Journal of Mathematical Psychology 98 (2020) 102402

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Parameter estimation approaches for multinomial processing tree models: A comparison for models of memory and judgment[☆]

Julia Groß [a,*], Thorsten Pachur [b]

[a] Universität Mannheim, Germany
[b] Max Planck Institute for Human Development, Berlin, Germany

ARTICLE INFO

Article history:
Received 14 June 2019
Received in revised form 24 April 2020
Accepted 20 May 2020
Available online 1 August 2020

Keywords:
Multinomial processing tree models
Simulation
Parameter recovery
Source memory
Hindsight bias

ABSTRACT

Multinomial processing tree (MPT) models are commonly used in cognitive psychology to disentangle and measure the psychological processes underlying behavior. Various estimation approaches have been developed to estimate the parameters of MPT models for a group of participants. These approaches are implemented in various programs (e.g., MPTinR, TreeBUGS) and differ with regard to how data are pooled across participants (no pooling, complete pooling, or partial pooling). The partial-pooling approaches differ with regard to whether correlations between individual-level parameters are explicitly modeled (latent-trait MPT) or not (beta-MPT). However, it is currently unclear whether the theoretical advantages of the partial-pooling approaches actually yield the best results in standard practice (i.e., with typical parameter values and amounts of data). We conducted parameter recovery analyses comparing the accuracy and precision of four estimation approaches for two MPT models: the source-monitoring model and the hindsight-bias model. For essential ("core") parameters of the two models, the partial-pooling approaches yielded the best results overall. Importantly, there were also model-specific differences between the approaches. For the source-monitoring model, the latent-trait approach achieved the best results. For more complex hindsight-bias model, the latent-trait approach appeared to be overparameterized for typical amounts of data; here, the beta-MPT approach was better. We derive recommendations for applications of the two MPT models.

1. Introduction

Multinomial processing tree (MPT) models are commonly used in various fields of psychological research (see Batchelder & Riefer, 1999; Erdfelder, Auer, Hilbig, Aßfalg, Moshagen, & Nadarevic, 2009; Hütter & Klauer, 2016, for theoretical and empirical reviews). As measurement tools, they help to disentangle the latent cognitive processes assumed to underlie people's responses in a particular experimental task. The interplay of these cognitive processes is represented as a processing tree, with each branch of the tree denoting a different sequence of processes that results in a response from a particular response category. The MPT model parameters represent the probability of each of these latent psychological processes. One or more sequences of latent processes can result in the same response category (many-to-one mapping).

As an illustrative example, consider the two-high-threshold model of recognition memory (2HTM; Snodgrass & Corwin, 1988). In recognition tasks, participants are asked to decide whether a test item is a previously studied item or a new distractor item. The model disentangles responses based on recognition memory from responses based on guessing. According to the 2HTM, an item can pass one of two detection thresholds (Fig. 1): With probability p_O, an old item is detected as old based on recognition memory, and the participant responds with "old". With probability p_N, a distractor item is detected as new based on recognition memory, and the participant responds with "new". If an item does not pass either threshold (i.e., if there is no detection), the participant guesses that it is old (new) with probability b ($1 - b$). Thus, the model measures and disentangles the probability of responses based on recognition memory (p_O,p_N) from that of responses based on guessing (b).

Due to the simple structure of MPT models, using them to formulate and test psychological theories is relatively straightforward. Moreover, the scope of MPT applications has widened with recent methodological developments. For example, various model extensions (Coolin, Erdfelder, Bernstein, Thornton, & Thornton, 2015; Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Smith & Batchelder, 2010) make it possible to account

[☆] This work was funded by Grant GR-4649/2-1 from the German Research Foundation (DFG)[I=59]. We thank Edgar Erdfelder for comments on a previous version of this manuscript. We thank Susannah Goss for editing the manuscript.

[*] Correspondence to: Department of Psychology, University of Mannheim, A5, Building C, D-68159 Mannheim, Germany.
E-mail address: julia.gross@psychologie.uni-mannheim.de (J. Groß).

https://doi.org/10.1016/j.jmp.2020.102402

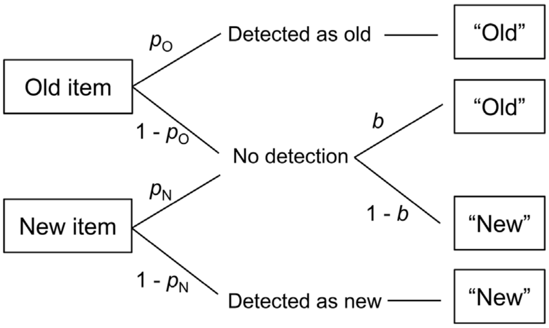

Fig. 1. The two-high-threshold model (2HTM) of recognition memory depicted as a processing tree. Rectangles represent observable events.

for heterogeneity in parameters across individuals (e.g., within different age groups; Groß & Pachur, 2019; Horn, Pachur, & Mata, 2015; Horn, Ruggeri, & Pachur, 2016; Kuhlmann, Bayen, Meuser, & Kornadt, 2016) or across both individuals and items (Matzke et al., 2015). Further, it is possible to capture variability in model parameters due to variability in external variables (Coolin et al., 2015; Coolin, Erdfelder, Bernstein, Thornton, & Thornton, 2016; Michalkiewicz, Arden, & Erdfelder, 2018) and to estimate correlations between individual-level parameters (e.g., to measure the stability of cognitive processes; Michalkiewicz & Erdfelder, 2016; Wulff & Kuhlmann, 2020).

To be useful as measurement models, MPT parameters must be estimated without bias (i.e., accurately) and with high precision (i.e., reliably). If there is bias, this may lead to incorrect conclusions about parameter differences between groups of participants or experimental conditions, particularly if the extent of bias varies across the parameter space (e.g., when low parameter values are consistently overestimated, but high parameter values are accurately estimated).[1] If parameter estimation is imprecise, this may limit the ability to detect differences between groups or conditions (i.e., type II errors may occur).

How well MPT parameters can be estimated depends on various factors. One is the parameterization of the model; that is, how the parameters are arranged in the tree and which of them are free to vary, set equal to each other, or set to a constant. Another is the estimation method used. As we explain in more detail in Section 2, estimation methods for MPT models can differ with regard to whether and how they incorporate heterogeneity in parameters (between participants or items) and correlations among model parameters. Depending on the presence and extent of parameter heterogeneity and correlations, the choice of estimation approach might crucially affect the accuracy and precision of parameter estimation.

The most common estimation approaches in the MPT literature are available to researchers in the form of programs (or packages) such as *TreeBUGS* (Heck, Arnold, & Arnold, 2018), *multiTree* (Moshagen, 2010), and *MPTinR* (Singmann & Kellen, 2013). But which approach should be used for a given data set and model? It is currently unclear whether the most sophisticated approaches—in particular, approaches that take into account heterogeneity and explicitly model parameter correlations—yield more accurate and more precise parameter estimates than simpler approaches under standard conditions (i.e., with a typical amount of data and typical parameter values), and how large the differences between the approaches actually are.

The goal of this article is to compare the ability of different approaches to MPT parameter estimation, as implemented in available programs, to accurately and precisely recover parameter values. We present a series of simulations that examine different estimation approaches in the context of two MPT models, namely, the source-monitoring model (Bayen, Murnane, & Erdfelder, 1996) and the hindsight-bias model (Erdfelder & Buchner, 1998). As described in Section 2.5, these models contain both "core" parameters, that represent key processes of theoretical interest, and "auxiliary" parameters, that are important for the model's architecture but are not of theoretical interest. Our focus is on the ability of the approaches to estimate the core parameters.

In the recovery analyses, we simulated sets of experiments in which the data for each sample of synthetic participants were generated by the respective MPT model with a set of (empirically informed) parameter values. The question is how well the estimation approaches estimated the parameter values that were used to generate the data. Further, we examined the extent to which parameter heterogeneity across the group of synthetic participants within an experiment affected recovery performance. Whenever we observed difficulties in the recovery of core parameters, we doubled the number of participants and/or of the items to test whether collecting higher (but still realistic) amounts of data would help to alleviate the problems.

In the following, we first describe the different estimation approaches considered (Section 2). We next introduce the MPT models of source monitoring and of hindsight bias (Section 3) and present the methods and results of various sets of parameter recovery studies for these two models (Sections 4 and 5). Finally, we discuss implications of our results for the application of the two models in particular, and for MPT model-based research in general (Section 6).

2. Parameter estimation for a group of participants: complete pooling, no pooling, and partial pooling of data

A researcher who has collected data from a group of participants and seeks to estimate MPT model parameters for this group can choose among various options. The researcher could (a) fit the model to the data pooled across participants, (b) fit the model separately to each participant and then average the individually estimated parameters across participants, or (c) use a hierarchical approach that estimates parameters on the individual and the group level simultaneously. We refer to these three approaches to MPT parameter estimation as complete-pooling, no-pooling, and partial-pooling approaches, respectively.

2.1. Complete pooling

The conventional approach for parameter estimation in MPT modeling is complete pooling (e.g., Batchelder & Riefer, 1999; Hu & Batchelder, 1994). Here, the MPT model is fit to the response frequencies aggregated across participants. A crucial assumption of the complete-pooling approach is that the group of participants provides a set of independent and identically distributed (*i.i.d.*) observations; that is, potential heterogeneity among participants is not taken into account. To the extent that the *i.i.d.* assumptions hold (see Smith & Batchelder, 2008, for model-free tests of heterogeneity in categorical data) and there are no (or negligible) correlations among individual-level parameters, the complete-pooling approach can yield an accurate description of a group of participants. Complete pooling has been the default approach in MPT modeling for many years (see, e.g., the research reviewed in Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016).

However, if the *i.i.d.* assumptions do not hold and there is heterogeneity among participants, estimates of confidence intervals based on the complete-pooling approach may be too small

[1] Bias may be less problematic if it is constant across groups or conditions (e.g., van Ravenzwaaij, Donkin, & Vandekerckhove, 2017).

(Batchelder & Riefer, 1999; Klauer, 2006; Rouder & Lu, 2005), potentially leading to incorrect inferences about differences in parameters between different groups or experimental conditions. If, in addition, the model parameters are correlated across individuals, parameter estimates based on complete pooling may be biased (Erdfelder, 2000, Chapter 5; Klauer, 2010).

2.2. No pooling

In the no-pooling approach, estimates for the group of participants are obtained by fitting the model separately to each person, and then averaging the individual-level parameter estimates to characterize the group. One obvious requirement for the no-pooling approach to produce reliable parameter estimates is that sufficient numbers of observations per participant are available (Cohen, Sanborn, & Shiffrin, 2008). In practice, however, this is often difficult to achieve, because long testing sessions may not be feasible. In addition, some types of behavioral responses occur only rarely, leading to low frequencies in the respective response category. When there are only few observations, the individual-level parameters are likely to be inaccurate and have large variances. As a consequence, the no-pooling approach leads to results that diverge from those of the complete-pooling approach (Batchelder & Riefer, 1986), with the complete-pooling approach being more accurate in the majority of cases (Chechile, 2009). Accordingly, use of the no-pooling approach is rather uncommon in MPT modeling (but see, e.g., Meissner & Rothermund, 2013).

2.3. Partial pooling

Partial pooling (Klauer, 2010; Matzke et al., 2015; Smith & Batchelder, 2010) represents a compromise between complete pooling and no pooling. Heterogeneity across individuals is incorporated by estimating parameters for each individual; at the same time, the individual-level estimates are informed by the aggregate by assuming that the individual-level parameters are drawn from a group-level distribution (which is estimated simultaneously). With partial pooling, the model thus has a two-level, hierarchical structure. One consequence of this approach is so-called shrinkage, whereby individual estimates that seem unlikely given the overall distribution are "pulled in" toward the group average. Shrinkage corrects for measurement error and can thus lead to more reliable estimates (e.g., Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Partial pooling is an increasingly common approach in MPT modeling (e.g., Arnold, Bayen, & Böhm, 2014; Arnold, Bayen, Kuhlmann, & Vaterrodt, 2013; Arnold, Bayen, & Smith, 2015; Groß & Pachur, 2019; Horn et al., 2015, 2016; Kuhlmann et al., 2016; Michalkiewicz & Erdfelder, 2016). Heterogeneity among participants can be incorporated in the model in various ways.

2.3.1. Independent beta distributions

A common way to implement a partial-pooling approach is to represent the individual-level parameters as independent beta distributions whose parameters are, in turn, assumed to be drawn from beta distributions on the group level. This *beta-MPT approach* (Smith & Batchelder, 2010) thus allows for variability of parameters across individuals, but parameter correlations are not part of the model.

2.3.2. Multivariate normal distribution

The *latent-trait approach* (Klauer, 2010) explicitly models correlations between individual-level parameters. Here, individual parameters are represented as the individual's displacement from a group-level mean, $\Phi^{-1}(\theta_i) = \mu + \delta_i$, with δ_i drawn from a multivariate normal distribution (in probit space) with mean 0 and a variance–covariance matrix. That way correlations between parameters can be estimated more freely and more accurately than with independent beta distributions (Klauer, 2010; Smith & Batchelder, 2010). However, due to the variance–covariance matrix, more parameters need to be estimated for the latent-trait model than for the beta-MPT model.

2.4. Estimation approaches implemented in the available MPT modeling programs

As mentioned in Section 1, the different estimation approaches have been implemented in a number of convenient software tools. For complete pooling and no pooling, the program *multiTree* (Moshagen, 2010) and the R package *MPTinR* (Singmann & Kellen, 2013) are currently the most commonly used tools. In these packages, the approaches are implemented using maximum-likelihood parameter estimation (Hu & Batchelder, 1994). The two partial-pooling approaches, in contrast, are implemented in the R package *TreeBUGS* (Heck et al., 2018) within a Bayesian framework (reflecting how hierarchical approaches are typically applied). Thus, different statistical frameworks are used for the complete-pooling and no-pooling approaches, on the one hand, and the partial-pooling approach, on the other. As our focus is to provide a practical guide to estimating MPT model parameters with the available software, we do not strive to disentangle the role of the estimation approach from that of the underlying statistical framework. In principle, it is also possible to implement the complete-pooling and no-pooling approaches in a Bayesian framework (see, e.g., Singmann, Heck, Barth, Groß, & Kuhlmann, 2019). In simple cases, it may also be possible to use maximum-likelihood estimation for a hierarchical MPT model (see Batchelder & Smith, 2010, for a discussion). However, we are not aware of MPT research that has done so.

2.5. The present analyses

As outlined in Section 2.1 to 2.3, the various approaches to MPT parameter estimation are based on different assumptions and have specific requirements. In terms of accuracy and precision, the partial-pooling approaches should be superior to other approaches due to their ability to correct for measurement error and to fully utilize the information available, even if this information is sparse. In principle, the latent-trait approach may be the most powerful approach as it also explicitly models parameter covariation. However, the number of covariance parameters in a latent-trait model increases exponentially with an increasing number of model parameters; therefore, even more data points on the individual level may be needed to harness this theoretical advantage in practice.

To date, the different approaches to estimating MPT model parameters have not been systematically compared. Previous parameter recovery analyses for MPT models focused on one estimation approach only (e.g., complete pooling; Hilbig, Erdfelder, & Pohl, 2010; or partial pooling, Klauer, 2010) and one MPT model only (Klauer, 2010; Matzke et al., 2015). Moreover, the analyses were conducted for MPT models with a relatively simple structure and few parameters (Hilbig et al., 2010; Klauer, 2010; Matzke et al., 2015). By contrast, we conducted parameter recovery analyses for four estimation approaches that differ in the degree of pooling and the modeling of parameter heterogeneity and individual-level parameter correlations. Further, we examined the results across different MPT models, including both a relatively simple MPT model and a more complex MPT model with a large number of parameters (some "core", some "auxiliary"). Specifically, we considered the MPT model of source monitoring (Bayen et al., 1996) and the MPT model of hindsight bias (Erdfelder & Buchner, 1998).

Whenever we observed problems in the recovery of core parameters, we doubled the number of participants and/or items to test whether collecting a greater (but still realistic) amount of data would help alleviate these problems. The results may thus be informative for researchers considering whether it might be worthwhile to increase the number of items or participants (see also Cohen et al., 2008).

In addition, we compared the results under parameter homogeneity—that is, assuming the same parameter for all participants—with the results under parameter heterogeneity, where individuals' parameters varied around a common mean. For the latter, we also examined the impact of correlations between the model parameters. We relied on empirically informed magnitudes of heterogeneity and parameter correlations.

Note that our analyses do not gauge how appropriate the different approaches are for characterizing the data in each experiment. Each estimation result represents a valid inference about the data for the specific estimation method. Instead, our goal is to evaluate how well the ground truth, which is only partly reflected in the data (which represents a sample of the ground truth), can be recovered given a specific amount of data.

As our examination focuses on empirically realistic situations, not all of our conclusions will necessarily generalize beyond these situations. General conclusions would require an exhaustive investigation of different combinations of parameter values, variances, and covariances, which is beyond the scope of this article.

In comparing the different estimation approaches, we focused on the recovery of those parameters that represent meaningful and theoretically interesting psychological processes (i.e., "core" parameters), as these are the ones on which insights from MPT modeling are based (e.g., concerning differences between experimental conditions or groups). In addition to these core parameters, some MPT models contain additional, "auxiliary" parameters. These are required for the architecture of the model but are usually not interpreted (or even reported). We base our conclusions on how well the different approaches are able to estimate the core parameters. We next describe the two MPT models in more detail before going on to present the recovery analyses.

3. Two MPT models

3.1. The MPT model of source monitoring

Source monitoring refers to judgments about the source of information (Johnson, Hashtroudi, & Lindsay, 1993). A typical source-monitoring experiment has two phases. In a learning phase, participants are presented with a number of items (e.g., words or sentences) from two different sources (source A and source B; e.g., two persons, lists, or modalities). In a subsequent recognition phase, participants are asked to identify items as having previously been presented by source A, by source B, or as new items.

Bayen et al. (1996) developed a multinomial model of source monitoring that disentangles the probabilities of item detection, source memory, and different forms of guessing (for related models, see Batchelder & Riefer, 1990; Batchelder, Riefer, & Hu, 1994; Klauer & Wegener, 1998; Meiser & Bröder, 2002). The model is shown in the left panel of Fig. 2. All parameters of source monitoring are considered "core" parameters, as each can be of theoretical interest in a source-monitoring study.

For items from source A, participants detect an item as old with probability D_1 (and fail to detect an item as old with probability 1 - D_1). Given item detection, participants correctly discriminate the source of an item (here: A) with probability d_1. If source discrimination fails (with probability 1 - d_1), participants guess that an item came from source A with probability a. Given an item-detection failure (with probability 1 - D_1), participants guess with probability b that an item is old. Given successful guessing (with probability b), participants guess with probability g that it came from source A.

The model has separate parameters for the detection of items from source A (D_1), items from source B (D_2), and new items (D_3). The model also distinguishes between memory for source A (d_1) and memory for source B (d_2). As there are eight parameters but only six response categories, restrictions are needed to make the model identifiable. Depending on the restrictions imposed, different submodels can be derived. The right panel of Fig. 2 shows identifiable submodels (Bayen et al., 1996). For example, submodel 4 assumes equal item-detection probabilities for source A, source B, and new items ($D_1 = D_2 = D_3$), equal source-discrimination probabilities for source A and source B ($d_1 = d_2$), and equal guessing probabilities for source A and source B ($a = g$). Which submodel is appropriate for a given application depends on the purpose of the research (e.g., testing for differences in source memory between the two sources or testing for differences in source guessing with vs. without item recognition) and on whether the model adequately describes the data.

In a series of experiments, Bayen et al. (1996) demonstrated the validity of the parameters by means of selective-influence manipulations. The source-monitoring model has been used to address various research questions (e.g., how schematic knowledge influences source-monitoring parameters; Bayen, Nakamura, Dupuis, & Yang, 2000; Spaniol & Bayen, 2002) in various populations, including clinical patients (Keefe, Arnold, Bayen, & Harvey, 1999) and older adults (Kuhlmann et al., 2016).

3.2. The MPT model of hindsight bias

Hindsight bias is a common phenomenon when people retrospectively assess their knowledge (Blank, Musch, & Pohl, 2007; Roese & Vohs, 2012). Specifically, people recall their prior judgments of facts (e.g., country populations, historical dates) or outcomes of events (e.g., football matches, historical events) with bias towards the true facts or outcomes that they have ascertained in the meantime. In the typical hindsight-bias paradigm, participants are first asked to provide numerical judgments (original judgments, OJs) in response to difficult questions (e.g., "In what year was Antonio Vivaldi born?"). After a retention interval, they are asked to recall their OJs (recall of original judgments, ROJs). Before or at recall, the correct answer (correct judgment, CJ; here: 1678) is presented for some of the questions (experimental items) but not for others (control items).

Two separate processes are assumed to underlie the hindsight-bias phenomenon (Erdfelder & Buchner, 1998). *Recollection bias* occurs when presentation of the CJ impairs recollection of the OJ, such that OJs of experimental items are recalled with a lower probability than OJs of control items. *Reconstruction bias* occurs when recall of the OJ fails and reconstruction of the OJ is biased towards the revealed CJ. Both biases are represented by parameters in Erdfelder and Buchner's (1998) MPT model of hindsight bias. Participants' responses (e.g., OJ: 1750, ROJ: 1720) are assigned to discrete response categories representing the rank order of OJ, CJ, and ROJ (e.g., the category 'CJ < ROJ < OJ'). Allowing for ties between ROJ and CJ, there are 10 possible rank order categories for experimental items and 10 for control items. The model includes 13 parameters, each reflecting a different

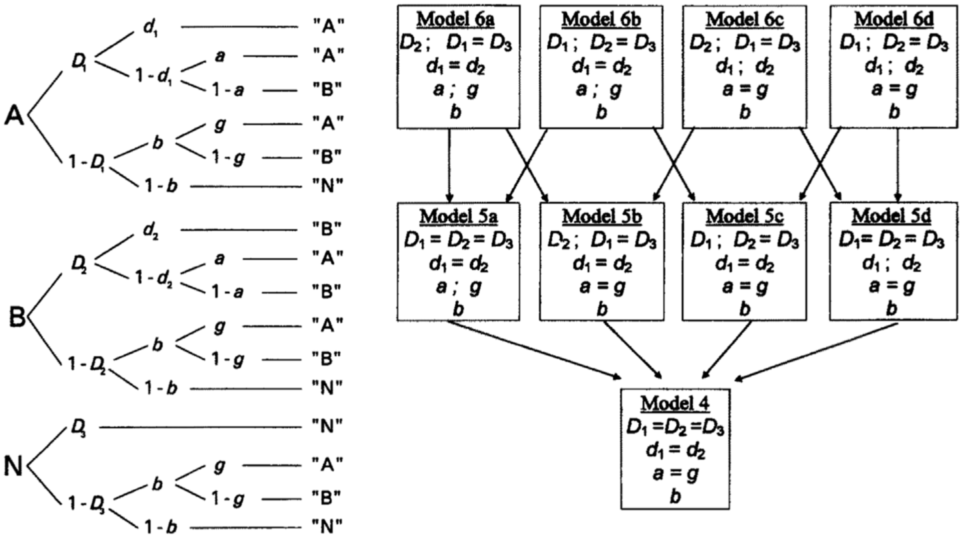

Fig. 2. Left panel: The multinomial model of source monitoring by Bayen et al. (1996). Right panel: Identifiable submodels of the source monitoring model. A = Source A item; B = Source B item; N = distractor item; D_1 = probability of detecting an item from Source A; D_2 = probability of detecting an item from Source B; D_3 = probability of detecting that a distractor is new; d_1 = probability of correctly discriminating the source of an item from Source A; d_2 = probability of correctly discriminating the source of an item from Source B; a = probability of guessing that a detected item is from Source A; b = probability of guessing that an item is old; g = probability of guessing that an undetected item is from Source A. From "Source discrimination, item detection, and multinomial models of source monitoring" by Bayen et al. (1996), *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, p. 202 (https://doi.org/10.1037//0278-7393.22.1.197). © 1996 by the American Psychological Association.

underlying cognitive process; it is therefore also termed HB13 (for a detailed model description, see Erdfelder & Buchner, 1998). For the purposes of the present article, it is sufficient to be familiar with the main assumptions of the model, as described in Fig. 3.

The three core parameters of the HB13 model are the following. For control items (i.e., for which the CJ is not revealed), participants recollect their OJ with probability r_C. If recollection fails (i.e., with probability 1 – r_C), reconstruction of the OJ is unbiased. For experimental items (i.e., for which the CJ is revealed), in contrast, participants recollect their OJ with probability r_E. If recollection of the OJ is impaired by the presentation of the CJ, then $r_C > r_E$ (recollection bias). If recollection fails (i.e., with probability 1 – r_E), reconstruction of the OJ is biased toward the CJ with probability b (reconstruction bias). With probability 1 – b, reconstruction is unbiased.[2]

In a series of experiments, Erdfelder and Buchner (1998) demonstrated parameter validity by means of selective-influence manipulations. Like the source-monitoring model, the hindsight-bias model has been applied to study young adults (e.g., Dehn & Erdfelder, 1998; Erdfelder, Brandt, & Bröder, 2007), older adults and children (Coolin et al., 2015; Groß & Bayen, 2015; Groß & Pachur, 2019; Pohl, Bayen, Arnold, Auer, & Martin, 2018; Pohl et al., 2010), as well as clinical populations (Groß & Bayen, 2017b; Ruoß & Becker, 2001). In Sections 4 and 5, we describe the parameter recovery simulations for the two models.

4. Simulations 1–3: source-monitoring model

The simulations with the source-monitoring model were conducted as follows. In a first step we generated 100 synthetic data sets, each representing a source-monitoring experiment with 36 participants, each responding to 96 items (32 items in each of the three conditions: source A, source B, new items). As a basis for the data generation (i.e., setting of realistic levels of the parameter values, level of heterogeneity among participants, and parameter correlations), we used a data set by Schaper, Kuhlmann, & Bayen, 2019, Exp. 1).[3] An overview of all parameter values as well as variances and covariances can be found in Appendix A. The data in the simulations were generated with TreeBUGS (Heck et al., 2018). The code is available in the Supplemental Material.

As mentioned in Section 3.1, there are different submodels for the source-monitoring MPT model, suited for different research questions or experimental settings. Most studies have applied either submodel 4 (the most parsimonious model) or submodel 5d (Schaper & Kuhlmann, 2019), with the latter containing two source-memory parameters, one for source A (d_1) and one for source B (d_2). Submodel 5d is called for when source-memory probabilities are expected to differ (e.g., for expected vs. unexpected item–source pairings, as in the reference data set; Schaper et al., 2019).

We applied four estimation approaches to the synthetic data sets: the complete-pooling (CP) and the no-pooling (NP) method (implemented in the MPTinR package; Singmann & Kellen, 2013), as well as the two Bayesian partial-pooling (PP) approaches: the beta-MPT (PP-B) and the latent-trait MPT (PP-LT) method (implemented in the TreeBUGS package; Heck et al., 2018). For the latter, we used the default priors implemented in TreeBUGS as prior distributions for the group-level parameters. Specifically, for the beta-MPT method, the shape parameters of the beta distributions, α and β, were defined as gamma distributions with shape 1 and rate 0.1, truncated at 1 (the truncation makes sampling more stable, but otherwise has no effects). For the latent-trait method, the priors for the group-level parameters were standard normal distributions ($\mu = 0$ and $\sigma^2 = 1$), implying uniform distributions in probability space. Samples were drawn until Gelman–Rubin statistic < 1.1 and number of effective samples > 500 for all core parameters.[4]

[2] In some applications, the parameter c has also been considered a "core" parameter. It denotes the probability that the CJ is adopted as the OJ (i.e., it represents the most extreme case of biased reconstruction) and is typically very small, except in young children (e.g., Pohl, Bayen, & Martin, 2010).

[3] We thank Marie Luisa Schaper, Beatrice Kuhlmann, and Ute J. Bayen for sharing their data.

[4] For each simulation, we initially ran three chains with 300,000 posterior samples (120,000 were discarded as burn-in) and a thinning factor of 10; if necessary, additional samples were drawn until the convergence criteria were reached.

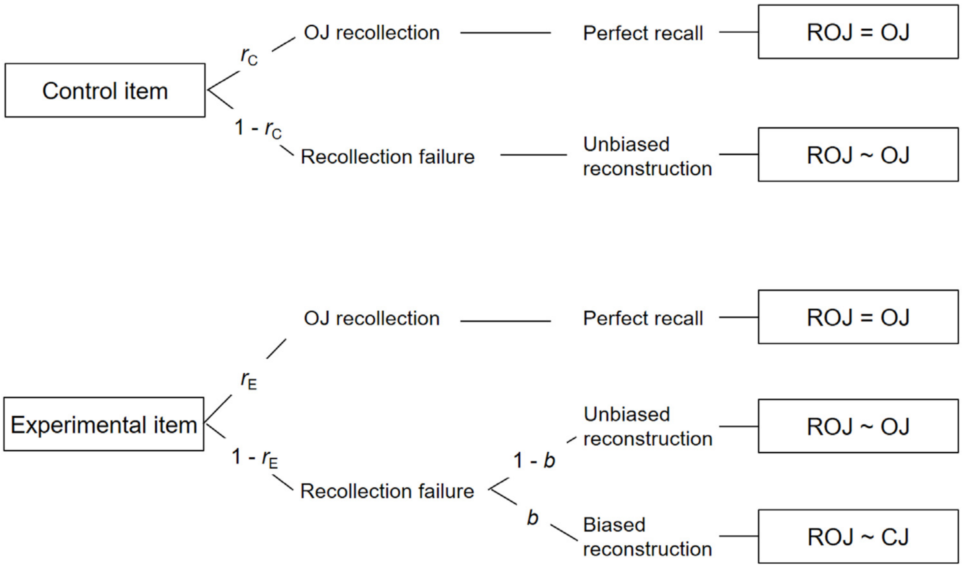

Fig. 3. Main assumptions of the multinomial model of hindsight bias by Erdfelder and Buchner (1998). Rectangles represent observable events. r_C and r_E =OJ recollection probabilities for control and experimental items, respectively; b = probability of a biased reconstruction given a failure to recollect the OJ; OJ = original judgment; ROJ = recall of original judgment; CJ = correct judgment. r_C > r_E indicates a recollection bias.

Our conclusions about the performance of the estimation approaches in recovering the generating parameters were based on the distribution of estimates obtained for the group of participants in each experiment. These estimates were determined as follows: For the complete-pooling approach, we aggregated the data across participants and then applied the approach to obtain a group estimate. For the no-pooling method, we estimated a model for each individual and then averaged across the individual parameter estimates in each experiment. Finally, for the partial-pooling approach (which simultaneously estimates individual-level and group-level parameters for each experiment), we used the means of the group-level posterior distribution.[5]

We assessed parameter recovery performance for each parameter using two measures. First, as an index of bias in the parameter estimation, we determined the average deviation, Δ, of the estimated from the true parameter values across the simulated experiments. Negative values of Δ indicate underestimation of the true parameter value; positive values of Δ indicate overestimation (where it is helpful, we also report the absolute deviation $|\Delta|$ in the text). Second, as an index of imprecision, we determined the standard deviation (*SD*) of parameter estimates.

Table 1 gives an overview of the simulations for the source-monitoring model. Simulation 1 examined the estimation approaches for submodel 4. In Simulation 1a, we assumed homogeneity among participants (and uncorrelated parameters).[6] In Simulation 1b, we introduced parameter heterogeneity; in Simulation 1c, we additionally introduced correlations among individual-level parameters. In Simulation 2, we examined the estimation approaches for submodel 5d. In Simulation 2a, we assumed homogeneity among participants (and uncorrelated parameters). In Simulation 2b, we introduced parameter heterogeneity; in Simulation 2c, we additionally introduced parameter correlations. In Simulation 3, we examined the estimation approaches for submodel 5d with twice as many participants and items ($N_P = 72$ and $N_I = 192$).

4.1. Overview of results

Before presenting the results of the individual simulations, we first summarize the findings that emerged consistently across the simulations. For submodel 4, all estimation approaches led to very good recovery performance, and this held for both parameter heterogeneity and parameter correlations. Overall, the partial-pooling approach with explicit modeling of parameter correlations (PP-LT) produced the most accurate results, but the other approaches were also very accurate.

The results for submodel 5d were different. Here, all estimation approaches led to very good recovery performance for the majority of parameters ($D_1 = D_2 = D_3, a = g, b$). However, all estimation approaches produced rather imprecise estimates of parameter d_1 (source memory for source A; $SD \leq .16$), and all estimation approaches severely overestimated d_1 under parameter heterogeneity ($\Delta \leq .18$). With twice the number of observations, this overestimation was reduced only for PP-LT (the other approaches remained inaccurate). In the following, we describe the results in more detail. The main results of Simulations 1–3 are displayed in Fig. 4 (submodel 4) and 5 (submodel 5d), which show the distribution of group estimates across the experiments in each simulation. A detailed description of recovery results (Δ and *SD*) for all model parameters can be found in Appendix B.

4.2. Simulation 1 (submodel 4): Homogeneity/heterogeneity among participants, (un)correlated parameters

In Simulation 1a, which assumed homogeneity across participants, all estimation approaches recovered all model parameters with very high accuracy, showing no or negligible bias ($|\Delta| \leq .01$) and negligible imprecision ($SD \leq .02$) (see also the solid lines in Fig. 4). In empirical data, however, it is often more realistic to assume parameter heterogeneity. Does this affect the results? The results of Simulation 1b (in which the data were generated with heterogeneity) and Simulation 1c (which additionally introduced parameter correlations) were very similar to each other; therefore, Fig. 4 shows the results for Simulation 1b only (dotted lines). As can be seen, all approaches continued to perform very well under heterogeneity and parameter correlations. For parameter b (guessing "old"), there was moderate overestimation and imprecision across the CP, NP, and PP-B approaches ($\Delta \leq .07$; $SD \leq .05$); only the estimates of PP-LT remained unbiased ($|\Delta| \leq .01$).

[5] To convey the uncertainty in the posterior estimates, we additionally report the average upper and lower bounds of the 95% highest density intervals of the posterior distributions across experiments for each simulation in the Supplemental Material (Tables S1 and S2).

[6] Note that even with homogeneity, the distribution of the frequencies across response categories will vary somewhat across experiments due to sampling error.

Table 1
Overview of model recovery analyses for the source-monitoring model.

	Simulation 1			Simulation 2			Simulation 3
	1a	1b	1c	2a	2b	2c	3
Submodel		SM-4			SM-5d		SM-5d
Parameter heterogeneity	No	Yes	Yes	No	Yes	Yes	Yes
Correlations between parameters	No	No	Yes	No	No	Yes	No
Number of participants N_P		36			36		72
Number of items N_I(A, B, N)		96 (32, 32, 32)			96 (32, 32, 32)		192 (64, 64, 64)
True parameter values		$D_1 = .71$ $d_1 = .53$ $a = g = .64$ $b = .23$			$D_1 = .55$ $d_1 = .20$; $d_2 = .75$ $a = g = .75$ $b = .24$		$D_1 = .55$ $d_1 = .20$; $d_2 = .75$ $a = g = .75$ $b = .24$

Note. We generated 100 synthetic data sets (each representing a source-memory experiment) for each simulation condition. SM = Source monitoring, N_P = number of participants, N_I = number of items, A = Source A items, B = source B items, N = distractor items.

4.3. Simulation 2 (submodel 5d): Homogeneity/heterogeneity among participants, (un)correlated parameters

In Simulation 2, we examined parameter recovery for submodel 5d, which has separate parameters for memory for source A (d_1) and source B (d_2). All estimation approaches except NP performed similarly well under parameter homogeneity (Simulation 2a) for the majority of parameters ($D_1 = D_2 = D_3$, $a = g$, b, d_2). The NP approach underestimated the source-guessing parameter, a ($\Delta = -.06$), and one of the two source-memory parameters, d_2 ($\Delta = -.08$). For the other source-memory parameter, d_1, all estimation approaches led to imprecise estimates under parameter homogeneity (Simulation 2a), with *SD* ranging from .06 (NP) to .15 (CP). Under heterogeneity (with or without correlations; Simulations 2b and 2c, respectively), all estimation approaches additionally led to severely biased estimates of d_1, with overestimation of up to .18 (PP-LT).

4.4. Simulation 3 (submodel 5d): Number of participants and items

To investigate whether increasing the number of observations would help alleviate the accuracy and precision problems observed for the estimation of parameter d_1, in Simulation 3 we doubled the number of participants and items relative to Simulation 2 (from $N_\mathrm{P} = 36$ and $N_\mathrm{I} = 96$ to $N_\mathrm{P} = 72$ and $N_\mathrm{I} = 192$; see also solid lines in Fig. 5), assuming heterogeneity (as in Simulation 2b; shown for comparison as gray lines in Fig. 5). For all estimation approaches and all parameters, this led to an increase in precision for the estimation of parameter d_1 ($.05 \leq SD \leq .11$). Overall, the PP-LT approach performed best. However, while PP-LT produced the smallest bias for d_1 ($\Delta = .04$), precision for d_1 remained relatively poor ($SD = .09$). By contrast, the NP approach produced the largest bias ($\Delta = .16$), but the level of precision was tolerable ($SD = .05$).

4.5. Discussion

Our first set of simulations compared the ability of the different estimation approaches (complete pooling, no pooling, and partial pooling) to recover core parameters of the source-monitoring model, specifically submodels 4 and 5d. We examined the results under homogeneity versus heterogeneity and for correlated versus uncorrelated individual-level parameters; we also tested whether difficulties in parameter recovery could be alleviated by increasing the numbers of participants and items.

For both submodels, PP-LT proved to be the most accurate estimation approach; this held for standard amounts of data as well as for a higher (but still realistic) amount of data. Recovery performance for submodel 4 was generally very good, irrespective of whether or not there was heterogeneity across participants and whether or not parameters were correlated. One exception is that parameter b (guessing "old") was overestimated when there was heterogeneity among participants; only the PP-LT approach estimated the parameter without bias. As parameter heterogeneity is probably the rule rather than the exception in empirical data, this result is highly relevant for practical applications of the source-monitoring MPT model.

Recovery for submodel 5d showed a high level of imprecision in the estimation of the d_1 parameter. This imprecision occurred with all estimation approaches with the exception of NP, where imprecision was only moderate; however, here the estimation of d_1 was substantially biased. When there was heterogeneity across participants, problems associated with the estimation of d_1 were amplified—we observed severe overestimation on top of imprecision for all estimation approaches. We address the structure of submodel 5d (and associated parameter interdependence) as one possible factor contributing to these problems in Section 6. Importantly, the use of PP-LT can substantially improve estimation of d_1, but only with rather large numbers of participants and items. For researchers who want to apply submodel 5d, it is thus important to collect a sufficient amount of data to keep overestimation of d_1 at a reasonable level. With the other approaches, overestimation of d_1 is likely to persist. Overall, to enable the detection of differences in source memory in the presence of parameter heterogeneity, researchers must be sure to use an experimental design providing a large amount of data.

In sum, for both submodels of the source-monitoring model, PP-LT emerged to be the best-suited approach. Next, we examine whether the same holds for the MPT model for hindsight bias, which has 13 parameters and is thus more complex.

5. Simulations 4–6: Hindsight-bias model

The simulations with the hindsight-bias model were conducted as follows. In a first step we generated 100 synthetic data sets, each representing a hindsight-bias experiment with 47 participants, each responding to 48 items (24 control and 24 experimental items).[7] For data generation, we used parameter values typically observed for young adults in previous research (parameters: b = .30, r_C = .38, r_E = .36; e.g., Bayen, Erdfelder, Bearden, & Lozito, 2006; Groß & Bayen, 2015, 2017b). For these simulations, we used variances and covariances from a data set by Groß and Bayen (2017a); see Appendix A for a list of all 13 generating parameter values, variances, and covariances.

Table 2 gives an overview of the simulations for the hindsight-bias model. Simulation 4a examined the estimation approaches

[7] The number of participants and items for Simulations 4 and 5 are taken from a study by Groß and Bayen (2017a); these numbers were doubled for Simulation 6.

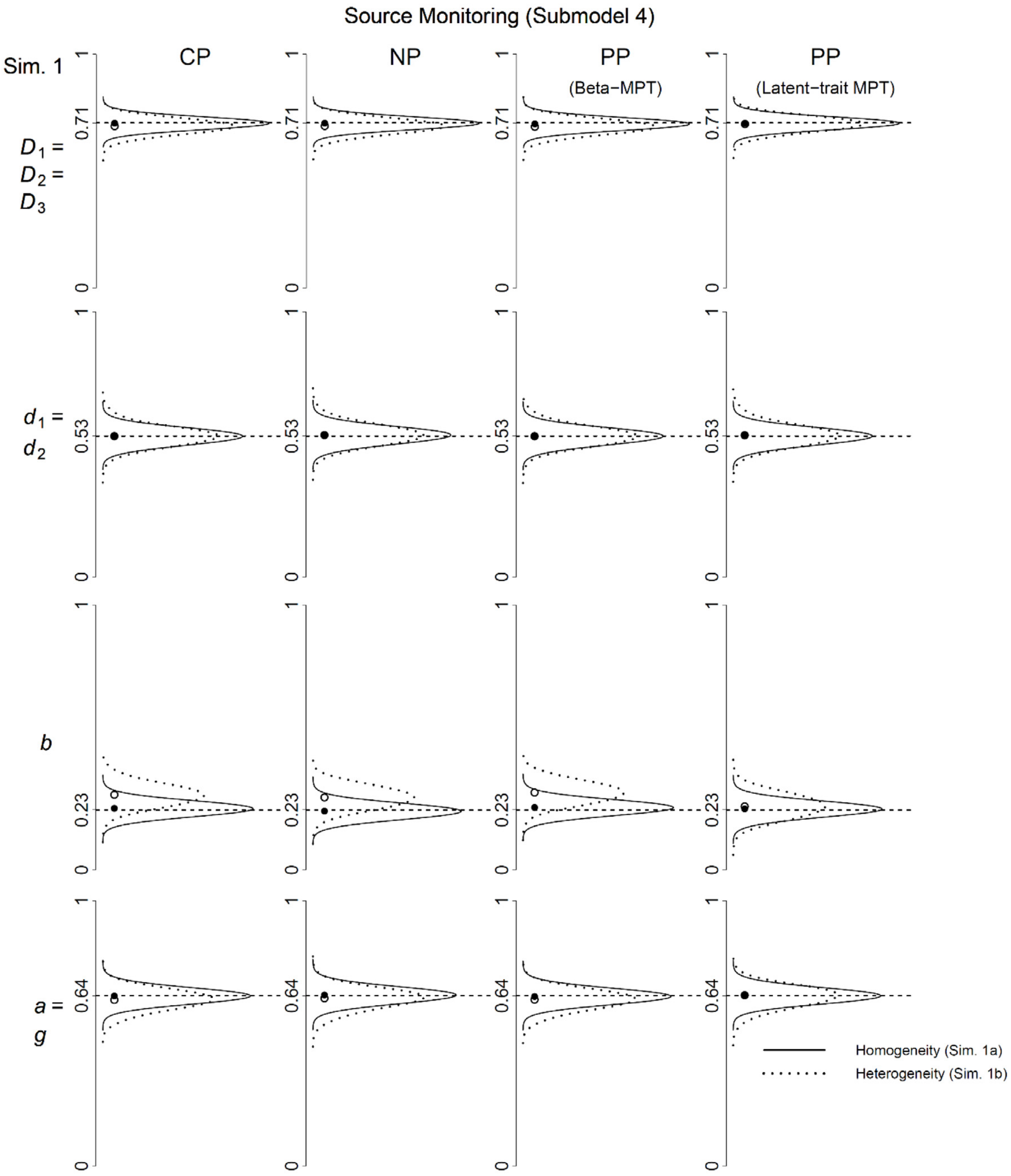

Fig. 4. Results of the parameter recovery for the parameters of submodel 4 of the source-monitoring model. The plots show the distributions of the group estimates across the simulated experiments, with circles representing the median of the distributions. CP = complete pooling; NP = no pooling; PP = partial pooling. For CP, there was one group estimate for each experiment; for NP, there were estimates for each individual, which we averaged for each experiment; for the two PP approaches, there was one group-level estimate (the posterior mean) for each experiment. D_1 = probability of detecting an item from Source A; D_2 = probability of detecting an item from Source B; D_3 = probability of detecting that a distractor is new; d_1 = probability of correctly discriminating the source of an item from Source A; d_2 = probability of correctly discriminating the source of an item from Source B; a = probability of guessing that a detected item is from Source A; b = probability of guessing that an item is old; g = probability of guessing that an undetected item is from Source A. Results are shown as a function of parameter heterogeneity.

assuming homogeneity among participants. In Simulations 4b and 4c, we introduced parameter heterogeneity and parameter correlations.

Ideally, an estimation approach should be equally accurate and precise across the entire parameter space. In Simulation 5, we therefore tested the robustness of conclusions across different values of the core parameters (b, r_C, r_E); here, a constant was either subtracted from or added to these values (see Section 5.3 for details). In Simulation 6, we repeated the analyses with twice as many items and/or participants.

5.1. Overview of results

Before presenting the results of the individual simulations, we summarize the findings for the core parameters of the hindsight-bias model (i.e., r_C, r_E and b) that emerged consistently

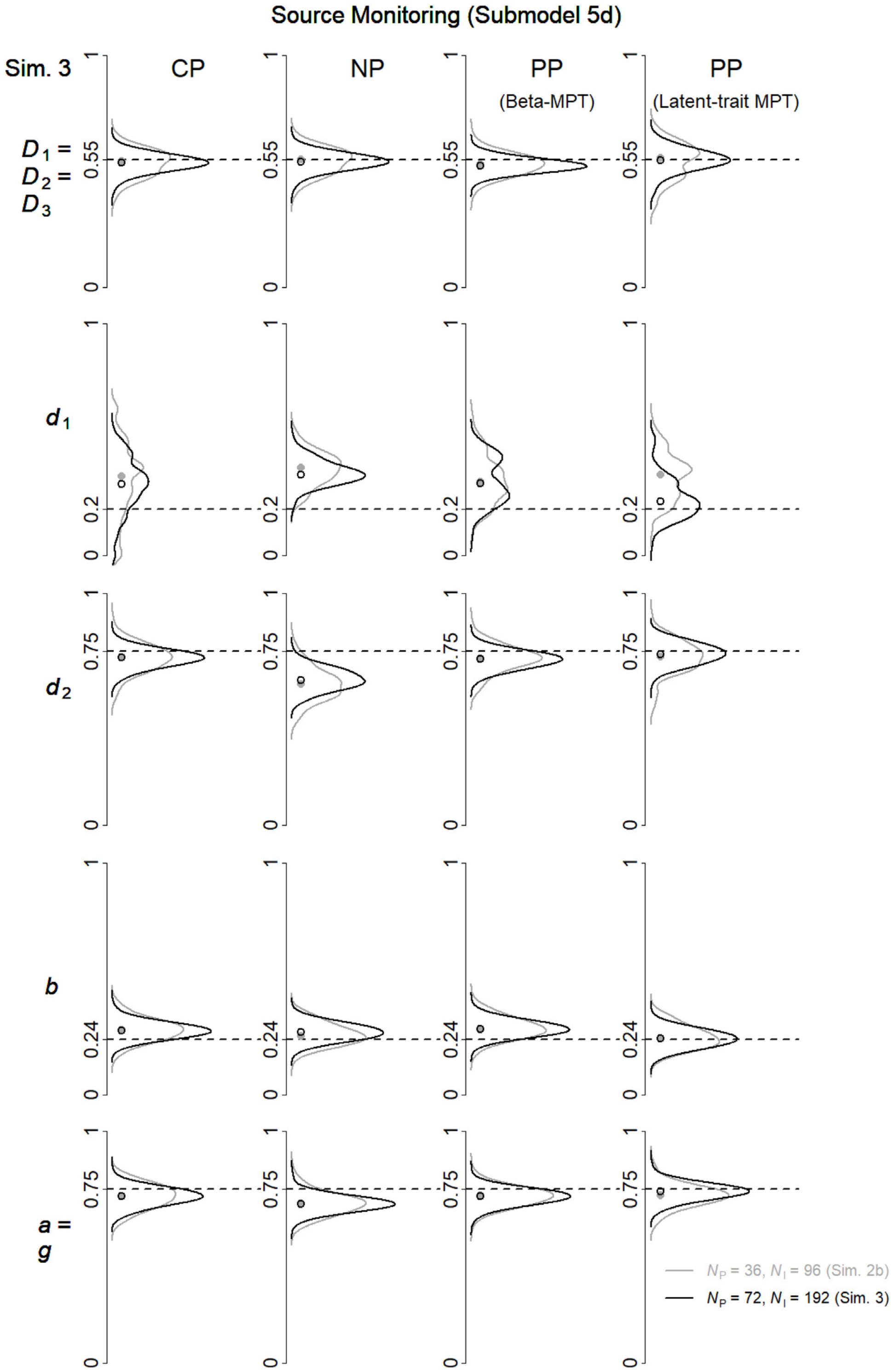

Fig. 5. Results of the parameter recovery for the parameters of submodel 5d of the source-monitoring model. The plots show the distributions of the group estimates across the simulated experiments, with circles representing the median of the distributions. CP = complete pooling; NP = no pooling; PP = partial pooling. D_1 = probability of detecting an item from Source A; D_2 = probability of detecting an item from Source B; D_3 = probability of detecting that a distractor is new; d_1 = probability of correctly discriminating the source of an item from Source A; d_2 = probability of correctly discriminating the source of an item from Source B; a = probability of guessing that a detected item is from Source A; b = probability of guessing that an item is old; g = probability of guessing that an undetected item is from Source A; N_P = number of participants; N_I = number of items. Results are shown as a function of number of observations.

across simulations. For the recollection parameters r_C and r_E, all estimation approaches showed very good recovery performance, with $|\Delta| \leq .03$ (bias) and $SD \leq .03$ (imprecision), with PP-LT showing the best results. For the reconstruction bias parameter b, by contrast, PP-B was the best approach, whereas NP proved particularly problematic due to substantial overestimation.

The key results of Simulations 4–6 are shown in Figs. 6 and 7, which display the distribution of group estimates across the simulated experiments in selected simulations. In addition, all recovery results (Δ and SD) for r_C, r_E, and b are reported in Appendix B. In the following, we describe the results of Simulations 4–6 in more detail.

Table 2
Overview of model recovery analyses for the hindsight-bias model.

	Simulation 4			Simulation 5			Simulation 6		
	4a	4b	4c	5a–5i			6a	6b	6c
Parameter heterogeneity	No	Yes	Yes		Yes			Yes	
Correlations between parameters	No	No	Yes		No			No	
Number of participants N_P		47			47		47	94	94
Number of items N_I		48			48		96	48	96
(control, experimental)		(24, 24)			(24, 24)		(48, 48)	(24, 24)	(48, 48)
True parameter values		b = .30		b = .15	r_C = .15	r_E = .13		b = .30	
		r_C = .38		b = .30	r_C = .38	r_E = .36		r_C = .38	
		r_E = .36		b = .55	r_C = .60	r_E = .58		r_E = .36	

Note. We generated 100 synthetic data sets (each representing a hindsight-bias experiment) for each simulation condition. N_P = number of participants, N_I = number of items.

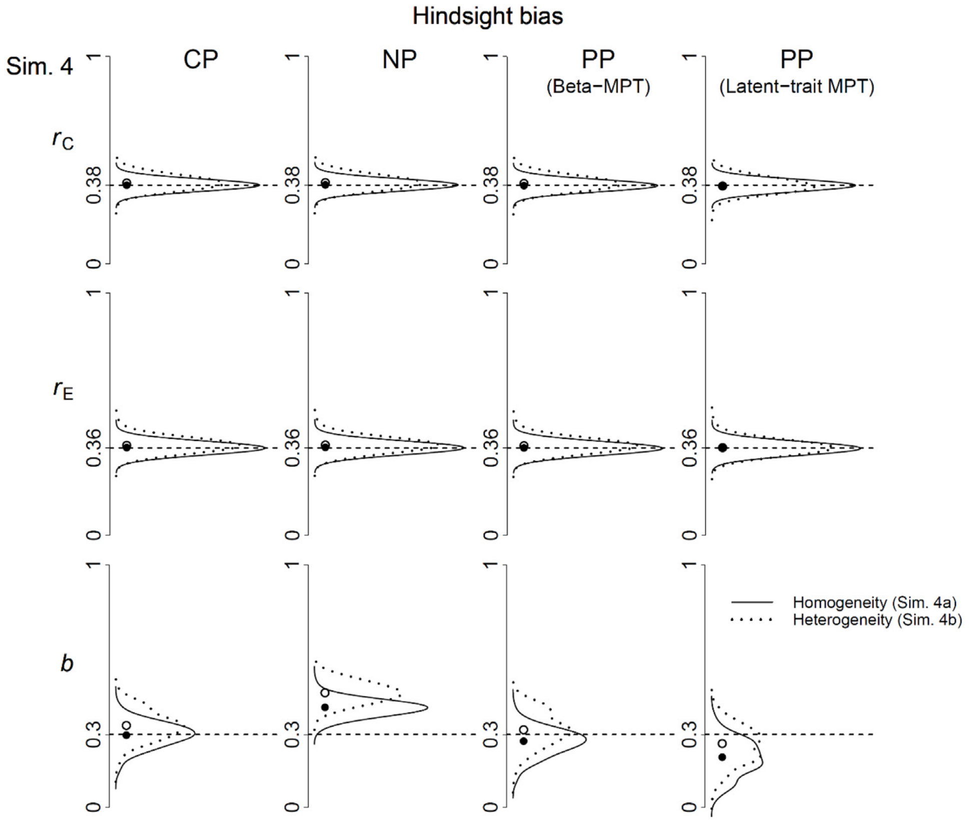

Fig. 6. Results of the parameter recovery for the core parameters of the hindsight-bias model, b, r_C, and r_E, respectively. The plots show the distributions of the group estimates across the simulated experiments, with circles representing the median of the distributions. CP = complete pooling; NP = no pooling; PP = partial pooling; r_C = probability of a recollection of the original judgment for control items; r_E = probability of a recollection of the original judgment for experimental items; b = probability of a biased reconstruction given a failure to recollect the original judgment. Results are shown as a function of parameter heterogeneity.

5.2. Simulation 4: Homogeneity/heterogeneity among participants, (un)correlated parameters

In Simulation 4a, the data were generated assuming homogeneity among participants. For the recollection parameters r_C and r_E, recovery performance was excellent: Irrespective of the estimation approach, there was no bias (all Δ = .00) and high precision (all $SD \leq .02$). For the reconstruction bias parameter b, however, there were notable differences between the estimation approaches. With NP, there was substantial overestimation, with a mean of Δ = .11. With PP-LT, there was substantial underestimation, with a mean of $\Delta = -.10$. Only CP (Δ = .00) and PP-B ($\Delta = -.03$) produced no (or negligible) bias.

In Simulations 4b and 4c, the data were generated with heterogeneity among participants and either without parameter correlations (Simulation 4b) or with parameter correlations (Simulation 4c). As the results for Simulations 4b and 4c were very similar to each other, Fig. 6 shows the results for Simulation 4b only (dotted lines). As in Simulation 4a, r_C and r_E were recovered very well in both simulations: Across all estimation approaches, there was very little bias ($|\Delta| \leq .01$) and little imprecision ($SD \leq .03$). Further, both PP-LT and NP again had difficulty recovering parameter b. PP-LT again produced underestimation, with $\Delta = -.05$ for both Simulations 4b and 4c—much smaller than in Simulation 4a. NP again produced overestimation, with Δs = .17 for both simulations—even larger than in Simulation 4a. CP (with Δs = .04) was more accurate, and PP-B (with Δs = .02) showed the best results.

5.3. Simulation 5: High versus low parameter values

If the amount of bias or precision varies across the parameter space, conclusions about differences between groups or conditions may be compromised (e.g., because the parameters of

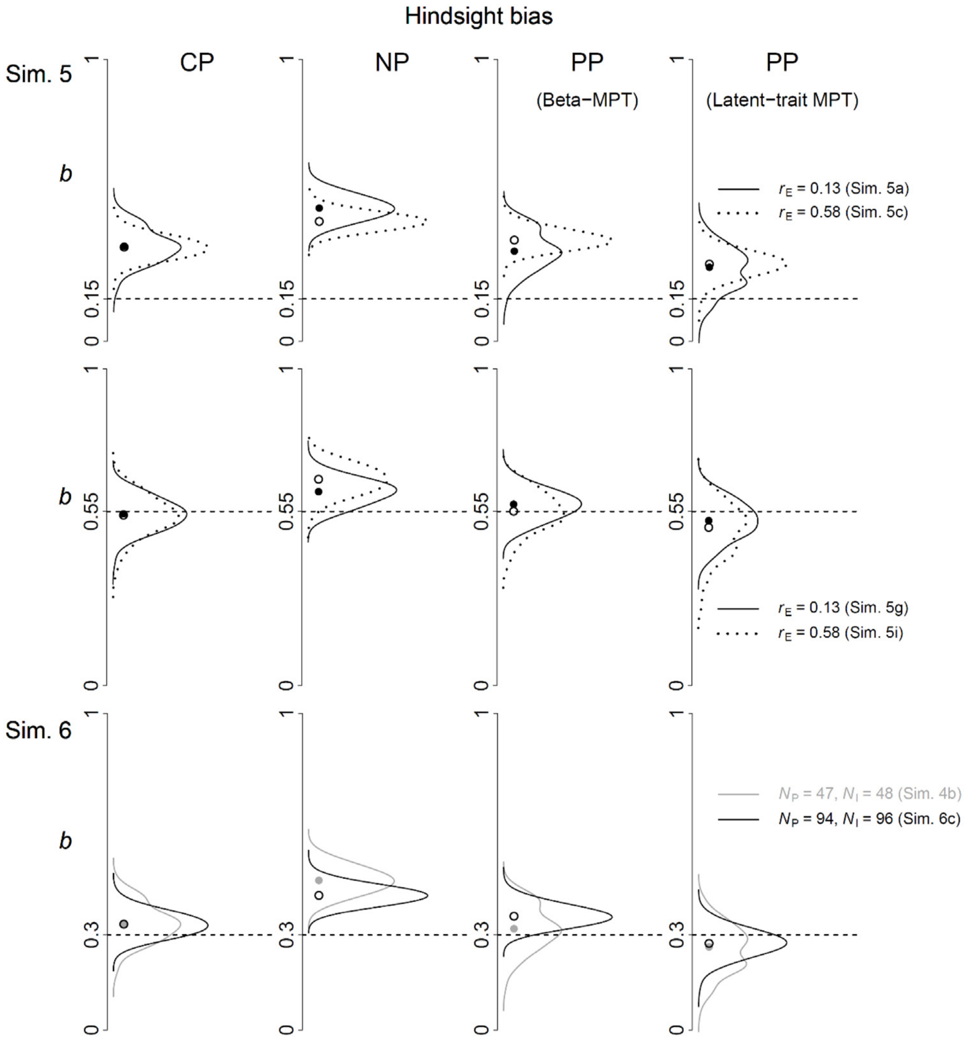

Fig. 7. Results of the parameter recovery for parameter *b* of the hindsight-bias model. CP = complete pooling; NP = no pooling; PP = partial pooling; r_C = probability of a recollection of the original judgment for control items; r_E = probability of a recollection of the original judgment for experimental items; *b* = probability of a biased reconstruction given a failure to recollect the original judgment; N_P = number of participants; N_I = number of items. Results are shown as a function of level of r_E for a low value of *b* (upper row) and a high value of *b* (middle row). Results are shown as a function of number of participants and items (lower row).

individuals with a low value on a parameter are measured less reliably than those of individuals with a high value). Simulation 5 therefore compared recovery performance for different levels of the parameter values (with the amount of heterogeneity among participants kept constant). Here, we varied the magnitude of the actual value of the parameters (i.e., that generated the data) and used low, medium, and high levels of the reconstruction-bias parameter *b* (b = .15, .30, or .55) and low, medium, and high values of both of the recollection parameters (r_C = .15 and r_E = .13, r_C = .38 and r_E = .36, r_C = .60 and r_E = .58, respectively). Combining each of these three levels yielded a total of nine simulations (5a–5i; see also Appendix B).

For the r_C and r_E parameters, all estimation methods again produced little bias ($|\Delta| \leq .03$) and little imprecision (*SD* < .03). This held across all value levels of r_C and r_E, with the exception that CP, NP, and PP-B showed slight overestimation for low values (.15, .13) of these parameters (up to Δ = .03). Only the PP-LT approach showed no bias at all (Δ = .00).

The results for the parameter *b* are shown in Fig. 7 (top and middle panel). As in Simulation 4, recovery of this parameter differed across estimation approaches. Overall, recovery was best for PP-B: estimation performance was least affected by the values of r_C, r_E, and *b*, and the estimates showed smaller bias ($\Delta \leq .04$) than in PP-LT, which again showed underestimation of *b* ($\Delta \leq$ -.07). NP showed poor recovery of *b* that, in addition, varied notably across the levels of r_C and r_E (with a bias as large as Δ = .26 for b = .15, r_C = .60, and r_E = .58).

5.4. Simulation 6: Number of observations

Simulations 4 and 5 indicated that recovery of parameter *b* was particularly poor in NP, but that the other approaches also showed some problems. Could these difficulties be alleviated by increasing the number of observations (i.e., participants and/or items)? In Simulation 6, we reran the analyses with twice as many participants as in Simulations 4 and 5 (N_I = 48, N_P = 94), or twice as many items (N_I = 96, N_P = 47), or both (N_I = 96, N_P = 94), and compared recovery performance.

The key results are shown in Fig. 7 (lower panel). For all approaches, the recovery performance for the *b* parameter indeed improved with more observations, specifically when the number of items was increased. Estimation of *b* was more precise for all approaches ($SD \leq .06$), but the effect on bias differed across the approaches. Whereas bias decreased to an acceptable level (Δ = −.03) with twice the number of observations in PP-LT, resulting in even better results than in PP-B (Δ = .06), it remained high in NP (Δ = .13).

5.5. Discussion

We compared the ability of different estimation approaches (complete pooling, no pooling, partial pooling) to recover the core parameters (b, r_C, and r_E) of the hindsight-bias model for different situations, including heterogeneity across participants, correlations between the model parameters, and levels of the parameter values. All approaches achieved very good recovery of the recollection parameters r_C and r_E. When the values of r_C and r_E were very low (i.e., .15 and .13, respectively), r_C and r_E were slightly overestimated in all approaches except PP-LT. However, the size of this bias ($|\Delta| \leq .03$) seems tolerable.

The picture for the b parameter was more complex. Given that b, which indicates the probability of biased reconstruction, is the key contributor to hindsight bias (e.g., Erdfelder et al., 2007; Erdfelder & Buchner, 1998), this result is of high practical importance. Generally, recoverability of the b parameter was rather poor, and this held for most of the estimation approaches. Nevertheless, PP-B produced the best overall recovery performance with a typical amount of data; in addition, estimation accuracy was similarly accurate across different levels of the parameter values (i.e., levels of b, r_C, and r_E).

Why did PP-LT not perform better, given that it represents the most sophisticated modeling approach and allows parameter correlations to be explicitly modeled? One reason could be the high number of covariance parameters that need to be estimated with this approach. The hindsight-bias model itself already has 13 parameters; in addition, some of the processes assumed in the model occur only rarely (e.g., adopting the correct answer as one's prior judgment, represented as parameter c in the model), meaning that some response categories typically have very low frequencies (e.g., 'ROJ = CJ < OJ'). Thus, in the context of data situations common in empirical studies, the model may be overparameterized.

The NP approach produced severely biased estimates of b, with overestimation of up to $\Delta = .26$. This bias likely occurs because individual-level frequencies in the response categories involved are very low, making it impossible to accurately estimate parameters on the level of the participant. The resulting average across individual-level parameters is inaccurate (see Section 2.2). Naturally, this type of bias disappears with a sufficient number of observations per participant, but how many observations would be needed? As shown in our simulations, increasing the number of items to a number that could still be realistically implemented in an experiment ($N_I = 96$) did not lead to a noteworthy reduction in bias. Additional simulations showed that with the NP approach a small bias in b remained even with an unrealistic number of items ($N_I = 1000$).

In sum, if a researcher's primary interest is in the reconstruction-bias parameter b, the choice of estimation approach needs to be based on the amount of data available. With a typical level of heterogeneity and a typical number of participants ($\sim$50) and items ($\sim$50), the PP-B approach produces the most accurate estimates—independent of whether core parameter values are high or low. The latter aspect is particularly important if the goal is to examine group differences in parameters (e.g., age-group differences) or differences between experimental conditions (Bayen et al., 2006; Erdfelder & Buchner, 1998; Groß & Bayen, 2015, 2017b; Pohl et al., 2018, 2010).

6. General discussion

A key goal in cognitive psychology is to disentangle and measure the cognitive processes underlying behavior. Cognitive models describe the operation and interaction of latent processes that are assumed to contribute to behavior, as represented by model parameters. MPT models, a class of cognitive models for categorical data, are widely used in different fields in psychology. Recent methodological developments in model formulation and estimation have considerably widened the scope of their applicability (Coolin et al., 2015, 2016; Klauer, 2010; Matzke et al., 2015; Smith & Batchelder, 2010). Yet the various estimation approaches available have not yet been rigorously compared against each other in settings that are typical of empirical investigations. Therefore, little is known about whether and under what conditions the theoretical advantages of recently developed partial-pooling approaches actually play out when those approaches are applied to estimate the parameters of existing MPT models.

To address these questions, we conducted two sets of parameter recovery studies to investigate how well different parameter estimation approaches (as implemented in MPT software packages) recovered the true parameters in simulated data for two commonly used MPT models: the (relatively simple) source-monitoring model and the (more complex) hindsight-bias model. We focused our comparison on parameter values and methodological features that reflect typical experimental settings; thus, the results of our simulations have direct bearing on applications of MPT models and allow recommendations to be derived for MPT modelers. We present these recommendations in Section 6.3.

For core parameters of the two models, partial-pooling approaches indeed showed the best estimation performance overall with a realistic number of observations. For the source-monitoring model, the latent-trait approach was more accurate than the beta-MPT approach across the simulations. However, across all estimation approaches, there was severe bias and imprecision in the estimation of the model's d_1 parameter (submodel 5d). This problem may be rooted in the architecture of the model (we address this possibility in Section 6.1). Importantly, when we used the latent-trait partial pooling approach with a very large (but still realistic) number of observations, bias in estimation was reduced to a tolerable level.

For the more complex hindsight-bias model, the latent-trait approach appeared to be overparameterized for situations with typical numbers of participants and items (we observed underestimation of the true parameter value of parameter b in all simulations); instead, the (less complex) beta-MPT approach was most accurate. The latent-trait approach yielded more accurate estimates than the beta-MPT approach only when the number of observations was very high. Thus, the latent-trait approach seems to require amounts of data that exceed what is commonly collected in empirical studies, if its theoretical advantages are to pay off in practice. The no-pooling approach seemed generally unsuited.

Overall, an important insight from our results is that, when working with typical amounts of data, which estimation approach provides the most accurate parameter estimation depends on the MPT model under consideration. This interaction calls for a systematic examination of how model characteristics affect the accuracy and precision of parameter estimation. A second important insight is that some core parameters of the MPT models for source monitoring and hindsight bias—models frequently used in empirical investigations—can be estimated only with some degree of bias. We next discuss one aspect of the model architecture that might affect how well its parameters can be estimated.

6.1. Interdependence between model parameters

With submodel 5d of the source-monitoring model, we observed marked difficulties with the estimation of parameter d_1 across all estimation approaches. Which properties of the model could have contributed to the difficulties observed? One may be parameter interdependence, producing trade-offs between

Table A.1
True parameter values, heterogeneity, and parameter correlations for source-monitoring submodel 4.

	b	$D_1 = D_2 = D_3$	$d_1 = d_2$	$a = g$
b	**.226** (0.764)	−.088	−.142	.034
$D_1 = D_2 = D_3$		**.706** (0.369)	.000	.035
$d_1 = d_2$			**.531** (0.441)	−.169
$a = g$				**.642** (0.474)

Note. The main diagonal shows true parameter values (bold) and heterogeneity (standard deviation on latent probit scale); outside the main diagonal are parameter correlations (latent probit scale). Values are the results of a PP-LT analysis of Schaper et al. (2019, Experiment 1, "JOL & JOS" condition).

Table A.2
True parameter values, heterogeneity, and parameter correlations for source-monitoring submodel 5d.

	b	$D_1 = D_2 = D_3$	d_1	d_2	$a = g$
b	**.241** (0.728)	−.399	−.067	−.250	.107
$D_1 = D_2 = D_3$	−.399	**.550** (1.096)	.216	.578	.202
d_1	−.067	.216	**.201** (1.658)	.163	.020
d_2	−.250	.578	.163	**.752** (0.577)	.184
$a = g$	.107	.202	.020	.184	**.750** (0.587)

Note. The main diagonal shows true parameter values (bold) and heterogeneity (standard deviation on latent probit scale); outside the main diagonal are parameter correlations (latent probit scale). Values are the results of a PP-LT analysis of Schaper et al. (2019, Experiment 1, "Post-only" condition).

parameters (e.g., Krefeld-Schwalb, Pachur, & Scheibehenne, 2020; Li, Lewandowsky, & DeBrunner, 1996). For some computational models, it has been observed that there is quite a high level of structural dependence between model parameters that arises from the mathematical architecture of the model (e.g., Spektor & Kellen, 2018; van Ravenzwaaij, Dutilh, & Wagenmakers, 2011; see also Heathcote, Brown, & Wagenmakers, 2015). For instance, in cumulative prospect theory (Tversky & Kahneman, 1992) there is a strong negative association between the model's choice sensitivity and outcome sensitivity parameters (Scheibehenne & Pachur, 2015). Such structural dependencies can produce distortions in parameter estimation (partial-pooling models included) and make it inappropriate to interpret the model parameters independently of each other.

To gauge the extent to which there are also structural dependencies in the MPT models considered here, we generated one synthetic data set per model (using the parameter values in Appendix A, without heterogeneity and without individual-level parameter correlations) and used the latent-trait partial-pooling approach to analyze the data. We inspected the joint group-level posterior distributions for each pair of parameters in each of the models (see Appendix C). For the source-monitoring submodel 5d, several parameters were indeed strongly correlated with each other. Specifically, parameter a ($= g$) showed a high correlation with both parameter d_1 (r = -.77) and parameter d_2 (r = .51). One implication of this high correlation is that obtaining a high estimate for a ($= g$) may be driven by d_2 being high or d_1 being low, rather than necessarily reflecting that the process a (guessing source A) has a high probability. It is possible that this trade-off contributed to the high level of bias and imprecision in the estimation of d_1 for this model. Similar dependencies, although less pronounced, were observed for parameter b of the hindsight-bias model, which was correlated with auxiliary parameters c and g_{12} ($r \leq$ .48) and also produced serious problems in recovery.[8]

To our knowledge, there is currently no research on the implications of parameter interdependencies in MPT models. In view of our findings that such interdependencies may contribute to difficulties in parameter estimation, further investigation is clearly needed.

6.2. Limitations

Before turning to our recommendations for applications of MPT models, we acknowledge several limitations of our analyses. First, our insights about the different estimation procedures were obtained in the context of two specific MPT models, and we cannot ensure that our conclusions are generalizable to other MPT models. In fact, it seems likely that the suitability of the different estimation procedures differs across MPT models. This may even hold across different model variants (e.g., submodels of the source-monitoring model). Second, we examined only a subset of possible conditions under which the two MPT models can, in principle, be applied. For instance, different results may have emerged for different degrees of heterogeneity, or for other, unexplored combinations of parameter values. Finally, our results are mute with regard to parameter recovery using less conventional estimation approaches, such as Bayesian variants of the no-pooling and complete-pooling approaches (e.g., Singmann et al., 2019). However, given that we focused on commonly employed approaches, empirically informed parameter values, and typical methodological settings, we believe that our conclusions are relevant for many practical contexts.

6.3. Recommendations

Our findings can guide the choice of parameter estimation approach for MPT models. In research on source monitoring, it is typically of interest to measure the probability of item recognition (D_1, D_2, D_3), source memory (d_1, d_2), and source guessing (a, g). All estimation approaches yielded good recovery of these parameters for submodel 4 (which has one source-memory parameter). The PP-LT approach showed excellent recovery; for the other approaches, parameter b (guessing "old") was estimated with some bias.

Submodel 5d has separate parameters for memory for the two sources, d_1 and d_2. Recovery of d_1 was highly imprecise across all estimation approaches and severely biased under parameter heterogeneity. For applications of the source-monitoring model, we propose the following recommendations:

[8] For the source-monitoring submodel 4 (which has four parameters), all six pairs of parameters showed correlations below .30 (i.e., small to moderate correlations).

Table A.3
True parameter values, heterogeneity, and parameter correlations for the hindsight-bias model.

	b	c	g_{g1}	g_{g2}	g_{g3}	g_{l1}	g_{l2}	g_{l3}	h	l_C	l_E	r_C	r_E
b	**.30** (0.706)	−.159	.116	.011	.218	−.028	.185	.022	.012	−.189	−.226	−.050	−.019
c		**.05** (0.630)	−.012	−.011	−.063	.011	−.066	−.031	−.024	.020	.030	−.045	−.038
g_{l1}			**.82** (0.436)	.016	−.003	−.112	−.075	.040	.010	−.171	−.075	−.059	−.076
g_{l2}				**.49** (0.254)	.004	−.096	.019	.141	−.009	.010	.029	−.176	−.185
g_{l3}					**.77** (2.813)	.084	.117	−.039	.044	−.263	−.280	.096	.090
g_{g1}						**.83** (0.351)	.019	−.139	−.015	−.097	−.029	.090	.114
g_{g2}							**.51** (0.147)	.036	.014	−.034	−.121	.037	.055
g_{g3}								**.75** (1.936)	.006	.013	.025	.104	.134
h									**.01** (0.331)	−.020	−.032	.039	.045
l_C										**.49** (0.118)	.272	−.038	−.063
l_E											**.49** (0.116)	−.017	−.066
r_C												**.38** (0.521)	.718
r_E													**.36** (.393)

Note. The main diagonal shows true parameter values (probability scale, bold) and standard deviations (latent probit scale); outside the main diagonal are parameter correlations (latent probit scale). Values are the results of a PP-LT analysis of (Groß & Bayen, 2017a, young adults).

Table B.1
Average deviation Δ for the parameters of source-monitoring model 4 in each data set (standard deviations in brackets).

Simulation		b		$D_1 = D_2 = D_3$		$d_1 = d_2$		$a = g$		b		$D_1 = D_2 = D_3$		$d_1 = d_2$		$a = g$	
		Complete pooling								No pooling							
1a	No heterogeneity, no correlations	.00	[.02]	.00	[.01]	.00	[.02]	.00	[.02]	−.01	[.02]	.00	[.01]	.00	[.02]	.00	[.02]
1b	Heterogeneity, no correlations	.06	[.04]	−.02	[.02]	.00	[.03]	−.02	[.03]	.05	[.04]	−.01	[.02]	.00	[.03]	−.01	[.04]
1c	Heterogeneity, correlations	.05	[.04]	−.01	[.02]	.00	[.04]	−.01	[.03]	.03	[.04]	−.01	[.02]	.00	[.04]	.00	[.03]
		Partial pooling (latent-trait MPT)								Partial pooling (beta-MPT)							
1a	No heterogeneity, no correlations	.00	[.02]	.00	[.01]	.00	[.02]	.00	[.02]	.01	[.02]	.00	[.01]	.00	[.02]	.00	[.02]
1b	Heterogeneity, no correlations	−.01	[.04]	−.01	[.02]	.00	[.04]	−.01	[.04]	.07	[.04]	−.02	[.02]	.00	[.03]	−.02	[.03]
1c	Heterogeneity, correlations	.00	[.05]	.00	[.03]	.00	[.04]	.00	[.04]	.06	[.04]	−.01	[.02]	.00	[.04]	−.01	[.03]

Table B.2
Average deviation Δ for the parameters of source-monitoring submodel 5d in each data set (standard deviations in brackets).

Simulation		b		$D_1 = D_2 = D_3$		d_1		d_2		$a = g$	
		Complete pooling									
2a	No heterogeneity, no correlations	.00	[.02]	.00	[.01]	.01	[.15]	.00	[.03]	.00	[.03]
2b	Heterogeneity, no correlations	.04	[.05]	−.02	[.06]	.12	[.16]	−.03	[.06]	−.03	[.05]
2c	Heterogeneity, correlations	−.03	[.03]	−.01	[.05]	.16	[.15]	−.02	[.05]	−.05	[.05]
3	$N_P = 72$, $N_I = 192$	.04	[.03]	−.01	[.03]	.10	[.11]	−.03	[.03]	−.03	[.03]
		No pooling									
2a	No heterogeneity, no correlations	−.01	[.02]	.00	[.01]	.11	[.06]	−.08	[.04]	−.06	[.03]
2b	Heterogeneity, no correlations	.02	[.04]	−.01	[.05]	.17	[.07]	−.14	[.07]	−.06	[.05]
2c	Heterogeneity, correlations	.02	[.04]	−.01	[.05]	.17	[.08]	−.14	[.07]	−.07	[.05]
3	$N_P = 72$, $N_I = 192$	.03	[.03]	−.01	[.03]	.16	[.05]	−.12	[.04]	−.06	[.03]
		Partial pooling (latent-trait MPT)									
2a	No heterogeneity, no correlations	.00	[.02]	.00	[.01]	.01	[.09]	.00	[.03]	−.01	[.03]
2b	Heterogeneity, no correlations	.01	[.05]	−.01	[.08]	.14	[.10]	−.03	[.07]	−.03	[.05]
2c	Heterogeneity, correlations	.00	[.04]	.00	[.07]	.18	[.10]	−.03	[.06]	−.04	[.05]
3	$N_P = 72$, $N_I = 192$	.00	[.03]	.00	[.04]	.04	[.09]	−.01	[.04]	−.01	[.03]
		Partial pooling (beta-MPT)									
2a	No heterogeneity, no correlations	.01	[.02]	0	[.01]	.09	[.09]	−.02	[.03]	−.03	[.03]
2b	Heterogeneity, no correlations	.05	[.05]	−.02	[.05]	.12	[.10]	−.03	[.06]	−.03	[.04]
2c	Heterogeneity, correlations	.01	[.04]	−.02	[.04]	.14	[.11]	−.02	[.04]	−.04	[.04]
3	$N_P = 72$, $N_I = 192$	.04	[.03]	−.02	[.02]	.12	[.10]	−.04	[.03]	−.03	[.03]

Note. N_P = number of participants; N_I = number of items.

Table B.3
Average deviation Δ for the parameters of the hindsight-bias model in each data set (standard deviations in brackets).

Simulation		b		r_C		r_E		b		r_C		r_E	
		Complete pooling						No pooling					
4a	No heterogeneity, no correlations	.00	[.05]	.00	[.02]	.00	[.01]	.11	[.03]	.00	[.02]	.00	[.01]
4b	Heterogeneity, no correlations	.04	[.06]	.01	[.03]	.01	[.02]	.17	[.04]	.01	[.03]	.01	[.02]
4c	Heterogeneity, correlations	.04	[.07]	.01	[.03]	.01	[.02]	.17	[.04]	.01	[.03]	.01	[.03]
5a	$b = .15$, $r_E = .13$	.06	[.05]	.03	[.02]	.02	[.01]	.20	[.03]	.03	[.02]	.02	[.02]
5b	$b = .15$, $r_E = .36$	.06	[.06]	.02	[.03]	.01	[.02]	.23	[.04]	.02	[.03]	.01	[.02]
5c	$b = .15$, $r_E = .58$	.04	[.06]	−.01	[.03]	−.01	[.03]	.26	[.04]	−.01	[.03]	−.01	[.03]
5d	$b = .30$, $r_E = .13$	.03	[.05]	.03	[.02]	.02	[.02]	.15	[.04]	.03	[.02]	.02	[.02]
5e	$b = .30$, $r_E = .36$	.04	[.06]	.01	[.03]	.01	[.02]	.17	[.04]	.01	[.03]	.01	[.02]
5f	$b = .30$, $r_E = .58$	.04	[.07]	−.01	[.03]	.00	[.02]	.21	[.04]	−.01	[.03]	.00	[.02]
5g	$b = .55$, $r_E = .13$	−.01	[.05]	.03	[.02]	.02	[.02]	.06	[.04]	.03	[.02]	.02	[.02]
5h	$b = .55$, $r_E = .36$	−.02	[.05]	.01	[.03]	.01	[.03]	.07	[.04]	.01	[.03]	.01	[.03]
5i	$b = .55$, $r_E = .58$	−.01	[.06]	−.01	[.03]	.00	[.03]	.10	[.04]	.00	[.03]	.00	[.03]
6a	$N_P = 47$, $N_I = 96$	.03	[.05]	.01	[.03]	.01	[.02]	.13	[.03]	.01	[.03]	.01	[.02]
6b	$N_P = 94$, $N_I = 48$	.03	[.04]	.01	[.02]	.01	[.02]	.17	[.03]	.01	[.02]	.01	[.02]
6c	$N_P = 94$, $N_I = 96$	.03	[.03]	.01	[.02]	.01	[.02]	.13	[.02]	.01	[.02]	.01	[.02]
		Partial pooling (latent-trait MPT)						Partial pooling (beta-MPT)					
4a	No heterogeneity, no correlations	−.10	[.07]	.00	[.02]	.00	[.01]	−.03	[.06]	.00	[.02]	.00	[.01]
4b	Heterogeneity, no correlations	−.05	[.07]	.00	[.03]	.00	[.03]	.02	[.07]	.01	[.03]	.01	[.02]
4c	Heterogeneity, correlations	−.05	[.09]	.00	[.03]	.00	[.03]	.02	[.08]	.01	[.03]	.01	[.02]
5a	$b = .15$, $r_E = .13$	−.03	[.05]	.00	[.02]	.00	[.02]	.03	[.06]	.03	[.02]	.02	[.01]
5b	$b = .15$, $r_E = .36$	−.04	[.06]	.00	[.03]	.00	[.03]	.02	[.06]	.01	[.03]	.01	[.02]
5c	$b = .15$, $r_E = .58$	−.06	[.04]	.00	[.03]	.00	[.03]	−.01	[.04]	−.01	[.03]	−.01	[.03]
5d	$b = .30$, $r_E = .13$	−.04	[.07]	.00	[.02]	.00	[.02]	.04	[.06]	.03	[.02]	.02	[.02]
5e	$b = .30$, $r_E = .36$	−.05	[.07]	.00	[.03]	.00	[.03]	.02	[.07]	.01	[.03]	.01	[.02]
5f	$b = .30$, $r_E = .58$	−.07	[.08]	.00	[.03]	.00	[.03]	−.01	[.08]	−.01	[.03]	−.01	[.02]
5g	$b = .55$, $r_E = .13$	−.03	[.06]	.00	[.02]	.00	[.02]	.02	[.05]	.03	[.02]	.02	[.02]
5h	$b = .55$, $r_E = .36$	−.04	[.07]	.00	[.03]	.00	[.03]	.01	[.05]	.01	[.03]	.01	[.03]
5i	$b = .55$, $r_E = .58$	−.06	[.08]	.00	[.03]	.00	[.03]	.00	[.06]	−.01	[.03]	−.01	[.03]
6a	$N_P = 47$, $N_I = 96$	−.03	[.06]	.00	[.03]	.00	[.02]	.06	[.05]	.01	[.03]	.01	[.02]
6b	$N_P = 94$, $N_I = 48$	−.05	[.05]	.00	[.02]	.00	[.02]	.05	[.05]	.01	[.02]	.01	[.02]
6c	$N_P = 94$, $N_I = 96$	−.03	[.04]	.00	[.02]	.00	[.02]	.06	[.03]	.01	[.02]	.01	[.02]

Note. N_P = number of participants; N_I = number of items.

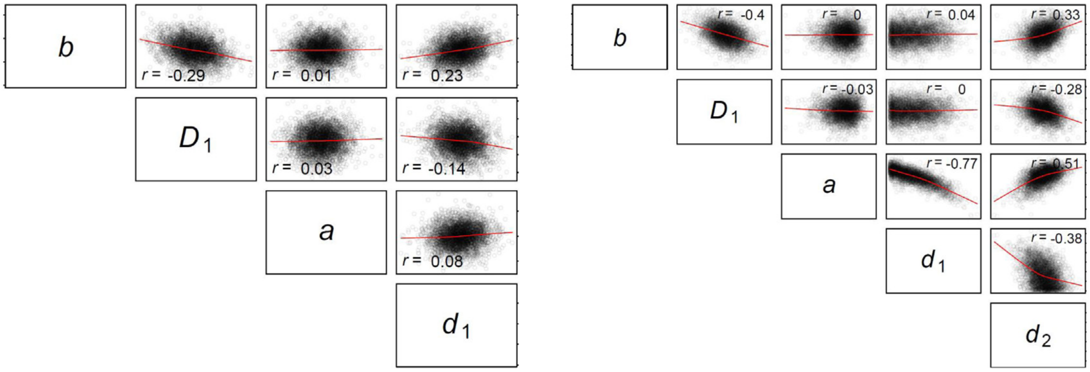

Fig. C.1. Joint group-level posterior distributions for all parameters of source-monitoring models 4 (left) and 5d (right), with the respective correlation coefficients. Plots and coefficients are based on 4000 samples.

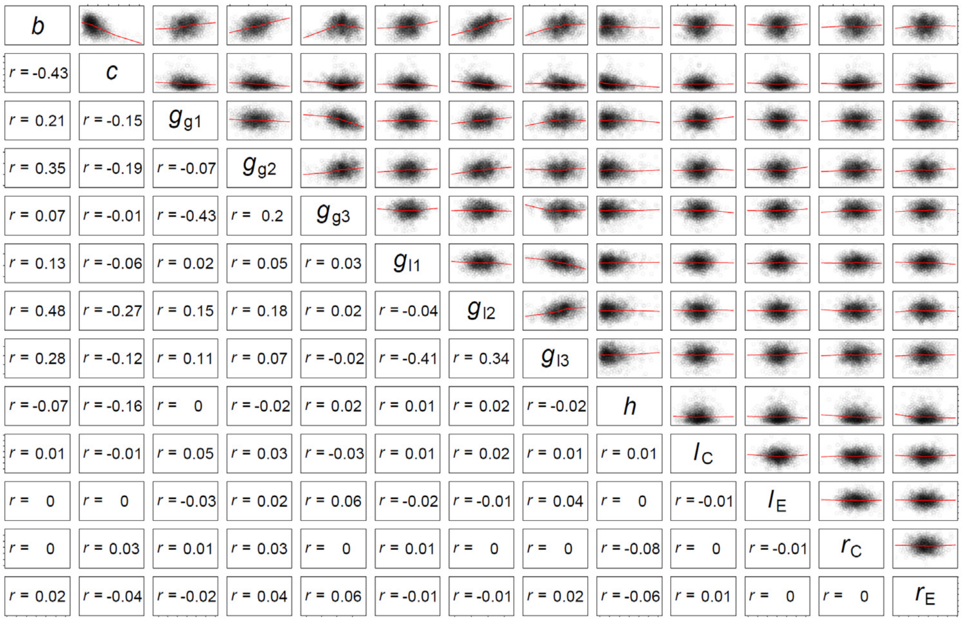

Fig. C.2. Joint group-level posterior distributions for all 13 parameters of the hindsight-bias model, with the respective correlation coefficients. Plots and coefficients are based on 1500 samples.

1. The PP-LT approach showed the best recovery performance overall for the models of source monitoring considered in our analyses (i.e., submodels 4 and 5d).
2. For applications of submodel 5d, we recommend using a very large number of observations (∼72 participants, ∼192 items) to ensure sufficiently accurate and precise estimates of d_1. However, even then, results on the difference in memory between the two sources should be interpreted with caution.

In research on hindsight bias, it is of interest to measure the probability that original judgments are recollected (represented by parameters r_C and r_E) as well as the probability that original judgments are reconstructed with bias towards the correct answer (represented by parameter b). In our analyses, recovery of the recollection parameters was generally very good, whereas recovery of the reconstruction-bias parameter b varied considerably. Based on our results, we propose the following recommendations for applications of the hindsight-bias model:

1. With typical numbers of participants (∼50) and items (∼50) and a typical degree of heterogeneity in parameter values (e.g., with a sample of young adults), the PP-B approach produces the most accurate estimates and therefore seems the most appropriate approach. The accuracy of the estimates obtained with this approach is, in addition, largely unaffected by whether the parameter values are high or low.
2. The PP-LT approach seems particularly useful with a very large number of observations (N_P >100, N_I >100), or when the main interest is in the recollection parameters, r_C and r_E, rather than in the reconstruction-bias parameter b.
3. The NP approach is unsuitable when the main interest is in the reconstruction-bias parameter b. If a researcher is interested in obtaining individual parameter estimates, a PP approach should be used.

Finally, it is important to reiterate that none of the approaches considered here seems generally suited for MPT modeling, at least in the context of typical experimental situations. To inform the choice of an estimation approach in a given setting, it therefore seems advisable to conduct parameter recovery analyses to test which estimation approach allows accurate inferences to be drawn for the data at hand. Fortunately, the development of various *R* packages has facilitated not only the application of more complex modeling approaches, but also the implementation of parameter recovery analyses (Heck et al., 2018; Singmann & Kellen, 2013).

7. Conclusions

Researchers seeking to estimate the parameters of an MPT model for empirical data can choose among a variety of approaches of varying complexity and statistical frameworks. Beyond the conventional estimation approach in the MPT literature, which relies on complete pooling, they can also draw on no-pooling and partial-pooling methods with or without modeling of parameter correlations (reflecting most recent developments). These different approaches to parameter estimation are freely available in easy-to-use software packages. While the latent-trait partial-pooling approach has several theoretical advantages, our results illustrate that, in practice, its application in not warranted for every MPT model—particularly given natural limitations on the amounts of data that can be collected. Moreover, even under ideal conditions, the accuracy of sophisticated estimation approaches may be constrained if a given model displays structural interdependencies between parameters. Our results thus indicate that insights about parameter recoverability should guide not only the choice of a method for parameter estimation, but also model development.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.jmp.2020.102402.

Appendix A. True parameter values, heterogeneity, and parameter correlations

See Tables A.1–A.3.

Appendix B. Simulation results

See Tables B.1–B.3.

Appendix C. Parameter interdependence

See Figs. C.1 and C.2.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmp.2020.102402.

References

Arnold, N. R., Bayen, U. J., & Böhm, M. F. (2014). Is prospective memory related to depression and anxiety? A hierarchical MPT modelling approach. *Memory*, *23*, 1215–1228. http://dx.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*, 326–333. http://dx.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*, 143–152. http://dx.doi.org/10.1027/1618-3169/a000287.

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

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564. http://dx.doi.org/10.1037/0033-295X.97.4.548.

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

Batchelder, W. H., Riefer, D. M., & Hu, X. (1994). Measuring memory factors in source monitoring: Reply to kinchla. *Psychological Review*, *191*, 172–176. http://dx.doi.org/10.1037/0033-295X.101.1.172.

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*, 1003–1018. http://dx.doi.org/10.1037/0278-7393.32.5.1003.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 197–215. http://dx.doi.org/10.1037//0278-7393.22.1.197.

Bayen, U. J., Nakamura, G. V., Dupuis, S. E., & Yang, C.-L. (2000). The use of schematic knowledge about sources in source-monitoring. *Memory & Cognition*, *28*, 480–500. http://dx.doi.org/10.3758/BF03198562.

Blank, H., Musch, J., & Pohl, R. F. (2007). Hindsight bias: On being wise after the event. *Social Cognition*, *25*, 1–9. http://dx.doi.org/10.1521/soco.2007.25.1.1.

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

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*, 692–712. http://dx.doi.org/10.3758/PBR.15.4.692.

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

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

Dehn, D. M., & Erdfelder, E. (1998). What kind of bias is hindsight bias?. *Psychological Research*, *61*, 135–146. http://dx.doi.org/10.1007/s004260050020.

Erdfelder, E. (2000). Multinomiale Modelle in der kognitiven Psychologie [Multinomial models in cognitive psychology]. Unpublished habilitation thesis, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany.

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*, 108–124. http://dx.doi.org/10.1027/0044-3409.217.3.108.

Erdfelder, E., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition*, *25*, 114–131. http://dx.doi.org/10.1521/soco.2007.25.1.114.

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

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

Groß, J., & Bayen, U. J. (2017a). Age differences in hindsight bias in a memory- and in a no-memory hindsight-bias task. [Unpublished raw data].

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

Groß, J., & Pachur, T. (2019). Age differences in hindsight bias: A meta-analysis. *Psychology and Aging*, *34*, 294–310. http://dx.doi.org/10.1037/pag0000329.

Heathcote, A., Brown, S. D., & Wagenmakers, E. J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann, & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, NY: Springer.

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

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*, 123–134. http://dx.doi.org/10.1037/a0017518.

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. http://dx.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*, 1470–1485. http://dx.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*, 21–47. http://dx.doi.org/10.1007/BF02294263.

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

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28. http://dx.doi.org/10.1037/0033-2909.114.1.3.

Keefe, R. S., Arnold, M. C., Bayen, U. J., & Harvey, P. D. (1999). Source monitoring deficits in patients with schizophrenia; a multinomial modelling analysis. *Psychological Medicine*, *29*, 903–914. http://dx.doi.org/10.1017/S0033291799008673.

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

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

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the who said what? paradigm. *Journal of Personality and Social Psychology*, *75*, 1155–1178. http://dx.doi.org/10.1037/0022-3514.75.5.1155.

Krefeld-Schwalb, A., Pachur, T., & Scheibehenne, B. (2020). Structural parameter interdependencies in computational models of cognition (submitted for publication).

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*, 875–889. http://dx.doi.org/10.1037/pag0000140.

Li, S.-C., Lewandowsky, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General*, *125*, 360–369. http://dx.doi.org/10.1037/0096-3445.125.4.360.

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*, 205–235. http://dx.doi.org/10.1007/s11336-013-9374-9.

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

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*, 45–69. http://dx.doi.org/10.1037/a0030734.

Michalkiewicz, M., Arden, K., & Erdfelder, E. (2018). Do smarter people employ better decision strategies? The influence of intelligence on adaptive use of the recognition heuristic. *Journal of Behavioral Decision Making*, *31*, 3–11. http://dx.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*, 454–468. http://dx.doi.org/10.3758/s13421-015-0567-6.

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

Pohl, R. F., Bayen, U. J., Arnold, N., Auer, T. S., & Martin, C. (2018). Age differences in processes underlying hindsight bias: A life-span study. *Journal of Cognition and Development*, *19*, 278–300. http://dx.doi.org/10.1080/15248372.2018.1476356.

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

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*, 411–426. http://dx.doi.org/10.1177/1745691612454303.

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*, 573–604. http://dx.doi.org/10.3758/BF03196750.

Ruoß, M., & Becker, H.-W. (2001). Der Hindsight Bias trägt zur Chronifizierung von Schmerzen bei [The Hindsight Bias contributes to the chronification of pain]. *Zeitschrift für Psychologie*, *209*, 316–342. http://dx.doi.org/10.1026/0044-3409.209.3.316.

Schaper, M. L., & Kuhlmann, B. G. (2019). A multiverse reanalysis of published multinomial processing tree modeling results: Models of source monitoring. *Paper presented at the 3rd DFG network meeting on hierarchical MPT modeling. Mannheim, Germany*.

Schaper, M. L., Kuhlmann, B. G., & Bayen, U. J. (2019). Metamemory expectancy illusion and schema-consistent guessing in source monitoring. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*, 470–496. http://dx.doi.org/10.1037/xlm0000602.

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

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*, 1248–1284. http://dx.doi.org/10.1080/03640210802414826.

Singmann, H., Heck, D., Barth, M., Groß, J., & Kuhlmann, B. G. A Bayesian and frequentist multiverse pipeline for MPT models: Applications to recognition memory. *Paper presented at the Tagung experimentell arbeitender Psychologen*. London, UK.

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

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. http://dx.doi.org/10.3758/PBR.15.4.713.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. http://dx.doi.org/10.1016/j.jmp.2009.06.007.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50. http://dx.doi.org/10.1037/0096-3445.117.1.34.

Spaniol, J., & Bayen, U. J. (2002). When is schematic knowledge used in source monitoring?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 631–651. http://dx.doi.org/10.1037//0278-7393.28.4.631.

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*, 2047–2068. http://dx.doi.org/10.3758/s13423-018-1446-5.

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

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*, 547–556. http://dx.doi.org/10.3758/s13423-016-1081.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E. J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*, 94–105. http://dx.doi.org/10.1016/j.jmp.2010.08.010.

Wulff, L., & Kuhlmann, B. G. (2020). Is knowledge reliance in source guessing a cognitive trait? Examining stability across time and domain. *Memory & Cognition*, *48*, 256–276. http://dx.doi.org/10.3758/s13421-019-01008-1.