	.07	.06
Arnold, Bayen, and Smith (2015)	Students	413	248	.00	.00	.06	.04
R. E. Smith et al. (2014), Experiment 2	Students	90	62	.01	.01	.04	.06
R. E. Smith et al. (2014), Experiment 3	Students	80	112	.02	.01	.08	.06

Note. See Table 2. C_1 = probability of detecting Type 1 ongoing-task items; C_2 = probability of detecting Type 2 ongoing-task items; M = retrospective component (target recognition); P = prospective component.

Table 6
Overview of Hindsight Bias Data Sets for Hindsight Bias (HB) Model

HB model				Mean absolute deviation			
Data set	Population	*N*	*K*	*b*	*c*	r_C	r_E
Bayen et al. (2006), Experiment 1	Older adults and students	52	54	.08	.03	.00	.00
Bayen et al. (2006), Experiment 2	Older adults and students	64	50	.07	.03	.01	.01
Bernstein et al. (2011)	Children, older adults and students	194	19	.17	.11	.01	.01
Coolin et al. (2015)	Older adults and students	124	50	.06	.01	.00	.01
Coolin et al. (2016)	Older adults	80	56	.09	.02	.01	.01
Erdfelder et al. (2007), Experiment 1	Students	20	54	.15	.16	.00	.00
Groß and Bayen (2015)	Older adults and students	128	96	.03	.02	.04	.03
Groß and Bayen (2017)	Students	142	60	.06	.01	.00	.00
Pohl et al. (2010)	Children and students	139	49	.15	.04	.00	.00

Note. See Table 2. b = probability of biased reconstruction; c = probability of adoption of the correct judgment; r_E = probability of recollection for experimental items; r_C = probability of recollection for control items.

Table 7
Overview of Recognition-Based Inference Data Sets for the r-Model

r-Model				Mean absolute deviation		
Data set	Population	*N*	*K*	*a*	*b*	*r*
Castela and Erdfelder (2017), Experiment 1, Domain 1	Students	73	190	.00	.00	.02
Castela and Erdfelder (2017), Experiment 1, Domain 2	Students	74	190	.00	.00	.02
Castela and Erdfelder (2017), Experiment 1, Domain 3	Students	72	190	.00	.01	.01
Castela and Erdfelder (2017), Experiment 2	Students	51	421	.00	.00	.01
Filevich et al. (2019), Domain 1	Other	99	70	.00	.02	.05
Filevich et al. (2019), Domain 2	Other	99	70	.00	.00	.08
Hilbig et al. (2015), Experiment 1	Students	44	190	.00	.01	.02
Hilbig et al. (2015), Experiment 2	Students	59	190	.00	.00	.02
Hilbig et al. (2015), Experiment 3	Students	95	84	.00	.00	.01
Hilbig and Richter (2011)	Students	28	190	.01	.01	.02
Hilbig, Erdfelder, and Pohl (2010), Experiment 6	Students	34	136	.00	.01	.03
Hilbig, Erdfelder, and Pohl (2010), Experiment 7b	Students	13	91	.01	.00	.01
Hilbig et al. (2011)	Students	63	182	.00	.01	.01
Hilbig et al. (2012), Experiment 1	Students	68	153	.00	.00	.01
Hilbig and Pohl (2009), Experiment 1	Students	24	190	.01	.01	.01
Hilbig and Pohl (2009), Experiment 2	Students	72	135	.00	.00	.03
Hilbig and Pohl (2009), Experiment 3	Students	62	91	.00	.00	.01
Hilbig and Pohl (2008), Experiment 5	Students	101	55	.00	.00	.02
Hilbig et al. (2009)	Students	78	91	.01	.01	.02
Hilbig, Scholl, and Pohl (2010), Experiment 1	Students	19	120	.00	.01	.01
Hilbig, Scholl, and Pohl (2010), Experiment 2	Students	36	120	.01	.01	.05
Horn et al. (2015)	Older adults and students	78	276	.00	.00	.03
Horn et al. (2016)	Children	115	153	.01	.01	.05
M&E, Experiment 1 (day group, Test 1)	Students	33	300	.00	.01	.01
M&E, Experiment 1 (day group, Test 2)	Students	33	300	.00	.00	.01
M&E, Experiment 1 (week group, Test 1)	Students	31	300	.00	.01	.01
M&E, Experiment 1 (week group, Test 2)	Students	31	300	.00	.01	.01
M&E, Experiment 2 (day group, Test 1)	Students	41	300	.00	.00	.01
M&E, Experiment 2 (day group, Test 2)	Students	41	300	.00	.00	.02
M&E, Experiment 2 (week group, Test 1)	Students	42	300	.00	.00	.01
M&E, Experiment 2 (week group, Test 2)	Students	42	300	.00	.00	.02
M&E, Experiment 3 (different group, islands)	Students	64	300	.00	.00	.02
M&E, Experiment 3 (different group, musicians)	Students	64	300	.01	.00	.02
M&E, Experiment 3 (related group, celebrities)	Students	68	300	.00	.00	.02
M&E, Experiment 3 (related group, movies)	Students	68	300	.00	.00	.02
M&E, Experiment 4 (names)	Students	87	300	.00	.00	.02
M&E, Experiment 4 (pictures)	Students	87	300	.00	.00	.02
Michalkiewicz et al. (2018)	Students	92	300	.00	.00	.04
Pohl (2017), Experiment 2	Students	191	66	.00	.01	.03
Pohl et al. (2017), Experiment 1	Students	118	300	.00	.00	.01
Pohl et al. (2017), Experiment 2	Students	30	300	.00	.00	.03
Pohl et al. (2013)	Students	60	105	.00	.01	.02

Note. See Table 2. M&E = Michalkiewicz and Erdfelder (2016); *a* = recognition validity; *b* = knowledge validity; *r* = probability of relying on recognition.

clustering model. In sum, as predicted, the absence of SAI is associated with larger divergence.

As we are interested in divergences that we do not expect a priori based on statistical theory, we included in the analyses reported in Parts 2 and 3 only those models for which SAI holds.

Patterns of Divergence. Beyond the overall reassuring pattern of similar estimates, Figure 4 shows that the spread around the main diagonal is noticeable and differs across method pairs. The RMSEs in the upper left corner of each panel (i.e., excluding the pair-clustering model) range from .026 to .109. Within the partial-pooling (PP) methods, divergence is comparatively small (all RMSE $< .05$). The overall smallest RMSE of .026 is observed for PP-LT-C and PP-LT-NC, which are very similar to each other relative to other methods (they only differ in whether parameter correlations across participants are explicitly included in the model). Likewise, the two bootstrap-based no-pooling methods NP-PB and NP-NPB also show a comparatively small divergence, RMSE = .035, and are also very similar.

Interestingly, the partial-pooling methods without correlation parameters (PP-B, PP-LT-NC) produced generally slightly lower RMSEs with other methods than the latent-trait variant with correlation parameters (PP-LT-C); the partial-pooling method with correlation parameters (PP-LT-C) produced a slightly larger divergence with other methods. This pattern of results suggests two possibilities: Either parameters are truly correlated, and not including correlation parameters in the model produces biased estimates; or parameters are truly uncorrelated, and including correlation parameters in the model produces biased estimates. As shown in Appendix C, most data sets have nonzero correlations among model parameters, making the first possibility somewhat more

Table 8
Overview of Implicit Attitude Tasks Data Sets for the Quad and ReAL Model

Quad model				Mean absolute deviation				
Data set	Population	*N*	*T*	AC_1	AC_2	*D*	*G*	OB
Beer et al. (2008)	Students	16	381	.01	.02	.01	.01	.14
Calanchini et al. (2013)	Students	236	120	.02	.02	.02	.02	.10
Calanchini et al. (2014), Experiment 1a (Asian-White)	Students	168	144	.02	.02	.01	.01	.18
Calanchini et al. (2014), Experiment 1a (Black–White)	Students	168	144	.02	.01	.01	.02	.12
Calanchini et al. (2014), Experiment 1b (Black–White)	Students	49	120	.02	.01	.01	.03	.17
Calanchini et al. (2014), Experiment 1b (flower-insect)	Students	49	120	.01	.02	.02	.02	.06
Calanchini et al. (2014), Experiment 1c (flower-insect)	Students	56	120	.02	.02	.01	.02	.15
Calanchini et al. (2014), Experiment 1c (threat stereotype)	Students	56	144	.01	.02	.01	.02	.08
Calanchini et al. (2014), Experiment 2a (Black–White)	Other	200	120	.03	.03	.02	.02	.12
Calanchini et al. (2014), Experiment 2a (skin tone)	Other	200	120	.03	.02	.01	.02	.11
Calanchini et al. (2014), Experiment 2b (sexuality)	Other	200	120	.03	.02	.01	.02	.28
Calanchini et al. (2014), Experiment 2a (disability)	Other	200	120	.03	.03	.02	.02	.25
Calanchini et al. (2014), Experiment 2c (age)	Other	200	120	.02	.02	.01	.03	.04
Calanchini et al. (2014), Experiment 2c (gender stereotype)	Other	200	120	.02	.03	.01	.01	.11
Gonsalkorale et al. (2011)	Other	72	120	.02	.01	.01	.02	.07
Jin et al. (2016), Experiment 1	Students	117	120	.01	.01	.01	.02	.15
Jin et al. (2016), Experiment 2	Students	52	120	.01	.01	.01	.02	.08
Lueke and Gibson (2015; age)	Students	71	80	.02	.03	.01	.02	.18
Lueke and Gibson (2015; race)	Students	72	80	.02	.02	.01	.02	.10
Wrzus et al. (2017; happy–unhappy)	Students	563	160	.03	.02	.02	.03	.18
Wrzus et al. (2017; number–letter)	Students	564	160	.01	.01	.01	.02	.06

ReAL model				Mean absolute deviation						
Data set	Population	*N*	*T*	A_1	A_2	L_1	L_2	L_3	L_4	Re
Koranyi and Meissner (2015)	Students	154	320	.01	.01	.01	.02	.01	.01	.03
Meissner and Rothermund (2013), Experiment 1–3	Students	160	320	.01	.01	.01	.01	.01	.01	.05
Meissner and Rothermund (2013), Experiment 4	Students	40	320	.01	.00	.02	.01	.01	.01	.05
Meissner and Rothermund (2013), Experiment 5	Students	40	320	.01	.01	.01	.01	.01	.02	.02
Meissner and Rothermund (2013), Experiment 6	Students	77	320	.01	.01	.01	.01	.01	.01	.03
Meissner and Rothermund (2013), Experiment 7	Students	85	320	.01	.01	.01	.01	.01	.01	.03
Meissner and Rothermund (2015), Experiment 1	Students	80	320	.01	.01	.01	.01	.01	.01	.04
Meissner and Rothermund (2015), Experiment 2	Students	77	384	.01	.01	.01	.01	.01	.01	.02
Van Dessel et al. (2017)	Students	40	189	.10	.09	.01	.02	.01	.00	.31

Note. See Table 2. For the Quad model: AC_1 = activated association between one target group and negative evaluations; AC_2 = activated association between the other target group and positive evaluations; *D* = detection of correct responses; *G* = guessing; OB = overcoming biased associations. For the ReAL model: A_1, A_2 = evaluative associations of the target categories; L_1, L_2, L_3, L_4 = label-based identification of the correct response for the target categories (L_1, L_2) and the attribute categories (L_3, L_4); Re = recoding.

likely. Another source of substantive divergences are differences in distributional assumptions—normal on probit scale versus beta distribution—as shown by the RMSE of .036 between PP-LT-NC and PP-B.

The most noticeable divergence with other methods occurred for NP-Bayes, both within the no-pooling methods and across all methods (RMSEs between .09 and .11). Additionally, NP-Bayes was the only method where the pattern of divergences systematically deviated from the main diagonal. Specifically, when NP-Bayes is shown on the *x*-axis in Figure 4 (sixth column), we see an s-shaped relationship with all other methods such that the NP-Bayes estimate is less extreme (i.e., pushed away from the boundary toward .50). We attribute this consistent pattern to the increased influence of the prior for this method. Specifically, for NP-Bayes, we employed independent flat priors on the parameter range from 0 to 1 for all parameters at the participant level; this assumption magnified the influence of the prior distribution on the posterior estimate at the group level. This pattern of results contrasts with the other Bayesian methods, where the prior is only applied once. NP-Bayes can thus be considered generally unsuitable for parameter estimation at the group level, and we will not consider it any further.

One further pattern that emerged was that the partial-pooling methods show, on average, the smallest divergence across different pooling levels. In addition to the small divergence within all partial-pooling methods, the partial-pooling methods also showed small divergence with complete-pooling MLE—the most traditional method—with RMSEs < .05. Likewise, the partial-pooling methods showed comparatively small divergence with NP-MLE and NP-NPB, with RMSEs between .054 and .065. Given that partial-pooling is a hybrid between complete-pooling and no-pooling, such a result is perhaps unsurprising and, at the same time, reassuring.

To investigate whether the value of the parameter estimates can be used for predicting divergence, Figure 4 shows a generalized linear additive model (GAM, in red; Baayen et al., 2017) predicting one estimate from the value of the other estimate. A GAM is a flexible regression model that can account for a nonlinear relationship using

Table 9
Overview of Process-Dissociation Data Sets for the Process-Dissociation (PD) Model

PD model				Mean absolute deviation	
Data set	Population	*N*	*K*	*a*	*r*
Bodner et al. (2000), Experiment 2	Students	30	80	.01	.02
Bodner et al. (2000), Experiment 4	Students	28	80	.01	.02
Caldwell and Masson (2001), Experiment 1	Older adults and students	60	48	.01	.01
Caldwell and Masson (2001), Experiment 2	Students	28	144	.01	.00
Rouder et al. (2008), Experiment 1	Students	66	96	.01	.06
Rouder et al. (2008), Experiment 2	Students	44	32	.02	.03
Stahl et al. (2015), Experiment 1	Students	181	211	.00	.02

Extended PD model				Mean absolute deviation		
Data set	Population	*N*	*K*	*a*	r_E	r_I
Klauer et al. (2015), Experiment 1	Students	40	287	.05	.05	.14
Klauer et al. (2015), Experiment 3	Students	40	335	.07	.07	.13
Klauer et al. (2015), Experiment 5	Students	162	694	.03	.11	.11

Note. See Table 2. *a* = automatic processes; *r* = controlled processes; r_E = controlled processes under exclusion conditions; r_I = controlled processes under inclusion conditions.

a thin-plate regression spline. Overall, we did not find any effects other than the predicted relationship along the main diagonal, with two exceptions already discussed. First, for NP-Bayes there was an s-shaped relationship (column 6). Second, estimates from the pair-clustering model (for which SAI does not hold) again produce large divergences. This can, for example, be seen in the first column as a dip at the right boundary driven by the blue pair-clustering estimates.

Cases With Large Divergence. Across all pairs (excluding the pairs involving NP-Bayes, but including the pair-clustering model), 19.6% of all absolute deviations were larger than .05, 9.3% were larger than .10, and 1.8% were larger than .25.[14]

We next considered the *worst cases*—the single largest absolute deviation for each of the 28 methods pairs. The largest worst case appears between CP-MLE and NP-PB with an absolute deviation of .97. The smallest worst case appears between PP-LT-C and PP-LT-NC with an absolute deviation of 0.26. In line with the results reported above, the pair-clustering model (with singletons excluded from analyses) features prominently among the worst cases (for 11 method pairs), but several worst cases involve models for which SAI holds, for example, the prospective memory model (seven method pairs), the ReAL model (three method pairs) and the Quad model (three method pairs). Furthermore, worst cases from the pair-clustering model are not necessarily larger than worst cases from other models. For example, both the largest and smallest worst case comes from the pair-clustering model and only two of the nine worst cases (all larger than .8) are from the pair-clustering model (the other seven are from the prospective memory model).

To summarize, despite these individual cases with large divergences, the set of all pairwise comparisons shown in Figure 4 is overall reassuring. We found large agreement between methods, and there was no clear or consistent bias (with the exception of NP-Bayes). Furthermore, the partial-pooling methods showed comparatively large agreement with all other methods. However, for each pair, we also saw at least a few data points for which the divergence is considerable. Even for the two methods with the strongest agreement, PP-LT-C and PP-LT-NC, we found divergences larger than .20 (even after excluding the pair-clustering model). In addition, even though certain models and data sets were more likely to show large divergence, nonnegligible divergence appeared for all models. In the next step, we will test the prediction of perfect agreement derived above.

Part 2: Testing the Prediction of Perfect Agreement

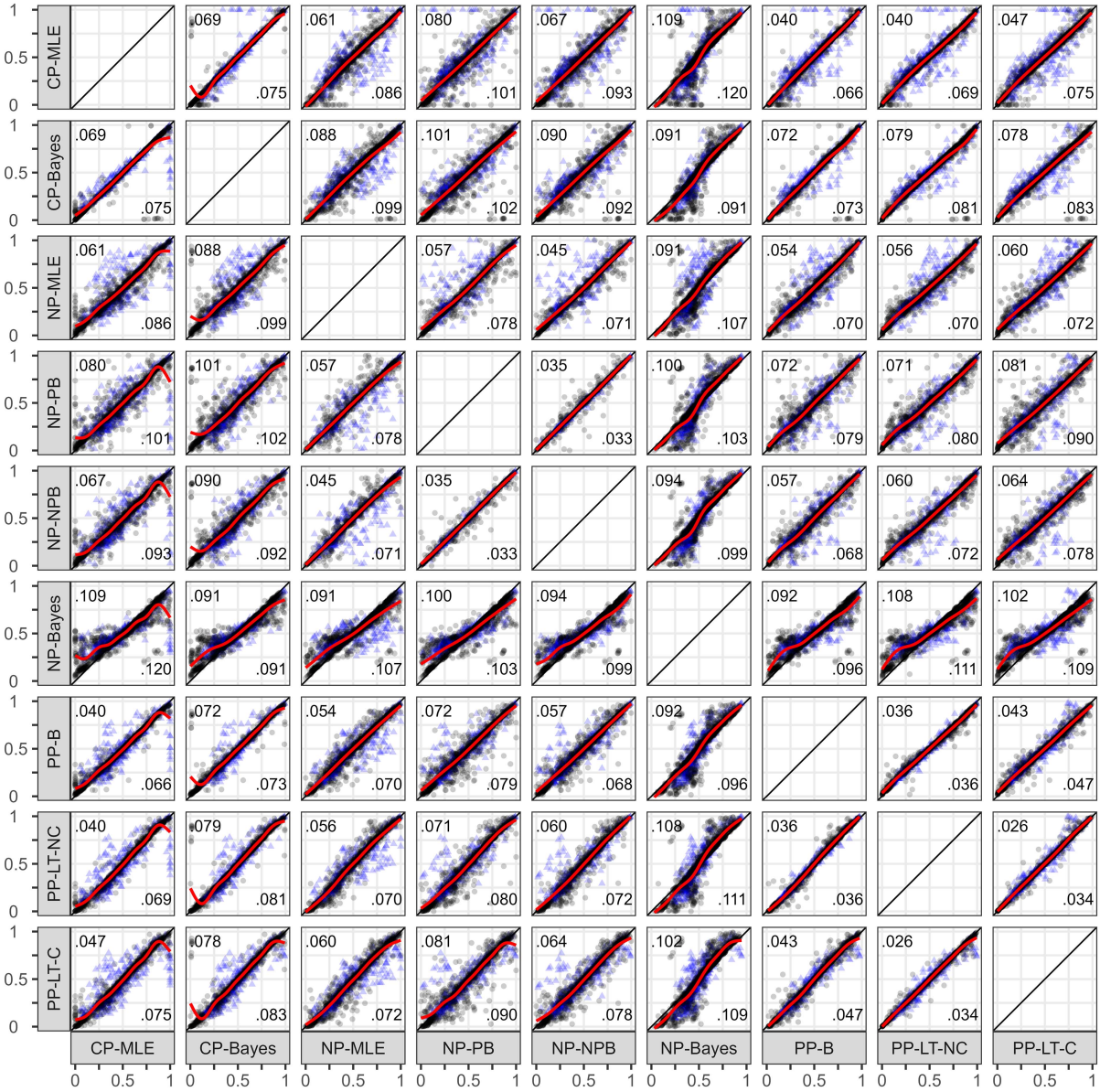
We predict perfect agreement between two estimation method pairs for group-level parameters—that is, a divergence of zero—if (a) a model holds for each participant (possibly with different parameters), (b) the correlations between parameters on the same branch approach zero, and (c) the *SE* approaches zero (details shown in Appendix B). Only with our large data set, we can identify empirical conditions where (b) and (c) hold. If there is perfect agreement between methods in these cases, then this is in line with the assumption that (a) holds, too. By contrast, if the prediction of perfect agreement would fail in these cases, then this would be in conflict with assumption (a). Our multiverse meta-analysis thus offers the opportunity to provide an empirical test of a central assumption of the MPT models class. To provide a compact test of the prediction, we focus on the five selected pairs of reference and comparison methods (the results are qualitatively the same with all comparison methods and can be found in the additional online material available at OSF: https://osf.io/waen6/).

First, we tested the prediction by looking at the absolute deviations in a descriptive manner by binning the data. The results are shown in the top row of Figure 5. In the lower left corner of each panel, the *SE* and correlations both approach zero. We predict that for these cases, the mean absolute deviation should also be close to zero. As can be seen, this held for most of the method pairs, as indicated by a green- or yellow-colored square in the leftmost corner. However, most empirical data sets had *SE*s and correlations notably larger than zero; and the descriptive analysis alone does not seem to permit a final judgment for our prediction.

[14] When excluding the pairs involving NP-Bayes and excluding the pair-clustering model, 15.2% of all absolute deviations were larger than .05, 6.7% were larger than .10, and 1.0% were larger than .25.

Figure 4

Parameter Divergence Across All Methods

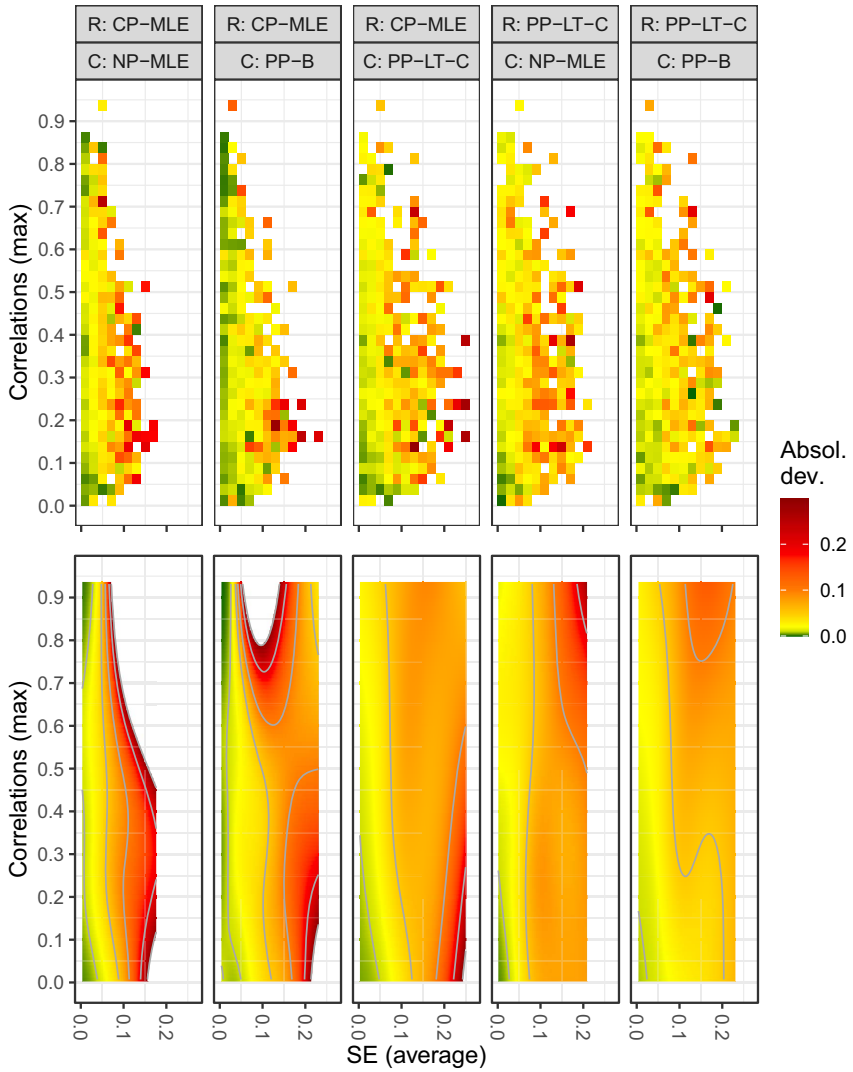


Note. Each data point shows the two estimates for one parameter across two methods. Grey circles show estimates from models which are structural-aggregation invariant, whereas blue triangles show estimates from the pair-clustering (PC) model for which structural-aggregation invariance does not hold. Data points are plotted semitransparently so that areas with more points appear darker. The value in the lower right corner of each panel gives the RMSE of the estimates for the two methods across all pairs, whereas the value in the upper left corner gives the RMSE after excluding the PC model. The red line shows the predicted relationship between both estimates from a generalized additive model using a thin-plate regression spline (using all data points shown). RMSE = root-mean-squared error; CP-MLE = complete-pooling, maximum likelihood estimation; CP-Bayes = complete-pooling, Bayesian parameter estimation; NP-MLE = no-pooling, maximum likelihood estimation; NP-PB = no-pooling, parametric bootstrap; NP-NPB = no-pooling, nonparametric bootstrap; NP-Bayes = no-pooling, Bayesian parameter estimation; PP-B = beta-MPT approach; PP-LT-NC = partial-pooling, latent-trait without correlation parameters; PP-LT-C = partial-pooling, latent-trait with correlation parameters; MPT = multinomial processing tree. See the online article for the color version of this figure.

We therefore also performed a second, model-based analysis to test the prediction. Specifically, we used a GAM to predict the absolute deviation for each method pair using a bivariate tensor smooth over both the correlation and the *SE* (Baayen et al., 2017). To constrain the model prediction to the positive range, we used a gamma distribution with a log link as the conditional distribution of the absolute deviation. Results are shown in the bottom row of Figure 5. In the area of interest where both the *SE* and the

Figure 5

Test of Prediction of Perfect Agreement



Note. For each combination of reference ("R:") and comparison method ("C:"), each panel shows the mean absolute deviation (Absol. dev.; top row) and the predicted mean absolute deviation from a generalized additive model using a bivariate tensor smooth (bottom row) as a function of the average *SE* on the *x*-axis and the maximum parameter correlations on the *y*-axis. The absolute deviation is indicated with a color scale such that green indicates a mean absolute deviation between 0 and .01. For the top panel, data are binned in bins of size .02 (*x*-axis) and .025 (*y*-axis), reflecting the different observed variable ranges. CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; *SE* = standard error; MPT = multinomial processing tree. See the online article for the color version of this figure.

correlations are zero, the GAM predictions and associated 95% CIs are very small, with the upper bounds of the CIs < .042. Note that the GAM predictions are solely data-informed (i.e., *semiparametric*); the model does not "know" that we are interested in this specific parameter region. These results therefore further support the prediction of perfect agreement.

Taken together, the prediction of perfect agreement appeared to hold. Across all considered method pairs, the pattern of observed absolute deviations supports—across methods, models, and data sets—the assumption that the considered MPT models hold approximately for each participant they were applied to. This result provides an important empirical validation of the assumptions underlying MPT model applications in psychology.

In the next analysis part, we will aim to identify moderators that might explain the degree of absolute divergence in cases of parameter correlations and *SE*s larger than zero.

Part 3: Sources of Divergence

In Part 1 (see Figure 4), we used the RMSE as our measure of divergence between two methods. In Part 3, we attempt to reduce the RMSE by including further information. To do so, we predict the absolute deviation between the estimates of two methods using linear regression. Note that the estimate of the residual standard deviation of such a regression model, which is a measure of model misfit, is equal to the RMSE. In other words, in this part, we test whether adding additional predictors to a regression model predicting the absolute deviation reduces the RMSE.

We started with a baseline model including only the absolute deviation between two methods (i.e., a regression model with only an intercept). The mean absolute deviations for the selected method pairs are shown in Figure 6. For example, for the pair of CP-MLE and NP-MLE, the observed mean absolute deviation is .032. Based on the value from a CP-MLE estimate, we predict that the NP-MLE estimate will be the same value ±.032.[15] Table 10 shows the RMSEs for these baseline models (in the row labeled "$\bar{x}$ [baseline] (1)"), which correspond to the mean absolute deviations given in Figure 6.[16] Our next question was whether we can reduce the RMSEs of the baseline models by adding the moderators described above.[17]

Table 10
RMSE for the Baseline Model and the Models With Categorical Moderators for Selected Method Pairs

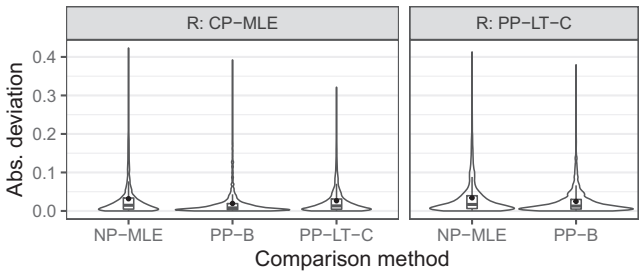
Predictor in the regression model	CP-MLE			PP-LT-C	
	NP-MLE	PP-B	PP-LT-C	NP-MLE	PP-B
$\bar{x}$ [baseline] (1)	.050	.035	.039	.049	.035
Model (8)	.049	.034	.038	.047	.034
Submodel (12)	.047	.033	.037	.046	.034
Parameter (49)	.037	.027	.032	.036	.030
Data set (145–147)	.047	.030	.037	.046	.031
Population (5)	.050	.035	.038	.048	.035
Goal (2)	.050	.035	.039	.049	.035

Note. The two header rows indicate the combination of reference method and comparison method. Values in parentheses in the first column are the number of levels for the corresponding categorical moderator (= number of coefficients of the linear model predicting the absolute deviation including the intercept). For data set, the number of levels differs somewhat across method pair as in some cases a method failed for a specific data set, in which case no estimates for this method were available for this data set. RMSE = root-mean-squared error; CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; MPT = multinomial processing tree.

Univariate Analyses: Model-Dependent Moderators

We first considered the model-dependent moderators. These represent idiosyncratic aspects of the set of models considered here and therefore do not allow generalization to other MPT models. In Table 10, rows two ("Model (8)") to row five ("Data set (145–147)") show the RMSEs after adding the corresponding predictors to the baseline model. In line with our earlier observation (see Part 1), we found the most substantial reduction in RMSE for the MPT model parameter (row five). That is, the MPT model parameter explains, to a substantial degree, the magnitude of divergence between methods. The reduction in RMSE for the two related moderators, model and submodel (rows three and four), was considerably smaller. Likewise, the reduction in RMSE for data set (row six) was also noticeably smaller. Taken together, these results suggest that the MPT model parameter had the largest influence on predicting the absolute deviation between two methods, compared to the other model-dependent moderators. To foreshadow the following results, none of the model-independent moderators alone could reduce the RMSE to a similar degree. Thus, we consider the reduction in RMSE provided by model parameter the benchmark for the remaining moderators. Note, however, that with values between .027 and .037 the RMSEs are nonnegligible even when accounting for the MPT model parameter.

Figure 6
Distribution of Absolute Deviation Across Selected Method Pairs



Note. For each combination of reference method ("R:") and comparison method (on *x*-axis), the plot shows the distribution of individual absolute deviation (Abs. deviation) values as violin plots. The boxplot in the background shows the 25% quantile, the 50% quantile (i.e., the median), and the 75% quantile. The solid point shows the mean which are (from left to right) .032, .019, .027, .032, and .025. Given the highly asymmetric distribution, the mean is similar to the 75% quantile in most cases. CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; MPT = multinomial processing tree.

[15] As we predict the absolute deviation, this prediction generally holds for both directions (i.e., does not depend on which of the two methods is designated as reference or comparison method), unless the independent variables are specific to one of the methods (e.g., value of parameter estimate for reference method in Figure 7). This predicted absolute deviation can be understood as creating a symmetric uncertainty band around the estimates of one method in which the estimates of the other method is expected to be. The smaller the predicted absolute deviation between two methods, the better the prediction. Thus, the uncertainty band would also be ±.032 if we were to predict the CP-MLE estimates from the NP-MLE.

[16] Because we do not make point predictions but consider the uncertainty of the predictions, the baseline RMSEs in Table 10 are lower than the ones in Figure 4 (upper left corner). For example, for the pair of CP-MLE and NP-MLE, the RMSE is reduced to .050 (compared to .061 in Figure 4).

[17] To simplify comparison with the multivariate results presented later, all results presented in this section are based on only those observations for which we have values for all potential moderators (e.g., excluding all observations from data sets for which the latent trait partial-pooling method failed as this method provided the values of the parameter trade-offs moderator). This approach reduces the total number of observations across considered pairs by 5.8% compared to the number of data points in Figure 4. Note that all results presented in this section also exclude the observations from the pair-clustering model for which structural aggregation invariance is violated.

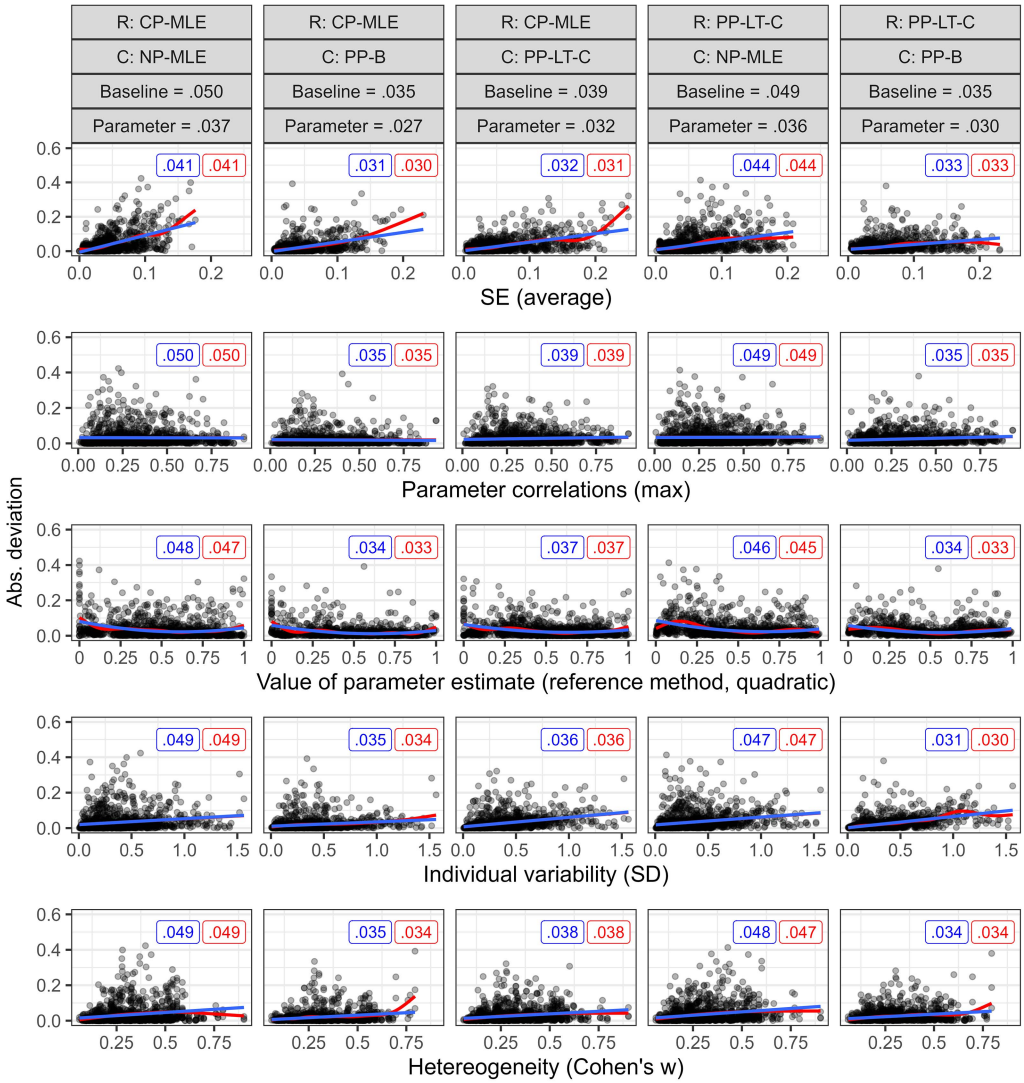
Univariate Analyses: Model-Independent Moderators

Next, we examined the model-independent moderators that, in principle, allow generalization beyond the set of MPT models considered here. The two model-independent categorical moderators—population of participants and scientific goal—showed no reduction in RMSE compared to the baseline model (Table 10).

We then considered the model-independent continuous moderators; results are shown in Figures 7 and 8. Each panel shows the relationship between the absolute deviation for one method pair and one moderator. The blue line and number show the linear relationship between the moderator and the absolute deviation (in some cases after a nonlinear transformation of the moderator described in the *x*-axis label) along with the corresponding RMSE;

Figure 7

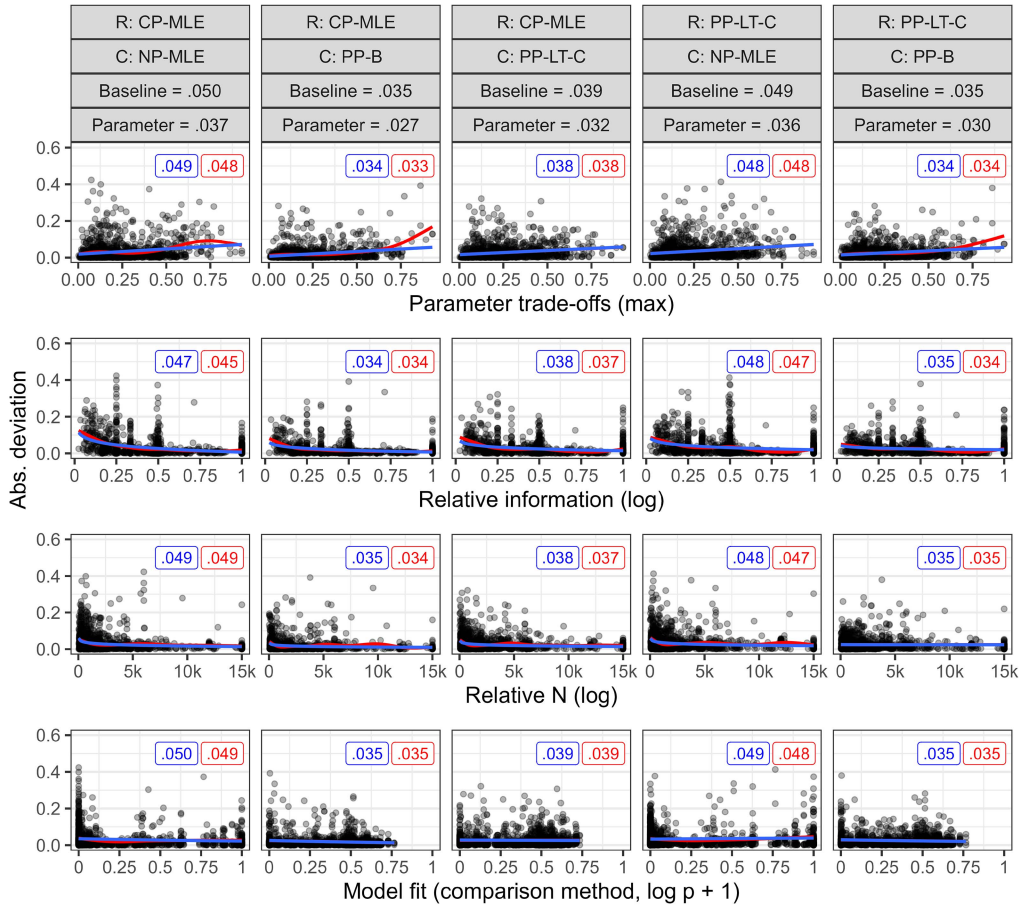
Univariate Relationship of Continuous Model-Independent Moderators With Absolute Deviation (Part 1)



Note. Each column shows one pair of reference method ("R:") and comparison method ("C:"). The *y*-axis depicts the absolute deviation (Abs. deviation) between the two estimation methods indicated in the header. The *x*-axis shows the value on one of the considered moderators. Data points are plotted semitransparently so that areas with more points appear darker. The blue line shows the linear (in the case of the value of the parameter estimates, the quadratic) relationship between the moderator and the absolute deviation and the blue number shows the corresponding RMSE. The red line shows the nonlinear relationship between based on a generalized additive model using a thin-plate regression spline and the red number the corresponding RMSE. To simplify comparisons, the RMSE of the baseline model and the model including the MPT parameter are given in the figure header. RMSE = root-mean-squared error; CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; *SE* = standard error; MPT = multinomial processing tree. See the online article for the color version of this figure.

Figure 8

Univariate Relationship of Continuous Model-Independent Moderators With Absolute Deviation (Abs. Deviation; Part 2)



Note. For the first row, the blue line shows the linear relationship, for rows two to four the blue line shows the linear relationship after transformation of the variable (log-transformation or log(p + 1)-transformation). For the third row, values are winsorized at 15,000 (15k). See Figure 7 for more details. CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; MPT = multinomial processing tree. See the online article for the color version of this figure.

the red line and number show a flexible nonlinear relationship based on a GAM along with the corresponding RMSE. For comparison, we show the RMSE for model parameter and baseline in the figure header. As mentioned above, no single model-independent continuous moderator achieved the same reduction in RMSE as MPT model parameter, neither with a linear nor a nonlinear relationship.

The first two rows of Figure 7 show the two potential moderators already considered in Part 2: the *SE*, and the across-participant correlations of the parameter with other parameters on the same branch. In line with the results described earlier, we found that including *SE* substantially reduces RMSE compared to the baseline model (first row in Figure 7). This is true for all but the PP-LT-C and PP-B method pair, for which the RMSE reduction is comparatively small. The relationship is such that within each panel, with the *SE* at the minimum, the predicted absolute deviation is around 0, and with the *SE* at the maximum, predicted deviation is between .08 and .18. In contrast to the *SE*, the parameter correlations (second row) showed essentially no univariate relationships with absolute deviation; the RMSEs were similar to those of the baseline model.

The moderators shown in row three and four of Figure 7 provided some reduction in RMSE compared to the baseline model. The value of the parameter estimate (using the value of the reference method) showed a slight u-shaped relationship, with larger predicted absolute deviations for small and large values (row three). The individual variability (*SD*), one of two measures of individual differences, showed a positive relationship with absolute deviation (row four). For this moderator, we observed an interesting pattern with respect to reduction in RMSE: For all pairs but PP-LT-C and PP-B, the reduction is only small; by contrast, for the PP-LT-C and PP-B pair, the reduction is rather substantial. In fact, for this method pair, *SD* provides an even larger reduction than *SE*. Finally, row five shows the relationship of absolute deviations with the second measure of individual differences, the heterogeneity of responses. Here, we

only found a minor reduction in RMSE for all method pairs. For none of the moderators shown in Figure 7 did the nonlinear relationships (in red) provide a substantial additional improvement in RMSE.

The results in Figure 8 show the four remaining moderators for which the reduction in RMSE compared to the baseline model is minor at best. Row one shows the parameter trade-offs for which we see essentially no reduction in RMSE compared to the baseline model. Rows two and three show the relative parameter information and the relative *N*, respectively, which both show a small negative relationship (on the log scale) with absolute deviation. Row four shows the model fit which also appears to provide no information that would allow predicting the absolute deviation between two methods.

Multivariate Analyses: Model-Independent Moderators

It is possible that a combination of moderators better predicts the absolute deviation between two methods than any single moderator alone. Therefore, we next considered different multiple regression models in which we entered the moderators as additive main effects. The RMSE results are shown in Table 11 (section "without interactions"). For comparison, the "parameter" model (with the largest overall reduction in RMSE so far), and the model with only *SE* (the model-independent moderator with the largest reduction in RMSE so far), are also shown.

The "all" model includes additive main effects of all model-independent moderators considered. As can be seen, this model performed similar to the "parameter" model and did not lead to a much larger reduction for most method pairs.

The next question was whether we need all the continuous model-independent moderators to achieve an RMSE similar to the "parameter" model, or whether a subset of relevant moderators suffices. To investigate this question, we estimated all possible multiple regressions containing either only two or only three moderators. Table 11 shows, in rows "best 2" and "best 3," respectively, the RMSE for the model with the lowest (i.e., best) RMSE among all these models for each method pair. Results show that a model with only the two or three best moderators performed similar to the model with all moderators. Importantly, we found that some moderators consistently appeared in the set of the two best or three best moderators across method pairs. Specifically, a parsimonious model with an RMSE that is comparable to the "all" model is one that includes *SE*, the value of the estimate (linear and quadratic), and *SD*.

Figure 9 shows the unstandardized multiple regression coefficients for the model with *SE*, parameter value, and *SD* across the method pairs considered here. The estimates of the regression coefficients were generally in agreement, with the exception of *SE*, which showed noticeable variability across method pairs. Larger *SE*s were associated with larger absolute deviation, with a mean estimate of around .50. This means that an *SE* that is .10 larger is associated with an absolute deviation that is on average approximately .05 larger, if all other predictors remain constant. To simplify interpretation of the coefficients relating to the parameter value, we subtracted .50 from the value before including them in the model. We found a slightly negative linear and clearly positive quadratic relationship indicating that the model predicts larger absolute deviations at the boundary of the parameter space (i.e., near 0 and, to a smaller degree, near 1). For example, if we use the mean estimates of the coefficients, compared to a parameter value of .50, a parameter value of .10 (i.e., a parameter that is .40 smaller) is associated with an absolute deviation that is on average $-0.40 \times -0.019 + (-0.40)^2 \times 0.16 = 0.03$ larger, if all other predictors remain constant. Finally, larger individual variability (*SD*) further increased the predicted absolute deviation with a regression coefficient of around .01–.02.

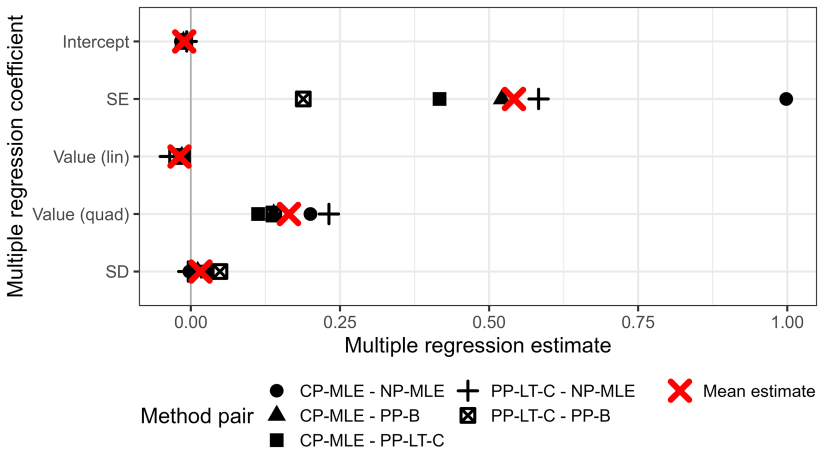
Table 11
RMSE for the Multiple Regression Models Predicting Absolute Deviation

	CP-MLE			PP-LT-C	
Predictors in the model	NP-MLE	PP-B	PP-LT-C	NP-MLE	PP-B
Parameter (53)	.037	.027	.032	.036	.030
SE	.041	.031	.032	.044	.033
Without interactions					
All (11)	.037	.027	.029	.039	.027
Best 2 (4)	.039	.029	.031	.041	.030
Best 3 (5)	.038	.028	.030	.040	.029
SE, value, *SD* (5)	.039	.029	.030	.041	.029
With two-way interactions					
All (55)	.032	.023	.026	.033	.023
RAMP (19–29)	.033	.023	.027	.034	.023

Note. The two header rows indicate the combination of reference method and comparison method. Values in parentheses after the model name in the first column are the number of regression coefficients (including the intercept). To simplify comparison, the parameter column is the same as in Table 10 and the *SE* column the same as in Figure 7. Best 2 and best 3 show the lowest RMSE among all possible models with only two or three moderators, respectively. The RAMP models are based on the method of Hao et al. (2018) and differ in their numbers of parameters. For models involving the value of the estimate of the reference method (e.g., all main), we included both a linear and quadratic coefficient. RMSE = root-mean-squared error; CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; RAMP = regularization algorithm under marginality principle; MPT = multinomial processing tree.

Figure 9

Multiple Regression Coefficients Predicting Absolute Deviation From Three Best Moderators



Note. For ease of interpretation, we subtracted .50 from the parameter value before entering it into the regression model. The mean (median, not shown) estimates are (from top to bottom): −0.011 (−0.010), 0.54 (0.52), −0.019 (−0.015), 0.16 (0.14), 0.016 (0.012). CP-MLE = complete-pooling, maximum likelihood estimation; PP-LT-C = partial-pooling, latent-trait with correlation parameters; NP-MLE = no-pooling, maximum likelihood estimation; PP-B = beta-MPT approach; *SE* = standard error; MPT = multinomial processing tree. See the online article for the color version of this figure.

In a final step, we considered whether adding two-way interactions among the model-independent moderators can further improve the predictive abilities of the multiple regression models (see Table 11, section "with two-way interactions"). First, we considered a model with all two-way interactions among the nine moderators considered here ("All"). As expected, compared to the "parameter" model, this model reduced the RMSE for most of the method pairs. Next, we asked: Is there a restricted subset of interactions that is mainly responsible for the improved predictive power? To do so we used a penalized model selection approach among all possible two-way interactions ("RAMP").[18] We found no consistent set of interactions across method pairs that could improve performance noticeably. Importantly, even with interactions included, a nonnegligible absolute deviation (ranging from .023 to .034 across pairs) remained unexplained.

Discussion

Applying a cognitive model requires the researcher to make a number of decisions. Two important decisions concern the statistical framework and level of pooling. Ideally, parameter estimates from cognitive models are robust across such modeling decisions. To examine how choice of estimation method affects modeling-based results, we conducted a multiverse meta-analysis in which we re-estimated empirical data from 164 published data sets that applied one of nine different cognitive MPT models.

Prior to exploring empirical divergence between results in these data sets, we identified conditions under which we would expect divergence to occur based on statistical theory. Our predictions for these conditions were confirmed. First, for the pair-clustering model for which one parameter occurs more than once in some of the branches of the model (i.e., for which structural-aggregation invariance does not hold), divergence was noticeably larger than for the other models included, for which structural-aggregation invariance holds. A researcher should check the structural-aggregation invariance property of their model in advance. If it does not hold, they should expect a larger amount of divergence and ideally explore the robustness of their results across the multiverse of estimation methods. Second, for models for which structural-aggregation invariance holds, the divergence between methods was near zero for cases in which all parameter correlations on the same branch were approximately zero and the standard error was approximately zero. This result is in line with the assumption that—across the multiverse of methods, models, and data sets—the considered MPT models hold for each participant they were applied to, an assumption common to all MPT applications. This result provides an important validation of MPT applications in psychological research.

These preconditions—zero correlations and zero standard error—occur only rarely for a specific empirical data set. We therefore asked, how large is the effect of parameter estimation method on parameter estimates in empirical data? Overall, our meta-analysis established a reassuringly high convergence across methods, with partial-pooling parameter estimates showing the most consistent

[18] We used the regularization algorithm under marginality principle (RAMP; Hao et al., 2018) approach which uses a LASSO penalty to select among all possible main effects and two way interactions while attempting to minimize the total number of nonzero coefficients. The RMSE of the resulting model ("RAMP") is very similar to the RMSE of the model with all interactions. However, the number of parameters and included interactions did differ quite dramatically across method pairs. Of the possible 45 interactions (RAMP also considers quadratic terms for each of the nine moderators), one appears in all five method pairs (quadratic effect of value), five appear in four pairs, seven appear in three pairs, 10 appear in two pairs, 10 appear in one pair only, and 12 interactions appear in none of the pairs.

results when compared with other methods. Nonetheless, there are single cases of high divergence even for such high-agreement method pairs. Although some model parameters are more likely to show high divergence (as we will discuss in more detail below), divergence can, in principle, occur for every MPT model. That is, despite our finding that agreement between methods is generally high, the possibility always remains that any given data set is an exceptional case that exhibit considerable divergence between estimation methods.

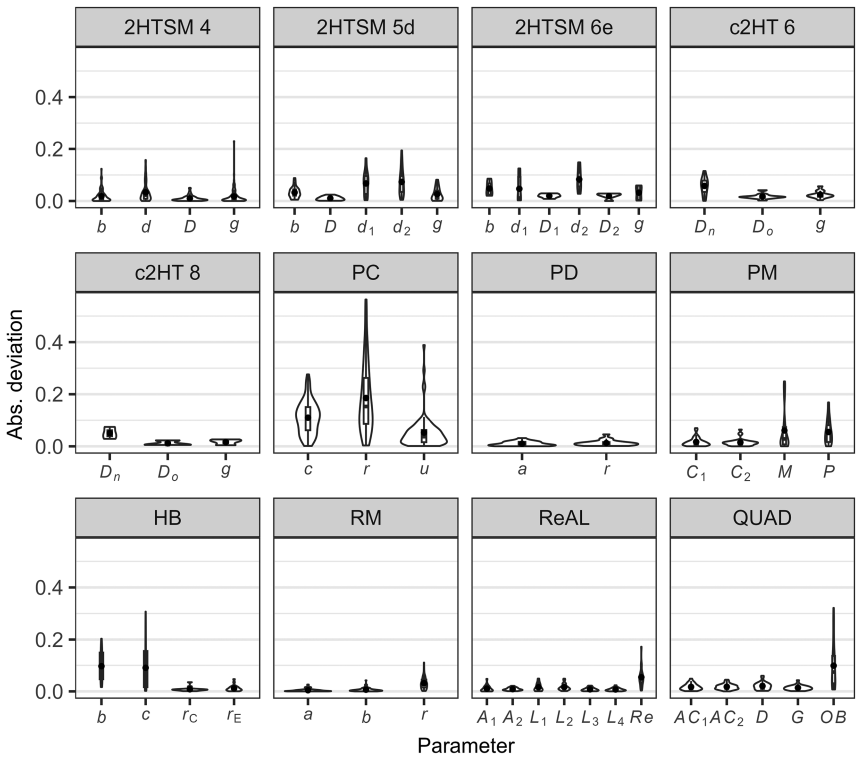
With regard to the estimation methods for MPT models we therefore propose two different recommendations; one pragmatic recommendation, and one for increased certainty. In the vast majority of cases—especially when effects of theoretical interest are not small—a pragmatic choice is to use only one method. Our results show that in this case, a partial-pooling method is the safest bet. Which partial-pooling method to use cannot be answered with our empirical approach, but we may nevertheless use our results as a starting point for some speculations. The partial-pooling methods without correlations (PP-B, PP-LT-NC) show smaller divergences with all other methods than the partial-pooling method with correlations; however, most data sets show nonzero correlations (Appendix C), which is also a plausible assumption about the underlying data-generating process. Consequently, using PP-LT-C seems like a sensible first choice (but see Groß & Pachur, 2020, and discussion below).

The downside of this pragmatic approach is that it does not allow assessing the uncertainty of one's results that would result from choosing a different method. If researchers want extra certainty in their results and wish to know how method-dependent their results are, a multiverse analysis using the `MPTmultiverse` package (Singmann et al., 2020) is a good choice. Occasions for requiring this extra certainty are, for example, when developing new models, when effects of theoretical interest are small, or when the parameter of interest is known to show large divergence (see Figure 10 discussed below). Such a multiverse analysis should be limited to the set of methods that are sensible for a given data set (and model); for instance, no-pooling methods should not be included if the number of observations per participant is low (no-pooling Bayes should generally be avoided).

In our analysis, we did not only document divergence, but also aimed at identifying the sources of divergence. To do so, we included several moderators that might be relevant for the choice of estimation method based on theoretical grounds and simulation studies. These moderators included properties that are specific to

Figure 10

Absolute Deviations Across MPT Submodels and Their Core Parameters



Note. Each *x*-axis tick shows the distribution of individual absolute deviation (Abs. deviation) scores for each core model parameter (see Table 1) for the complete-pooling MLE and partial-pooling latent-trait with correlations method pair. We omitted the second PD variant (PD_E), as it only contained data from three studies. The distribution is shown as violin plots. The boxplot in the background shows the 25% quantile, the 50% quantile (i.e., the median), and the 75% quantile. The solid point shows the mean. MPT = multinomial processing tree; 2HTSM = two-high-threshold model of source monitoring; c2HT = confidence-rating two-high-threshold; PC = pair-clustering; PD = process-dissociation; PM = prospective memory; HB = hindsight-bias; RM = r-model; ReAL = ReAL model; QUAD = Quad model.

MPT models and their tree structure (e.g., the MPT model, model parameter, or the relative information available to estimate a parameter), as well as aspects that are universal to parameter estimation of cognitive models (e.g., the magnitude of interindividual differences, or the standard error of the estimate). Somewhat surprisingly, divergence was not well explained by most of these moderators—neither in univariate nor multivariate regression models. For example, neither parameter correlations, nor parameter trade-offs, nor heterogeneity of the observed response frequencies—aspects that are differently captured by the different estimation methods—explained substantial divergence between the methods. The absence of effects cannot be explained by a restriction of range of these variables (see Figures 7 and 8). It appears that, contrary to earlier theoretical considerations (e.g., J. B. Smith & Batchelder, 2008), these variables are not major sources of divergence between parameter estimates for empirical data. The only way for these variables to have a substantial effect on divergence would be in terms of three-way interactions (e.g., heterogeneity might affect divergence only when both *SE* and correlations are high), interactions of even higher orders, or as effects that are specific to the MPT model parameter. An exploration of such higher order effects is beyond the scope of the present work.

Among the moderators that did explain divergence between estimation methods to a noticeable degree, the MPT model parameter was the best single predictor. This result implies that researchers should thoroughly familiarize themselves with the respective MPT model and check if the parameters of main interest show large divergence across estimation methods. In Figure 10, we provide an exemplary overview of the divergence between complete-pooling MLE (the most traditional method) and partial-pooling latent-trait with correlations (the most advanced method) for the models' core parameters. As can be seen, the pair-clustering (PC) model, which is not structurally aggregation invariant, appears to be an exceptional case; here, mean divergence for the *c* and *r* parameters is high, and divergence spans across a large range (up to .50). However, among other models for which structural-aggregation invariance holds, divergence for selected parameters can also be large (e.g., up to .32 for the OB parameter of the Quad model).

The MPT model parameter is the most idiosyncratic aspect of the model and carries no general information that is universal for all MPT models. We therefore aimed to find model-independent moderators that provide a similar reduction in RMSE; however, no single model-independent moderator predicted divergence to a similar degree. In combination, three model-independent moderators explained a similar amount of RMSE as the MPT model parameter: the standard error of the estimate, the parameter value, and the standard deviation of the individual estimates. This result has implications for researchers, because these moderators can (to some extent) be controlled prior to data collection. According to our results, the most important aspect affecting divergence is the standard error, which captures estimation uncertainty (with a higher standard error being associated with larger divergence between estimation methods). To reduce the standard error and associated divergence, the researcher can make sure to collect a large amount of data or to improve the experimental design (Heck & Erdfelder, 2019).[19] Although this conclusion is not new to researchers who apply statistical and cognitive modeling (e.g., Calanchini et al., 2021; Meissner & Rothermund, 2013), it deserves to be reiterated here once more.

In addition, the value of the parameter estimate showed a u-shaped (quadratic) relationship with divergence, such that values near the boundaries of the parameter space (i.e., 0 and 1) were associated with larger divergences. To reduce divergence between methods, the researcher can make adjustments to the study design to avoid parameters near the boundaries. For example, if a parameter denotes a recognition probability, parameter values near the boundaries 0 and 1 (indicating no recognition or perfect recognition memory, respectively) can be avoided by making the recognition task easier or more difficult.

Finally, individual variability (i.e., the standard deviation of the individual parameter estimates) appeared consistently as a predictor of divergence across method pairs, although only with comparatively less predictive power, and—contrary to our expectation—across all method pairs and not only for method pairs that do (vs. do not) account for individual variability (i.e., complete-pooling vs. other methods). Compared to the other two moderators, individual variability is more difficult to control by the researcher. One possibility is to recruit homogeneous samples (e.g., with a similar-age and educational background). However, this approach might conflict with the research question at hand.

Yet, even when taking into account these model-independent aspects, we found that a large part of the divergence remained unexplained. Our analysis thus shows the importance of examining sources of divergence in a large-scale empirical data set, in addition to identifying and addressing these factors in simulation studies. By using an empirical data set can we identify factors that actually explain why some methods yield different results than others.

The Role of Priors in Bayesian Estimation

One feature that is specific to Bayesian estimation is the requirement of a prior distribution which is updated in light of the data, resulting in the posterior distribution. But how much uncertainty regarding the parameter estimates results from the choice for the specific priors used?

Our multiverse meta-analysis relied on the noninformative or weakly informative priors that are typically used with these methods (as implemented in `TreeBUGS`). These priors are expected to have a limited effect on parameter estimates (Gelman et al., 2013) and appear to be the universal choice among researchers using MPT models. The very few exceptions that we know of are cases that focus on hypothesis testing using Bayes factors, which is much more dependent on the prior than parameter estimation (Gronau et al., 2019; Heck & Wagenmakers, 2016; Sarafoglou et al., 2023). However, as exemplified in the case of Bayesian no-pooling in our analysis, it appears that there might

[19] In general, for partial-pooling methods increasing the number of participants increases power, and therefore decreases the standard error, more than increasing the number of observations per participant (Rouder & Haaf, 2018; Westfall et al., 2014). However, in our data, the number of participants and the average number of observations per participant were correlated with the value of the standard error to a similar degree ($r = -.22$ and $r = -.31$, respectively), with the latter correlation even being somewhat larger in magnitude. At the same time, these two indicators of the amount of data were largely uncorrelated with each other ($r = .05$). This suggests that either increasing the number of items or the number of participants (or both) could help in decreasing the standard error for MPT models. When in doubt, we recommend performing a brief simulation study for establishing which aspect of the study design has a larger effect on the standard error for a given model and situation.

be unusual cases where noninformative priors can have a surprisingly large effect on the group-level estimates that are ultimately obtained.

To explore whether such exceptional cases can be expected more frequently, we explored the degree of uncertainty that emerges when deviating from the standard choice of priors. To this end, we re-estimated parameters of the PP-LT-C method for three selected data sets, each associated with a different MPT model. The models were fit using a series of different informative priors, some of which we consider to be quite extreme.[20] We then obtained the distribution of absolute deviations across priors by comparing the estimates of the PP-LT-C model with informative priors to the estimates of the PP-LT-C model with standard priors. We then compared this distribution of deviations across priors to the distribution of absolute deviations across methods obtained in our main analysis (e.g., Tables 2, 3, and 6). For each of the 11 core parameters considered, the mean and median absolute deviations were much smaller across priors than across methods. For 10 of the 11 core parameters, the maximum absolute deviation was much smaller across priors than across methods.[21] Altogether, this additional analysis showed that even when using rather extreme priors that are unlikely to be used in practice, the effect of priors appears to be considerably smaller than the effect of estimation method, both on average and for the maximum absolute deviation (i.e., the worst case).

Strengths and Limitations

With a total of 164 data sets from 142 empirical experiments (corresponding to 61% of the 232 experiments identified as eligible), we were able to gather a large, representative body of empirical MPT data—thanks to the effort of our scientific network and support of fellow MPT researchers. Over and above a traditional meta-analysis, our multiverse meta-analysis required the availability of data at the individual level. Due to this rich individual-level database, we were well positioned to compare the results of all estimation methods considered, which is a major strength of our approach.

Furthermore, our analysis considered a broad selection of MPT models from different paradigms and different subareas of cognitive and social psychology (e.g., memory, decision making, implicit-attitude measurement). We also included data collected from different populations (e.g., clinical patients, college students, older adults). Whereas the selection encompassed a sizeable number of applications—we included all major MPT models that see regular use—this selection is not comprehensive, and other MPT models and populations could also be considered. Given our broad selection, however, it seems unlikely that models not considered here would yield dramatically different results let alone change the picture obtained here.

There are also limitations to our approach. For instance, although our results inform about divergence of results from the different estimation methods, they do not provide information about the accuracy of results, as the true data-generating process remains unknown. From our multiverse analysis alone, we cannot infer that the partial-pooling methods, which show the highest agreement both within and across methods, are also accurate (i.e., unbiased). For example, including correlation parameters in a partial-pooling method (i.e., PP-LT-C) may result in biased estimates for specific, complex MPT models when using typical amounts of data (Groß & Pachur, 2020). We propose that, for cognitive modeling, a joint consideration of the results of simulation studies (to learn about important regularities) and empirical studies (to examine if the identified regularities are relevant for empirical data) will yield the most valuable insight. Our analysis showed that the moderators identified in theoretical considerations and simulation results do not explain divergence as much as expected in empirical applications. Only an approach based on empirical data sets allows us to gain such an insight.

Another shortcoming is that our analysis focused on parameter estimation only and did not consider the convergence or divergence of statistical inferences about parameters (e.g., is the parameter value significantly or credibly larger than zero, or different from .50?) or sets of parameters (e.g., do parameter values differ significantly or credibly between conditions?). Due to considerable variability in our data sets with regard to models and research questions (including within- and between-subjects comparisons, or no comparisons at all), examining the impact on inferences is beyond the scope of this article. We consider this an important follow-up topic that can only be adequately addressed in a model- or paradigm-specific fashion. As discussed above, when interested in statistical inference, the effect of the choice of prior distribution is likely to be larger than when only interested in parameter estimation; here, the use of different priors would need to be considered more systematically.

Implications for Cognitive Modeling

Our meta-analysis complements research in the field of "meta-science"; specifically, research that examines the effect of decisions in the research process (e.g., operationalization, data processing, data analysis) on results (e.g., Baribault et al., 2018; Boehm, Annis, et al., 2018; Dutilh et al., 2019; Landy et al., 2020; Starns et al., 2019; Steegen et al., 2016). Although some of these decisions can be arbitrary in some situations, they can nevertheless affect the results to a substantial degree. As a consequence, the research community has recognized the importance of documenting the impact of these decisions on results to increase transparency and reproducibility, and to counteract questionable research practices.

Here, we provide an example of a multiverse analysis in the field of cognitive modeling. Cognitive modeling has become increasingly popular and, just like regular data-analysis and statistical modeling, involves a number of decisions. In the scarce existing literature on the effect of cognitive-modeling decisions, teams of researchers were asked to use their fitting routines (Boehm, Annis, et al., 2018; Dutilh et al., 2019), and then the results between the different approaches

[20] The three data sets were from the c2HT model (Jaeger et al., 2012), the 2HTSM (Bayen & Kuhlmann, 2011, Experiment 1), and the HB model (Coolin et al., 2016). To maximize the potential effect of the prior, we selected data sets for which the size of the data set was on the smaller side and the absolute deviations reported above (i.e., across estimation methods) was at least medium. Across the different runs, we changed the priors for the group-level parameters for which the default is a standard normal distribution ($\mu = 0$, $\sigma = 1$) on the probit scale, which corresponds to a noninformative (uniform) prior distribution on the probability scale. We independently manipulated both the mean (in three steps, −1, 0, and 1) and standard deviation (in three steps, 0.33, 1, and 3) of the normal distribution resulting in a total of eight models in addition to the variant with the default priors. For priors with $\sigma > 1$, the implied prior on the probability scale is bimodal at the edges of the parameter space.

[21] The only exception was parameter *b* from the HB model for which the maximum absolute deviation across priors was .18, whereas it was .17 across methods.

were compared. By contrast, our work took a more systematic approach and investigated—for the field of MPT modeling—the effect of two important modeling decisions, which resulted in a set of different estimation methods. In the final part of this article, we therefore revisit our main results and speculate on how they might contribute to the broader literature on cognitive modeling.

One of our key results was that the divergence between estimation methods was, on average, small. This is a reassuring result for the class of MPT models, but we recognize that this finding might not necessarily generalize to other cognitive models, such as response time models or models of decision making under risk. For example, there is some evidence from decision making models whose parameters are estimated on a largely unconstrained space (as opposed to MPT parameters which are restricted to the range from 0 to 1) that sparse individual-level data can lead to severely inflated no-pooling estimates (Nilsson et al., 2011). The average of such no-pooling estimates can then diverge quite substantially from partial-pooling estimates. Furthermore, it is well known that for certain model classes, such as learning models, aggregation artifacts are to be expected, so complete-pooling has to be avoided at all costs (Estes, 1956; Evans et al., 2018). Hence, whether or not divergence occurs between different estimation methods should be assessed for each model class separately.

Another important set of our results was that (a) partial-pooling struck the best balance between complete-pooling and no-pooling and that (b) the most important model-independent predictor for divergence was the standard error of the parameter. From the standpoint of statistical theory, these two results seem almost trivial: Partial-pooling was developed to strike a balance between complete-pooling and no-pooling. And, what other than the standard error—the statistical measure of estimation uncertainty—should mainly predict divergence? Thus, while our results might have been expected to some degree, their empirical confirmation provide strong evidence for the validity of the underlying statistical theory and for the practical usability of the considered (partial-pooling) methods. One common problem is that methodological researchers are often more interested in developing ever new methods, instead of extensively testing existing methods (Heinze et al., 2022). By contrast, our study turned out to provide an extensive real-world test of partial-pooling methods, which they passed with flying colors. We hope this empirical success will encourage researchers who work with other model classes to seriously consider using partial-pooling methods. Additionally, our results highlight again that if the standard error of a parameter estimate is large, then the evidence we can derive from such a result might be severely limited.[22] Researchers, check your standard errors.

Our systematic exploration of different estimation methods provided some further insights that are likely relevant for other model classes. As mentioned above, we found systematic differences for partial-pooling methods that do (vs. do not) explicitly model parameter correlations. This is one important, but often overlooked, decision for all partial-pooling applications (for some exceptions, see Bates et al., 2015; Groß & Pachur, 2020). Deciding on whether or not to include correlation parameters in partial-pooling models not only requires considering whether there likely are substantial parameter correlations across participants, but also considering if including correlation parameters might affect parameter estimates given the number of observations. If partial-pooling becomes the de facto standard in the field, which we view as likely, whether to explicitly model parameter correlations needs to be addressed both empirically and with simulation studies in a model-class specific manner.

For two of the methods which are not commonly used we found concerning patterns of results. First, in a Bayesian no-pooling approach, the prior seems to have considerable—and perhaps undue—influence. Given how a Bayesian no-pooling approach works, this is likely a general problem that cannot easily be avoided (e.g., changing the prior does not seem to provide a straightforward solution as long as the prior is to remain both general and proper). Thus, a Bayesian no-pooling approach should probably be avoided for all model classes, unless it can be shown that the problem identified here does not occur. Second, we found that the Bayesian complete-pooling method produced a few estimates that dramatically differed from all other methods (see Figure 4, second column). Recall that this method uses a custom MCMC sampler and is not commonly used in cognitive-modeling applications (although it features prominently in introductions to Bayesian cognitive modeling, e.g., Lee & Wagenmakers, 2013). Here, we used it mainly to complement our multiverse. It is therefore unclear whether these few divergences amount to a serious issue in practice.

Taken together, these two concerning results suggest that if a researcher uses an estimation method that is relatively new or not yet well-established, we advise them to use at least a "minimal" multiverse approach and compare the results with a well-established method. Any noticeable divergence that might emerge should encourage a further exploration of the differences, and how they can be explained.

Conclusions

We examined the robustness of parameter estimation in MPT models across nine different methods that emerge from the combination of two important modeling decisions—the level of data pooling and the statistical framework. A reanalysis of individual-level data from 164 published data sets involving nine popular MPT models showed a reassuringly high convergence across methods. These results indicate that previous parameter estimates based on one of the considered methods would likely be retained if a different method had been applied. Furthermore, we found that partial-pooling methods produced the smallest divergence across methods, making them our recommended default method. Although individual cases with non-negligible deviations were found for all models and method pairs, there was one major exception to the generally positive picture: estimates for an MPT model for which structural-aggregation invariance did not hold. Further analyses revealed that the best predictors for the degree of divergence were the specific MPT model parameter (Figure 10) and the standard error of the estimate. Other predictors considered previously, such as parameter correlations, had limited predictive ability. These results highlight the importance of supplementing simulation studies with large-scale analyses of empirical data.

Recent calls for methodological reforms have brought a number of issues to researchers' attention. Among these is the impact of the decisions made throughout the data-analytic process on the final

[22] In this context, *large* can generally be understood relative to what the parameter estimate is to be compared with. For example, if we are interested if a specific parameter estimate is larger than 0, then the value of the estimate should be at least twice as large as the corresponding standard error. If we are interested in comparing two parameter estimates across conditions, then their difference should generally be much larger than the corresponding standard errors.

outcome (Gelman & Loken, 2014; Simmons et al., 2011; Steegen et al., 2016). The present work is a large-scale attempt to address this issue in the context of the MPT model class—a particularly relevant testbed given its breadth of applications and its methodological maturity (Batchelder & Riefer, 1999; Erdfelder et al., 2009). The positive outcome of our analyses vindicates the long-standing efforts made by MPT modelers on matters concerning parameter estimation and the influence of different sources of variability (e.g., Chechile, 1998; Hu & Batchelder, 1994; Klauer, 2006, 2010; Matzke et al., 2015; Riefer & Batchelder, 1988, 1991; J. B. Smith & Batchelder, 2010). More broadly, this project as a whole provides a template for future efforts to address the challenge of "researcher degrees of freedom" in the context of cognitive modeling.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Alexander, G. E., Satalich, T. A., Shankle, W. R., & Batchelder, W. H. (2016). A cognitive psychometric model for the psychodiagnostic assessment of memory-related deficits. *Psychological Assessment*, *28*(3), 279–293. https://doi.org/10.1037/pas0000163

*Arnold, N. R., Bayen, U. J., & Böhm, M. F. (2015). Is prospective memory related to depression and anxiety? A hierarchical MPT modelling approach. *Memory*, *23*(8), 1215–1228. https://doi.org/10.1080/09658211.2014.969276

*Arnold, N. R., Bayen, U. J., Kuhlmann, B. G., & Vaterrodt, B. (2013). Hierarchical modeling of contingency-based source monitoring: A test of the probability-matching account. *Psychonomic Bulletin & Review*, *20*(2), 326–333. https://doi.org/10.3758/s13423-012-0342-7

*Arnold, N. R., Bayen, U. J., & Smith, R. E. (2015). Hierarchical multinomial modeling approaches: An application to prospective memory and working memory. *Experimental Psychology*, *62*(3), 143–152. https://doi.org/10.1027/1618-3169/a000287

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Ravenzwaaij, D. v., White, C. N., Boeck, P. D., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. https://doi.org/10.1016/j.jmp.2013.12.002

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free-recall of clusterable pairs. *Psychological Review*, *87*(4), 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*(2), 129–149. https://doi.org/10.1111/j.2044-8317.1986.tb00852.x

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. https://doi.org/10.3758/BF03210812

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015, July 10). *Parsimonious mixed models*. arXiv:1506.04967. https://doi.org/10.48550/arXiv.1506.04967

*Bayen, U. J., Erdfelder, E., Bearden, J. N., & Lozito, J. P. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1003–1018. https://doi.org/10.1037/0278-7393.32.5.1003

*Bayen, U. J., & Kuhlmann, B. G. (2011). Influences of source—Item contingency and schematic knowledge on source monitoring: Tests of the probability-matching account. *Journal of Memory and Language*, *64*(1), 1–17. https://doi.org/10.1016/j.jml.2010.09.001

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*(1), 197–215. https://doi.org/10.1037/0278-7393.22.1.197

*Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage*, *43*(4), 775–783. https://doi.org/10.1016/j.neuroimage.2008.08.033

Bell, R., & Buchner, A. (2010). Valence modulates source memory for faces. *Memory & Cognition*, *38*(1), 29–41. https://doi.org/10.3758/MC.38.1.29

*Bell, R., Mieth, L., & Buchner, A. (2015). Appearance-based first impressions and person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 456–472. https://doi.org/10.1037/xlm0000034

*Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/a0031849

*Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 378–391. https://doi.org/10.1037/a0021971

*Besken, M., & Gülgöz, S. (2008). Reliance on schemas in source memory: Age differences and similarity of schemas. *Aging, Neuropsychology, and Cognition*, *16*(1), 1–21. https://doi.org/10.1080/13825580802175650

*Bodner, G. E., Masson, M. E. J., & Caldwell, J. I. (2000). Evidence for a generate–recognize model of episodic influences on word-stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 267–293. https://doi.org/10.1037/0278-7393.26.2.267

Boehm, U., Annis, J., Frank, M., Hawkins, G., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T., Ravenzwaaij, D. v., Servant, M., Singmann, H., Starns, J., Voss, A., Wiecki, T., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the diffusion decision model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*(4), 1614–1631. https://doi.org/10.3758/s13428-018-1054-3

*Bröder, A., Herwig, A., Teipel, S., & Fast, K. (2008). Different storage and retrieval deficits in normal aging and mild cognitive impairment: A multinomial modeling analysis. *Psychology and Aging*, *23*(2), 353–365. https://doi.org/10.1037/0882-7974.23.2.353

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944. https://doi.org/10.1080/09658211.2013.767348

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General*, *124*(2), 137–160. https://doi.org/10.1037/0096-3445.124.2.137

*Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, *43*(5), 321–325. https://doi.org/10.1002/ejsp.1941

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*(4), Article e0250068. https://doi.org/10.1371/journal.pone.0250068

*Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin*, *40*(10), 1285–1296. https://doi.org/10.1177/0146167214540723

*Caldwell, J. I., & Masson, M. E. (2001). Conscious and unconscious influences of memory for object location. *Memory & Cognition*, *29*(2), 285–295. https://doi.org/10.3758/bf03194922

*Castela, M., & Erdfelder, E. (2017). Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences. *Judgment and Decision Making*, *12*(6), 537–552. https://doi.org/10.1017/S1930297500006677

Chechile, R. A. (1998). A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology*, *42*(4), 432–471. https://doi.org/10.1006/jmps.1998.1210

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*(6), 562–576. https://doi.org/10.1016/j.jmp.2009.06.005

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. 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*(2), 328–348. https://doi.org/10.3758/s13423-014-0691-5

*Coolin, A., Erdfelder, E., Bernstein, D. M., Thornton, A. E., & Thornton, W. L. (2016). Inhibitory control underlies individual differences in older adults' hindsight bias. *Psychology and Aging*, *31*(3), 224–238. https://doi.org/10.1037/pag0000088

Dienes, Z. (2008). *Unterstanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

*Dodson, C. S., & Shimamura, A. P. (2000). Differential effects of cue dependency on item and source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 1023–1044. https://doi.org/10.1037//0278-7393.26.4.1023

*Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151. https://doi.org/10.1037/a0024957

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., … Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127. https://doi.org/10.1038/scientificamerican0577-119

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie*. MADOC. https://madoc.bib.uni-mannheim.de/63897/

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 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*(1), 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*(2), 387–414. https://doi.org/10.1037//0278-7393.24.2.387

Erdfelder, E., Quevedo Pütter, J., & Schnuerch, M. (2023, February 6). *On aggregation invariance of multinomial processing tree models*. PsyArXiv. https://doi.org/10.31234/osf.io/59csk

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. https://doi.org/10.1037/h0045156

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403–408. https://doi.org/10.3758/BF03193784

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592–605. https://doi.org/10.1037/rev0000105

*Filevich, E., Horn, S. S., & Kühn, S. (2019). Within-person adaptivity in frugal judgments from memory. *Psychological Research*, *83*(3), 613–630. https://doi.org/10.1007/s00426-017-0962-7

*Francis, W. S., Taylor, R. S., Gutiérrez, M., Liaño, M. K., Manzanera, D. G., & Penalver, R. M. (2018). The effects of bilingual language proficiency on recall accuracy and semantic clustering in free recall output: Evidence for shared semantic associations across languages. *Memory*, *26*(10), 1364–1378. https://doi.org/10.1080/09658211.2018.1476551

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), Article 460. https://doi.org/10.1511/2014.111.460

*Giang, T., Bell, R., & Buchner, A. (2012). Does facial resemblance enhance cooperation? *PLOS ONE*, *7*(10), Article e47809. https://doi.org/10.1371/journal.pone.0047809

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Chapmann & Hall.

Golz, D., & Erdfelder, E. (2004). Effekte von l-dopa auf die speicherung und den abruf verbaler informationen bei schlaganfallpatienten [Effects of L Dopa on storage and retrieval of verbal information in stroke patients]. *Zeitschrift für Neuropsychologie*, *15*(4), 275–286. https://doi.org/10.1024/1016-264X.15.4.275

*Gonsalkorale, K., Sherman, J. W., Allen, T. J., Klauer, K. C., & Amodio, D. M. (2011). Accounting for successful control of implicit racial bias: The roles of association activation, response monitoring, and overcoming bias. *Personality and Social Psychology Bulletin*, *37*(11), 1534–1545. https://doi.org/10.1177/0146167211414064

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using warp-III bridge sampling. *Psychometrika*, *84*(1), 261–284. https://doi.org/10.1007/s11336-018-9648-3

*Groß, J., & Bayen, U. J. (2015). Adult age differences in hindsight bias: The role of recall ability. *Psychology and Aging*, *30*(2), 253–258. https://doi.org/10.1037/pag0000017

*Groß, J., & Bayen, U. J. (2017). Effects of dysphoria and induced negative mood on the processes underlying hindsight bias. *Cognition & Emotion*, *31*(8), 1715–1724. https://doi.org/10.1080/02699931.2016.1249461

Groß, J., & Pachur, T. (2020). Parameter estimation approaches for multinomial processing tree models: A comparison for models of memory and judgment. *Journal of Mathematical Psychology*, *98*, Article 102402. https://doi.org/10.1016/j.jmp.2020.102402

Hao, N., Feng, Y., & Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, *113*(522), 615–625. https://doi.org/10.1080/01621459.2016.1264956

*Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition*, *34*(4), 826–838. https://doi.org/10.3758/BF03193430

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

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., & Wagenmakers, E.-J. (2016). Adjusted priors for Bayes factors involving reparameterized order constraints. *Journal of Mathematical Psychology*, *73*, 110–116. https://doi.org/10.1016/j.jmp.2016.05.004

Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., & White, I. R. (2022, September 27). *Phases of methodological research in biostatistics—Building the evidence base for new methods*. arXiv. https://doi.org/10.48550/arXiv.2209.13358

*Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2010). One-reason decision making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 123–134. https://doi.org/10.1037/a0017518

*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*(4), 827–839. https://doi.org/10.1037/a0022638

*Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2012). A matter of time: Antecedents of one-reason decision making based on recognition. *Acta Psychologica*, *141*(1), 9–16. https://doi.org/10.1016/j.actpsy.2012.05.006

*Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition*, *43*(4), 659–671. https://doi.org/10.3758/s13421-014-0493-z

*Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology*, *55*(6), 394–401. https://doi.org/10.1027/1618-3169.55.6.394

*Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- versus evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1296–1305. https://doi.org/10.1037/a0016565

*Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making*, *22*(5), 510–522. https://doi.org/10.1002/bdm.644

*Hilbig, B. E., & Richter, T. (2011). Homo heuristicus outnumbered: Comment on Gigerenzer and Brighton (2009). *Topics in Cognitive Science*, *3*(1), 187–196. https://doi.org/10.1111/j.1756-8765.2010.01123.x

*Hilbig, B. E., Scholl, S. G., & Pohl, R. F. (2010). Think or blink—Is the recognition heuristic an "intuitive" strategy? *Judgment and Decision Making*, *5*(4), 300–309. https://doi.org/10.1017/S1930297500003533

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (2023). Many-analysts religion project: Reflection and conclusion. *Religion, Brain & Behavior*, *13*(3), 356–363. https://doi.org/10.1080/2153599X.2022.2070263

*Horn, S. S., Bayen, U. J., Smith, R. E., & Boywitt, C. D. (2011). The multinomial model of prospective memory: Validity of ongoing-task parameters. *Experimental Psychology*, *58*(3), 247–255. https://doi.org/10.1027/1618-3169/a000091

*Horn, S. S., Pachur, T., & Mata, R. (2015). How does aging affect recognition-based inference? A hierarchical Bayesian modeling approach. *Acta Psychologica*, *154*, 77–85. https://doi.org/10.1016/j.actpsy.2014.11.001

*Horn, S. S., Ruggeri, A., & Pachur, T. (2016). The development of adaptive decision making: Recognition-based inference in children and adolescents. *Developmental Psychology*, *52*(9), 1470–1485. https://doi.org/10.1037/dev0000181

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*(1), 21–47. https://doi.org/10.1007/BF02294263

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*(2), 187–211. https://doi.org/10.2307/1942661

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*(1), 116–159. https://doi.org/10.1080/10463283.2016.1212966

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

*Jaeger, A., Cox, J. C., & Dobbins, I. G. (2012). Recognition confidence under violated and confirmed memory expectations. *Journal of Experimental Psychology: General*, *141*(2), 282–301. https://doi.org/10.1037/a0025687

*Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306. https://doi.org/10.1037/a0015525

*Jin, Z., Rivers, A. M., Sherman, J. W., & Chen, R. (2016). Measures of implicit gender attitudes may exaggerate differences in underlying associations among Chinese urban and rural women. *Psychologica Belgica*, *56*(1), 13–22. https://doi.org/10.5334/pb.308

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*, Article 102400. https://doi.org/10.1016/j.jmp.2020.102400

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422–431. https://doi.org/10.1037/0033-2909.99.3.422

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*(1), 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*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

*Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

*Klauer, K. C., & Meiser, T. (2000). A source-monitoring analysis of illusory correlations. *Personality and Social Psychology Bulletin*, *26*(9), 1074–1093. https://doi.org/10.1177/01461672002611005

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning *Psychological Review*, *107*(4), 852–884. https://doi.org/10.1037//0033-295X.107.4.852

*Koen, J. D., Aly, M., Wang, W.-C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1726–1741. https://doi.org/10.1037/a0033671

*Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1536–1542. https://doi.org/10.1037/a0020448

*Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory*, *18*(8), 519–522. https://doi.org/10.1101/lm.2214511

*Koranyi, N., & Meissner, F. (2015). Handing over the reins: Neutralizing negative attitudes toward dependence in response to reciprocal romantic liking. *Social Psychological and Personality Science*, *6*(6), 685–691. https://doi.org/10.1177/1948550615580169

Krefeld-Schwalb, A., Scheibehenne, B., & Pachur, T. (2022). Structural parameter interdependencies in computational models of cognition. *Psychological Review*, *129*(2), 313–339. https://doi.org/10.31234/osf.io/pxmnw

*Kroneisen, M., & Bell, R. (2018). Remembering the place with the tiger: Survival processing can enhance source memory. *Psychonomic Bulletin & Review*, *25*(2), 667–673. https://doi.org/10.3758/s13423-018-1431-z

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

*Kuhlmann, B. G., Bayen, U. J., Meuser, K., & Kornadt, A. E. (2016). The impact of age stereotypes on source monitoring in younger and older adults. *Psychology and Aging*, *31*(8), 875–889. https://doi.org/10.1037/pag0000140

*Kuhlmann, B. G., & Touron, D. R. (2017). Relate it! Objective and subjective evaluation of mediator-based strategies for improving source memory in younger and older adults. *Cortex*, *91*, 25–39. https://doi.org/10.1016/j.cortex.2016.11.015

*Kuhlmann, B. G., Vaterrodt, B., & Bayen, U. J. (2012). Schema bias in source monitoring varies with encoding conditions: Support for a probability-matching account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1365–1376. https://doi.org/10.1037/a0028147

*Küppers, V., & Bayen, U. J. (2014). Inconsistency effects in source memory and compensatory schema-consistent guessing. *Quarterly Journal of Experimental Psychology*, *67*(10), 2042–2059. https://doi.org/10.1080/17470218.2014.904914

Landy, J. F., Jia, M. (L.), Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

*Lueke, A., & Gibson, B. (2015). Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding. *Social Psychological and Personality Science*, *6*(3), 284–291. https://doi.org/10.1177/1948550614559651

*Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*(1), 205–235. https://doi.org/10.1007/s11336-013-9374-9

*Meiser, T. (2003). Effects of processing strategy on episodic memory and contingency learning in group stereotype formation. *Social Cognition*, *21*(2), 121–156. https://doi.org/10.1521/soco.21.2.121.21318

*Meiser, T., & Hewstone, M. (2001). Crossed categorization effects on the formation of illusory correlations. *European Journal of Social Psychology*, *31*(4), 443–466. https://doi.org/10.1002/ejsp.55

*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*(1), 45–69. https://doi.org/10.1037/a0030734

*Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the implicit association test. *Social Psychological and Personality Science*, *6*(7), 740–748. https://doi.org/10.1177/1948550615580381

*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*(1), 3–11. https://doi.org/10.1002/bdm.2040

*Michalkiewicz, M., & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, *44*(3), 454–468. https://doi.org/10.3758/s13421-015-0567-6

*Mieth, L., Bell, R., & Buchner, A. (2016a). Cognitive load does not affect the behavioral and cognitive foundations of social cooperation. *Frontiers in Psychology*, *7*, Article 1312. https://doi.org/10.3389/fpsyg.2016.01312

*Mieth, L., Bell, R., & Buchner, A. (2016b). Memory and disgust: Effects of appearance-congruent and appearance-incongruent information on source memory for food. *Memory*, *24*(5), 629–639. https://doi.org/10.1080/09658211.2015.1034139

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Nestler, S., & Erdfelder, E. (2023). Random effects multinomial processing tree models: A maximum-likelihood approach. *Psychometrika*, *88*(3), 809–829. https://doi.org/10.1007/s11336-023-09921-w

Nilsson, H., Rieskamp, J., & Wagenmakers, E. J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*(1), 84–93. https://doi.org/10.1016/j.jmp.2010.08.006

*Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General*, *139*(2), 341–364. https://doi.org/10.1037/a0018926

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

*Pavawalla, S. P., Schmitter-Edgecombe, M., & Smith, R. E. (2012). Prospective memory after moderate-to-severe traumatic brain injury: A multinomial modeling approach. *Neuropsychology*, *26*(1), 91–101. https://doi.org/10.1037/a0025866

Pawitan, Y. (2014). *In all likelihood statistical modelling and inference using likelihood*. Oxford University Press.

Plummer, M. (2003, March 20–22). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* [Conference session]. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria, United States.

*Pohl, R. F. (2017). Measuring age-related differences in using a simple decision strategy: The case of the recognition heuristic. *Zeitschrift für Psychologie*, *225*(1), 20–30. https://doi.org/10.1027/2151-2604/a000283

*Pohl, R. F., Bayen, U. J., & Martin, C. (2010). A multiprocess account of hindsight bias in children. *Developmental Psychology*, *46*(5), 1268–1282. https://doi.org/10.1037/a0020209

*Pohl, R. F., Erdfelder, E., Hilbig, B. E., Liebke, L., & Stahlberg, D. (2013). Effort reduction after self-control depletion: The role of cognitive resources in use of simple heuristics. *Journal of Cognitive Psychology*, *25*(3), 267–276. https://doi.org/10.1080/20445911.2012.758101

*Pohl, R. F., Michalkiewicz, M., Erdfelder, E., & Hilbig, B. E. (2017). Use of the recognition heuristic depends on the domain's recognition validity, not on the recognition validity of selected sets of objects. *Memory & Cognition*, *45*(5), 776–791. https://doi.org/10.3758/s13421-017-0689-0

*Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224–232. https://doi.org/10.1037/a0017682

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org/

Read, T., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–335). 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*(2), 184–201. https://doi.org/10.1037/1040-3590.14.2.184

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

*Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General*, *137*(2), 370–389. https://doi.org/10.1037/0096-3445.137.2.370

*Rummel, J., Boywitt, C. D., & Meiser, T. (2011). Assessing the validity of multinomial models using extraneous variables: An application to prospective memory. *Quarterly Journal of Experimental Psychology*, *64*(11), 2194–2210. https://doi.org/10.1080/17470218.2011.586708

Sarafoglou, A., Kuhlmann, B. G., Aust, F., & Haaf, J. M. (2023). *Theory-informed refinement of Bayesian hierarchical MPT modeling*. PsyArXiv. https://doi.org/10.31234/osf.io/kvyt5

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, *22*(2), 391–407. https://doi.org/10.3758/s13423-014-0684-4

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000561

*Schnitzspahn, K. M., Horn, S. S., Bayen, U. J., & Kliegel, M. (2012). Age effects in emotional prospective memory: Cue valence differentially affects the prospective and retrospective component. *Psychology and Aging*, *27*(2), 498–509. https://doi.org/10.1037/a0025021

*Schütz, J., & Bröder, A. (2011). Signal detection and threshold models of source memory. *Experimental Psychology*, *58*(4), 293–311. https://doi.org/10.1027/1618-3169/a000097

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284. https://doi.org/10.1080/03640210802414826

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

*Simons, J. S., Verfaellie, M., Galton, C. J., Miller, B. L., Hodges, J. R., & Graham, K. S. (2002). Recollection-based memory in frontotemporal dementia: Implications for theories of long-term memory. *Brain*, *125*(11), 2523–2536. https://doi.org/10.1093/brain/awf247

Singmann, H., Groß, J., & Kuhlmann, B. G. (2024). *MPT multiverse meta-analysis*. https://osf.io/waen6

Singmann, H., Heck, D. W., & Barth, M. (2020). *MPTmultiverse: Multiverse analysis of multinomial processing tree models* (R package) [Computer software]. https://CRAN.R-project.org/package=MPTmultiverse

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. https://doi.org/10.3758/s13428-012-0259-0

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Psychology Press.

*Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 615–625. https://doi.org/10.1037/0278-7393.30.3.615

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*(4), 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*(1), 167–183. https://doi.org/10.1016/j.jmp.2009.06.007

Smith, R. E., & Bayen, U. J. (2004). A multinomial model of event-based prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 756–77. https://doi.org/10.1037/0278-7393.30.4.756

*Smith, R. E., & Bayen, U. J. (2005). The effects of working memory resource availability on prospective memory: A formal modeling approach. *Experimental Psychology*, *52*(4), 243–256. https://doi.org/10.1027/1618-3169.52.4.243

*Smith, R. E., Bayen, U. J., & Martin, C. (2010). The cognitive processes underlying event-based prospective memory in school-age children and young adults: A formal model-based study. *Developmental Psychology*, *46*(1), 230–244. https://doi.org/10.1037/a0017100

*Smith, R. E., & Hunt, R. R. (2014). Prospective memory in young and older adults: The effects of task importance and ongoing task load. *Aging, Neuropsychology, and Cognition*, *21*(4), 411–431. https://doi.org/10.1080/13825585.2013.827150

*Smith, R. E., McConnell Rogers, M. D., McVay, J. C., Lopez, J. A., & Loft, S. (2014). Investigating how implementation intentions improve non-focal prospective memory tasks. *Consciousness and Cognition*, *27*, 213–230. https://doi.org/10.1016/j.concog.2014.05.003

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin & Review*, *25*(6), 2047–2068. https://doi.org/10.3758/s13423-018-1446-5

*Stahl, C., Barth, M., & Haider, H. (2015). Distorted estimates of implicit and explicit learning in applications of the process-dissociation procedure to the SRT task. *Consciousness and Cognition*, *37*, 27–43. https://doi.org/10.1016/j.concog.2015.08.003

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., … Wilson, J. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

*Süssenbach, P., Gollwitzer, M., Mieth, L., Buchner, A., & Bell, R. (2016). Trustworthy tricksters: Violating a negative social expectation affects source memory and person perception when fear of exploitation is high. *Frontiers in Psychology*, *7*, Article 2037. https://doi.org/10.3389/fpsyg.2016.02037

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.1007/BF00122574

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*, Article 1. https://doi.org/10.1038/s43586-020-00001-2

*Van Dessel, P., Mertens, G., Smith, C. T., & De Houwer, J. (2017). The mere exposure instruction effect: Mere exposure instructions influence liking. *Experimental Psychology*, *64*(5), 299–314. https://doi.org/10.1027/1618-3169/a000376

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

Vandekerckhove, J., White, C. N., Trueblood, J. S., Rouder, J. N., Matzke, D., Leite, F. P., Etz, A., Donkin, C., Devezer, B., Criss, A. H., & Lee, M. D. (2019). Robust diversity in cognitive science. *Computational Brain & Behavior*, *2*(3), 271–276. https://doi.org/10.1007/s42113-019-00066-7

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, *112*(4), 862–880. https://doi.org/10.1037/0033-295X.112.4.862

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

*Wrzus, C., Egloff, B., & Riediger, M. (2017). Using implicit association tests in age-heterogeneous samples: The importance of cognitive abilities and quad model processes. *Psychology and Aging*, *32*(5), 432–446. https://doi.org/10.1037/pag0000176

(*Appendices follow*)

Appendix A

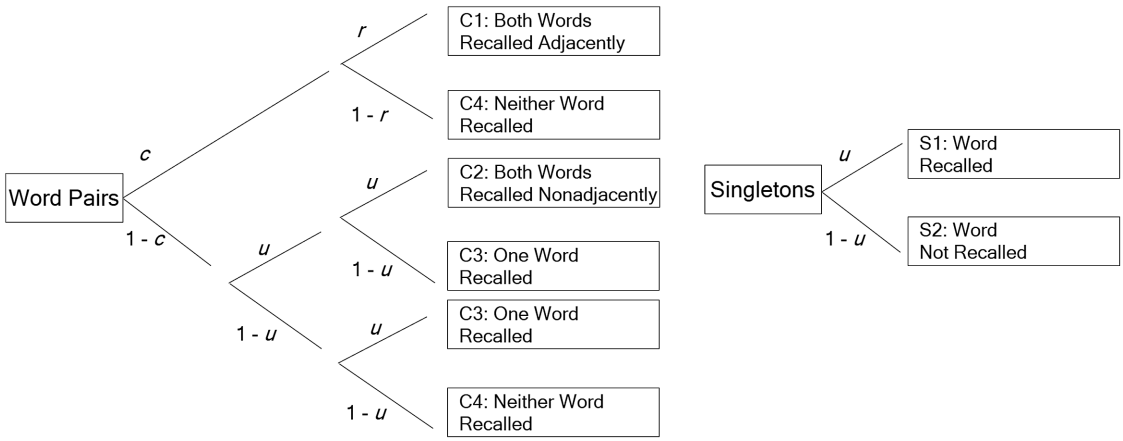
General Model Equation Structure of MPT Models

MPT models aim at explaining probabilities of observable responses as a function of latent probabilities that represent outcomes of underlying cognitive processes. Let us provide a formal introduction to the MPT model class using the pair-clustering model of Batchelder and Riefer (1986) shown in Figure A1 (for a more general tutorial on MPT modeling, see Schmidt et al., 2023). The pair-clustering model refers to a memory experiment in which participants are asked for free recall of a list of sequentially presented words. Some of these words are semantically related to another word in the list (i.e., word-pair items) while others are not (i.e., singleton items). The model aims at measuring three cognitive process outcomes, namely, storage of a word-pair as a cluster in memory (latent probability c), retrieval of a stored cluster from memory (latent probability r), and successful storage and retrieval of an unclustered single word (latent probability u). Each of the two item types is modeled in a separate processing tree. In the general case, MPT models include K trees with J_k response categories C_{kj} in tree k, $k = 1, \ldots, K$, $j = 1, \ldots, J_k$, and $S \leq \sum_{k=1}^{K}(J_k - 1)$ parameters θ_s, $s = 1, \ldots, S$, each of which is an element of [0,1]. In the pair-clustering model, we have $K = 2$, $J_1 = 4$, $J_2 = 2$, and $S = 3$, and the three parameters are $\theta_1 = c$, $\theta_2 = r$ and $\theta_3 = u$, respectively. The general structure of the MPT model equation that describes how response probabilities $p(C_{kj})$ depend on the S parameters collected in the parameter vector $\boldsymbol{\theta}$ is given by

$$p(C_{kj}|\boldsymbol{\theta}) = \sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} \theta_s^{a_{skji}} (1 - \theta_s)^{b_{skji}}, \tag{A1}$$

where I_{kj} is the number of branches that end up in category C_{kj} while a_{skji} and b_{skji} indicate how often a parameter θ_s and its complement $(1 - \theta_s)$, respectively, appear on the i-th branch of category C_{kj} (Hu & Batchelder, 1994). To illustrate, there is only a single branch for category C_{11} of the pair clustering model, namely, the first branch in Figure A1. This branch represents successful cluster storage with probability $\theta_1 = c$ followed by successful cluster retrieval with probability $\theta_2 = r$. No other parameter is involved in this branch. Hence, for the first branch, $a_{1111} = a_{2111} = 1$, whereas all other a_{s111} and b_{s111} are zero. The a_{skji} and b_{skji} count variables for other branches can be derived accordingly.

Figure A1
MPT Model of Pair-Clustering



Note. C1 = word-pair Category 1; C2 = word-pair Category 2; C3 = word-pair Category 3; C4 = word-pair Category 4; S1 = singleton Category 1; S2 =singleton Category 2; c = probability that a word-pair is stored as a cluster; r = probability that a previously stored cluster is successfully retrieved as a cluster; u = probability that a single word is stored and retrieved; MPT = multinomial processing tree.

(*Appendices continue*)

Appendix B

Aggregation Invariance Properties and Implications for Parameter Estimates

In this appendix, we first introduce the notions of structural and empirical aggregation invariance of MPT models and then discuss implications of these invariance properties for divergence between parameter estimates.

Structural and Empirical Aggregation Invariance

Equation A1 refers to a single participant with parameter vector $\boldsymbol{\theta}$. When we assume that an MPT model holds for each participant but participants may differ in their parameter values, the parameter vector $\boldsymbol{\theta}$ becomes a random vector $\boldsymbol{\Theta}$, with $\boldsymbol{\theta}_n$ being the value of this random vector for participant n, $n = 1, \ldots, N$. By implication, the response category probabilities $p(C_{kj})$ also become random variables $P(C_{kj})$:

$$P(C_{kj}|\boldsymbol{\Theta}) = \sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} \Theta_s^{a_{skji}} (1 - \Theta_s)^{b_{skji}}. \tag{B1}$$

As shown by Erdfelder et al. (2023), the expected frequencies $N_k \cdot E(P(C_{kj}|\boldsymbol{\Theta}))$ predicted by an MPT model for category C_{kj} when parameters may vary between participants are easily derived when this model satisfies the conditions of structural and empirical aggregation invariance defined as follows:

Definition 1 (Structural Aggregation Invariance)

An MPT model is called structurally aggregation invariant (SAI), when none of its parameters Θ_s (or parameter complements $(1 - \Theta_s)$) occurs repeatedly in any branch of the model, and if a parameter and its complement never co-occur in the same branch.

Definition 2 (Empirical Aggregation Invariance)

We call MPT models empirically aggregation invariant (EAI), when all parameters (or parameter complements) that co-occur in the same branch are stochastically independent.

Based on these two definitions, it can be shown analytically that any MPT model that satisfies the SAI and EAI properties and holds for each individual must also hold for the category frequencies aggregated across individuals. Moreover, the parameters of the model for the aggregate data are just the expectations (i.e., the means $E(\boldsymbol{\Theta})$) of the individual parameters (cf. Erdfelder et al., 2023). Specifically, making use of the SAI and EAI properties one can derive the expected category frequencies for the aggregate data as follows:

$$\begin{aligned} N_k \cdot E(P(C_{kj}|\boldsymbol{\Theta})) &= N_k \cdot E\left(\sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} \Theta_s^{a_{skji}} (1 - \Theta_s)^{b_{skji}}\right) \\ &= N_k \cdot \left(\sum_{i=1}^{I_{kj}} \prod_{s=1}^{S} (E(\Theta_s))^{a_{skji}} (1 - E(\Theta_s))^{b_{skji}}\right). \end{aligned} \tag{B2}$$

Notably, for typical applications in the MPT context, a weaker concept of empirical aggregation invariance suffices to derive the same result.

Definition 3 (Weak Empirical Aggregation Invariance)

We call MPT models weakly empirically aggregation invariant (WEAI), when the covariance between any two parameters (or parameter complements) that co-occur in the same branch is zero.

While WEAI is only a necessary and not a sufficient condition for EAI, MPT models that are both SAI and WEAI also imply Equation B2 under either of two conditions: (a) We consider MPT models with no more than two parameters (or parameter complements) per branch, such as the pair-clustering model or the process-dissociation model discussed in this article; or (b) we consider MPT models with any number of parameters per branch but restrict their possible distributions to those were deviations from stochastic independence—if any—are fully described by pairwise covariances (e.g., the multivariate normal distribution on a probit scale). The hierarchical MPT models considered in our multiverse analysis are all in line with the second requirement, so that the combination of SAI according to Definition 1 and WEAI according to Definition 3 suffices to derive Equation B2.

Note, however, that the derivation of Equation B2 does not work anymore when either SAI or WEAI or both properties are violated. The pair-clustering model, for example, violates SAI because parameter u and its complement $(1 - u)$ occur repeatedly in the branches 3–6 of the word-pair tree in Figure A1. Hence, some of the exponents a_{skji} and b_{skji} are larger than 1 for this model and derivation of the last line in Equation B2 is not anymore possible. The only way to make the pair-clustering model consistent with the last line is to assume that (a) parameter u is a constant that does not vary between individuals and (b) the covariance between the c and the r parameters across individuals is zero.

Convergence of Parameter Estimates

The derivation in the previous subsection refers to the expectations of the parameters, that is, to their true means at the population level. However, this has direct implications for all estimation methods of group-level parameters that are consistent. If the sample size gets very large and the respective MPT model that satisfies SAI and EAI holds for each participant, all these methods must converge against the true expectations. Since all estimation methods considered in this multiverse analysis are consistent, divergence between group-level parameter estimates derived with different methods should asymptotically vanish when both the number of participants and the number of responses per participant[B1] are sufficiently large. Alternatively, we can refer to standard errors of parameter estimates that must convergence against zero when sample sizes and the number of responses converge against infinity. It follows that group-level mean parameter estimates should be identical across estimation methods if their standard errors approach zero. Or in other words, the more precisely we can estimate a specific MPT parameter, the lower we expect

[B1] Note that a large number of responses per participant is especially crucial for all no-pooling methods considered in this multiverse analysis. If the number of responses is small, no-pooling estimates (including single-participant ML) may suffer from systematic estimation bias that does not vanish (but rather stabilizes) when only the number of participants increases.

(*Appendices continue*)

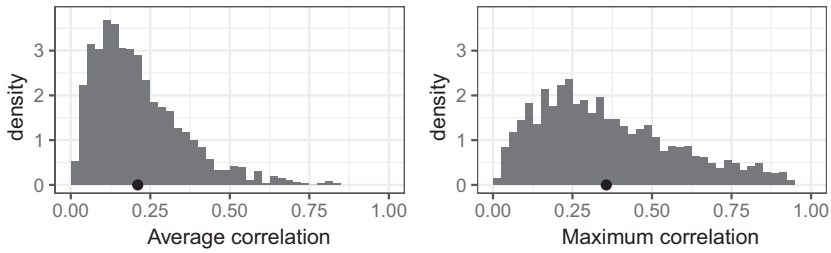
divergences to be, as long as the invariance properties of a valid MPT model hold.

Taken together, we can derive the following empirical predictions for MPT models for which the SAI property holds. *The agreement between any pair of estimation methods considered in our multiverse analysis should be essentially perfect (i.e., no disagreement) as long as the correlations with other parameters on the same branch is approximately zero and the standard error approaches zero*.

This relationship is expected to hold for any consistent estimation method under the distribution models considered in our multiverse analysis. Importantly, however, to derive this result we need to assume that the corresponding model indeed holds for each participant. However, this assumption is shared by all methods addressed in our multiverse analysis. Were we to see noticeable divergences from the empirical prediction, this could be taken as evidence that this assumption does not hold to a satisfactory degree.

Appendix C

Distribution of Parameter Correlations



Note. Distributions of parameter correlations across all parameters and models, as estimated from the partial-pooling, latent-trait with correlation parameters model. In each panel, the black dot indicates the mean. The left panel shows the distribution of average correlations for each parameter ($M = 0.21$) and the right panel the distribution of maximum correlations for each parameter ($M = 0.36$).

Received March 9, 2023
Revision received January 8, 2024
Accepted March 8, 2024 ■