Processing Tree Models for Discrete and Continuous Variables

DANIEL W. HECK

Inaugural Dissertation
Submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences (GESS) at the University of Mannheim.

Thesis Defence:
September 5th, 2017

Supervisor:
Prof. Dr. Edgar Erdfelder

Dean of the Faculty of Social Sciences:
Prof. Dr. Michael Diehl

Academic Director of the CDSS:
Prof. Dr. Edgar Erdfelder

Thesis Reviewers:
Prof. Dr. Arndt Bröder
Prof. Dr. Thorsten Meiser

Defense Committee:
Prof. Dr. Arndt Bröder
Prof. Dr. Edgar Erdfelder
Prof. Dr. Thorsten Meiser

"Computational models help in understanding how things might work. When a system is too complex to understand, it often helps to understand a simpler system with analogous behavior."

— W. Daniel Hillis, *Why physicists like models and why biologists should*

Contents

Für Rina.

Abstract

In psychology, multinomial processing tree (MPT) models explain how qualitatively different processes determine observed response behavior. Even though they have successfully been used in many applications, MPT models are inherently limited to discrete data such as choice frequencies. In my thesis, I therefore propose two new approaches that extend MPT models to continuous variables such as response times (RTs), fine-grained response scales, or process-tracing measures. Both approaches assume that continuous variables follow a finite mixture distribution with mixture weights determined by the processing tree structure and state-specific, continuous component distributions.

In the first approach, RT-extended versions of MPT models are obtained by categorizing continuous observations into a finite number of intervals. Thereby, the RT component distributions can be estimated by histograms, allowing researchers to test the relative speed of different latent processes as demonstrated for the two-high-threshold model of recognition memory. In a theoretical paper, the new method is used to develop an RT-extended MPT model of recognition-based decisions and test competing process models. Even though the two theoretical accounts under consideration differ strongly and assume serial and parallel information integration, respectively, empirical tests have proved to be difficult without the new measurement model.

As a second approach, I developed the class of generalized processing tree (GPT) models that assume parametric families instead of histograms for the continuous component distributions. The main advantages of GPT models are that they reduce the flexibility of the component distributions, can result in more precise estimates for the processing-tree parameters compared to MPT models, and allow for modeling one or more continuous variables jointly. In a first empirical application, a GPT model provided a good account of mouse-tracking data in a semantic-categorization paradigm. In sum, my thesis lays the foundations for new empirical tests of discrete-state theories of cognition by jointly modeling discrete and continuous variables.

1 Articles

This cumulative thesis is based on two published articles and one manuscript submitted for publication that will be discussed in the chronological order of development. The main text provides an overview and an encompassing theoretical framework for these three papers, whereas statistical and empirical details are found in the original articles appended to this thesis.

Heck, D. W. & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Heck, D. W. & Erdfelder, E. (in press). Linking process and measurement models of recognition-based decisions. *Psychological Review*. doi:10.1037/rev0000063

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2017). *Generalized processing tree models: Jointly modeling discrete and continuous variables*. Manuscript submitted for publication.

During my dissertation, I have also worked on other topics related to multinomial processing tree models and statistical modeling in general. These articles are not included in the thesis, because they focus either on practical applications or on statistical aspects such as model selection, Bayesian inference, or software implementations. Nonetheless, several of these papers are referred to in the main text, since they laid the foundations for the novel developments presented in my thesis.

Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, *60*, 29–34. doi:10.1016/j.jmp.2014.06.002

Erdfelder, E., Castela, M., Michalkiewicz, M., & Heck, D. W. (2015). The advantages of model fitting compared to model simulation in research on preference construction. *Frontiers in Psychology*, *6*, 140. doi:10.3389/fpsyg.2015.00140

Heck, D. W., Wagenmakers, E.-J., & Morey, R. D. (2015). Testing order constraints: Qualitative differences between Bayes factors and normalized maximum likelihood. *Statistics & Probability Letters*, *105*, 157–162. doi:10.1016/j.spl.2015.06.014

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

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. *Judgment and Decision Making*, *11*, 527–536. Retrieved from http://journal.sjdm.org/16/16613/jdm16613.html

Heck, D. W., Hilbig, B. E., & Moshagen, M. (2017). From information processing to decisions: Formalizing and comparing probabilistic choice models. *Cognitive Psychology*, *96*, 26–40. doi:10.1016/j.cogpsych.2017.05.003

Heck, D. W., Hoffmann, A., & Moshagen, M. (2017). *Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model*. Manuscript submitted for publication.

Heck, D. W., Overstall, A., Gronau, Q. F., & Wagenmakers, E.-J. (2017). *Quantifying uncertainty in transdimensional Markov chain Monte Carlo using discrete Markov models*. Manuscript submitted for publication.

Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*. doi:10.3758/s13428-017-0869-7

Heck, D. W. & Moshagen, M. (in press). RRreg: An R package for correlation and regression analyses of randomized response data. *Journal of Statistical Software*.

Klein, S. A., Hilbig, B. E., & Heck, D. W. (in press). Which is the greater good? A social dilemma paradigm disentangling environmentalism and cooperation. *Journal of Environmental Psychology*. doi:10.1016/j.jenvp.2017.06.001

Miller, R., Scherbaum, S., Heck, D. W., Goschke, T., & Enge, S. (in press). On the relation between the (censored) shifted Wald and the Wiener distribution as measurement models for choice response times. *Applied Psychological Measurement*.

2 Introduction

2.1 Multinomial Processing Tree Models

Many theories in psychology state that human behavior, cognition, and emotion are each determined by qualitatively distinct processes. For instance, theories of memory distinguish between storage and retrieval of information (e.g., Batchelder & Riefer, 1986), theories of judgment and decision making assume that people rely on distinct, ecologically adaptive strategies (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999), and dual-process theories of categorization and reasoning distinguish between similarity-based and rule-based processes (e.g., Sloman, 1996). Despite their success and popularity, testing such theories empirically is often difficult, because observable behavior can in principle emerge from more than one of the hypothesized processes. Therefore, it is rarely warranted to treat observed responses as one-to-one indicators of latent processes.

As a remedy, psychological measurement models define how latent cognitive processes determine observable behavior. Compared to verbal theories, such mathematical models are more precise, and thereby allow for stronger tests of theoretical predictions (Erdfelder, Castela, Michalkiewicz, & Heck, 2015). After establishing the validity of a measurement model, its parameter estimates can be used as process-pure measures of the hypothesized processes to test psychological theories and predictions (Erdfelder, 2000). The present thesis is concerned with a specific modeling framework, multinomial processing tree (MPT) models (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Riefer & Batchelder, 1988), that accounts for frequencies of discrete responses by assuming a finite number of latent cognitive states. More specifically, MPT models assume that observed response frequencies follow multinomial distributions, in which the expected category probabilities are modeled by a processing tree. In such a psychologically motivated probability tree, conditional probabilities describe the prevalence of the hypothesized latent processes.

In the following, I will rely on the two-high-threshold (2HT) model of recognition memory (Snodgrass & Corwin, 1988) as a running example to outline the definition and structure of MPT models. Note that this model also serves as an example in my first and third paper (Heck & Erdfelder, 2016; Heck, Erdfelder, & Kieslich, 2017) to illustrate the new approaches of extending MPT models in general to continuous variables. The 2HT model shown in Figure 2.1 is one of the most simple and prominent instances of the MPT model class and represents the building block of many more complex memory models (e.g., the source-monitoring model; Batchelder

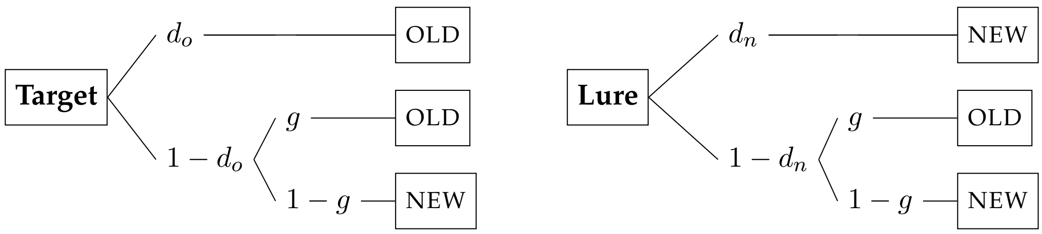

FIGURE 2.1: The two-high threshold (2HT) model of recognition memory (Snodgrass & Corwin, 1988).

& Riefer, 1990; Bayen, Murnane, & Erdfelder, 1996). In standard recognition-memory paradigms, participants first have to learn a list of words or pictures, and later have to judge studied and new stimuli (often called targets and lures) as OLD or NEW, respectively. Without an explicit model, it is unclear how memory performance relates to the frequencies of correct OLD responses to targets (hits) and incorrect OLD responses to lures (false alarms). The 2HT addresses this issue by making explicit assumptions about the underlying memory and response processes that can be evaluated for theoretical plausibility and, in turn, tested empirically.

The 2HT model assumes that hits are due to two qualitatively distinct processes: With probability d_o, participants recognize the item as being old with certainty and thus respond OLD. Alternatively, with probability $1 - d_o$, participants are in an uncertainty state and do not have any memory of the presented item. In this case, the 2HT model assumes that an OLD response is given with guessing probability g based on a general preference for one of the two options. Overall, since the two processing paths are disjoint, the probability of OLD responses is thus given by the sum of both branches

$$P(\text{OLD} \mid \text{target}) = d_o + (1 - d_o)g. \tag{2.1}$$

Similarly, the 2HT model assumes that lures are detected with probability d_n and that the guessing probability g is identical to the guessing probability for targets, resulting in the probability $(1 - d_n)g$ for false alarms. Statistically, the observed frequencies of hits and false alarms follow independent binomial distributions with rate parameters defined by these model probabilities (for a formal definition, see Hu & Batchelder, 1994).

Many advantages of MPT models can be illustrated for the case of the 2HT model. First, goodness-of-fit tests allow for testing whether a model fits empirical data in absolute terms — which is the case for the 2HT model (Bröder & Schütz, 2009). Second, the validity of MPT models can be established by demonstrating that experimental manipulations selectively influence specific parameters (Erdfelder & Buchner, 1998; Klauer & Wegener, 1998): In case of the 2HT model, memory-strength manipulations (e.g., presentation duration or frequency) affect only the memory parameters d_o and d_n, whereas base-rate manipulations (i.e., changing the proportion of targets in the test list) affect only the guessing parameter g. Selective influence is an important prerequisite for the validity of MPT models, since it ensures that parameters can be interpreted in terms of the hypothesized latent processes. Once the validity of an MPT model has been established, it can be used as a measurement tool to test how experimental manipulations affect different cognitive processes. Moreover, the framework of *cognitive psychometrics* (Batchelder, 1998) relies on MPT models to test whether specific populations of participants differ in information processing. For instance, Riefer, Knapp, Batchelder, Bamber, and Manifold (2002) applied the MPT model developed by Batchelder and Riefer (1980) to disentangle storage and retrieval in free recall to isolate cognitive processes that cause lower memory performance in schizophrenics and alcoholics.

During the last decades, MPT models have gained increasing popularity and provided new insights in many fields of psychology, a success that can be attributed to several factors (for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016). Most importantly, the processing-tree structure establishes a formal link between psychological theory and observable responses and thereby provides better explanations of observed behavior compared to ad hoc measures (Bröder & Meiser, 2007; Hilbig, 2010a). Second, MPT models are mathematically tractable, which facilitates model development and conceptual understanding by substantive researchers. Third, statistical inference is well developed and understood. Besides standard maximum-likelihood estimation algorithms (Hu & Batchelder, 1994) and software (e.g., Moshagen, 2010; Singmann & Kellen, 2013), Bayesian hierarchical approaches have been developed (Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Smith & Batchelder, 2010) and implemented in user-friendly software (Heck, Arnold, & Arnold, in press). Finally, the scope and usefulness of MPT models has continually been increased by new methodological developments. For instance, order constraints of the form $\theta_1 > \theta_2$ are easily implemented by reparameterizations resulting in a new, less flexible MPT model (Klauer, Singmann, & Kellen, 2015; Knapp & Batchelder, 2004) that can be compared statistically to the original model (Heck, Wagenmakers, and Morey, 2015; Wu, Myung, and Batchelder, 2010a, 2010b; but see Heck, Moshagen, and Erdfelder, 2014 for limitations). Another new area of applications concerns individual differences, where core concepts of MPT models have been combined with item response theory (Klauer, 2010; Matzke et al., 2015), for instance, to measure participants' response styles (Plieninger & Heck, 2017). In my thesis, I further expand the toolbox of MPT methods by developing novel approaches for modeling discrete and continuous variables jointly.

2.2 Jointly Modeling Continuous and Discrete Variables

MPT models gain much of their power from their flexibility, since they can easily be adapted to more complex experimental paradigms. For instance, the 2HT model can be generalized from binary response data to confidence ratings (Bröder, Kellen, Schütz, & Rohrmeier, 2013) or ranking tasks (Malejka, Heck, & Erdfelder, 2017) while retaining the core assumption that discrete memory states determine participants' accuracy. Nevertheless, MPT models are inherently limited in scope, because they can only account for discrete data such as choice frequencies. This is apparent from the statistical structure of the underlying multinomial distribution, which models the probabilities of counts falling into a finite number of categories. In principle, this is not a limitation as long as psychological theories merely make predictions about choice frequencies and other discrete outcome variables.

However, the objective of theories in cognitive psychology is to explain general mechanisms underlying human information processing. A good theory should have a large scope and predict a wide range of aspects of observable human behavior (Glöckner & Betsch, 2011). Hence, theories should generalize beyond response frequencies and also allow for testable predictions concerning continuous variables. For instance, much theorizing in psychology has focused on response times (RTs; Luce, 1986; Townsend & Ashby, 1983). By measuring the speed of participants' responses to various stimuli under different conditions, researchers aim to explain basic information processing of visual or acoustic stimuli (Donders, 1868), describe the retrieval of information from memory (Ratcliff & Murdock, 1976), or test serial versus parallel processing architectures (Townsend, 1990).

Since statistical methods should not constrain theory building, the question arises how to extend MPT models to continuous variables. Concerning the 2HT model for recognition memory, for instance, additional assumptions are required to derive testable predictions for RTs (Dube, Starns, Rotello, & Ratcliff, 2012; Heck & Erdfelder, 2016). An important question is whether the processing tree in Figure 2.1 is interpreted literally as a sequence of underlying processing steps as proposed by Hu (2001). According to this view, guessing responses must be stochastically slower than detection responses (Heck & Erdfelder, 2016). In general, however, the processing tree structure of the 2HT model can also be maintained merely as a structure of expected probabilities for response frequencies while assuming that guessing responses can actually be faster than detection responses (similar to the fast-guess model by Ollman, 1966). Besides these issues of the relative speed of memory and decision processes, the question arises whether researchers want to commit to parametric constraints when modeling RT distributions (e.g., assuming that RTs follow a log-normal distribution). Moreover, statistical problems of model identifiability, parameter estimation, and model comparison need to be addressed before testing whether any RT-extended version of the 2HT model can explain recognition judgments and speed. Obviously, these issues are not limited to the 2HT model but apply to MPT models in general.

Similar questions also arise when modeling other continuous variables than RTs. With the increasing availability of computers and electronic devices in general, data collection and preprocessing of various types of continuous variables has been greatly facilitated. For instance, software packages for creating experiments allow researchers to implement continuous response scales such as fine-graded sliders for confidence ratings (Province & Rouder, 2012) or circular scales measuring spatial memory for the location or color of presented items (Harlow & Donaldson, 2013; Oberauer & Lin, 2017). Moreover, in many paradigms such as the classical anchoring task (Tversky & Kahneman, 1974) or hindsight-bias paradigm (Pohl, 2007), participants are asked to directly estimate magnitudes instead of choosing between a limited number of predefined response options.

Besides more fine-grained response scales, behavioral or physiological variables are often recorded during the time course of each trial (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011). For instance, eye tracking is a prominent process-tracing method that is used to measure cognitive processes such as information search or attention allocation (e.g., Just & Carpenter, 1976). More recently, researchers have also started to analyze movements of participants during experimental tasks, for instance, by recording the trajectories of the computer mouse when choosing one of two options (Kieslich & Henninger, in press; Spivey, Grosjean, & Knoblich, 2005). Finally, continuous variables are recorded in many neuropsychological studies that measure electro-related potentials or use functional magnetic resonance imaging. The relevance of jointly modeling continuous neurophysiological and discrete behavioral variables was recently highlighted in a special issue of the *Journal of Mathematical Psychology* (Palmeri, Love, & Turner, 2017).

Overall, continuous variables are widely used in psychology and provide a rich source of information for our understanding of cognitive processes. Currently, however, no general-purpose methods exist to test psychological theories that assume qualitatively distinct states and make predictions for both discrete and continuous data. In my thesis, I address this gap by developing novel methods that extend the scope of MPT models to continuous variables. In the first paper (Heck & Erdfelder, 2016), I present a simple approach of extending MPT models to RTs and provide solutions to statistical issues such as model identifiability. In the second paper (Heck & Erdfelder, in press), I focus on the theoretical derivation of a new, RT-extended MPT model for recognition-based decisions, thereby linking process and measurement models that have previously been applied in isolation. Finally, in the third paper (Heck, Erdfelder, & Kieslich, 2017), I define the new class of generalized processing tree models which provide an alternative approach for modeling continuous variables in a discrete-state framework.

Besides these theoretical, methodological, and statistical contributions, I also address the pragmatic issue that new methods are often not used unless corresponding software is available. In cognitive modeling, user-friendly software needs to facilitate the specification, estimation, and testing of models. For instance, the R package `TreeBUGS` (Heck et al., in press) facilitates the analysis of Bayesian hierarchical MPT models that take heterogeneity across participants into account (Klauer, 2010; Smith & Batchelder, 2010). Similarly, the third paper of my thesis does not only define generalized processing tree models but also provides an implementation in the R package `gpt` using a simple modeling syntax. In the following, I summarize the core ideas and developments presented in these three papers. For technical details and empirical results, the reader is referred to the original articles in Appendix D.

3 Extending MPT Models

3.1 Measuring the Relative Speed of Cognitive Processes

Heck, D. W. & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

At their core, discrete-state theories of cognition predict that behavior depends on a finite number of latent states, and that in each trial of a study, one of these states determines observable outcomes independent of the other states (Luce, 1986; Townsend & Ashby, 1983; Yantis, Meyer, & Smith, 1991). Since people can be in different cognitive states across multiple observations, discrete-state theories predict a mixture distribution of state-specific outcomes for the observed data. For instance, dual-process models distinguish two systems of information processing: An automatic, associative System I and a deliberative, rule-based System II (e.g., Evans, 2008; Sloman, 1996). Predictions for observable behavior depend on the system used in a specific situation. Among other qualitative differences, System I responses are predicted to be rather fast, whereas System II responses are predicted to be rather slow.

Statistically, this dual-process account implies that observed RTs t follow a mixture distribution with probability density function

$$f(t) = \alpha g_1(t) + (1 - \alpha)\, g_2(t), \tag{3.1}$$

where α is the proportion of System I responses, and $g_1(t)$ and $g_2(t)$ are the probability density functions of RTs from System I and II, respectively. The statistical representation of dual-process theories as mixture models has several advantages. First, a mathematical model is more exact than a verbal theory and thus enables a stronger test of the core assumption of two systems of processing (Dixon, 2012; Falmagne, 1968; Miller, 2006; Province & Rouder, 2012; Yantis et al., 1991). Second, the model in Eq. 3.1 can be used as a measurement tool to estimate the proportion α of System I responses and the state-specific component distributions $g_i(t)$. Thereby, one can assess how these parameters change across experimental manipulations or between different populations of participants.[1]

[1] Without additional constraints, the model in Eq. 3.1 is not identifiable. Thus, additional constraints as those discussed in Heck, Erdfelder, and Kieslich (2017) are necessary to ensure unique parameter estimates, which is a prerequisite for its use as a measurement model.

Similar to dual-process theories, MPT models assume that observable behavior is governed by a finite number of cognitive states. For instance, the 2HT model of recognition memory assumes three distinct states (target detection, lure detection, and uncertainty) to explain old-new judgments. Correspondingly, the model predicts that observed RT distributions follow a finite mixture distribution similar to that in Eq. 3.1 (Province & Rouder, 2012). However, the 2HT also places strong constraints on the mixture weights (e.g., the parameter α in Eq. 3.1), because the expected probabilities that OLD and NEW responses are due to detection or guessing are defined by the processing tree structure in Figure 2.1. Based on these theoretical considerations, the question arises how to combine both assumptions — a mixture distribution for RTs and a processing-tree structure for choice frequencies — into a single statistical model.

As a solution, Heck and Erdfelder (2016) proposed to extend the processing tree structure of MPT models by adding latent component distributions to all processing paths. This approach is illustrated graphically in Figure 3.1 for the 2HT model. Overall, the model assumes six distinct processing paths that are mapped to four observable response categories. Each of these paths represents one possible way of processing a presented test item and is thus associated with a separate RT distribution. For instance, the illustration in Figure 3.1 shows that hits due to target detection are relatively fast as indicated by the first component distribution that is reached with probability d_o. In contrast, hits due to guessing are slower as indicated by component distributions in the uncertainty state that are shifted towards slower RTs, in line with the assumption of serial processing (Hu, 2001).

Formally, the extension of the 2HT model in Figure 3.1 results in a probability density function for hits that is defined by the mixture

$$f(t \mid \text{hit}) = \frac{d_o}{d_o + (1-d_o)g}\, g_1(t) + \frac{(1-d_o)g}{d_o + (1-d_o)g}\, g_2(t), \tag{3.2}$$

where g_1 and g_2 are the state-specific component densities for target detection and guessing OLD, respectively. Importantly, the mixture weights are constrained by the MPT structure and not estimated freely as in standard finite-mixture models. In a similar way, the RT distribution of correct rejections is modeled as a mixture of lure detection and guessing NEW. Note that the 2HT model assumes that incorrect responses (i.e., false alarms and misses) only emerge from incorrect guessing in the uncertainty state. Therefore, the observed RT distributions of false alarms and misses are identical to the corresponding component distributions.

In general, RT-extended MPT models define a joint likelihood of discrete choices and continuous RTs by making predictions similar to those in Eq. 3.2. Thereby, they improve upon two-step procedures, in which MPT parameters are first estimated and then used to test predictions about RTs (e.g., Hu, 2001; Province & Rouder, 2012). By modeling choices and RTs jointly, a single statistical model is obtained that can be fitted, tested, and compared against competing models within a coherent statistical

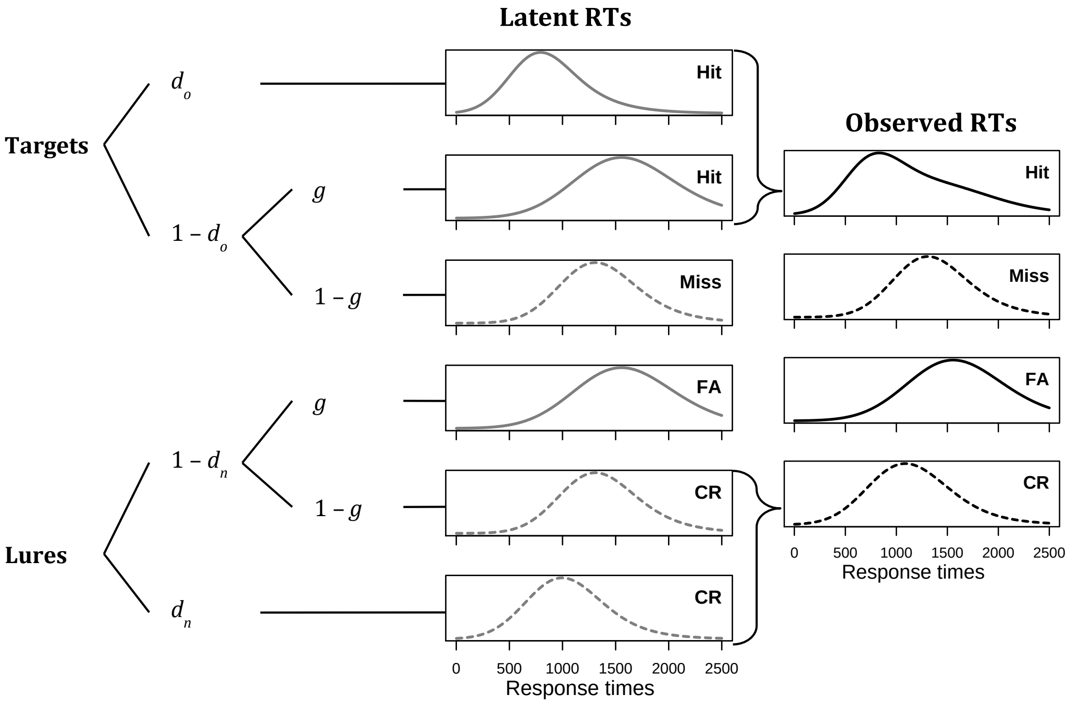

FIGURE 3.1: The 2HT model implies that observed RTs follow a mixture distribution, in which the mixture weights are constrained by the branch probabilities of the MPT structure.

framework (e.g., using maximum likelihood or Bayesian inference). However, before fitting such an RT-extended MPT model, it is necessary to decide how to model the component distributions $g_i(t)$. In the present section, I will outline the method proposed in Heck and Erdfelder (2016) that models the component distributions without assuming specific parametric shapes. In Section 3.3, I present an alternative approach that models the component densities by specific parametric families (e.g., Gaussian distributions). However, both approaches are built on the same core structure: The processing tree of an MPT model is expanded by assuming separate component distributions for each processing path (cf. Figure 3.1).

To model the component distributions $g_i(t)$ without parametric assumptions, the continuous distributions in Figure 3.1 can be approximated by histograms of the RTs (Yantis et al., 1991). Thereby, instead of estimating the component density functions $g_i(t)$, the method estimates the bin probabilities of the corresponding histograms. Formally, the component distributions are modeled by parameters L_{jb} defined as the probabilities that responses from processing branch j fall into the b-th RT interval as illustrated in Figure 3.2. This flexible approach allows researchers to model RT distributions more or less fine-grained by adjusting the number of RT intervals B. For instance, if the number of observations is small, responses can be categorized merely as "fast" or "slow" using $B = 2$ RT bins. In this case, the parameters L_{j1} are defined as the probabilities that responses from process j are "fast" and thus provide direct estimates of the relative speed of the latent processes.

The use of histograms has the benefit that the RT-extended MPT model is also

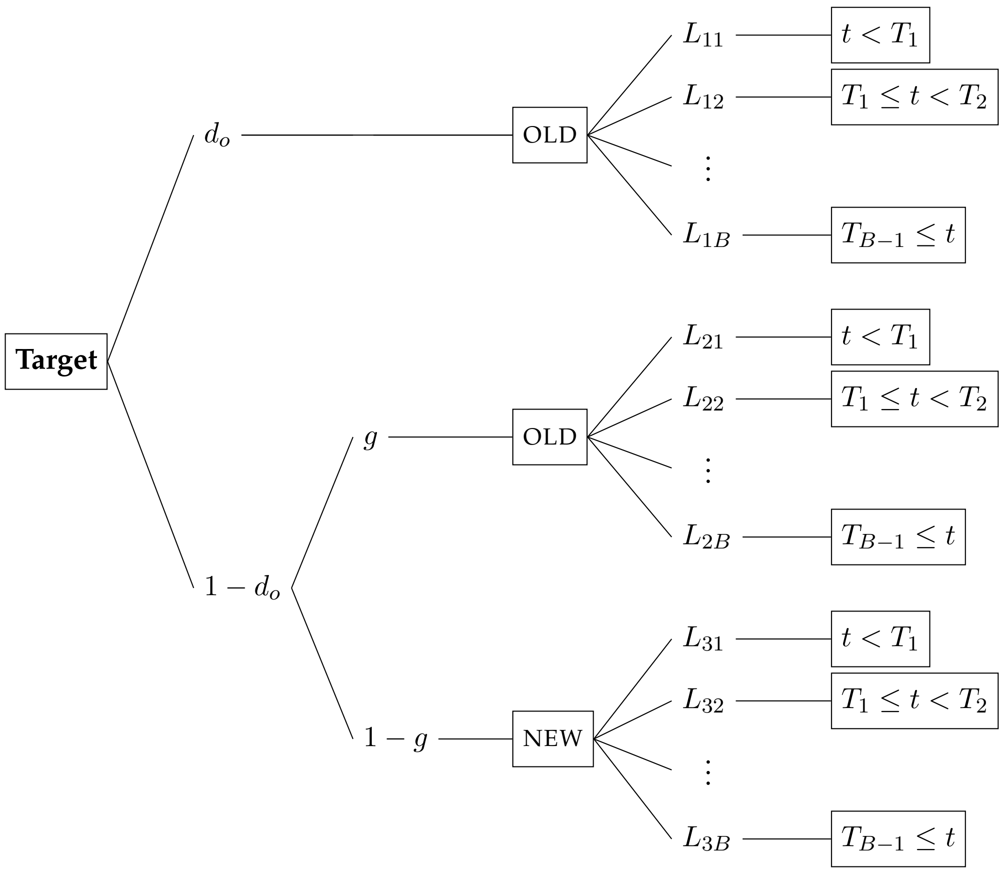

FIGURE 3.2: In the RT-extended 2HT model, the latent component distribution of branch j are modeled by the parameters L_{jb}, defined as the probabilities that RTs t are in the b-th interval.

included in the class of MPT models (Heck & Erdfelder, 2016). This is illustrated in Figure 3.2 for the 2HT model: Essentially, the original response categories OLD and NEW are split into more fine-grained categories depending on the observed RTs t per trial. In such a discretization of the RT distributions, the RT boundaries T_b define the intervals of the histograms and are identical across all response categories of a tree model. For instance, the parameter L_{11} gives the probability that target-detection responses are faster than the lowest RT boundary (i.e., $t < T_1$). Since the new model is an MPT model, its parameters can be estimated with existing software for MPT modeling (Heck et al., in press; Moshagen, 2010; Singmann & Kellen, 2013). Similarly, sophisticated model-selection methods originally implemented for standard MPT models can be directly used for any RT-extended MPT model (e.g., minimum description length or the Bayes factor; Heck et al., 2014; Heck & Wagenmakers, 2016; Vandekerckhove, Matzke, & Wagenmakers, 2015; Wu et al., 2010a).

Even though the reliance on histograms solves the issue of how to estimate the component distributions, it directly leads to a new question, namely, how to select the RT boundaries T_b. In the first paper of my thesis (Heck & Erdfelder, 2016), several strategies and their advantages and disadvantages are discussed. According to the most promising approach that is used for the empirical analyses in Heck and

Erdfelder (2016) and Heck and Erdfelder (in press), RT boundaries are computed separately for each participant as a function of the observed RTs. Thereby, inter-individual differences in absolute speed of the hypothesized cognitive processes are accounted for, whereas the relative speed can still be estimated. For instance, when using $B = 2$ RT intervals, the boundary T_1 is defined as the geometric mean RT per participant. Correspondingly, the parameters L_{j1} are interpreted as the probability that responses from processing path j are faster than the geometric mean RT per participant. Note that different strategies of defining RT boundaries (e.g., using the arithmetic mean or the median per person) lead to identical substantive conclusions in the empirical analysis, which shows that the method is robust with respect to this choice.

An important statistical issue that has to be solved before fitting and evaluating specific RT-extended MPT models concerns their identifiability, that is, whether unique parameter estimates can be obtained. Technically, an MPT model is identifiable, if the function that maps the parameters θ to the predicted category probabilities $p(\theta)$ is one-to-one (i.e., $p(\theta) = p(\theta')$ implies $\theta = \theta'$; Bamber & van Santen, 2000). In general, proving the identifiability of MPT models is difficult. Whereas analytical solutions are available for some models (e.g., Batchelder & Riefer, 1990; Meiser, 2005), heuristics and numerical strategies are often used to check identifiability for more complex models (Moshagen, 2010; Schmittmann, Dolan, Raijmakers, & Batchelder, 2010). However, given an identifiable MPT model, we proved that a simple counting rule or a more general matrix approach can be used to check whether a corresponding RT-extended MPT model is identifiable (see Appendix A in Heck & Erdfelder, 2016).

In a first application of the novel method, we proposed an identifiable, RT-extended version of the 2HT model (Heck & Erdfelder, 2016). In addition to the core structure illustrated in Figure 3.2, the model constrains the RT distribution of guessing OLD to be identical for targets and lures (the same constraint holds for the RT distribution of guessing NEW). The new model provided a good fit to recognition-memory data and allows researchers to test which component distributions are affected by experimental manipulations of memory strength and response bias. Note that this model applies to responses with an emphasis on accuracy (and not on speed) and thus needs to be adapted to account for speed-accuracy trade-offs, for instance, by adding an additional process of fast guessing that increases the speed and simultaneously decreases the accuracy of responses (Ollman, 1966; Yellott, 1971).

A major strength of the new RT-extended 2HT model is its ability to estimate and compare the relative speed of detection and guessing. According to a serial interpretation of the 2HT model, guesses are only made conditional on unsuccessful detection attempts and thus must be stochastically slower than responses due to item detection (Dube et al., 2012; Erdfelder, Küpper-Tetzel, & Mattern, 2011). Formally, this prediction implies that the cumulative density of detection RTs is above that of guessing RTs,

$$F_{\text{detect}}(t) > F_{\text{guess}}(t) \quad \text{for all } t \in \mathcal{R}^+, \tag{3.3}$$

a condition termed stochastic dominance. Note that stochastic dominance is stronger than the corresponding inequality on mean RTs (Heathcote, Brown, Wagenmakers, & Eidels, 2010). To test the assumption of serial detection and guessing states, Heck and Erdfelder (2016) fitted the RT-extended 2HT model in a Bayesian framework using $B = 8$ RT bins. Thereby, the Bayes factor can be computed to quantify the evidence in favor of stochastic dominance based on the method of Heathcote et al. (2010). The empirical results showed that target detection was stochastically faster than guessing, whereas results were ambiguous concerning the relative speed of lure detection. Note that the formal definition of stochastic dominance is relevant for all psychological theories that make predictions about the relative speed of latent processes. This includes dual-process theories that assume faster System I than System II responses (Evans, 2008; Sloman, 1996), but also "fast and frugal heuristics" (Gigerenzer et al., 1999; Gigerenzer & Gaissmaier, 2011) that assume a strictly serial sequence of processing steps and are the topic of the second paper of my thesis discussed in the next section.

3.2 Linking Process and Measurement Models

Heck, D. W. & Erdfelder, E. (in press). Linking process and measurement models of recognition-based decisions. *Psychological Review*. doi:10.1037/rev0000063

Often, people have to make decisions based on incomplete knowledge. For instance, when being asked which of two cities is more populous, people might not know the exact city populations. In such cases, one has to draw inferences based on probabilistic cues, which are features of the choice alternatives that are informative with respect to the criterion of interest (Brunswik, 1955). Regarding population size, one might know that a city has an airport, a metro, and a university and therefore conclude that the city is larger than cities that do not have these cues (Gigerenzer, Hoffrage, & Kleinbölting, 1991). The accuracy of judgments depends on the cue validities defined as the probability that an option having the cue has a higher criterion value than an option not having it. However, theories of judgment and decision making differ in their assumptions how multiple probabilistic cues are integrated to arrive at an overall judgment.

The recognition heuristic (RH; Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002) is a process model of memory-based decisions that applies in the above scenario when one of the presented objects is recognized and the other is not. In such cases, the RH states that the recognized object is chosen without consideration of any further knowledge. Since larger cities are more likely to be recognized, this heuristic will often result in the correct answer while requiring only the recognition status of the presented objects (Pachur, 2010). Therefore, the RH adheres to the principle of ecological rationality (Gigerenzer & Gaissmaier, 2011), because it exploits an existing correlation in the environment (i.e., larger cities are recognized with higher likelihood) to draw accurate inferences while reducing the effort of integrating available knowledge. In the following, I outline how the novel method presented in Section 3.1 (Heck & Erdfelder, 2016) provides a strong test of the RH as a process model of recognition-based decisions (Heck & Erdfelder, in press). Essentially, the critical test concerns the relative speed of relying on the RH, which is estimated by extending an MPT model for the standard RH paradigm to RTs.

A consensus has been reached that recognition is a very valid cue in many natural environments and domains (e.g., when judging city size; Pachur, 2010) and that it affects decision making to a substantial degree (i.e., recognized cities are usually chosen more often than unrecognized cities; Glöckner & Bröder, 2014; Goldstein & Gigerenzer, 2011). However, a controversy has evolved concerning the question whether the RH is a good process model of the underlying cognitive mechanism of how people integrate recognition information with further knowledge. As one of its core properties, the RH is noncompensatory, because it assumes that judgments rely only on recognition and are not affected by any further knowledge that might be available (Pachur, Bröder, & Marewski, 2008). In contrast, competing theories assume that recognition is integrated with further knowledge in a compensatory way where recognition only serves as one probabilistic cue among many others (Glöckner & Bröder, 2011; Hilbig & Pohl, 2009; Newell & Shanks, 2004). Whereas compensatory theories differ slightly how multiple cues representing recognition and knowledge are weighted and combined, they all assume that several less valid cues can in principle overrule one highly valid cue (whether this is possible depends on the environmental cue structure; Glöckner, Hilbig, & Jekel, 2014). Thereby, compensatory theories directly explain why the unrecognized option is sometimes chosen in memory-based decisions (Hilbig & Richter, 2011). If a participant has cue knowledge indicating that the recognized city has a small population (e.g., the city has no airport and no metro), these cues can overrule the recognition cue (which indicates that the recognized city is larger). Overall, a trade-off of this conflicting information can therefore result in the judgment that the unrecognized option is the larger city (Heck & Erdfelder, in press).

To test whether recognition information is integrated in a noncompensatory or compensatory way, researchers have manipulated both recognition and further knowledge to test how both factors influence judgments (e.g., Glöckner & Bröder, 2011; Hilbig, 2010b; Hochman, Ayal, & Glöckner, 2010; Oppenheimer, 2003; Richter & Späth, 2006). Many of these studies showed that further knowledge about the objects under consideration does indeed influence choices, decision times, and confidence ratings. However, much of this evidence has been disregarded by proponents of the RH arguing that the heuristic applies only to judgments in natural environments and domains — a requirement that is violated when recognition or further knowledge is experimentally manipulated (Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011).

To measure RH use in natural domains (e.g., for existing cities without presentation of further cues), Hilbig, Erdfelder, and Pohl (2010) proposed the r-model, an MPT model for estimating the probability r that choices are due to noncompensatory use of recognition. Essentially, the r-model requires information from two experimental phases. In the recognition phase, participants judge cities as being recognized or unknown; and in the decision phase, participants are presented with all pairwise combinations of cities and have to decide which one is larger. Based on this paradigm, the parameters of the r-model have been validated experimentally (Hilbig et al., 2010) and then used to estimate the probability of using recognition in a noncompensatory way across experimental manipulations (e.g., time pressure or availability of further information; Hilbig, Erdfelder, & Pohl, 2012; Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015; Pohl, Michalkiewicz, Erdfelder, & Hilbig, in press) or specific populations (e.g., older and younger adults; Horn, Pachur, & Mata, 2015; Michalkiewicz & Erdfelder, 2016).

Despite its usefulness as a measurement model, the r-model cannot distinguish between process models that assume different mechanisms of how information is integrated (Heck & Erdfelder, in press). This is due to the fact that process models of compensatory information integration make the same predictions as the RH when no further knowledge is available, in which case the core question about the integration of multiple cues becomes inconsequential. However, if experimental manipulations of the two factors of interest (i.e., recognition and further knowledge) violate the principle of ecological validity, and if the r-model cannot distinguish between different process models, the question arises whether the RH still has empirical content and makes predictions that can in principle be falsified (Glöckner & Betsch, 2011; Pohl, 2011).

To test competing process models underlying recognition-based decisions, an RT-extended r-model is proposed in the second paper of my thesis (Heck & Erdfelder, in press) using the novel method outlined in Section 3.1 (Heck & Erdfelder, 2016). Essentially, the new model allows researchers to estimate the relative speed of information integration when (a) recognition is used in isolation (R-only), (b) recognition is integrated with further knowledge indicating that the recognized city is larger (R-congruent), and (c) recognition is integrated with conflicting knowledge indicating that the recognized city is smaller (R-incongruent). Conditional on these three latent states, a serial-process interpretation of the RH implies that decisions are stochastically faster when relying only on recognition (R-only) compared to decisions in which recognition is integrated with further knowledge (R-congruent or R-incongruent). This prediction directly follows from the fact that serial information integration requires additional elementary processing steps to evaluate available knowledge and combine it with recognition (Heck & Erdfelder, in press; Hilbig & Pohl, 2009; Payne, Bettman, & Johnson, 1993).

In contrast to the serial RH, competing theories assume that information from multiple cues is integrated automatically and in parallel (Betsch & Glöckner, 2010; Glöckner & Betsch, 2008) and predict that R-congruent responses are actually faster than R-only responses. This follows directly from process models such as parallel constraint satisfaction (PCS) theory (Glöckner & Betsch, 2008) that provide an exact mathematical description of information integration (i.e., by searching for state of maximal coherence in a neural network). Correspondingly, judgments are predicted to be faster and more confident when additional information also indicates that the recognized city is larger (e.g., if one knows that it is a capital, has an airport and a metro) compared to situations in which only recognition is available to draw inferences (Betsch & Glöckner, 2010).

Importantly, data from the standard RH paradigm, which includes only a recognition and a decision phase, are sufficient to fit and test the RT-extended r-model. Hence, the new measurement model circumvents the necessity to manipulate recognition or knowledge experimentally and thereby allows for testing the RH in natural domains based on pre-experimental knowledge. Moreover, the RT-extended r-model can be directly linked to process models of the RH by simulations (Heck & Erdfelder, in press). First, recognition judgments and binary decisions are generated either by the serial RH or by PCS. In a second step, the RT-extended r-model is fitted to the simulated RTs and choices. In such simulations, both process models — the serial RH and PCS — can generate similar choice frequencies resulting in equivalent parameter estimates for the standard r-model of choice frequencies irrespective of the data-generating model. However, the RT-extended r-model has the power to discriminate data generated by the two process models by comparing the estimates for the relative speed of the latent processes. When simulating data with the serial RH model, R-only responses are estimated to be stochastically faster than those based on integration of further knowledge (i.e., R-congruent or R-incongruent). In contrast, when simulating data using PCS, R-only responses are estimated to be stochastically *slower* than R-congruent responses for which further knowledge is coherent with the recognition cue.

Based on these theoretical derivations and simulation results, we applied the RT-extended r-model in an empirical reanalysis of 29 data sets including approximately 400,000 recognition-based decisions. To increase the validity and robustness of the results, the new model was fitted and tested in three ways. First, a hierarchical Bayesian MPT model was fitted using only two RT bins, thereby taking the nested data structure into account (i.e., decisions nested in individuals, and individuals nested in data sets; Heck et al., in press). Second, stochastic dominance of RTs was tested on a more fine-grained scale using 40 RT bins and assuming independent and identically distributed trials across participants and data sets. Finally, again using two RT bins, each participant was classified as RH or PCS user based on model-selection methods that take order constraints into account (i.e., minimum description length; Heck, Hilbig, & Moshagen, 2017; Hilbig & Moshagen, 2014). All of these analyses consistently showed that decisions based on further recognition-congruent knowledge are stochastically faster compared to those based on recognition only. Hence, the results clearly support theories of compensatory information integration as implemented by the PCS model and falsify the serial RH account.

Overall, the second paper of my thesis (Heck & Erdfelder, in press) highlights the importance of including RTs in MPT models. By considering the relative speed of cognitive processes, novel predictions can be derived to test competing theoretical accounts. More generally, this approach shows how process and measurement models can be linked to advance our understanding of cognitive mechanisms. Whereas process models such as the serial RH or PCS provide exact predictions and describe the underlying mechanisms in great detail, measurement models such as the RT-extended r-model provide estimates of parameters of interest given limited information (e.g., without knowing exact cue patterns). Within such an integrative view, process and measurement models that have been used in isolation can in principle be combined to arrive at a deeper understanding of cognitive processes without the necessity of experimental manipulations of the processes of interest.

3.3 Generalized Processing Tree Models

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2017). *Generalized processing tree models: Jointly modeling discrete and continuous variables*. Manuscript submitted for publication.

The third paper of my thesis presents the new class of *generalized processing tree* (GPT) models that account for discrete responses and continuous variables jointly and thereby extend MPT models.[2] Similar to the method presented in Section 3.1 (Heck & Erdfelder, 2016), GPT models also define a processing tree that determines the probabilities of observed responses and assume that each processing path is associated with a separate component distribution of one or more continuous variables. However, GPT models make the additional constraint that parameterized distributions instead of unconstrained histograms are defined for the latent component distributions. For instance, RTs may be modeled by ex-Gaussian distributions (defined as the sum of a normal and an independent exponential random variable; Matzke & Wagenmakers, 2009), whereas other variables such as continuous confidence ratings may be better described by normal or beta distributions. Note that some or all of the corresponding parameters can be constrained across latent states to test theoretical predictions or to reduce the number of nuisance parameters (e.g., by assuming that the variance of confidence ratings is constant across latent states).

Irrespective of the specific parametric distributions, the likelihood function of GPT models depends on two sets of parameters. First, the parameters $\boldsymbol{\theta}$ are defined as conditional or unconditional probabilities of the latent processes and are thus equivalent to parameters in standard MPT models. Second, $\boldsymbol{\eta}$ includes all parameters of the latent component distributions (e.g., means and variances of one or more normal distributions). Note that this decomposition into two sets of parameters assumes a specific type of functional independence. Essentially, the parameters $\boldsymbol{\theta}$ determine

[2] Hu and Batchelder (1994) previously used the abbreviation GPT for *general processing tree* models to emphasize the precise statistical and mathematical representation of MPT models. However, in the recent literature, the latter abbreviation is used almost exclusively.

only the probabilities of following one of the processing paths (i.e., the weights of the finite mixture distribution). In contrast, the parameters $\boldsymbol{\eta}$ determine only the density function of the continuous variables conditional on the latent states (i.e., the component distributions of the mixture distribution). The assumption of functional independence constrains the statistical model, is often substantively reasonable, and facilitates the derivation of an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) for maximum-likelihood estimation of GPT models (Heck, Erdfelder, & Kieslich, 2017).[3]

The use of parameterized component distributions instead of histograms as in Heck and Erdfelder (2016) has several benefits. First, it does not require categorization of continuous observations into intervals, an approach that results in a loss of information and thus in a decrease of statistical efficiency. Second, the parameters $\boldsymbol{\theta}$ of a GPT model can be identifiable even if the corresponding MPT model for response frequencies only is not identifiable. This is illustrated in the empirical example in Heck, Erdfelder, and Kieslich (2017), where explicit assumptions about the parametric component distributions allow to obtain unique parameter estimates for a GPT model with more parameters $\boldsymbol{\theta}$ than free categories. Third, the inclusion of continuous variables can result in smaller standard errors of the GPT estimates compared to those of an equivalent MPT model. This effect, shown in a simulation in the third paper of my thesis, can be intuitively explained within the expectation-maximization algorithm (EM; Dempster et al., 1977). To fit a model, the EM algorithm estimates the latent states for all observations and maximizes the model parameters conditional on these latent-state estimates in alternating order. If the continuous component distributions strongly differ across states, the continuous variables are very informative for estimating the latent states, which results in more precise parameter estimates. For instance, in case of the 2HT model, detection and guessing responses can be better discriminated if the corresponding RTs are very fast and very slow, respectively, compared to the case that RTs are identically distributed.

Besides these statistical advantages, GPT models allow researchers to model multiple continuous variables by using multivariate component distributions. For instance, an extension of MPT models to RTs and confidence ratings is obtained by assuming independence of the continuous variables conditional on the latent states (Heck, Erdfelder, & Kieslich, 2017). Based in this assumption, it is sufficient to define univariate component distributions for each of the variables in isolation (e.g., ex-Gaussian and beta distributions for RTs and confidence ratings, respectively). As a consequence, this assumption implies that correlations between two or more continuous variables emerge only from aggregating across multiple latent states. For instance, in recognition memory, observed RTs are often positively correlated with confidence ratings (Ratcliff & Starns, 2013; Weidemann & Kahana, 2016). In a corresponding GPT version of the 2HT model, both variables can be assumed to be

[3] The MPT-RT approach (Heck & Erdfelder, 2016) also makes this assumption of functional independence.

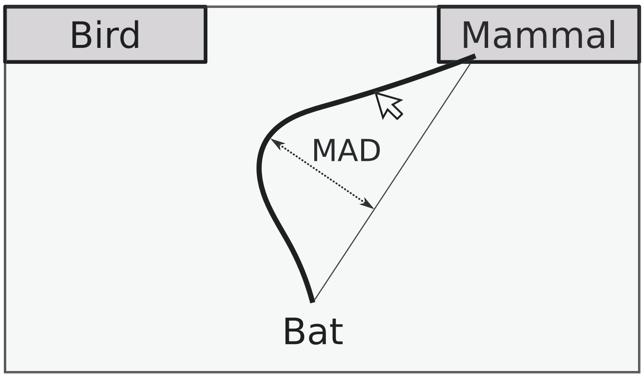

FIGURE 3.3: In the mouse-tracking version of a semantic categorization tasks, participants' mouse movements are recorded during the decision process. The maximum absolute deviation (MAD) is defined as the maximum perpendicular distance between the observed and a direct trajectory.

independent conditional on the detection and guessing state. However, if detection responses are both faster and more confident, and guessing responses are both slower and less confident, aggregation across these two latent states results in a positive correlation of RTs and confidence ratings.

A major pragmatic advantage of GPT models is that they are easily adapted to different experimental paradigms similar to MPT models. As a consequence of their mathematical and conceptual simplicity, GPT models can be defined in a text file by listing all of the processing paths and the parameterization of the corresponding component distributions. Importantly, this intuitive modeling syntax developed in Heck, Erdfelder, and Kieslich (2017) is sufficient to completely specify any GPT model. This is due to the fact that the complexity of GPT models is constrained by simplifying assumptions such as the separation into two disjoint sets of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ for discrete and continuous variables, respectively. Based on this syntax, the R package `gpt` implements the EM algorithm proposed in Heck, Erdfelder, and Kieslich (2017) and thereby facilitates model development, fitting, and assessment.

As a first empirical application of the new model class, Heck, Erdfelder, and Kieslich (2017) tested a GPT model for semantic categorization. In the study by Kieslich and Henninger (in press), participants had to judge the category membership of animals (e.g., whether a *bat* is a *bird* or a *mammal*). In addition, the trajectory of the mouse cursor was measured during the decision process. As illustrated in Figure 3.3, the position of the mouse cursor from its start at the bottom center to its end at the chosen alternative (either at the upper left or at the upper right corner) was recorded with a constant sampling rate. The fundamental assumption underlying this mouse-tracking methodology is that the underlying cognitive processes directly affect the curvature of the observed trajectories (Dale, Kehoe, & Spivey, 2007). In case of semantic categorization of animals, mouse movements should be attracted more strongly towards the incorrect category when processing atypical compared to typical animals even when the final choice is correct.

Usually, data from mouse-tracking studies are analyzed by (a) summarizing each trajectory by a descriptive statistic (e.g., the MAD, defined as the maximum absolute deviation of a trajectory from a straight line connecting its start and end points) and (b) comparing means of these statistics across experimental conditions (e.g., performing a t-test of the MADs of correct responses for typical vs. atypical animals; Freeman & Ambady, 2010; Kieslich & Henninger, in press). However, such an analysis ignores valuable information that is contained in the accuracy of judgments, the MADs of incorrect responses, and the shape of the MAD distribution. As a remedy, Heck, Erdfelder, and Kieslich (2017) proposed a GPT model that provides a complete account of all data based on the feature comparison model (FCM; Smith, Shoben, & Rips, 1974). Essentially, the FCM assumes that two qualitatively different processes can be used for categorization. On some trials, a similarity-based overall comparison process provides clear evidence in favor of category membership, and thus responses are given relatively fast and direct, resulting in small MADs of the mouse trajectories. In contrast, on trials where this process does not allow for a clear judgment, only the defining features of animals and categories are compared, which is a more time-consuming and less direct process, resulting in larger MADs.

The proposed GPT model for semantic categorization had a very good fit to the accuracy of judgments and the MAD distributions. Importantly, the GPT model also allowed for testing several theoretical predictions of the FCM. For instance, the accuracy of the first, overall comparison process should be higher for typical than for atypical animals, whereas accuracy of the second, defining-feature comparison process should not be affected by typicality (Smith et al., 1974). Empirically, the parameter estimates were in line with these predictions, thereby contributing to the psychological validity of the proposed GPT model. Note that such process-specific predictions cannot be directly tested using ad hoc measures (e.g., the mean difference in MADs between typical and atypical animals). Therefore, this empirical application of GPT models to the mouse-tracking paradigm highlights one direction of linking psychological theory to observable outcomes (i.e., discrete responses and MADs). Moreover, the new model may serve as a prototype for testing dual-process theories (Evans, 2008; Sloman, 1996), which are related to the FCM both conceptually and historically (Smith & Sloman, 1994).

In sum, GPT models extend MPT models to continuous variables and thereby provide an alternative to the approach presented in Section 3.1 (Heck & Erdfelder, 2016). By specifying parametric distributions conditional on each processing path, GPT models make more specific predictions, can provide more precise parameter estimates than MPT models, and facilitate the inclusion of multiple continuous variables. However, as for all statistical models, GPT models may provide biased estimates and invalid conclusions if the model's assumptions are violated. Besides the core assumption of a finite number of cognitive states, GPT models add strong constraints regarding the distributional shape of continuous variables conditional on the latent states. Ideally, substantive conclusions should not be affected when fitting a GPT model with misspecified parametric distributions (e.g., assuming log-normal instead of ex-Gaussian distributions). Therefore, future work should address the question how strongly these parametric assumptions affect parameter estimates. Similarly, more theoretical and empirical work is necessary to test whether GPT models are useful and provide valid explanations of psychological phenomena. As a first step, the formalization of GPT models in Heck, Erdfelder, and Kieslich (2017) lays out the statistical foundations for future developments of this new model class.

4 Discussion

4.1 Advantages, Limitations, and Open Questions

In my thesis, I developed and tested two new methods that extend the scope of multinomial processing tree models to jointly account for discrete and continuous data. In the first approach, RTs are categorized into a finite number of intervals, thereby defining a new, RT-extended MPT model that allows for estimating histograms of the latent component distributions (Heck & Erdfelder, 2016). As a first application, we proposed an RT-extended version of the 2HT model of recognition memory to assess the relative speed of target and lure detection and guessing. Based on this method, an RT-extended version of the r-model (Hilbig et al., 2010) was used in the second paper to test two competing process models of recognition-based decisions (Heck & Erdfelder, in press). As an alternative, I defined the new class of *generalized* processing tree (GPT) models that assume parametric component distributions instead of histograms, and applied this approach to account for semantic categorization in a mouse-tracking paradigm (Heck, Erdfelder, & Kieslich, 2017). Together, the two new methods provide a novel theoretical and statistical framework for testing discrete-state models of cognition.

Given the conceptual similarity of RT-extended MPT and GPT models, both approaches share several advantages. First, they are mathematically tractable, can easily be adapted to different paradigms, and are conceptually similar to MPT models, thereby reducing the obstacles for researchers to extend existing multinomial models to continuous variables. Second, both approaches define a joint likelihood function of discrete and continuous variables. Therefore, statistical inference is based on a single model instead of ad hoc methods such as separately fitting an MPT model and analyzing the continuous variable (e.g., Hu, 2001; Province & Rouder, 2012). Most importantly, the usefulness of RT-extended MPT and GPT models has already been demonstrated in several empirical applications across different substantive domains including recognition memory (Heck & Erdfelder, 2016), remember-know judgments (Li, 2015), memory-based decisions (Heck & Erdfelder, in press), and semantic categorization (Heck, Erdfelder, & Kieslich, 2017).

An open question concerns the statistical efficiency and power of RT-extended MPT and GPT models, both in absolute terms and relative to each other. Some preliminary simulations showed that the former class of models has sufficient power to detect differences in latent component distributions using realistic sample sizes (Heck & Erdfelder, 2016). To obtain a better understanding of the two frameworks, systematic simulations are required to provide guidelines for the number of trials necessary to obtain sufficiently precise parameter estimates. For instance, simulated data can be generated using a GPT model with specific parametric component distributions (e.g., ex-Gaussians) and for different effect sizes (e.g., by manipulating the means of the underlying distributions). By fitting RT-extended MPT and GPT models to these data, the statistical power to detect differences in the latent distributions can be estimated and compared for both approaches. However, it is very likely that the statistical efficiency and power is idiosyncratic to specific MPT structures, parametric assumptions, and parameter values. Therefore, the most promising solution is to perform new simulations for each model and empirical scenario of interest taking specific details of the model and the paradigm into account. Technically, the implementation of such simulations is straightforward because RT-extended MPT and GPT models are easily specified and fitted by user-friendly software such as the R package `gpt` (Heck, Erdfelder, & Kieslich, 2017).

Simulations also facilitate the assessment of the robustness of both methods against misspecification. In the first and third papers of my thesis (Heck & Erdfelder, 2016; Heck, Erdfelder, & Kieslich, 2017), RT-extended MPT and GPT models were fitted assuming independent and identically distributed observations. However, this assumption is often violated in practice due to heterogeneity across participants, stimuli, or data sets resulting in biased statistical inferences (Klauer, 2006). As a remedy, Bayesian hierarchical models account for such differences explicitly (Lee, 2011). For RT-extended MPT models, hierarchical extensions can be directly fitted using the R package `TreeBUGS` (Heck et al., in press) as demonstrated in Heck and Erdfelder (in press). In principle, it is straightforward to develop similar hierarchical extensions of GPT models by assuming a group-level distribution of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Once a hierarchical structure and a set of prior distributions has been defined, the GPT model can be implemented and fitted in software such as JAGS (Plummer, 2003). An important issue specific to GPT models is their robustness with respect to misspecification of the latent component distributions. Ideally, the validity of substantive conclusions (e.g., whether two component distributions are identical) should not hinge on specific parametric distributions as long as they share some core features (e.g., left-skeweness as for the ex-Gaussian, Wald, and other standard RT distributions; Heathcote, 2004). Similarly, it is desirable that substantive conclusions based on RT-extended MPT models are robust with respect to the number of RT bins and the exact locations of the RT boundaries. Even though preliminary simulations and robustness checks indicate that this is indeed the case (Heck & Erdfelder, 2016), future work should address this question more systematically.

Besides considerations about statistical power and robustness, the psychological research question should determine whether to use an RT-extended MPT or a GPT model. If the focus is on testing the core assumption that RTs follow a mixture distribution (Yantis et al., 1991) or on comparing the relative speed of latent processes (Heck & Erdfelder, in press), RT-extended MPT models are advantageous, because they do not require auxiliary assumptions about the parametric shape of the component distributions. This is especially important for MPT models, since they may often include latent component distributions that cannot be observed in isolation. For instance, in the 2HTM (cf. Figure 3.1), the latent RT distribution of detection cannot be observed directly, because the observed RT distribution of correct responses is contaminated by guessing RTs. In such cases, it is not possible to directly test parametric assumptions by comparing the observed against the hypothesized distribution. As a remedy, observed RT distributions can be compared against the fitted mixture distribution of a GPT model, thereby testing parametric assumptions and the mixture structure simultaneously. However, if psychological theory predicts specific distributional shapes for one or more continuous variables, the GPT approach allows to include this additional information in the model. A GPT model allows for a stricter test of the theory, provides a more concise summery of the data, and is likely to result in an increase of statistical efficiency. Moreover, GPT models can provide unique parameter estimates in paradigms in which a corresponding MPT or RT-extended MPT model would not be identifiable (Heck, Erdfelder, & Kieslich, 2017).

A possible limitation of both RT-extended MPT and GPT models concerns their conceptualization as measurement models. Essentially, both approaches can be interpreted as tools that allow researchers to estimate and compare component distributions conditional on latent states. Put differently, Morey (2017) summarized RT-extended MPT models as "mathematical tools they use for unmixing the response time distributions." This objective of the two modeling approaches differs from that of process models which describe the exact mechanism of information processing. In future work, GPT models can be extended in this direction by making stronger assumptions about the underlying mechanisms.

For instance, Donkin, Nosofsky, Gold, and Shiffrin (2013) proposed a discrete-slots model of visual working memory for responses and RTs that is closely related to the 2HT model. At its core, this model assumes that response behavior is determined either by a detection or by a guessing process that are modeled by two separate evidence accumulation processes. More precisely, the underlying linear-ballistic accumulator model assumes that evidence for the two response options is accumulated continuously in time, and that a choice is made once the corresponding accumulator reaches a response threshold (e.g., participants respond OLD when the corresponding accumulator is faster than that of the NEW response; Brown & Heathcote, 2008). Thereby, evidence-accumulation models predict the joint distribution of a discrete and a continuous variable based on a shared set of parameters. As a consequence, a change in one parameter affects both observable variables simultaneously, for instance, a lower response threshold results in less accurate responses and faster RTs (Donkin et al., 2013). In contrast, GPT models assume discrete latent states and rely on separate sets of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ to model the discrete and continuous variables, respectively. In future work, evidence-accumulation and GPT models could be combined in a general framework similar to the model of Donkin et al. (2013) or competitively tested against each other. Either way, such endeavors will add to our understanding about the exact mathematical representation of latent cognitive processes.

Alternatively, a serial interpretation of MPT models implies that the nodes in the processing tree form a sequence of serial stages (Hu, 2001). In such a process model, observed RTs emerge as the sum of the latent processing times of the traversed stages. In case of the 2HT model in Figure 3.1, this approach assumes four stage-specific parameters: In the first stage, these are the processing times for successful and failed detection attempts and, conditional on being in the uncertainty state, the processing times of guessing OLD and NEW. To obtain a fully specified statistical model, these processing times can be assumed to be exponentially distributed (Klauer, 2015). Interestingly, the resulting model will be a special case of a GPT model with specific component distributions (i.e., general gamma distributions, defined as the sum of independent random variables that are exponentially distributed with different rate parameters; McGill, 1963).[4] By explicitly defining the structure of the component distributions, the resulting model makes more specific predictions (e.g., about the stochastic dominance of the underlying distributions) compared to a model without these constraints. However, if such a model is rejected by empirical data, it is not clear whether this is due to the core assumption of serial processing stages or due to auxiliary assumptions such as the parametric shapes or independence of the latent processing times. In contrast, measurement models such as the RT-extended 2HT model provide a distribution-free test of stochastic dominance of latent processes (Heck & Erdfelder, 2016) that requires less assumptions and can thereby falsify complete classes of process models (Malmberg, 2008), for instance, any process model assuming a specific order of serial processing.

4.2 Future Directions

The extension of existing MPT models to continuous variables allows for new empirical tests of discrete-state theories of cognition. The new methods are especially relevant for modeling RTs, one of the most important variables in cognitive psychology (Luce, 1986). In several previous applications of multinomial models, RTs were left aside or analyzed separately, because the conceptual and statistical foundations for jointly modeling discrete and continuous variables were missing. For instance, several MPT models have been proposed for paradigms that usually focus on RTs rather than choice frequencies as in the implicit association test (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Greenwald, McGhee, & Schwartz, 1998;

[4] Similar to other RT models, an independent and additive nondecision time can be added to the model (Luce, 1986). The corresponding GPT model will have component distributions that are defined by the convolution of a Student's t-distribution for nondecision time with the general gamma distribution for the latent processing times (Klauer, 2015).

Meissner & Rothermund, 2013). Based on the novel approaches presented in my thesis, both accuracy and RTs can be modeled jointly in such paradigms.

Moreover, MPT models have been developed to examine how truth judgments of true and false statements are affected by knowledge and processing fluency, defined as the meta-cognitive experience of the ease of processing (e.g., Fazio, Brashier, Payne, & Marsh, 2015; Hilbig, 2012; Unkelbach & Stahl, 2009). Whereas some models assume that truth judgments only depend on fluency conditional on the absence of conclusive knowledge (Hilbig, 2012; Unkelbach & Stahl, 2009), Fazio et al. (2015) recently proposed that the opposite relation holds, that is, knowledge is only used if truth judgments are not based on processing fluency. By now, these predictions have only been tested by means of MPT models for choice frequencies. However, when assuming that fluency and knowledge are used in a strictly serial manner, the two opposite theories predict different orders of stochastic dominance for the latent RT distributions of the two processes. Since the MPT models of Fazio et al. (2015) and Hilbig (2012) closely resemble the 2HT model, the corresponding RT-extension presented in Heck and Erdfelder (2016) can be adapted to test whether the use of knowledge is stochastically faster than the use of fluency.

Besides RTs, other types of continuous variables such as fine-grained judgments on circular scales can be included in GPT models. A promising candidate for future developments is the two-high threshold model of source monitoring (2HTSM; Batchelder & Riefer, 1990; Bayen et al., 1996). The model is shown in Figure 4.1 and disentangles item memory (whether an item was studied or not), source memory (e.g., whether studied words were presented in blue or red), and guessing. More specifically, the parameters D_A, D_B, and D_N are the probabilities of detecting targets and lures, whereas d_A and d_B are the probabilities of correctly retrieving the source conditional on item memory. Moreover, a is the probability of choosing Source A conditional on item detection and source uncertainty, whereas the parameters b and g jointly define the probabilities of guessing A, B, or N conditional on recognition uncertainty.

Recently, Harlow and Donaldson (2013) tested participants' memory for continuous source information by presenting words jointly with a cross that was located randomly on a circle. When cued with these words in the test phase, participants' memory for the location was best described by a mixture model of a uniform guessing distribution and a symmetric detection distribution concentrated at the correct location. In a follow-up study, Harlow and Yonelinas (2016) further compared participants' confidence ratings for the probability of retrieving the source and for the precision of the location judgments.

In both studies, however, only studied words were presented in the test phase. Therefore, participants had to engage only in cued recall of continuous source information and not in the recognition of the words themselves. To examine the relation of item and source memory, lures have to be included in the test phase, in which case location judgments are only made conditional on OLD responses. The resulting

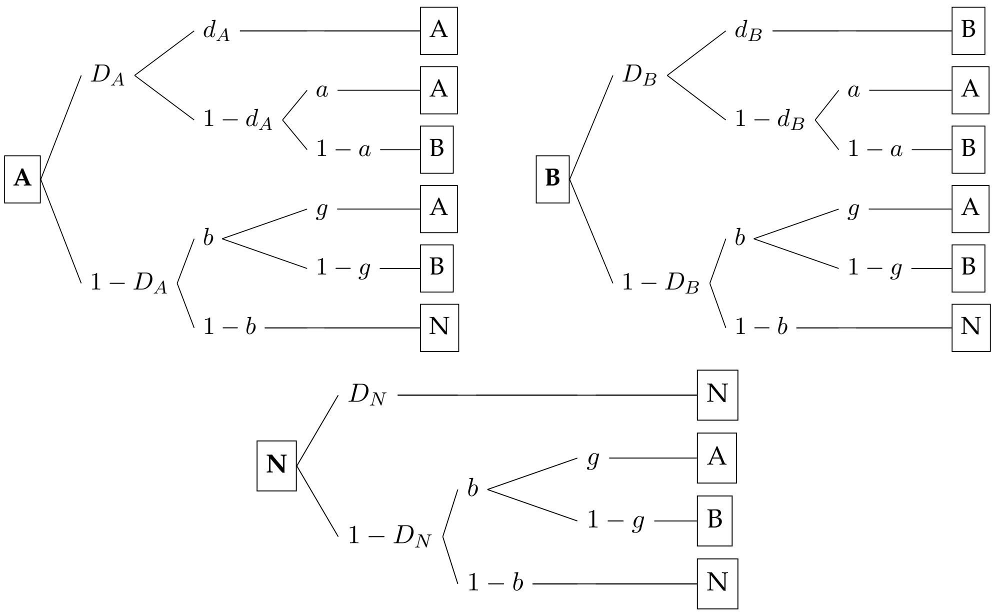

FIGURE 4.1: The two-high threshold model of source monitoring (Bayen, Murnane, & Erdfelder, 1996).

joint distribution of old-new and location judgments can then be modeled by the GPT model for continuous source monitoring (CSM) shown in Figure 4.2.[5] In this model, the continuous variable is defined as the angle between location judgment and correct value measured in radians on the interval $(-\pi, \pi]$. Conditional on item detection D and source recollection d, the CSM model replaces the certainty state for recognizing Source A or B of the 2HTSM in Figure 4.1 by a symmetric distribution with a peak at the correct location. For instance, as illustrated in Figure 4.2, the von Mises distribution for a circular variable can be used with the free parameter κ that determines the concentration around the mean $\mu = 0$ (Oberauer, Stoneking, Wabersich, & Lin, 2017). In contrast, source guessing results in a uniform distribution of location judgments on the interval $(-\pi, \pi]$, because any response biases (e.g., a preference towards the right side of the circle) cancel out when locations are sampled randomly from a uniform distribution.

The CSM model in Figure 4.2 defines the probabilities D for item memory, d for source memory, and the parameter κ for the precision of source memory. If the model holds in empirical validation studies, effects of different memory-strength manipulations on these parameters can be examined. For instance, it can be tested whether the inclusion of old-new judgments affects the probability d and precision κ of source memory compared to the original cued-recall task of Harlow and Donaldson (2013). Moreover, a similar model can be used if location judgments are also made conditional on NEW responses to both targets and lures. In this case, the CSM model predicts thats these judgments are uniformly distributed, because the observable

[5] Technically, the CSM in Figure 4.2 is not a GPT model because continuous observations y_k are missing if $x_k =$ NEW. However, an equivalent GPT model is obtained by assuming a uniform distribution conditional on NEW responses and using dummy values $y_k = 0$. This will only add a constant to the log likelihood and thus not affect parameter estimation.

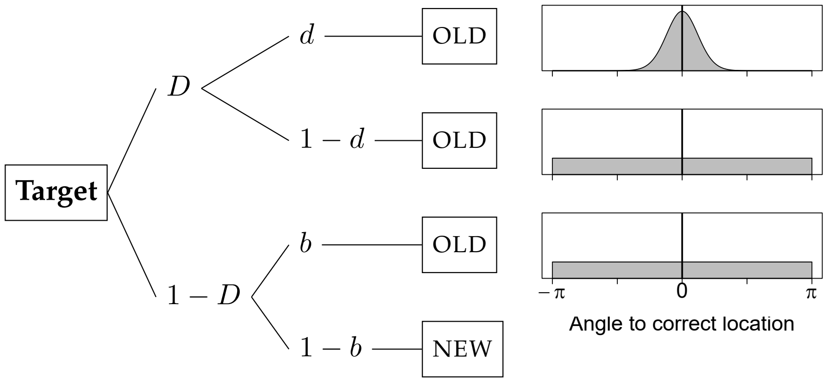

FIGURE 4.2: The two-high threshold model of continuous source monitoring (CSM). The angles between location judgments and correct locations are modeled as a circular variable using the von Mises and uniform distributions on the interval $(-\pi, \pi]$. The processing tree for lures is similar to that in Figure 4.1 with a uniform distribution for location judgments of OLD guesses.

NEW responses indicate non-detection of both the item and the source. This provides a strong test of discrete-state theories, since above-chance source memory for NEW responses is predicted by competing accounts that assume continuous memory signals (Malejka & Bröder, 2016; Starns, Hicks, Brown, & Martin, 2008).

The CSM model can further be extended to examine feature binding across multiple source dimensions (Boywitt & Meiser, 2013; Meiser & Bröder, 2002). For instance, both color and location of studied words can be manipulated independently on continuous scales. To model this task, the probabilities of recognizing the two sources are defined similarly as in other multidimensional source-monitoring models (e.g., by disentangling joint and independent retrieval of color and location; Meiser, 2014). Moreover, conditional on source detection and uncertainty, the two continuous distributions are modeled by von Mises and uniform distributions, respectively, similar to the CSM model in Figure 4.2. By appropriate constraints on the concentration parameters κ for the two source dimensions, it may be possible to test whether feature binding improves source memory for color and location. In sum, the example of the CSM model highlights the usefulness of GPT models for modeling continuous variables such as fine-grained judgments on circular scales.

4.3 Conclusion

In my thesis, I have proposed new methods to model the joint distribution of discrete and continuous variables assuming a finite number of processing states. These RT-extended MPT and GPT models provide the conceptual and statistical foundations for novel empirical tests of discrete-state theories. Both approaches naturally extend MPT models, and their application has already provided new insights into memory, decision-making, and semantic-categorization processes. Given the popularity of discrete-state theories in psychology, future applications of RT-extended MPT and GPT models have the potential to contribute to our understanding of cognitive processes.

5 References

Bamber, D. & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. doi:10.1006/jmps.1999.1275

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. doi:10.1037/1040-3590.10.4.331

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

Batchelder, W. H. & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129–149. doi: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. doi: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. doi:10.3758/BF03210812

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. doi:10.1037/0278-7393.22.1.197

Betsch, T. & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry*, *21*, 279–294. doi:10.1080/1047840X.2010.517737

Boywitt, C. D. & Meiser, T. (2013). Conscious recollection and binding among context features. *Consciousness and Cognition*, *22*, 875–886. doi:10.1016/j.concog.2013.05.010

Bröder, A. & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606. doi:10.1037/a0015279

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 1–29. doi:10.1080/09658211.2013.767348

Bröder, A. & Meiser, T. (2007). Measuring source memory. *Zeitschrift für Psychologie / Journal of Psychology*, *215*, 52–60. doi:10.1027/0044-3409.215.1.52

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

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217. doi:10.1037/h0047470

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*, 469–487. doi:10.1037/0022-3514.89.4.469

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*, 15–28. doi:10.3758/BF03195938

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Dixon, P. (2012). Assessing the evidence for response time mixture distributions. *Behavior Research Methods*, *44*, 706–724. doi:10.3758/s13428-012-0244-7

Donders, F. C. (1868). Die Schnelligkeit psychischer Processe. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, *30*, 657–681. Retrieved September 19, 2015, from http://www.sciencedirect.com/science/article/pii/0001691869900651

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*, 873–902. doi:10.1037/a0034247

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language*, *67*, 389–406. doi:10.1016/j.jml.2012.06.002

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie*. Friedrich-Wilhelms University, Bonn: Unpublished habilitation thesis.

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

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. doi:10.1037/0278-7393.24.2.387

Erdfelder, E., Castela, M., Michalkiewicz, M., & Heck, D. W. (2015). The advantages of model fitting compared to model simulation in research on preference construction. *Frontiers in Psychology*, *6*, 140. doi:10.3389/fpsyg.2015.00140

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*, 7–22. Retrieved from http://ideas.repec.org/a/jdm/journl/v6y2011i1p7-22.html

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. doi:10.1146/annurev.psych.59.103006.093629

Falmagne, J. C. (1968). Note on a simple fixed-point property of binary mixtures. *British Journal of Mathematical and Statistical Psychology*, *21*, 131–132. doi:10.1111/j.2044-8317.1968.tb00403.x

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. doi:10.1037/xge0000098

Freeman, J. B. & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226–241. doi:10.3758/BRM.42.1.226

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482. doi:10.1146/annurev-psych-120709-145346

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. doi:10.1037/0033-295X.103.4.650

Gigerenzer, G. & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*, 100–121. Retrieved from http://journal.sjdm.org/11/rh15/rh15.html

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528. doi:10.1037/0033-295X.98.4.506

Glöckner, A. & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*, 215–228. Retrieved from http://journal.sjdm.org/bn3/bn3.html

Glöckner, A. & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*, 711–721. Retrieved from http://journal.sjdm.org/11/m38/m38.html

Glöckner, A. & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making*, *6*, 23–42. Retrieved July 25, 2015, from http://journal.sjdm.org/11/rh4/rh4.html

Glöckner, A. & Bröder, A. (2014). Cognitive integration of recognition information and additional cues in memory-based decisions. *Judgment and Decision Making*, *9*, 35–50. Retrieved July 25, 2015, from http://journal.sjdm.org/13/13912/jdm13912.html

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133*, 641–666. doi:10.1016/j.cognition.2014.08.017

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90. doi:10.1037/0033-295X.109.1.75

Goldstein, D. G. & Gigerenzer, G. (2011). The beauty of simple models: Themes in recognition heuristic research. *Judgment and Decision Making*, *6*, 392–395. Retrieved from http://journal.sjdm.org/11/rh23/rh23.html

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. doi:10.1037/0022-3514.74.6.1464

Harlow, I. M. & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review*, *20*, 318–325. doi:10.3758/s13423-012-0340-9

Harlow, I. M. & Yonelinas, A. P. (2016). Distinguishing between the success and precision of recollection. *Memory*, *24*, 114–127. doi:10.1080/09658211.2014.988162

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*, 678–694. doi:10.3758/BF03206550

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology*, *54*, 454–463. doi:10.1016/j.jmp.2010.06.005

Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*. doi:10.3758/s13428-017-0869-7

Heck, D. W. & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Heck, D. W. & Erdfelder, E. (in press). Linking process and measurement models of recognition-based decisions. *Psychological Review*. doi:10.1037/rev0000063

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2017). *Generalized processing tree models: Jointly modeling discrete and continuous variables*. Manuscript submitted for publication.

Heck, D. W., Hilbig, B. E., & Moshagen, M. (2017). From information processing to decisions: Formalizing and comparing probabilistic choice models. *Cognitive Psychology*, *96*, 26–40. doi:10.1016/j.cogpsych.2017.05.003

Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, *60*, 29–34. doi:10.1016/j.jmp.2014.06.002

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

Heck, D. W., Wagenmakers, E.-J., & Morey, R. D. (2015). Testing order constraints: Qualitative differences between Bayes factors and normalized maximum likelihood. *Statistics & Probability Letters*, *105*, 157–162. doi:10.1016/j.spl.2015.06.014

Hilbig, B. E. (2010a). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making*, *5*, 272–284. Retrieved July 25, 2015, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.1293&rep=rep1&type=pdf

Hilbig, B. E. (2010b). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, *17*, 923–930. doi:10.3758/PBR.17.6.923

Hilbig, B. E. (2012). How framing statistical statements affects subjective veracity: Validation and application of a multinomial model for judgments of truth. *Cognition*, *125*, 37–48. doi:10.1016/j.cognition.2012.06.009

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. doi:10.1037/a0017518

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

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

Hilbig, B. E. & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, *21*, 1431–1443. doi:10.3758/s13423-014-0643-0

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

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

Hochman, G., Ayal, S., & Glöckner, A. (2010). Physiological arousal in processing recognition information: Ignoring or integrating cognitive cues? *Judgment and Decision Making*, *5*, 285–299. Retrieved from http://journal.sjdm.org/10/rh9/rh9.html

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. doi:10.1016/j.actpsy.2014.11.001

Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, *45*, 603–634. doi:10.1006/jmps.2000.1340

Hu, X. & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/bf02294263

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

Just, M. A. & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480. doi:10.1016/0010-0285(76)90015-3

Kieslich, P. J. & Henninger, F. (in press). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*. doi:10.3758/s13428-017-0900-z

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31. doi:10.1007/s11336-004-1188-3

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

Klauer, K. C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in multinomial processing tree models: An extension of Knapp and Batchelder (2004). *Journal of Mathematical Psychology*, *64*, 1–7. doi:10.1016/j.jmp.2014.11.001

Klauer, K. C. (2015). Multinomial processing tree models and response time distributions. Freiburg, Germany: Social Psychology and Methodology.

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

Knapp, B. R. & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*, 215–229. doi:10.1016/j.jmp.2004.03.002

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

Li, Y. (2015). *Response time modeling for the remember-know paradigm* (Unpublished Master Thesis, Universität Mannheim, Germany).

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Malejka, S. & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition*, *44*, 63–72. doi:10.3758/s13421-015-0549-8

Malejka, S., Heck, D. W., & Erdfelder, E. (2017). Testing core properties of recognition-memory models: Ranking judgments are consistent with two-high-threshold theory. Manuscript submitted for publication.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive psychology*, *57*, 335–384. doi:10.1016/j.cogpsych.2008.02.004

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. doi:10.1007/s11336-013-9374-9

Matzke, D. & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817. doi:10.3758/pbr.16.5.798

McGill, W. J. (1963). Stochastic latency mechanisms. In D. Luce (Ed.), *Handbook of Mathematical Psychology* (pp. 1–309). John Wiley & Sons.

Meiser, T. (2005). A hierarchy of multinomial models for multidimensional source monitoring. *Methodology*, *1*, 2–17. doi:10.1027/1614-1881.1.1.2

Meiser, T. (2014). Analyzing stochastic dependence of cognitive processes in multidimensional source recognition. *Experimental Psychology*, *61*, 402–415. doi:10.1027/1618-3169/a000261

Meiser, T. & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 116–137. doi: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. doi:10.1037/a0030734

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. doi:10.3758/s13421-015-0567-6

Miller, J. (2006). A likelihood ratio test for mixture effects. *Behavior Research Methods*, *38*, 92–106. doi:10.3758/BF03192754

Morey, R. D. (2017). Stitching time onto a tree: Joining models of accuracy with response times to understand cognition. *Psychonomic Society: Featured Content*. Retrieved from http://www.psychonomic.org/news/news.asp?id=331895

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

Newell, B. R. & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 923–935. doi:10.1037/0278-7393.30.4.923

Oberauer, K. & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*, 21–59. doi:10.1037/rev0000044

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H.-Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memoryOberauer, Stoneking, Wabersich, & Lin. *Journal of Vision*, *17*, 11–11. doi:10.1167/17.5.11

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155–156. doi:10.3758/BF03328004

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition*, *90*, B1–B9. doi:10.1016/S0010-0277(03)00141-0

Pachur, T. (2010). Recognition-based inference: When is less more in the real world? *Psychonomic Bulletin & Review*, *17*, 589–598. doi:10.3758/PBR.17.4.589

Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making*, *21*, 183–210. doi:10.1002/bdm.581

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Cognitive Science*, *2*, 147. doi:10.3389/fpsyg.2011.00147

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 59–64. doi:10.1016/j.jmp.2016.10.010

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Plieninger, H. & Heck, D. W. (2017). A new model for acquiescence at the interface of psychometrics and cognitive psychology. Manuscript submitted for publication.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, p. 125). Vienna, Austria.

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition*, *25*, 14–31. doi:10.1521/soco.2007.25.1.14

Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making*, *6*, 423–438. Retrieved from http://journal.sjdm.org/11/rh19/rh19.html

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

Province, J. M. & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*, 14357–14362. doi:10.1073/pnas.1103880109

Ratcliff, R. & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190. Retrieved September 23, 2015, from http://psycnet.apa.org/journals/rev/83/3/190/

Ratcliff, R. & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*, 697–719. doi:10.1037/a0033152

Richter, T. & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 150–162. doi:10.1037/0278-7393.32.1.150

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

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201. doi:10.1037/1040-3590.14.2.184

Schmittmann, V. D., Dolan, C. V., Raijmakers, M. E., & Batchelder, W. H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. doi:10.3758/BRM.42.3.836

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making*, *6*, 733–739. Retrieved from http://journal.sjdm.org/11/m36/m36.html

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. doi:10.1037/0033-2909.119.1.3

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241. doi:10.1037/h0036351

Smith, E. E. & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, *22*, 377–386. doi:10.3758/BF03200864

Smith, J. B. & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi: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. doi:10.1037/0096-3445.117.1.34

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 10393–10398. doi:10.1073/pnas.0503903102

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition*, *36*, 1–8. doi:10.3758/MC.36.1.1

Townsend, J. T. & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Townsend, J. T. (1990). Serial vs. Parallel Processing: Sometimes They Look Like Tweedledum and Tweedledee but They Can (and Should) Be Distinguished. *Psychological Science*, *1*, 46–54. doi:10.1111/j.1467-9280.1990.tb00067.x

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*, 1124–1131. doi:10.1126/science.185.4157.1124

Unkelbach, C. & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. doi:10.1016/j.concog.2008.09.006

Vandekerckhove, J. S., Matzke, D., & Wagenmakers, E. (2015). Model comparison and the principle of parsimony. In *Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–319). New York, NY: Oxford University Press. Retrieved July 28, 2015, from http://escholarship.org/uc/item/9j47k5q9.pdf

Weidemann, C. T. & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*, 150670. doi:10.1098/rsos.150670

Wu, H., Myung, J. I., & Batchelder, W. H. (2010a). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, *17*, 275–286. doi:10.3758/PBR.17.3.275

Wu, H., Myung, J. I., & Batchelder, W. H. (2010b). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, *54*, 291–303. Retrieved from ://000277482400001

Yantis, S., Meyer, D. E., & Smith, J. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, *110*, 350–374. doi:10.1037/0033-2909.110.2.350

Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology*, *8*, 159–199.

A Acknowledgements

I would like to thank many persons who supported me during my dissertation, and without whom this thesis would not have been possible.

First of all, I would like to thank Edgar, who has been the best supervisor I could wish for. He provided advice and support in times when I was doubtful about my ideas, but also gave me the freedom to work on other projects unrelated to my dissertation. From him, I learned about the importance of precise psychological theories, and how to think formally and write clearly. Most of all, I have always been motivated and impressed by Edgar's intrinsic and endless spirit for science. With his optimism and strong belief in me and my work, he encouraged me to be brave and explore new ideas.

I also would like to thank Morten, who often provided valuable advise and from whom I learned a lot about psychological methodology and the practical aspects of science. I am grateful to Benni for many enlightening discussions about theories of judgment and decision making, and to E.J. for showing me the beauty of Bayesian statistics and his invitation to Amsterdam. Moreover, I would like to thank my fellow students, who provided both intellectual and moral support. I am grateful to Isa, who always lifted my spirit and encouraged me; to Hansjörg, who has been a great office neighbor and taught me much about test theory; to Simone, who often supported me with valuable advice; to Marta, Martha, and Max, who helped me a lot during the first years of my dissertation; to Nina and Denis, who invited me to contribute to TreeBUGS; and Pascal, Yiqi, Sophie, and Lili, who supported and encouraged me during the last year. I also would like to thank all members of the Psycho III chair, especially Annette and Rahel for their support with organizational things. Moreover, I am indebted to my grad school for intellectual and organizational support, especially to Daniel Stegmüller, Thomas Gschwend, Marion Lehnert, and Milanka Stojkovic.

I am very grateful to all of my friends for moral and emotional support. In the words of a famous Hobbit: "I don't know half of you half as well as I should like, and I like less than half of you half as well as you deserve." I would also like to thank my family, especially my parents Christel and Klaus and my sister Johanna, who always supported me during my dissertation and made this path possible in the first place. Most of all, I am thankful to Rina, who was always on my side, encouraged me during difficult times, and shared my happiness during good times. Thank you.

B Statement of Originality

I affirm in lieu of oath that the following statements are to the best of my knowledge true and complete.

1. I hereby affirm that the presented doctoral dissertation with the title *Processing Tree Models for Discrete and Continuous Variables* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not yet present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby confirm the accuracy of the affirmation above.
5. I am aware of the significance of this affirmation and the legal consequences in case of untrue or incomplete statements.

Signed: ____________________________________

Date: ____________________________________

C Co-Author's Statements

Co-Author: Edgar Erdfelder

It is hereby confirmed that the following articles included in the thesis *Processing Tree Models for Discrete and Continuous Variables* were primarily conceived and written by Daniel W. Heck, Ph. D. candidate at the Center for Doctoral Studies in Social and Behavioral Sciences of the Graduate School of Economic and Social Sciences, University of Mannheim.

Heck, D. W. & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Heck, D. W. & Erdfelder, E. (in press). Linking process and measurement models of recognition-based decisions. *Psychological Review*. doi:10.1037/rev0000063

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2017). *Generalized processing tree models: Jointly modeling discrete and continuous variables*. Manuscript submitted for publication.

I sign this statement to the effect that Daniel W. Heck is credited as the primary source of ideas and the main author of all three articles as he derived the theoretical and statistical background, collected the data, implemented the new approaches in statistical software, analyzed the data, wrote the first drafts of the articles, and also contributed to improving the manuscripts. I contributed to the conceptual foundations of the new modeling approaches, suggested the basic ideas for the specific models employed in the three papers (2HT-RT model, RT-version of the r-model, and GPT-version of the FCM), and provided recommendations for improving the manuscripts and making the message of the articles clearer.

Mannheim, June 2017 Prof. Dr. Edgar Erdfelder

Co-Author: Pascal Kieslich

It is hereby confirmed that the following article included in the thesis *Processing Tree Models for Discrete and Continuous Variables* was primarily conceived and written by Daniel W. Heck, Ph. D. candidate at the Center for Doctoral Studies in Social and Behavioral Sciences of the Graduate School of Economic and Social Sciences, University of Mannheim.

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2017). *Generalized processing tree models: Jointly modeling discrete and continuous variables*. Manuscript submitted for publication.

I sign this statement to the effect that Daniel W. Heck is credited as the primary source of ideas and the main author of the article as he derived and implemented the new class of statistical models, analyzed the data, and wrote the article. I provided the empirical mouse-tracking data for the reanalysis and contributed to the interpretation of the analyses.

Mannheim, June 2017 Pascal Kieslich

D Copies of Articles

Psychon Bull Rev (2016) 23:1440–1465
DOI 10.3758/s13423-016-1025-6

THEORETICAL REVIEW

Extending multinomial processing tree models to measure the relative speed of cognitive processes

Daniel W. Heck[1] **· Edgar Erdfelder**[1]

Published online: 16 June 2016

Abstract Multinomial processing tree (MPT) models account for observed categorical responses by assuming a finite number of underlying cognitive processes. We propose a general method that allows for the inclusion of response times (RTs) into any kind of MPT model to measure the relative speed of the hypothesized processes. The approach relies on the fundamental assumption that observed RT distributions emerge as mixtures of latent RT distributions that correspond to different underlying processing paths. To avoid auxiliary assumptions about the shape of these latent RT distributions, we account for RTs in a distribution-free way by splitting each observed category into several bins from fast to slow responses, separately for each individual. Given these data, latent RT distributions are parameterized by probability parameters for these RT bins, and an extended MPT model is obtained. Hence, all of the statistical results and software available for MPT models can easily be used to fit, test, and compare RT-extended MPT models. We demonstrate the proposed method by applying it to the two-high-threshold model of recognition memory.

Keywords Cognitive modeling · Response times · Mixture models · Processing speed

Electronic supplementary material The online version of this article (doi:10.1007/s13423-016-1025-6) contains supplementary material, which is available to authorized users.

✉ Daniel W. Heck
dheck@mail.uni-mannheim.de

[1] Department of Psychology, School of Social Sciences, University of Mannheim, Schloss EO 254, 68131 Mannheim, Germany

Many substantive psychological theories assume that observed behavior results from one or more latent cognitive processes. Because these hypothesized processes can often not be observed directly, measurement models are important tools to test the assumed cognitive structure and to obtain parameters quantifying the probabilities that certain underlying processing stages take place or not. Multinomial processing tree models (MPT models; Batchelder & Riefer, 1990) provide such a means by modeling observed, categorical responses as originating from a finite number of discrete, latent processing paths. MPT models have been successfully used to explain behavior in many areas such as memory (Batchelder & Riefer, 1986, 1990), decision making (Erdfelder, Castela, Michalkiewicz, & Heck, 2015; Hilbig, Erdfelder, & Pohl, 2010), reasoning (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007), perception (Ashby, Prinzmetal, Ivry, & Maddox, 1996), implicit attitude measurement (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Nadarevic & Erdfelder, 2011), and processing fluency (Fazio, Brashier, Payne, & Marsh, 2015; Unkelbach & Stahl, 2009). Batchelder & Riefer (1999) and Erdfelder et al. (2009) reviewed the literature and showed the usefulness and broad applicability of the MPT model class. In the present paper, we introduce a simple but general approach to include information about response times (RTs) into any kind of MPT model.

As a running example, we will use one of the most simple MPT models, the two-high-threshold model of recognition memory (2HTM; Bröder & Schütz, 2009; Snodgrass & Corwin, 1988). The 2HTM accounts for responses in a binary recognition paradigm. In such an experiment, participants first learn a list of items and later are prompted to categorize old and new items as such. Hence, one obtains frequencies of hits (correct *old*), misses (incorrect *new*), false alarms (incorrect *old*), and correct rejections (correct

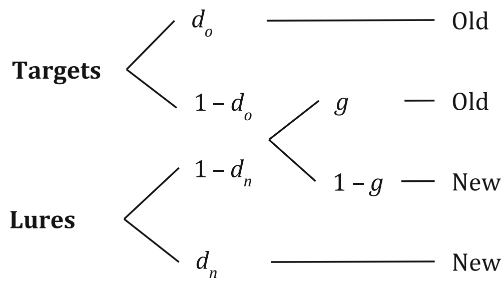

Fig. 1 The two-high threshold model of recognition memory

new responses). The 2HTM, shown in Fig. 1, assumes that hits emerge from two distinct processes: Either a memory signal is sufficiently strong to exceed a high threshold and the item is recognized as *old*, or the signal is too weak, an uncertainty state is entered, and respondents only guess *old*. The two processing stages of target detection and guessing conditional on the absence of detection are parameterized by the probabilities of their occurrence d_o and g, respectively. Given that the two possible processing paths are disjoint, the overall probability of an *old* response to an old item is given by the sum $d_o+(1-d_o)g$. Similarly, correct rejections can emerge either from lure detection with probability d_n or from guessing *new* conditional on nondetection with probability $1 - g$. In contrast, incorrect *old* and *new* responses always result from incorrect guessing.

The validity of the 2HTM has often been tested in experiments by manipulating the base rate of learned items, which should only affect response bias and thus the guessing parameter g (Bröder & Schütz, 2009; Dube, Starns, Rotello, & Ratcliff, 2012). If the memory strength remains constant, the model predicts a linear relation between the probabilities of hits and false alarms (i.e., a linear receiver-operating characteristic, or ROC, curve; Bröder & Schütz, 2009; Kellen, Klauer, & Bröder, 2013). The 2HTM is at the core of many other MPT models that account for more complex memory paradigms such as source memory (Bayen, Murnane, & Erdfelder, 1996; Klauer & Wegener, 1998; Meiser & Böder, 2002) or process dissociation (Buchner, Erdfelder, Steffens, & Martensen, 1997; Jacoby, 1991; Steffens, Buchner, Martensen, & Erdfelder, 2000). These more complex models have a structure similar to the 2HTM because they assume that correct responses either result from some memory processes of theoretical interest or from some kind of guessing.

Whereas MPT models are valuable tools to disentangle cognitive processes based on categorical data, they lack the ability to account for response times (RTs). Hence, MPT models cannot be used to test hypotheses about the speed of the assumed cognitive processes, for example, whether one underlying process is faster than another one. However, modeling RTs has a long tradition in experimental psychology, for instance, in testing whether cognitive processes occur serially or in parallel (Luce, 1986; Townsend & Ashby, 1983). Given that many MPT models have been developed for cognitive experiments that are conducted with the help of computers under controlled conditions, recording RTs in addition to categorical responses comes at a small cost. Even more importantly, substantive theories implemented as MPT models might readily provide predictions about the relative speed of the hypothesized processes or about the effect of experimental manipulations on processing speeds. For instance, the 2HTM can be seen as a two-stage serial process model in which guessing occurs only after unsuccessful detection attempts (see, e.g., Dube et al., 2012; Erdfelder, Küpper-Tetzel, & Mattern, 2011). Given this assumption of serial processing stages, the 2HTM predicts that, for both target and lure items, responses based on guessing are slower than responses based on detection (Province & Rouder, 2012). This hypothesis, however, cannot directly be tested because it concerns RT distributions of unobservable processes instead of directly observable RT distributions.

To our knowledge, there are mainly three general approaches to use information about RTs in combination with MPT models. First, Hu (2001) developed a method, based on an approach of Link (1982), that decomposes the mean RTs of the observed categories based on the estimated parameters of an MPT model. This approach assumes a strictly serial sequence of processing stages. Each transition from one latent stage to another is assigned with a mean processing time that is assumed to be independent of the original core parameters of the MPT model. Within each branch, all of the traversed mean processing times sum up to a mean observed RT. Despite its simplicity, this approach has not been applied often in the literature. One reason might be that the assumption of a strictly serial sequence of processing stages is too restrictive for many MPT models. Moreover, the method does not allow for testing different structures on the latent processing times because MPT parameters and mean processing speeds are estimated separately in two steps and not jointly in a single model.

The second approach of combining RTs with MPT models involves directly testing qualitative predictions for specific response categories (e.g., Dube et al., 2012; Erdfelder et al., 2011; Hilbig & Pohl, 2009). For instance, Erdfelder et al. (2011) and Dube et al. (2012) derived the prediction for the 2HTM that the mean RT of guessing should be slower than that of detection if guesses occur serially after unsuccessful detection attempts.[1] Moreover, a stronger

[1]Note that this order constraint on mean RTs of guessing and detection is implied by the stronger assumption of stochastic dominance discussed below.

response bias towards *old* items will result in a larger proportion of slow *old*-guesses. Assuming that response bias manipulations do not affect the speed of detection and the speed of guessing itself, we would thus expect an increase of the mean RT of hits with increasing guessing bias towards *old* responses (Dube et al., 2012). Whereas such indirect, qualitative tests may help clarify theoretical predictions, they often rest on assumptions of unknown validity and might require many paired comparisons for larger MPT models.

A third approach was used by Province and Rouder (2012) and Kellen et al. (2015). Similar to the two-step method by Hu (2001), the standard MPT model is fitted to the individual response frequencies to obtain parameter estimates first. Next, each response and the corresponding RT is assigned to the cognitive process from which it most likely emerged. For example, if a participant with 2HTM parameter estimates $\hat{d}_o = .80$ and $\hat{g} = .50$ produces a hit, then the recruitment probability that this hit was due to detection is estimated as $.80/(.80 + .20 \cdot .50) = .89$ (cf. Fig. 1). Hence, this response and its RT are assigned to the detection process rather than to the guessing process for which the recruitment probability estimate is only $(.20 \cdot .50)/(.80 + .20 \cdot .50) = .11$. Note that although assignments of responses to latent processes may vary between participants with different MPT parameter estimates, they are necessarily identical within participants. Hence, all RTs of a participant corresponding to a specific response type are always assigned to the same process. Classification errors implied by this procedure are likely to be negligible if recruitment probabilities are close to 1 as in our example but may be substantial if the latter are less informative (i.e., close to .50). Despite this problem, the method provides an approximate test of how experimental manipulations affect process-specific mean RTs. For example, both Province and Rouder (2012) and Kellen et al. (2015) found that the overall mean RTs across processes decreased when the number of repetitions during the learning phase increased. However, the mean RTs for the subsets of responses that were assigned to detection and guessing processes, respectively, were not affected by this experimental manipulation. Instead, the faster overall mean RTs for higher repetition rates could solely be explained by a larger proportion of fast detection responses (i.e., an increase of d_o with repetitions), or equivalently, by a smaller proportion of slow guesses.

All of these three approaches rely on separate, two-step analyses of categorical responses and RTs and do not allow to account for response frequencies and RTs in a single statistical model. Therefore, we propose a novel method that directly includes information about RTs into any MPT model. As explained in the following sections, the method rests on the fundamental idea that MPT models imply finite mixture distributions of RTs for each response category because they assume a finite number of underlying cognitive processes. Second, instead of modeling RTs as continuous variables, each observed response category is split into discrete bins representing fast to slow responses, similar to a histogram. Based on these more fine-grained categories, the relative speed of each processing branch is represented by probability parameters for the discrete RT bins, resulting in an RT-extended MPT model. Importantly, the underlying, unobservable RT distributions are modeled in a distribution-free way to avoid potentially misspecified distributional assumptions. Third, we introduce a strategy how to impose restrictions on the latent RT distributions in order to ensure their identifiability. Moreover, we discuss how to test hypotheses concerning the ordering of latent processes. Note that the RT-extended MPT model can be fitted and tested using the existing statistical tools for MPT models. We demonstrate the approach using the 2HTM of recognition memory.

MPT models imply mixtures of latent RT distributions

One core property of MPT models is their explicit assumption that a finite number of discrete processing sequences determines response behavior. In other words, each branch in an MPT model constitutes a different, independent cognitive processing path. It is straightforward to assume that each of these possible processes results in some (unknown) distribution of RTs. Given a category that is reached only by a single branch, the distribution of RTs for the corresponding process is directly observable. In contrast, if at least two branches lead to the same category, the observed distribution of RTs within this category will be a finite mixture distribution (Luce, 1986; Hu, 2001; Townsend & Ashby, 1983). This means that each RT of any response category of the MPT model originates from exactly one of the latent RT distributions corresponding to the branches of the MPT model, with mixture probabilities given by the core MPT structure. Mathematically, an observed RT distribution is a mixture of J latent RT distributions with mixture probabilities α_j if the observed density $f(x)$ is a linear combination of the latent densities $f_j(x)$,

$$f(x) = \sum_{j=1}^{J} \alpha_j f_j(x). \tag{1}$$

Because MPT models necessarily include more than a single response category, separate mixture distributions are assumed for the RT distributions corresponding to the different response categories, where the proportions α_j are defined by the branch probabilities of the core MPT model.

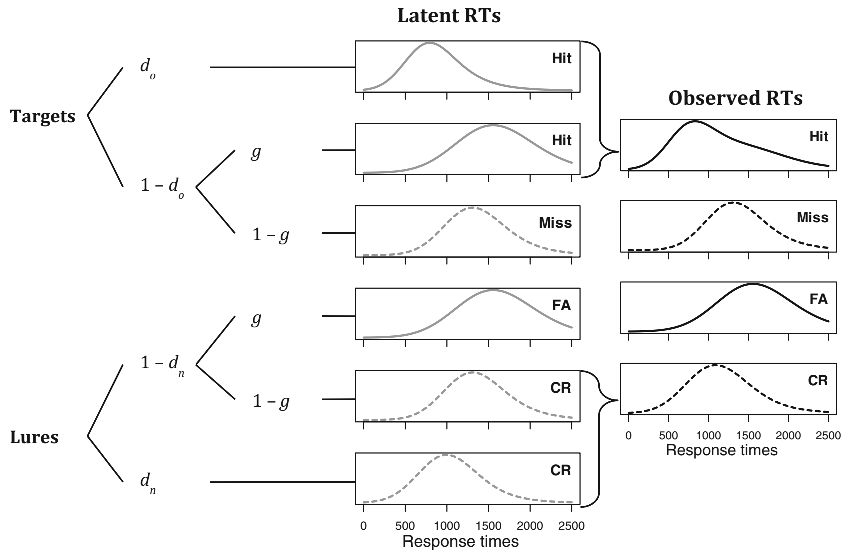

Fig. 2 The six hypothesized processing paths of the 2HTM imply six separate latent RT distributions. The observed RT distributions of hits and correct rejections are mixture distributions of the corresponding latent RT distributions. Solid and dashed densities represent *old* and *new* responses, respectively

This basic idea is illustrated in Fig. 2 for the 2HTM, where the six branches — and thus the six processing sequences — are assumed to have separate latent RT distributions. We use the term *latent* for these distributions because they cannot directly be observed in general. Consider, for instance, the RT distribution of hits: For a single *old* response to a target and the corresponding RT, it is impossible to decide whether it resulted from target detection (first branch) or guessing *old* in the uncertainty state (second branch). However, we know that a single observed RT must stem from one of two latent RT distributions, either from target detection with probability d_o or from guessing *old* with probability $(1-d_o)g$. Hence, the RTs for hits follow a two-component mixture distribution where the mixture probabilities are proportional to these two MPT branch probabilities.[2] In contrast to RTs of hits, the observed RT distributions of misses and false alarms are identical to the latent RT distributions of guessing *new* for targets and guessing *old* for lures, respectively, because only a single branch leads to each of these two categories. Note that, later on, some of these latent RT distributions may be restricted to be identical for theoretical reasons or in order to obtain a model that provides unique parameter estimates.

[2]Note that in order to obtain proper mixture probabilities that sum up to one, these path probabilities have to be normalized within each response category (e.g., for hits in the 2HTM, $P(\text{detect } old \mid \text{hit}) = d_o/(d_o + (1 - d_o)g)$ and $P(\text{guess } old \mid \text{hit}) = (1 - d_o)g/(d_o + (1 - d_o)g)$).

Whereas this mixture structure of latent RT distributions emerges directly as a core property of the model class, MPT models do not specify the temporal order of latent processes, that is, whether processes occur in parallel, in partially overlapping order, or in a strictly serial order (Batchelder & Riefer, 1999). In general, both parallel and serial processes can be represented within an MPT structure (Brown, 1998). Whereas the assumption of a serial order of latent processing stages might be plausible for the 2HTM (guessing occurs after unsuccessful detection attempts; cf. Dube et al., 2012; Erdfelder et al., 2011), other MPT models represent latent processes without the assumption of such a simple time course. Moreover, many MPT models can be reparameterized into equivalent versions with a different order of latent processing stages (Batchelder & Riefer, 1999), which prohibits simple tests of the processing order. In some cases, however, factorial designs allow for such tests using only response frequencies (Schweickert & Chen, 2008).

To avoid auxiliary assumptions about the serial or parallel nature of processing stages, our approach is more general and does not rely on an additive decomposition of RTs into processing times for different stages as proposed by Hu (2001). Instead, we directly estimate the latent RT distributions shown in Fig. 2 separately for each process (i.e., for each branch of the MPT model). Nevertheless, once the latent RT distributions are estimated for each branch, the results can be useful to exclude some of the possible processing sequences. For instance, we can test the assumption that detection occurs strictly before guessing by testing whether detection-responses are stochastically faster than guessing-responses (see below).

Another critical assumption concerns the exact shapes of the hypothesized latent RT distributions. These unobservable distributions could, for instance, be modeled by typical, right-skewed distributions for RTs such as ex-Gaussian, log-normal, or shifted Wald distributions (Luce, 1986; Matzke & Wagenmakers, 2009; Van Zandt & Ratcliff, 1995). However, instead of testing only the core structure of an MPT model, the validity of such a parametric model will also rest on its distributional assumptions. Since MPT models exist for a wide range of experimental paradigms, it is unlikely that a single parameterization of RTs will be appropriate for all applications. Moreover, many substantive theories might only provide predictions about the relative speed of cognitive processes instead of specific predictions about the shape of RTs. Given that most MPT models deal with RT distributions that are not directly observable, it is difficult to test such auxiliary parametric assumptions. Therefore, to avoid restrictions on the exact shape of latent RT distributions, we propose a distribution-free RT model.

Categorizing RTs into bins

Any continuous RT distribution can be approximated in a distribution-free way by a histogram (Van Zandt, 2000). Given some fixed boundaries on the RT scale, the number of responses within each bin is counted and displayed graphically. Besides displaying empirical RT distributions, discrete RT bins can also be used to test hypotheses about RT distributions without parametric assumptions. For instance, Yantis, Meyer, and Smith (1991) used discrete bins to test whether observed RT distributions can be described as mixtures of a finite number of basis distributions corresponding to different cognitive states (e.g., being in a prepared or unprepared cognitive state, Meyer, Yantis, Osman, & Smith, 1985). Yantis et al.'s (1991) method is tailored to situations in which all basis distributions can directly be observed in different experimental conditions. Using discrete RT bins, one can then test whether RT distributions in additional conditions are mixtures of these basis distributions. The method is mathematically tractable because the frequencies across RT bins follow multinomial distributions if identically and independently distributed RTs are assumed. Moreover, likelihood ratio tests allow for testing the mixture structure without the requirement of specifying parametric assumptions.

The model has two sets of parameters, the mixture probabilities α_{ij} that responses in condition i emerge from the cognitive state j, and the probabilities L_{jb} that responses of the state j fall into the b-th RT bin. Importantly, the latency parameters $L_{j1}, \ldots, L_{jB}$ model the complete RT distribution of the j-th state without any parametric assumptions. Using this discrete parametrization of the basis RT distributions, the mixture density in Eq. 1 gives the probability that responses in the i-th condition fall into the b-th RT bin,

$$p_{ib} = \sum_{j=1}^{J} \alpha_{ij} L_{jb}. \tag{2}$$

Given a sufficient amount of experimental conditions and bins, the model can be fitted and tested within the maximum likelihood framework.

Besides the advantage of avoiding arbitrary distributional assumptions, Yantis et al. (1991) showed that the method has sufficient power to detect deviations from the assumed mixture structure given realistic sample sizes between 40 to 90 responses per condition. However, the method is restricted to RT distributions for a single type of response, and it also requires experimental conditions in which the basis distributions can directly be observed. MPT models, however, usually account for at least two different types of responses and do not necessary allow for direct observations of the underlying component distributions. For instance, in the 2HTM (Fig. 2), the latent RT distributions of target and lure detection cannot directly be observed, because hits and false alarms also contain guesses. Moreover, MPT models pose strong theoretical constraints on the mixture probabilities α_{ij}, which are assumed to be proportional to the branch probabilities.

We retain the benefits of Yantis et al.'s (1991) method and overcome its limitations regarding MPT models by using

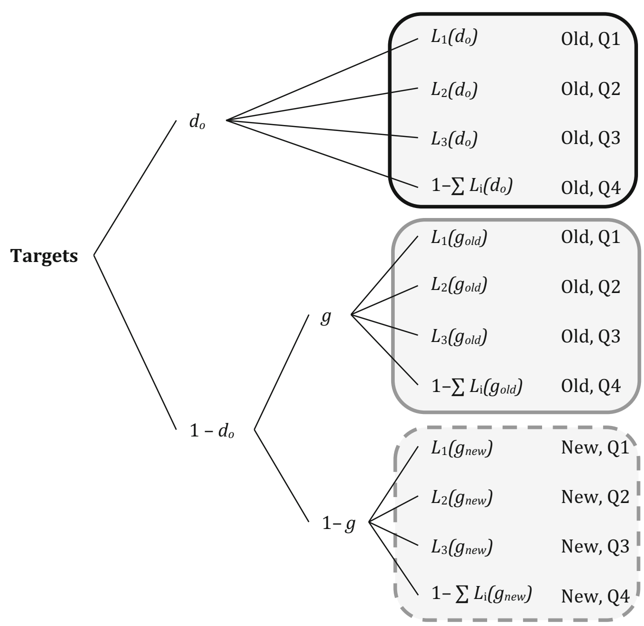

Fig. 3 The RT-extended two-high threshold model uses response categorizes from fast to slow (Q1 to Q4) to model latent RT distributions. The *gray boxes* represent the latent RT distributions for target detection, guessing *old*, and guessing *new* (each parameterized by three latency parameters)

discrete RT bins and probability parameters L_{jb} to model the latent RT distributions in Fig. 2. Thereby, instead of assuming a specific distributional shape, we estimate each latent RT distribution directly. This simple substitution produces a new, RT-extended MPT model. To illustrate this, Fig. 3 shows the RT-extended 2HTM for old items. In this example, each of the original categories is split into four bins using some fixed boundaries on the observed RTs. In the 2HTM, this results in a total of $4 \cdot 4 = 16$ categories for the RT-extended MPT model. Once the RTs have been used to obtain these more fine-grained category frequencies, the new MPT model can easily be fitted, tested, and compared to other models by means of existing software for MPT models (e.g., Moshagen, 2010; Singmann & Kellen, 2013).

Note that the number of RT bins can be adjusted to account for different sample sizes. Whereas experiments with many responses per participant allow for using many RT bins, the method can also be applied with small samples by using only two RT bins. The latter represents a special case in which responses are categorized as either fast or slow based on a single RT boundary. As a result, each latent RT distribution j is described by one latency parameter L_j, the probability that RTs of process j are below the individual RT boundary, which is direct measure of the relative speed of the corresponding process.

Choosing RT boundaries for categorization

If the latent RT distributions are modeled by discrete RT bins, the question remains how to obtain RT boundaries to categorize responses into bins. In the following, we discuss the benefits and limitations of three different strategies. One can use (1) fixed RT boundaries, (2) collect a calibration data set, or (3) rely on data-dependent RT boundaries. We used the last of these strategies in the present paper for reasons that will become clear in the following.

Most obviously, one can use fixed RT boundaries that are chosen independently from the observed data. For instance, to get eight RT bins, we can use seven equally-spaced points on a range typical for the paradigm under consideration, for example, from 500 ms to 2,000 ms in steps of 250 ms. Obviously, however, this strategy can easily result in many zero-count cells if the RTs fall in a different range than expected. This is problematic, because a minimum expected frequency count of five per category is a necessary condition to make use of the asymptotic χ^2 test for the likelihood ratio statistic, which is used to test the proposed mixture structure. For category counts smaller than five, the asymptotic test may produce misleading results (Agresti, 2013). Even more importantly, the resulting frequencies are not comparable across participants because of natural differences in the speed of responding (Luce, 1986). Hence, frequencies cannot simply be summed up across participants, which further complicates the analysis.

As a remedy, it is necessary to obtain RT boundaries that result in comparable bins across participants. If comparable RT bins are defined for all participants on a-priori grounds, it is in general possible to analyze the RT-extended MPT model for the whole sample at once, either by summing individual frequencies or by using hierarchical MPT extensions (Klauer, 2010; Smith & Batchelder, 2010). Hence, as a second strategy, one could use an independent set of data to calibrate the RT boundaries separately for each participant. Often, participants have to perform a learning task similar to the task of interest. We can then use any summary statistics of the resulting RT distribution to define the RT boundaries for the subsequent MPT analysis. For instance, the empirical 25 %-, 50 %-, and 75 %-quantiles ensure that the resulting four RT bins are a-priori equally likely (for B RT bins, one would use the b/B-quantiles with $b = 1, \ldots, B-1$). Given that the RT boundaries are defined for each participant following this rule, the resulting RT bins share the same interpretation and can be analyzed jointly. Besides minimizing zero-count frequencies, the adjusted RT boundaries for each participant eliminate individual RT differences in the RT-extended MPT model.

Often, a separate calibration data set might not be available or costly to obtain. In such a situation, it is possible to use the available data twice: First, to obtain some RT boundaries according to a principled strategy, and second, to analyze the same, categorized data with an RT-extended MPT model. For instance, Yantis et al. (1991) proposed to vary the location and scale of equal-width RT intervals to ensure a minimum of five observations per category. At first view, however, choosing RT boundaries as a function of the same data is statistically problematic. If the boundaries for categorization are fixed as in our first strategy, the multinomial sampling theory is correct and the standard MPT analysis is valid. In contrast, if the RT boundaries are stochastic and depend on the same data, it is not clear whether the resulting statistical inferences are still correct. However, the practical advantages of obtaining RT boundaries from the same data are obviously very large. Therefore, to justify this approach, we provide simulation studies in the Supplementary Material showing that the following strategy results in correct statistical inferences with respect to point estimates, standard errors, and goodness-of-fit tests.

According to the third strategy, RT boundaries are obtained from the overall RT distribution across all response categories as shown in Fig. 4. For each person, the distribution of RTs across all response categories is approximated by a right-skewed distribution, for instance, a log-normal distribution (Fig. 4a). This approximation serves as a reference distribution to choose reasonable boundaries that result

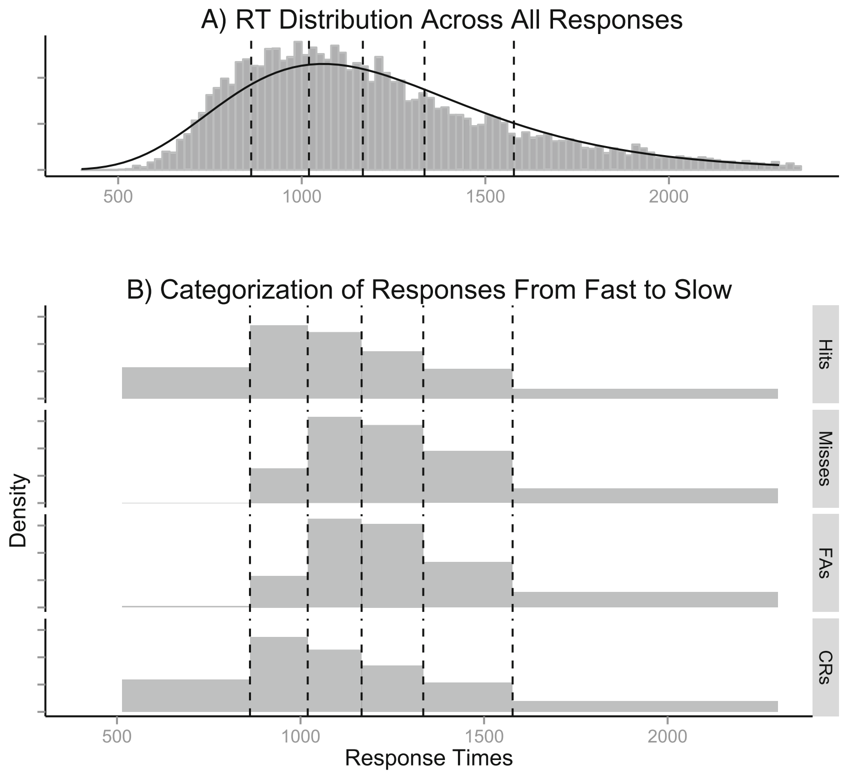

Fig. 4 A principled strategy to obtain comparable RT bins across individuals. **a** Computation of quantiles from a log-normal approximation to the RT distribution across all response categories. **b** Categorization of hits, misses, false alarms, and correct rejections from fast to slow

in RT bins with a precise interpretation, for example, the b/B-quantiles (the 25 %-, 50 %-, 75 %-quantiles for four RT bins). These individual RT boundaries are then used to categorize responses into B subcategories from fast to slow (Fig. 4b). By applying the same approximation strategy separately for each individual, the new category system has an identical interpretation across individuals. Moreover, the right-skewed approximation results in a-priori approximately equally likely RT bins. Importantly, this approximation is only used to define the RT boundaries. Hence, it is neither necessary that it fits the RT distribution, nor does it constrain the shape of the latent RT distributions because the latency parameters L_{jb} of the RT-extended MPT model can still capture any distribution.

We decided to use a log-normal approximation because it is easy to apply and uses less parameters than other right-skewed distributions such as the ex-Gaussian or the shifted Wald distribution (Van Zandt, 2000). However, in our empirical example below, using these alternative distributions resulted in similar RT boundaries, and hence, in the same substantive conclusions. The detailed steps to derive individual RT boundaries are (1) log-transform all RTs of a person, (2) compute mean and variance of these log-RTs, (3) get b/B-quantiles of a normal distribution with mean and variance of Step 2, (4) re-transform these log-quantiles t_b back to the standard RT scale to get the boundaries required for categorization (i.e., $\text{RT}_b = \exp(t_b)$).[3]

Note that there is an alternative, completely different approach of using RT boundaries based on the overall RT distribution. For instance, with two RT bins, we can categorize a response as fast or slow if the observed RT is below or above the overall median RT, respectively. However, instead of using an MPT model with latent RT distributions as in Fig. 3, it is possible to use the standard MPT tree in Fig. 1 twice but with separate parameters for fast and slow responses. Hence, this approach allows for testing whether the core parameters (d_o, d_n, and g in case of the 2HTM) are invariant across fast and slow responses. Since our interest here is in direct estimates of the relative speed of the hypothesized processes, we did not pursue this direction any further.

Identifiability: constraints on the maximum number of latent RT distributions

Up to this point, we assumed that each processing path is associated with its own latent RT distribution. However,

[3]Mathematically more concise, $\text{RT}_b = \exp\left[\text{SD}_{\log \text{RT}} \cdot (\Phi^{-1}(b/B) + \overline{\log \text{RT}})\right]$, where Φ^{-1} is the inverse cumulative density function of the standard-normal distribution.

assuming a separate latent RT distribution for each processing path will in general cause identifiability problems. For example, Fig. 2 suggests that it is not possible to obtain unique estimates for six latent RT distributions based on only four observed RT distributions (for hits, misses, false alarms, and correct rejections, respectively). Hence, it will usually be necessary to restrict some of these latent distributions in one way or another to obtain an identifiable model (Hu, 2001). For instance, in the 2HTM, one could assume that any type of guessing response is equally fast, thereby reducing the number of latent RT distributions to three. In practice, these restrictions should of course represent predictions based on psychological theory. However, before discussing restrictions on RTs in case of the 2HTM, we will first present a simple strategy how to check that a chosen set of restricted latent RT distributions can actually be estimated.

Before using any statistical model in practice, it has to be shown that fitting the model to data results in unique parameter estimates. A sufficient condition for unique parameter estimates is the global identifiability of the model, that is, identical category probabilities $\boldsymbol{p}(\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{\theta}')$ imply identical parameter values $\boldsymbol{\theta} = \boldsymbol{\theta}'$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$ in the parameter space Ω (Bamber & van Santen, 2000). For larger, more complex MPT models, it can be difficult to proof this one-to-one mapping of parameters to category probabilities analytically (but see Batchelder & Riefer, 1990; Meiser, 2005 for examples of analytical solutions). Therefore, numerical methods based on simulated model identifiability (Moshagen, 2010) or the rank of the Jacobian matrix Schmittmann, Dolan, Raijmakers, & Batchelder, (2010) have been developed to check the local identifiability of MPT models, which ensures the one-to-one relation only in the proximity of a specific set of parameters. All of these methods can readily be applied to RT-extended MPT models.

However, when checking the identifiability of an RT-extended MPT model, the problem can be split into two parts: (1) the identifiability of the core parameters $\boldsymbol{\theta}$ from the original MPT model and (2) the identifiability of the additional latency parameters $\boldsymbol{L}$ of the RT-extension given that the original MPT model is identifiable. The first issue has a simple solution. If the original MPT model is globally identifiable, then, by definition, the core parameters $\boldsymbol{\theta}$ of the extended MPT model are also globally identifiable. This holds irrespective of the hypothesized structure of latent RT distributions. This simple result emerges from the fact that the original category frequencies are easily recovered by summing up the frequencies across all RT bins within each category (see Appendix A).

Given a globally identifiable original MPT model, the identifiability of the latency parameters $\boldsymbol{L}$ in the RT-extended model can be checked using a simple strategy. First, all categories are listed that are reached by a single branch. The latency parameters of the respective latent RT distributions of these categories are directly identifiable. Intuitively, the RTs in such a category can be interpreted as process-pure measures of the latencies corresponding to the processing path (similar to the basis distributions of Yantis et al., 1991). Second, for each of the remaining categories, their associated latent RT distributions are listed. If for one of these categories, all but one of the latent RT distributions are identifiable (from the first step), then the remaining latent RT distribution of this category is also identifiable. This second step can be applied repeatedly using the identifiability of RT distributions from previous steps. If all of the hypothesized latent RT distributions are rendered identifiable by this procedure, the whole model is identifiable. Note, however, that this simple strategy might not always provide a definite result. For instance, if one or more latent RT distributions might not be rendered identifiable by this strategy, it is still possible that the model is identifiable. Therefore, the successful application of this simple strategy provides a sufficient but not a necessary condition for the identifiability of an RT-extended MPT model.

In case this simple strategy does not ensure the identifiability of all of the latency parameters, a different, more complex approach can be used. For this purpose, a matrix $P(\boldsymbol{\theta})$ is defined which has as many rows as categories and as many columns as latent RT distributions. A matrix entry $P(\boldsymbol{\theta})_{kj}$ is defined as the branch probability $p_{ki}(\boldsymbol{\theta})$ if the i-th branch of category k is assumed to generate RTs according to the j-th latent RT distribution; otherwise, it is zero (see Appendix A for details). The extended MPT model will be globally identifiable if and only if the rank of this matrix $P(\boldsymbol{\theta})$ is equal to or larger than the number of latent RT distributions. This general strategy directly implies that an identifiable RT-extended model cannot have more latent RT distributions than observed categories. Otherwise, the rank of the matrix $P(\boldsymbol{\theta})$ will necessarily be smaller than the number of latent RT distributions. For example, consider the 2HTM as shown in Fig. 2. For this model, the matrix $P(\boldsymbol{\theta})$ has four rows for the four types of responses and six columns for the latent RT distributions. Hence, this model is not identifiable. Below, we discuss theoretical constraints to solve this issue.

Note that these results about the identifiability of the latency parameters only hold if the core parameters are in the interior of the parameter space, that is, $\boldsymbol{\theta} \in (0, 1)^S$, where S is the number of parameters. In other words, if one of the processing paths can never be reached because some of the core parameters are zero or one, it will be

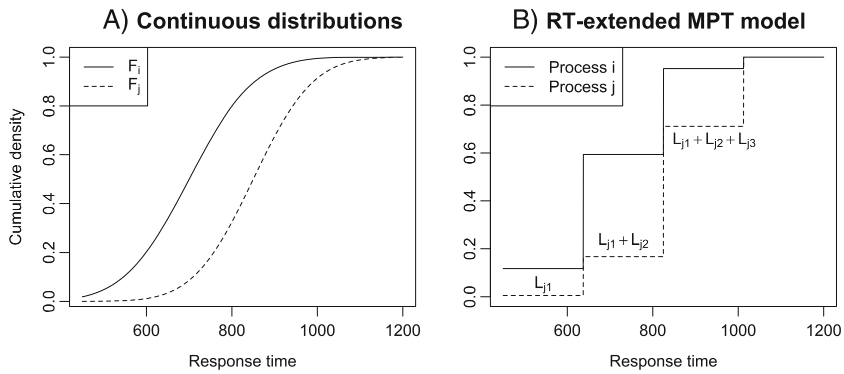

Fig. 5 **a** The continuous RT distribution F_j stochastically dominates F_i. **b** In an RT-extended MPT model, the latent RT distribution of process j is slower (i.e., RTs are larger) than that of process i

impossible to estimate the corresponding RT distribution if it does not occur in another part of the MPT model. Therefore, in experiments, it is important to ensure that all of the hypothesized cognitive processes actually occur. Otherwise, it is possible that some of the latent RT distributions are empirically not identified (Schmittmann et al., 2010). One solution to this issue is the analysis of the group frequencies aggregated across participants. Summing the individual frequencies of RT-extended MPT models is in general possible because only the relative speed of processing is represented in the data due to the individual log-normal approximation. However, despite the comparable data structure, the assumption of participant homogeneity is questionable in general (Smith & Batchelder, 2008) and might require the use of hierarchical MPT models (e.g., Klauer, 2010; Smith & Batchelder, 2010).

Testing the ordering of latent processes

There are two main types of hypotheses that are of interest with regard to RT-extended MPT models. On the one hand, psychological theory might predict which of the latent RT distributions corresponding to cognitive processes are identical or different. Substantive hypotheses of interest could be, for instance, whether RTs due to target and lure detection follow the same distribution, or whether the speed of detection is affected by response-bias manipulations. These hypotheses about the equality of latent RT distributions i and j can easily be tested by comparing an RT-extended MPT model against a restricted version with equality constraints on the corresponding latency parameters, that is, $L_{ib} = L_{jb}$ for all RT bins $b = 1, .., B - 1$. In other words, testing such a hypothesis is similar to the common approach in MPT modeling to test whether an experimental manipulation affects the parameters (Batchelder & Riefer, 1999; Erdfelder et al., 2009).

On the other hand, psychological theory might predict that RTs from one cognitive process are faster than those from another one (Heathcote, Brown, Wagenmakers, & Eidels, 2010). For instance, if guesses occur strictly after unsuccessful detection attempts in the 2HTM (Erdfelder et al., 2011), RTs resulting from detection should be faster than RTs resulting from guessing.[4] Such a substantive hypothesis about the ordering of cognitive processes translates into the statistical property of stochastic dominance of RT distributions (Luce, 1986; Townsend & Ashby, 1983). The RT distribution of process j stochastically dominates the RT distribution of process i if their cumulative density functions fulfill the inequality

$$F_j(t) \leq F_i(t) \text{ for all } t \in \mathcal{R}, \tag{3}$$

which is illustrated in Fig. 5a. Note that a serial interpretation of the 2HTM directly implies the stochastic dominance of guessing RTs over detection RTs. If the RTs corresponding to the cognitive processes i and j are directly observable, this property can be tested using the empirical cumulative density functions. Note that several nonparametric statistical tests for stochastic dominance use a finite number of RT bins to test Eq. 3 based on the empirical cumulative histograms (see Heathcote et al., 2010, for details).

In RT-extended MPT models, substantive hypotheses concern the ordering of *latent* RT distributions. Moreover, the continuous RTs t are replaced by a finite number of RT bins. However, Eq. 3 can directly be translated to an RT-extended MPT model considering that the term $\sum_{k=1}^{b} L_{jk}$ gives the probability of responses from process j falling into an RT bin between 1 and b. In other words, it gives the underlying cumulative density function for RTs of process

[4] In terms of serial stage models (Roberts & Sternberg, 1993), we are comparing the random variable of observed detection RTs T_d against the random variable of observed guessing RTs $T_g = T_d + T'$, where T' is a positive random variable resembling pure guessing duration.

j. Hence, the statistical property of stochastic dominance of process j over process i results in the following constraint on the latency parameters of an RT-extended MPT model:

$$\sum_{k=1}^{b} L_{jk} \leq \sum_{k=1}^{b} L_{ik} \text{ for all } b = 1, \ldots, B-1. \quad (4)$$

This constraint is illustrated in Fig. 5b, where process i is faster than process j.

When using only two RT bins, this relation directly simplifies to the linear order restriction $L_{j1} \leq L_{i1}$ because each latent RT distribution has a single latency parameter only. In MPT models, such linear order constraints can be reparameterized using auxiliary parameters η, for example, $L_{j1} = \eta L_{i1}$ (Klauer & Kellen, 2015; Knapp & Batchelder, 2004). To test this kind of restrictions, model selection techniques that take the diminished flexibility of order-constrained MPT models into account have become available (e.g., Klauer & Kellen, 2015; Vandekerckhove, Matzke, & Wagenmakers, 2015; Wu, Myung, & Batchelder, 2010). Hence, the ordering of latent processes can directly be tested using existing software and methods as shown in the empirical example below.

Even though the use of only two RT bins is attractive in terms of simplicity, it will also limit the sensitivity to detect violations of stochastic dominance. Specifically, the cumulative densities might intersect at the tails of distributions. In such a case, an analysis based on two RT bins might still find support for stochastic dominance even with large sample sizes (Heathcote et al., 2010). Therefore, it might be desirable to use more than two RT bins if the sample size is sufficiently large. However, when using more than two RT bins, a statistical test of Eq. 4 becomes more complex because the sum on both sides of the inequality allows for trade-offs in the parameter values. Importantly, a test of the set of simple linear order constraints $L_{jb} \leq L_{ib}$ for all bins b captures only a sufficient but not a necessary condition for stochastic dominance. In other words, the set of restrictions $L_{jb} \leq L_{ib}$ can be violated even though two latent RT distributions meet the requirement for stochastic dominance. Note that this problem transfers to any binary reparameterization of an extended MPT model with more than two RT bins (see Appendix B). As a consequence, standard software for MPT models that relies on a binary representations (e.g., Moshagen, 2010; Singmann & Kellen, 2013) can in general not be used to test stochastic dominance of latent RT distributions using more than two bins.

As a solution, Eq. 4 can be tested using Bayes factors based on the encompassing prior approach (Klugkist, Laudy, & Hoijtink, 2005), similarly as in Heathcote et al. (2010). Essentially, this approach quantifies the evidence in favor of a restriction by the ratio of prior and posterior probability mass over a set of order-constraints on the parameters. To test stochastic dominance of latent RT distribution, the RT-extended MPT model is fitted without any order constraints. Next, the posterior samples are used to estimate the prior and posterior probabilities that Eq. 4 holds. The ratio of these two probabilities is the Bayes factor in favor of stochastic dominance (Klugkist et al., 2005). We show how to apply this approach in the second part of our empirical example below.

The RT-extended two-high-threshold model

In this section, we discuss theoretical constraints on the number of latent RT distributions for the 2HTM. Moreover, we demonstrate how the proposed model can be applied to test hypotheses about latent RT distributions using data on recognition memory.

Theoretical constraints on the latent RT distributions of the 2HTM

As discussed above, the 2HTM with six latent RT distributions in Fig. 2 is not identifiable. However, assuming complete-information loss of the latent RT distributions of guessing, the model becomes identifiable. Here, complete-information loss refers to the assumption that guessing, conditional on unsuccessful detection, is independent of item type. In other words, the guessing probability g is identical for lures and targets because no information about the item type is available in the uncertainty state (Kellen & Klauer, 2015). This core property of discrete-state models should also hold for RTs: Given that an item was not recognized as a target or a lure, the RT distributions for guessing *old* and *new* should not differ between targets and lures. Based on a serial-processing interpretation of the 2HTM (Erdfelder et al., 2011), this actually includes two assumptions: Not only the time required for actual guessing but also the preceding time period from test item presentation until any detection attempts are stopped are identically distributed for targets and lures.[5] Both assumptions are reasonable. They are in line with the idea that participants do not have additional information about the item type that might affect their motivation to detect a test item. In

[5]As noted by a reviewer, this assumption is mandatory for the 2HTM when applied to *yes-no* recognition tasks. However, it might be dispensable in other models (e.g. the one-high threshold model) or when using other memory tasks (e.g. the two-alternative forced choice task).

sum, complete-information loss implies that RT distributions involving guessing are indistinguishable for targets and lures.

The assumption of complete-information loss restricts the six latent RT distributions in Fig. 2 to only four latent RT distributions; those of target and lure detection, and those of guessing *old* and *new*, respectively. This restriction renders the RT-extended 2HTM identifiable given that the core parameters d_o, d_n, and g are identified. Identification of the core parameters can be achieved, for example, by imposing the equality constraint $d_o = d_n$ (e.g., Bayen et al., 1996; Klauer & Wegener, 1998). To check the identifiability of the RT-extended model, we apply the simple strategy introduced above. First, we observe that misses and false alarms both result from single branches. Hence, the corresponding RT distributions for guessing *new* and *old* are identifiable. Second, because RTs for hits either emerge from the identifiable distribution of guessing *old* or from the process of target detection, the latter RT distribution is also identifiable. The same logic holds for correct rejections and lure detection.

In sum, all of the four latent RT distributions are identifiable within a single experimental condition. Therefore, we can estimate four separate latent RT distributions for several experimental manipulations. Here, we want to test whether the assumptions of complete-information loss and fast detection hold across different memory-strength and base-rate conditions, and whether the speed of detection is affected by the base-rate manipulation.

Sensitivity simulations

Before applying the proposed method to actual data, we show that the approach is sufficiently sensitive to decompose observed mixture distributions. First, we estimate the statistical power to detect a discrepancy between the latent RT distributions of detection and guessing using only two RT bins, and second, we show that our distribution-free approach can recover even atypical RT distributions using eight RT bins.

To assess the statistical power, we simulated data for a single experimental condition with true, underlying detection probabilities of $d_o = d_n = .7$ and a symmetric guessing probability of $g = .5$. To generate RTs from the latent RT distributions of detection and guessing, we used two ex-Gaussian distributions (i.e., RTs were sampled from the sum of a normal and an independent exponential random variable). Both of these distributions shared the standard deviation $\sigma = 100$ ms of the normal component and the mean $\nu = 300$ ms of the exponential component. Differences between the two distributions were induced by manipulating the mean of the normal component. Whereas the normal component for the guessing RTs had a constant mean of $\mu_g = 1,000$ ms, we manipulated this mean in the range of $\mu_d = 1,000$ ms, …, 700 ms in steps of 50 ms for the detection RTs. Note that the resulting mean RT differences correspond to effect sizes in terms of Cohen's d between $d = 0$ and $d = 0.95$.

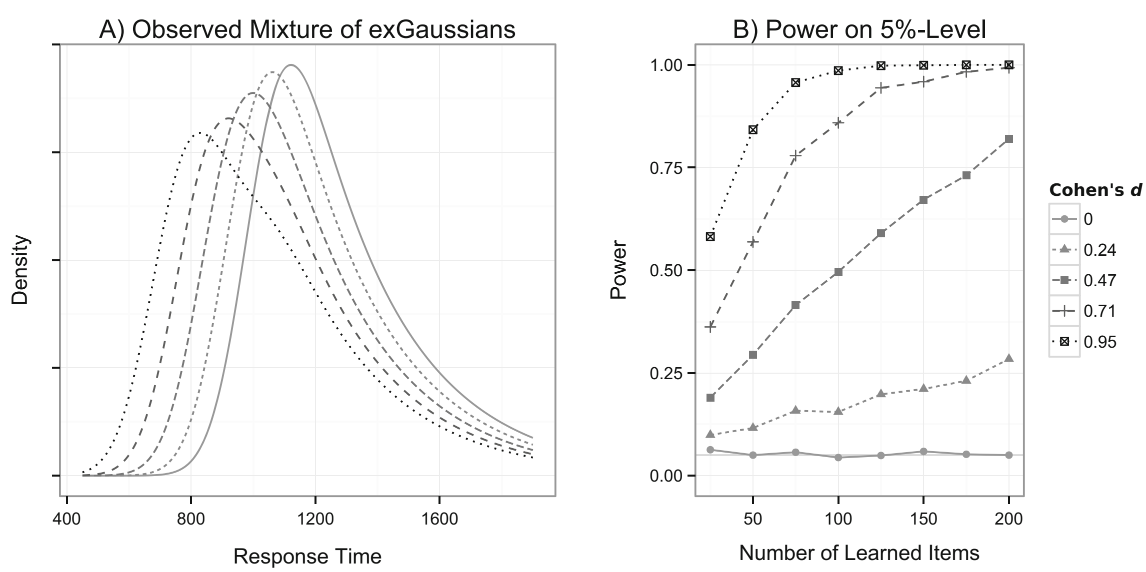

Fig. 6 **a** The theoretically expected mixture distributions for observed RTs of hits and correct rejections, depending on the (standardized) difference in means of the latent RT distributions of detection and guessing. **b** Power of the RT-extended 2HTM to detect the discrepancy between the latent RT distributions of detection and guessing

Figure 6a shows the theoretically expected mixture distributions for the observed RTs of hits and correct rejections for the five simulated conditions. Note that these mixture distributions are all unimodal and are thus statistically more difficult to discriminate than bimodal mixtures. For each of the five conditions, we generated 1,000 data sets, categorized RTs into two bins based on the log-normal approximation explained above, and fitted two RT-extended MPT models. The more general model had two latency parameters L_d and L_g for the relative speed of detection and guessing, respectively. The nested MPT model restricted these two parameters to be identical and was compared to the more general model by a likelihood-ratio test using a significance level of $\alpha = .05$.

Figure 6b shows the results of this simulation. When the latent RT distributions for detection and guessing were identical ($d = 0$), that is, the null hypothesis did hold, the test adhered to the nominal α-level of 5 % for all sample sizes. In this simple setting, medium and large effects were detected with sufficient power using realistic sample sizes. For instance, the power was 85.9 % to detect an effect of $d = 0.71$ based on a sample size of 100 learned items. As can be expected for a model that does not rely on distributional assumptions, the statistical power to detect a small difference in mean RTs ($d = 0.24$) was quite low even for 200 learned items.

In principle, the distribution-free approach allows for modeling nonstandard RT distributions, for instance, those emerging in experimental paradigms with a fixed time-window for responding. In such a scenario, latent RT distributions could potentially be right-skewed (fast detection), left-skewed (truncation by the response deadline), or even uniform (guessing). To test the sensitivity of the proposed method in such a scenario, we therefore sampled RT values in the fixed interval [0, 1] from beta distributions with these different shapes.

In contrast to the power study, we used a slightly more complex setting with different parameters for target and lure detection ($d_o = .65$, $d_n = .4$), two response bias conditions and parameters ($g_A = .3$, $g_B = .7$), and separate beta distributions to generate RTs for target detection (right-skewed), lure detection (left-skewed), and guessing in condition A and B (uniform and bimodal, respectively). Figure 7 shows the latent, data-generating RT distributions (black curves) along with the mean estimates (black histogram) across 500 replications (light gray), each based on 150 responses per item type and condition. The results clearly indicate a good recovery of all four distributions.

Overall, we conclude that our distribution-free approach is sufficiently sensitive to estimate latent RT distributions even if the observed mixture distributions are unimodal or if the latent distributions differ markedly in shape. Note that, as in all simulation studies, these results depend on the specific model and the scenarios under consideration. They do not necessarily generalize to more complex situations. Nevertheless, our simulations show that, in principle, the

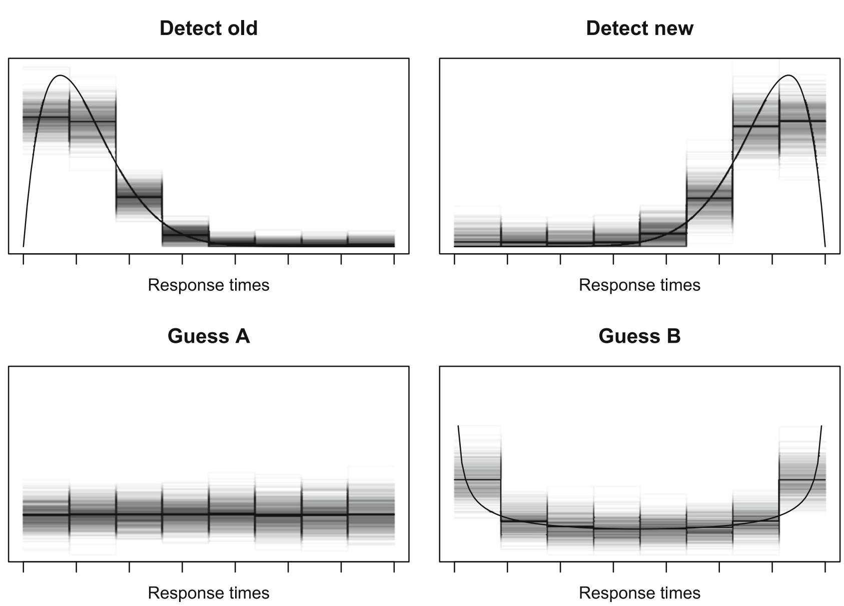

Fig. 7 Latent, data-generating beta distributions are shown by solid curves, whereas individual and mean estimates of the RT-extended 2HTM are shown as light gray and black histograms, respectively

categorization of RTs into bins allows for a distribution-free estimation of latent RT distributions.

Methods

As in previous work introducing novel RT models (e.g., Ollman, 1966; Yantis et al., 1991), we use a small sample with many responses per participant to illustrate the application of the proposed method. The data are from four students of the University of Mannheim (3 female, mean age = 24) who responded to 1,200 test items each. The study followed a 2 (proportion of old items: 30 % vs. 70 %) × 2 (stimulus presentation time: 1.0 vs. 2.2 seconds) within-subjects design. To obtain a sufficient number of responses while also maintaining a short testing duration per session, the study was split into three sessions that took place on different days at similar times. The four conditions were completed block-wise in a randomized order in each of the three sessions. We selected stimuli from a word list with German nouns by Lahl, Gritz, Pietrowsky, & Rosenberg, (2009). From this list, words shorter than three letters and longer than ten letters were removed. The remaining items were ordered by concreteness to select the 1464 most concrete words for the experiment. From this pool, words were randomly assigned to sessions, conditions, and item types without replacement.

The learning list contained 72 words in each of the four conditions including one word in the beginning and the end of the list to avoid primacy and recency effects, respectively. Each word was presented either for one second in the low memory strength condition or for 2.2 seconds in the high memory strength condition with a blank inter-stimulus interval of 200 ms. After the learning phase, participants worked on a brief distractor task (i.e., two minutes for finding differences in pairs of pictures). In the testing phase, 100 words were presented including either 30 % or 70 % learned items. Participants were told to respond either *old* or *new* as accurate and as fast as possible by pressing the keys ‘A’ or ‘K.’ Directly after each response, a blank screen appeared for a random duration drawn from a uniform distribution between 400 ms and 800 ms. This random inter-test interval was added to prevent the participants from responding in a rhythmic manner that might result in statistically dependent RTs (Luce, 1986). After each block of ten responses, participants received feedback about their current performance to induce base-rate-conform response behavior. The experiment was programmed using the open-source software OpenSesame (Mathôt, Schreij, & Theeuwes, 2012).

Analyses based on two RT bins

We first tested our three hypotheses (complete-information loss, fast detection, and effects of base-rates on detection RTs) based on only two RT bins using standard software and methods. In the next step, we reexamined our conclusions regarding stochastic dominance of detection RTs over guessing RTs using more RT bins.

Competing models

We compared six RT-extended MPT models to test our hypotheses about the latent RT distributions. All of these models share the same six core parameters of the basic 2HTM. In line with previous applications of the 2HTM (e.g., Bayen et al., 1996; Bröder & Schütz, 2009; Klauer & Wegener, 1998), we restricted the probability of detecting targets and lures to be identical, $d_o = d_n$, separately for the two memory strength conditions to obtain a testable version of the standard 2HTM. In addition to the two detection parameters, we used four parameters for the guessing probabilities, separately for all four experimental conditions. Note that we included separate response bias parameters for the memory strength conditions because this factor was manipulated between blocks, which can result in different response criteria (Stretch & Wixted, 1998). Whereas all of the six RT-extended MPT models shared these six core parameters, the models differed in their assumptions about the latent RT distributions.

The most complex substantive model (‘CI loss’) assumes four separate latent RT distributions for each of the four experimental conditions. This model allows for any effects of experimental manipulations on the relative speed of the underlying processes. The core assumptions of this model are (a) two-component mixtures for RTs of correct responses to targets and lures and (b) complete-information loss in the uncertainty state. Specifically, the latter assumption implies that guesses are equally fast for targets and lures. This holds for both *old* and *new* guesses, respectively.

The second substantive model (‘Fast detection’) is a submodel of the ‘CI loss’ model. It tests a necessary condition for a serial interpretation of the 2HTM (Dube et al., 2012; Erdfelder et al., 2011). According to the hypothesis that guesses occur strictly after unsuccessful detection attempts, responses due to detection must be faster than responses due to guessing. Specifically, we assumed that this hypothesis of stochastic dominance holds within each response category, that is, within *old*-responses ($L_{do} \geq L_{go}$) and within *new*-responses ($L_{dn} \geq L_{gn}$). Imposing these two constraints in all four experimental conditions results in a total of eight order constraints. Note that adding the corresponding constraints across response categories (i.e., $L_{do} \geq L_{gn}$ and $L_{dn} \geq L_{go}$) would result in the overall constraint

$$\min(L_{do}, L_{dn}) \geq \max(L_{go}, L_{gn}) \quad (5)$$

within each condition. However, this constraint requires the additional assumption that there is no difference in the overall speed of *old* and *new* responses. Since our interest is only in the core assumptions of the RT-extended 2HTM, we tested stochastic dominance only within response categories.

The third substantive model ('Invariant L_{do}') is nested in the 'Fast detection' model. It tests the hypothesis that the speed of the actual recognition process is not affected by response bias, or in other words, the restriction that target detection is similarly fast across base rate conditions, $L_{do}^{30\,\%} = L_{do}^{70\,\%}$. Since the constraint applies in both memory strength conditions, this reduces the number of free parameters by two. The fourth substantive model ('Invariant L_d') adds the corresponding constraint for lure detection to the model 'Invariant L_{do},' thus implying invariance of both L_{do} and L_{dn} against base-rate manipulations.

These four substantive RT-extended MPT models are tested against two reference models that should be rejected if the theoretical assumptions underlying the RT-extended 2HT model hold. The first of these models ('No mix') does *not* assume RT-mixture distributions for correct responses. Instead, it estimates $4 \cdot 4 = 16$ separate RT distributions, one for each response category of all experimental conditions. This model has no additional restrictions in comparison to the standard 2HTM without RTs because the categorization of responses from fast to slow is perfectly fitted within each response category. The second reference model ('Null') tests whether the proposed method is sufficiently sensitive to detect any differences in latent RT distributions at all. For this purpose, it assumes that all observed RTs stem from a single RT distribution by constraining all latency parameters to be equal, $L_j = L_i$ for all i, j.

Model selection results

Before testing any RT-extended MPT model, it is necessary to ensure that the basic MPT model fits the observed responses. Otherwise, model misfit cannot clearly be attributed either to the assumed structure of latent RT distributions or to the MPT model itself. In our case, the standard 2HTM with the restriction of identical detection probabilities for targets and lures ($d_o^{\text{strong}} = d_n^{\text{strong}}, d_o^{\text{weak}} = d_n^{\text{weak}}$) fitted the individual data of all participants well (all $G^2(2) \leq 3.6$, $p \geq .16$, with a statistical power of $1 - \beta = 88.3\,\%$ to detect a small deviation of $w = 0.1$). Figure 8 shows the fitted ROC curves (gray solid lines) and indicates that both experimental manipulations selectively influenced the core parameters as expected.[6] This visual impression was confirmed by likelihood-ratio tests showing that the d-parameters differed between memory strength conditions (all $\Delta G^2(1) > 21.1$, $p < .001$) and that the g-parameters differed between the base-rate conditions (all $\Delta G^2(2) > 11.3$, $p < .004$). Note that we did not constrain the guessing parameters to be identical across memory strength conditions because of possibly different response criteria per block (Stretch & Wixted, 1998). Both Fig. 8 and a likelihood-ratio test showed that this was indeed the case for Participant 3 ($G^2(2) = 15.7$, $p < .001$).

In addition to a good model fit, it is important that the core parameters of the MPT model are not at the boundary of the parameter space. If one of the core parameters is close to zero, the corresponding cognitive process is assumed not to occur, and thus, the corresponding latent RT distribution can neither be observed nor estimated if it does not occur in another part of the tree. Hence, if some of the core parameters are at the boundaries, it is possible that some of the latent RT distributions are empirically not identified. This issue did not arise in our data because all core parameter estimates varied between .10 and .81. Note that core parameter estimates and standard errors were stable across the different RT-extended 2HT models below with mean absolute differences of .013 and .001, respectively (maximum absolute differences of .079 and .008, respectively).

To select between the six competing RT-extended 2HT models, we used the Fisher information approximation (FIA; Rissanen, 1996), an information criterion based on the minimum description length principle (Grünwald, 2007; Rissanen, 1978). Compared to standard goodness-of-fit tests and other model section measures, FIA has a number of advantages. Resembling the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), FIA minimizes errors in predicting new data by making a trade-off between goodness of fit and model complexity. However, whereas AIC and BIC penalize the complexity of a model only by the number of free parameters, FIA takes additional information into account such as order constraints on parameters and the functional form of the competing models (Klauer & Kellen, 2011). This is important in the present case, because three of the models do not differ in the number of parameters. Whereas the substantive model 'CI loss' has the same number of parameters as the reference model 'No mix', it has a smaller complexity because of the structural assumption of conditional independence of RTs. The model 'Fast detection' is even less complex since it adds eight order constraints that decrease the flexibility even more without affecting the number of free parameters.

Besides accounting for such differences in functional flexibility, FIA has the benefits that it is consistent (i.e., it asymptotically selects the true model if it is in the set

[6] The same pattern of selective influence tests also emerged if d_o and d_n were not constrained to be identical (see Supplementary Material).

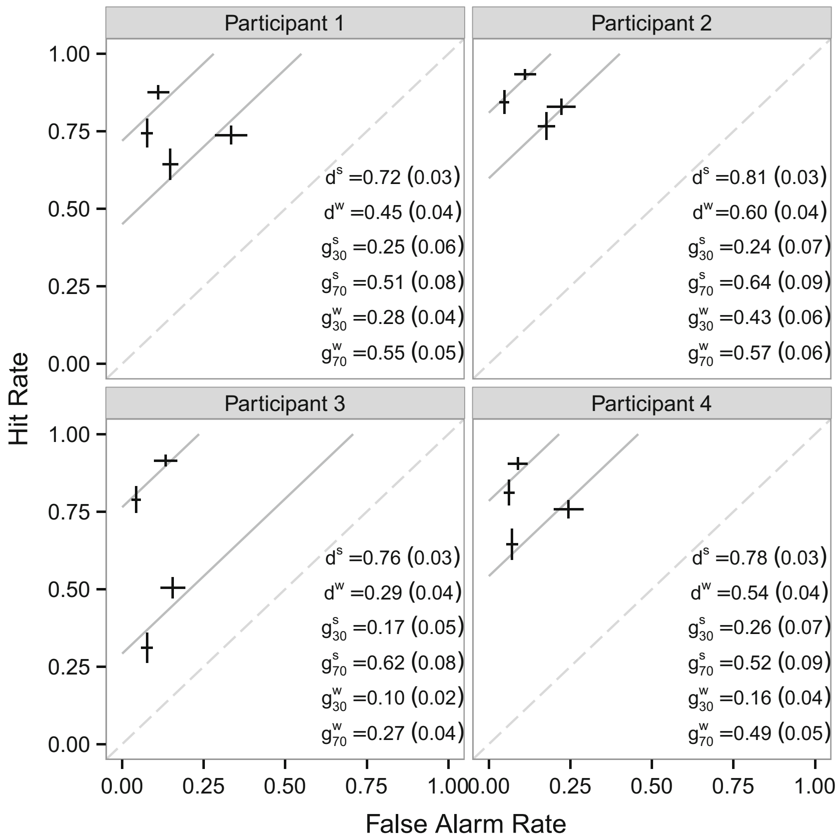

Fig. 8 Fit of the standard 2HTM (*solid gray lines*) to the observed frequencies (shown with standard errors). ML parameter estimates (and standard errors) are displayed for each participant

of competing models) and that it can efficiently be computed for MPT models (Wu et al., 2010). Despite these advantages, FIA can be biased in small samples and should therefore only be used if the empirical sample size exceeds a lower bound (Heck, Moshagen, & Erdfelder, 2014). In the present case, we can use FIA for model selection on the individual level because the lower bound of $N' = 190$ is clearly below 1,200, the number of responses per participant.

Table 1 shows the ΔFIA values for all participants and models, that is, the difference in FIA values to the preferred model. Moreover, the likelihood-ratio tests for all models are shown to assess their absolute goodness of fit. For all participants, both reference models were rejected. Despite its larger complexity, the model 'No mix' did not outperform the substantive models. Concerning the other extreme, the null model which assumes no RT differences at all

Table 1 Model selection for the RT-extended 2HTM models using two RT bins

		Participant 1			Participant 2			Participant 3			Participant 4		
Model	df	G^2	p	ΔFIA	G^2	p	ΔFIA	G^2	p	ΔFIA	G^2	p	ΔFIA
No mix	2	3.6	.16	13.2	0.3	.88	14.5	2.6	.27	13.0	2.0	.37	12.5
CI loss	2	3.6	.16	6.2	0.3	.88	7.4	2.6	.27	5.9	2.0	.37	5.5
Fast detection	2	3.6	.16	0.7	0.3	.88	1.9	2.6	.27	0.4	2.0	.37	**0.0**
Invariant L_{do}	4	6.7	.15	**0.0**	0.8	.93	**0.0**	8.1	.09	0.9	27.5	<.01	10.5
Invariant L_d	6	23.4	<.01	6.5	6.4	.38	0.9	10.0	.13	**0.0**	37.3	<.01	13.5
Null	17	128.7	<.01	45.0	111.3	<.01	39.2	90.0	<.01	25.8	103.1	<.01	32.3

Note. 1,200 responses per participant, which exceeds the lower-bound $N' = 190$ for the application of FIA (Heck et al., 2014) performed worst as indicated by the large ΔFIA values. Hence, the categorization of RTs into two bins preserved information about the relative speed of the underlying processes, which can be used to test our substantive hypotheses.

Concerning the substantive model, detection responses were faster than guessing responses for all participants as indicated by the low ΔFIA values of the model 'Fast detection.' Note that this model fitted just as well as the two more complex models 'No mix' and 'CI loss' and is therefore preferred by FIA due to its lower complexity. Regarding the effect of different base rates on the speed of detection, the results were mixed. For Participant 4, this manipulation clearly affected the detection speed as indicated by the large FIA and G^2 values for the 'Invariant L_{do}' and 'Invariant L_d' models. In contrast, different base rates did not affect the speed of target detection for the other three participants. Moreover, the model 'Invariant L_d,' which additionally restricts the speed of lure detection to be identical across base rates, was selected only for Participant 3.[7]

Overall, our results support the hypotheses of complete-information loss and fast detection. The results were ambiguous with respect to effects of the base-rate manipulation on detection speed. For some participants, the speed of target detection was invariant with respect to different base rates, for others (Participant 4, in particular) detection tended to be faster when the correct response was congruent with the base-rate conditions.

Plotting the relative speed of latent processes

To facilitate comparisons of the relative speed of the assumed cognitive processes, the parameter estimates $\hat{L}_j$ can directly be compared in a single graph for several processes and participants. Since L_j is defined as the probability that responses of process j are faster than the RT boundary, larger estimates directly correspond to faster cognitive processes. Therefore, the latency estimates in Fig. 9, based on the model 'CI loss,' represent a parsimonious and efficient way of comparing the relative speed of the hypothesized processes across experimental conditions and participants.

In line with the model-selection results, target and lure detection were always estimated to be faster than guessing, with the exception of Participant 2 in the base-rate conditions with 70 % targets. Note that many of the differences are quite substantial, especially when comparing the relative speed of target detection, which was clearly faster than guessing across all of our four experimental conditions. Moreover, for Participants 1 and 4, the time required to detect an item and give the correct response was not invariant under different base rates, in line with the rejection of the model 'Invariant L_{d}': These two participants were faster at detecting *old* items when base rates of targets were high (i.e., 70 % targets), and they were faster at rejecting *new* items when base rates of lures were high (i.e., 30 % targets). However, based on the present experiment, we cannot disentangle whether this effect resulted from actual differences in the speed of memory retrieval or from a generally increased preparedness towards the more prevalent response.

Testing stochastic dominance using more than two RT bins

In the following, we compute the Bayes factor based on the encompassing-prior approach (Hoijtink, Klugkist, & Boelen, 2008; Klugkist et al., 2005 to quantify the evidence in favor of the hypothesis that responses due to detection are stochastically faster than those due to guessing. The Bayes factor is defined as the odds of the conditional probabilities of the data $\boldsymbol{y}$ given the models $\mathcal{M}_0$ and $\mathcal{M}_1$ (Kass & Raftery, 1995),

$$\mathrm{B}_{01} = \frac{p(\boldsymbol{y} \mid \mathcal{M}_0)}{p(\boldsymbol{y} \mid \mathcal{M}_1)}, \tag{6}$$

and represents the factor by which the prior odds are multiplied to obtain posterior odds. Moreover, the Bayes factor has a direct interpretation of the evidence in favor of model $\mathcal{M}_0$ compared to model $\mathcal{M}_1$ and can be understood as a weighted average likelihood ratio (Wagenmakers et al., 2015). Note that model selection based on the Bayes factor takes the reduced functional complexity of order-constrained models into account and is asymptotically identical to model selection by FIA under some conditions (Heck, Wagenmakers, & Morey, 2015). In the present case, the model $\mathcal{M}_0$ is the order-constrained RT-extended 2HTM that assumes stochastic dominance within each response category (the model 'Fast detection' above). Moreover, the model $\mathcal{M}_1$ is the unconstrained model that only assumes complete-information loss ('CI loss'). Hence, in the present case, observed Bayes factors substantially larger than one provide evidence in favor of the hypothesis that detection is faster than guessing.

To compute the Bayes factor, we rely on a theoretical result showing that the Bayes factor for order-constrained models is identical to a simple ratio of posterior to prior probability if the prior distributions of the parameters are proportional on the constrained subspace

[7] However, without the constraint $d_o = d_n$, the model 'Fast detection' was selected for Participant 3 (see Supplementary Material).

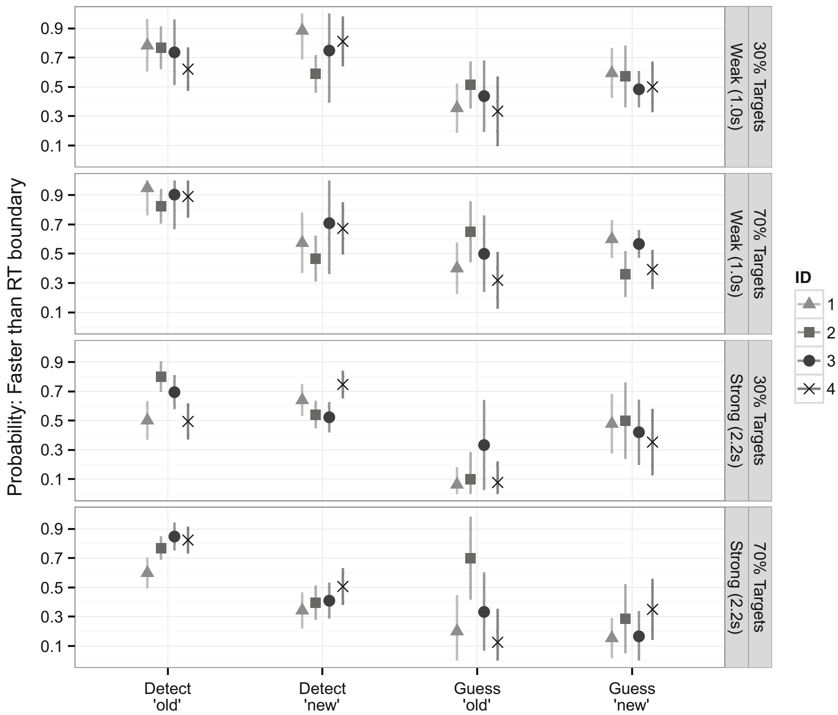

Fig. 9 Estimated probabilities $\hat{L}_{j1}$ that responses generated by each of four processes are faster than the individual RT boundaries, based on the model 'CI loss,' including 95 % confidence intervals

(Klugkist et al., 2005). For instance, the Bayes factor in favor of the simple order constraint $\theta_1 \leq \theta_2$ is identical to

$$B_{01} = \frac{p(\theta_1 \leq \theta_2 \mid \boldsymbol{y}, \mathcal{M}_1)}{p(\theta_1 \leq \theta_2 \mid \mathcal{M}_1)}, \tag{7}$$

where the numerator and the denominator are the posterior and prior probabilities, respectively, that the constraint holds within the unconstrained model $\mathcal{M}_1$. In our case, the simple order constraint $\theta_1 \leq \theta_2$ is replaced by the order constraint representing stochastic dominance in Eq. 4. In practice, this result facilitates the computation of the Bayes factor because one only needs to fit the unconstrained model using standard Markov chain Monte Carlo (MCMC) sampling.

In detail, computing the Bayes factor for the present purpose requires the following steps: (1) obtain RT-extended frequencies using the log-normal approximation or any other strategy as explained above, (2) estimate the full model 'CI loss' by sampling from the posterior distribution using MCMC (e.g., using the software JAGS, Plummer, 2003), (3) count the MCMC samples that fulfill the constraint of stochastic dominance in Eq. 4, (4) compute the Bayes factor as the ratio of this posterior probability vs. the prior probability that the constraint holds

Table 2 Bayesian factors in favor of stochastic dominance

Participant	4 RT bins			8 RT bins		
	B_{Ldo}	B_{Ldn}	p_{T1}	B_{Ldo}	BF_{Ldn}	p_{T1}
1	179.22	50.00	0.29	696.52	2.48	0.12
2	17.45	1.99	0.42	33.34	0.65	0.18
3	201.52	43.85	0.39	1269.57	94.65	0.22
4	174.79	44.81	0.37	288.45	51.88	0.23

The Bayes factors B_{Ldo} and B_{Ldn} quantify the evidence that target detection is faster than guessing *old* and that lure detection is faster than guessing *new*, respectively. The posterior predictive p-value p_{T1} provides an absolute measure of fit for the encompassing model 'CI Loss'

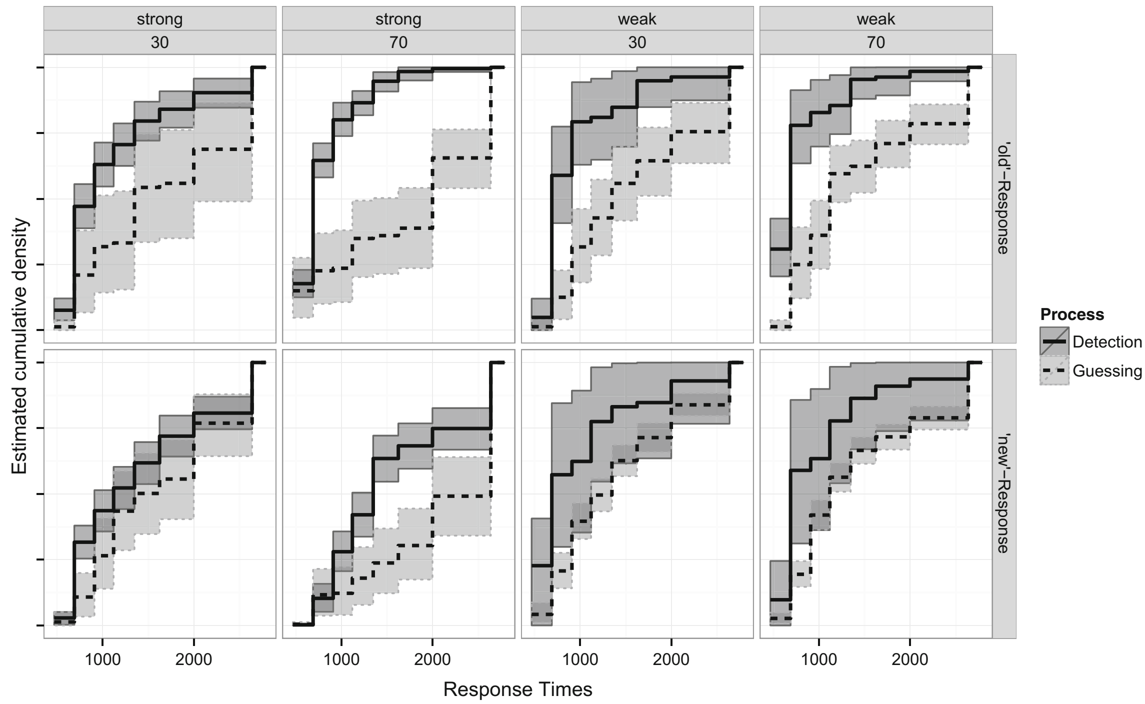

Fig. 10 Estimated cumulative density functions of the latent RT distributions of target detection and guessing *old* (including 80 % credibility intervals) for Participant 3

(Klugkist et al., 2005). In the present case, we estimated the prior probability that the constraint holds based on parameters directly sampled from the prior distribution. Note that we used uniform priors on the core parameters of the 2HTM and symmetric, uninformative Dirichlet priors on each set of the multinomial latency parameters, that is, $(L_{j1}, \ldots, L_{jB}) \sim \mathrm{Dir}(\alpha, \ldots, \alpha)$ with parameters $\alpha = 1/B$, similarly as in Heathcote et al. (2010). Note that the core parameter estimates for d and g differed by less than .02 compared to fitting the RT-extended 2HTM using maximum likelihood.

For testing stochastic dominance, we differentiated between the hypotheses that target detection is faster than guessing *old* and that lure detection is faster than guessing *new*. Table 2 shows the resulting Bayes factors using four and eight RT bins based on 50 million MCMC samples. The table also includes posterior predictive p-values p_{T1} that represent an absolute measure of fit similar to p-values associated with the χ^2 test (Klauer, 2010). The large p_{T1}-values imply that the model did not show substantial misfit for any participant and or number of RT bins. More interestingly, the large Bayes factors $\mathrm{B}_{\mathrm{Ldo}}$ in Table 2 represent substantial evidence that target detection was faster than guessing *old*. In contrast, the corresponding hypothesis that lure detection was faster than guessing *new* was only supported for Participants 3 and 4. For the other two participants, the Bayes factors $\mathrm{B}_{\mathrm{Ldn}}$ was close to one and does therefore neither provide evidence for nor against stochastic dominance.

In addition to computing the Bayes factor, the Bayesian approach allows to directly compute point estimates and posterior credibility intervals for the cumulative density functions of the latent RT distributions sketched in Fig. 5b. To illustrate the substantive meaning of large Bayes factors in favor of stochastic dominance, Fig. 10 shows the mean of the estimated cumulative densities for Participant 3 including 80 % credibility intervals. To facilitate the comparison, the cumulative densities are shown separately for *old* and *new* responses (rows) and for the four experimental conditions (columns). Note that the latency estimates are based on the full model ‘CI loss’ and are therefore not constrained to fulfill stochastic dominance necessarily. However, across all experimental conditions, the cumulative densities of RTs due to detection (dark gray area) are clearly above the cumulative densities of RTs due to guessing (light gray). Moreover, stochastic dominance is more pronounced for target detection and guessing *old* (first row) than for lure detection and guessing *new* (second row). Note that this larger discrepancy for target detection contributes to the larger Bayes factors $\mathrm{B}_{\mathrm{Ldo}}$ compared to $\mathrm{B}_{\mathrm{Ldn}}$.

Conclusion

In sum, the empirical example showed (1) how to test hypotheses about the relative speed of cognitive processes using two RT bins and standard MPT software and (2) how to test stronger hypotheses about the ordering of latent processes in a Bayesian framework using more than two RT bins. With respect to the 2HTM, we found support for the hypothesis that RTs of correct responses emerge as two-component mixtures of underlying detection and guessing processes. Moreover, both modeling approaches supported the hypothesis of slow guesses that occur strictly after unsuccessful detection attempts (Erdfelder et al., 2011). However, the latent RT distributions of detecting targets and lures differed across base-rate conditions for two participants, whereas those of the other two participants were not affected by this manipulation. Hence, further studies are necessary to assess the effects of base-rate manipulations on the speed of detecting targets and lures in recognition memory.

Discussion

We proposed a new method to measure the relative speed of cognitive processes by including information about RTs into MPT models. Basically, the original response categories are split into more fine-grained subcategories from fast to slow. To obtain inter-individually comparable categories, the individual RT boundaries for the categorization into bins are based on a log-normal approximation of the distribution of RTs across all response categories. Using these more informative frequencies that capture both response-type and response-speed information, an RT-extended MPT model accounts for the latent RT distributions of the underlying processes using the latency parameters L_{jb} defined as the probability that a response from the j-th processing path falls into the b-th RT bin. Importantly, this approach does not pose any a-priori restrictions on the unknown shape of the latent RT distributions. Our approach allows for testing whether two or more of the latent RT distributions are identical and whether some latent processes are faster than others. Such hypotheses concerning the ordering of latent processes can be tested by simple order constraints of the form $L_i < L_j$ in the standard maximum-likelihood MPT framework when using two RT bins and within a Bayesian framework when using more RT bins.

Substantively, our empirical example supported the hypothesis of complete-information loss, that is, responses due to guessing are similarly fast for targets and lures. Moreover, we found strong evidence that responses due to target detection are faster than those due to guessing *old*, but weaker evidence that the same relation holds for the speed of lure detection and guessing *new*. Moreover, RTs due to target detection were less affected by base rates than those due to lure detection. Both of these results indicate an important distinction. Whereas target detection seems to occur relatively fast and largely independent of response bias, lure detection seems to be a slower and response-bias dependent process.

Advantages of RT-extended MPT models

The proposed framework provides simple and robust methods to gain new insights. It allows for estimating and comparing the relative speed of cognitive processes based only on the core assumption of MPT models, that is, a finite number of latent processes can account for observed responses. Most importantly, we showed that using discrete RT bins preserves important information in the data that can be used to test substantive hypotheses. Both the power simulation and the empirical example showed that the approach is sufficiently sensitive to detect differences in latent RT distributions. Importantly, this indicates that categorizing RTs into bins does preserve structural information in the data.

The proposed approach also has several practical benefits. Its application is straightforward and based on a principled strategy to categorize RTs, in which the number of RT bins can be adjusted to account for different sample sizes. In practice, the choice of the number of RT bins will typically depend on the substantive question. If the interest is only in a coarse measure of the relative speed of cognitive processes, two RT bins might suffice. However, if sufficient sample sizes are available, using more than two RT bins might be beneficial to obtain more powerful tests that are more sensitive to violations of stochastic dominance.

An additional advantage of the proposed distribution-free framework is that modeling histograms of RTs reduces the sensitivity of the method towards outliers. Moreover, RT-extended MPT models can be fitted using existing software such as multiTree (Moshagen, 2010) or MPTinR (Singmann & Kellen, 2013). Alternatively, inter-individual differences can directly be modeled by adopting one of the Bayesian hierarchal extensions for MPT models proposed in recent years (Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Smith & Batchelder, 2010). Especially in cases with only a small to medium number of responses per participant, these methods are preferable to fitting the model to individual data separately. Last but not least, MPT models can easily be reparameterized to test order constraints (Klauer et al., 2015; Knapp & Batchelder, 2004), thereby facilitating tests of stochastic dominance.

Parametric modeling of latent RT distributions

Our approach aims at providing a general method to include RTs in any kind of MPT model in a distribution-free manner. However, for some substantive questions, more specific models that predict exact parametric shapes for RTs might be preferable. For instance, Donkin, Nosofsky, Gold, and Shiffrin (2013) extended the discrete-slots model for visual working memory, which closely resembles the 2HTM of recognition memory, to RTs by assuming a mixture of two evidence-accumulation processes. The discrete-slots model assumes that in each trial of a change-detection task, a probe either occupies one of the discrete memory slots or it does not. Depending on this true, latent state, RTs either emerge from a memory-based evidence-accumulation process or from a guessing-based accumulation process. Specifically, the time course of both accumulation processes is modeled by separate linear ballistic accumulators, which assume a race between two independent linear processes reaching a boundary (see Brown & Heathcote, 2008 for details).[8] Note that the resulting predicted RT distributions resemble specific instantiations of the latent RT distributions sketched in Fig. 2.

The model by Donkin et al. (2013) makes precise predictions how RT distributions change across different experimental conditions. For instance, the model predicts that the RT distribution of false alarms (i.e., *change*-responses in no-change trials) is invariant with respect to set-size manipulations, that is, how many stimuli are presented. Moreover, the model establishes a functional relationship between the guessing parameter g and the corresponding RT distribution for guessing responses. Given that the response threshold for *change*-guessing is much lower than that for *no-change*-guessing, the linear ballistic accumulator predicts both more and faster *change*-responses. Similarly, the model constraints the structure of RTs across different levels of set size and base rates. Donkin et al. (2013) showed that the relevant patterns in the data followed these predictions of the discrete-slots model. Moreover, the model performed better than competitor models assuming continuous memory strength.

Overall, this example illustrates the advantages of process models that directly transfer psychological theory into precise predictions for responses and RTs. Given such hypotheses about the mechanisms how RTs emerge from the underlying processes, these hypotheses should of course be included in the corresponding statistical model in the form of exact parametric constraints as in Donkin et al. (2013). If the data are in line with such a precise model for a broad range of context conditions, this provides strong support for the theory and clearly more evidence than fitting a less precise, distribution-free model. Compared to such computational models, our focus on measurement models is more general and useful for other types of applications. For instance, if the substantive questions concern only the relative speed or the ordering of latent processes irrespective of the precise nature of the underlying processes, the proposed distribution-free approach allows inclusion of RTs in any MPT model without the necessity to make auxiliary parametric assumptions.

In sum, this shows that both approaches — parametric computational models and measurement models such as RT-extended MPT models — have their advantages and disadvantages, many of which are typical for parametric and nonparametric models in general. Most importantly, the substantive conclusions concerning the underlying cognitive processes should eventually converge irrespective of the approach used.

RT extensions of more complex MPT models

The 2HTM used as an example in the present paper is one of the most simple MPT models. Of course, more complex MPT models might directly benefit from including RTs as well. For instance, in the field of judgment and decision making, the inclusion of RTs in MPT models might be useful in testing between different theoretical accounts. For instance, Hilbig et al. (2010) proposed and validated the r-model, which disentangles use of the recognition heuristic from decisions based on integration of further knowledge. According to the heuristic account, a recognized option is chosen in a noncompensatory manner, that is, further information does not influence the decision at all (Goldstein & Gigerenzer, 1999). As a consequence, decisions based on such a process should be faster compared to responses based on integration of further knowledge. In contrast, global information-integration models assume that a single, underlying process integrates all of the available information in order to make a decision (Glöckner & Betsch, 2008). If further congruent knowledge is used in addition to the actual recognition cue, decisions are predicted to be faster compared to decisions solely based on the recognition cue (Glöckner & Bröder, 2011). This prediction directly contradicts the hypothesis that choice of the recognized option is always faster than information integration. Obviously, an RT-extension of the r-model

[8] Note that this model allows for incorrect responses in the detection state (Donkin et al., 2013, p. 895) and thus differs from the model shown in Fig. 2. However, the corresponding predicted probability of incorrect responses in the detection state appeared to be very small due to a large drift rate towards the correct response.

could help in discriminating between these two theoretical accounts.

Social psychology is another field in which MPT models are often used, for instance, to measure implicit attitudes (e.g., Conrey et al., 2005; Nadarevic & Erdfelder, 2011) or effects of processing fluency (e.g., Fazio et al., 2015; Unkelbach & Stahl, 2009). Concerning measures of implicit attitudes, several authors proposed to rely on MPT models to analyze response frequencies instead of analyzing mean RTs in the implicit association test (IAT) and the go/no-go association task (e.g., Conrey et al., 2005; Meissner & Rothermund, 2013; Nadarevic & Erdfelder, 2011). These MPT models disentangle cognitive processes such as activation of associations and stimulus discrimination (Conrey et al., 2005) or recoding of response categories (Meissner & Rothermund, 2013). However, to do so, these models sacrifice information about the speed of processes and only rely on response frequencies. Obviously, the inclusion of RTs would allow for assessing the selective speed of and testing the temporal order of the hypothesized processes. For instance, one could test whether responses due to implicit associations are actually faster than those due to stimulus discrimination as is often assumed.

Conclusion

We proposed a novel and general method to estimate latent RT distribution of the hypothesized processes in MPT models in a distribution-free way. The approach enables researchers to directly test hypotheses about the relative speed of these processes. Thereby, it is possible to better understand the nature of the cognitive processes accounted for by MPT models. In sum, the proposed method follows the spirit of MPT modeling in general (Batchelder & Riefer, 1999): It provides a simple way for psychologists to test substantive hypotheses about the relative speed of cognitive processes without the need to rely on many auxiliary assumptions.

Acknowledgments We thank Chris Donkin, David Kellen, and Christoph Klauer for valuable comments and suggestions.

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation.

Appendix A: Identifiability of latent RT distributions

In the following, we assume that an MPT model is given with T trees, K_t categories in tree t, and I_{tk} branches that lead to a category C_{tk}. The model is parameterized by the core parameter vector $\boldsymbol{\theta} \in (0,1)^S$ to model the expected branch probabilities $p_{tki}(\boldsymbol{\theta})$ and the resulting category probabilities $p_{tk}(\boldsymbol{\theta}) = \sum_{i=1}^{I_{tk}} p_{tki}(\boldsymbol{\theta})$. For an exact definition of the branch probabilities and other details of the MPT model class, see Hu and Batchelder (1994).

To model latent RT distributions, each branch is further split into B bins with branch probabilities $q_{tkib}(\boldsymbol{\theta}, \boldsymbol{L}) = p_{tki}(\boldsymbol{\theta}) L_{tkib}$, where $L_{tkib} \in (0,1)$ are the latency parameters of interest (collected in the vector $\boldsymbol{L}$). Note that for each branch one of these latency parameters is fixed since $\sum_{b=1}^{B} L_{tkib} = 1$. For notational convenience, we will use the abbreviation $L_{tkiB} = 1 - \sum_{b=1}^{B-1} L_{tkib}$ and thus list B latency parameters for each branch, even though only $B-1$ of these are actually defined as free parameters. The expected category probabilities of the extended MPT model are then given by summing across the original I_{tk} branches of category C_{tk} separately for each RT bin, i.e., $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = \sum_{i=1}^{I_{tk}} q_{tkib}(\boldsymbol{\theta}, \boldsymbol{L})$.

Moreover, it is assumed that some of the latent RT distributions are restricted to be equal based on theoretical arguments. In Observations 3 and 4, we therefore refer to subsets of J RT distributions to facilitate notation. The j-th of these RT distribution is then modeled by the vector of latency parameters $\boldsymbol{L}_j = (L_{j1}, \ldots, L_{jB})'$, which replaces the corresponding latency parameters $(L_{tki1}, \ldots, L_{tkiB})'$ in the appropriate branches.

Observations 1 to 4 assume that the basic MPT model is globally identifiable and that the core parameters are in the interior of the parameter space, $\boldsymbol{\theta} \in (0,1)^S$.

Observation 1 (Core Parameters) If all of the original branches and categories of the basic identifiable MPT model are split into B bins to model RT distributions, the core parameters $\boldsymbol{\theta}$ of this extended MPT model are also globally identifiable. This holds independently of the number or exact parameterization of the latent RT distributions.

Proof We have to show that $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{tkb}(\boldsymbol{\theta}', \boldsymbol{L}')$ for all t, k, and b implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$. From the definition of $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L})$, it follows that

$$\sum_{i=1}^{I_{tk}} p_{tki}(\boldsymbol{\theta}) L_{tkib} = \sum_{i=1}^{I_{tk}} p_{tki}(\boldsymbol{\theta}') L'_{tkib}. \tag{8}$$

Summarizing across all RT bins from $b = 1, \ldots, B$ within each category C_{tk} yields

$$\sum_{i=1}^{I_{tk}} \left(p_{tki}(\boldsymbol{\theta}) \sum_{b=1}^{B} L_{tkib} \right) = \sum_{i=1}^{I_{tk}} \left(p_{tki}(\boldsymbol{\theta}') \sum_{b=1}^{B} L'_{tkib} \right). \tag{9}$$

Since $\sum_{b=1}^{B} L_{tkib} = 1$ and $p_{tk}(\boldsymbol{\theta}) = \sum_{i=1}^{I_{tk}} p_{tki}(\boldsymbol{\theta})$, it follows that $p_{tk}(\boldsymbol{\theta}) = p_{tk}(\boldsymbol{\theta}')$. The identifiability of the original model directly implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$. □

Observation 2 (Single-Branch Identifiability) If a category C_{sm} is only reached by a single branch in a globally identifiable MPT model (i.e., $I_{sm} = 1$), the corresponding latency parameters $\boldsymbol{L}_{sm1} = (L_{sm11}, \ldots, L_{sm1B})'$ of this branch in the RT-extended MPT model are globally identifiable.

Proof We have to show that $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{tkb}(\boldsymbol{\theta}', \boldsymbol{L}')$ for all t, k, and b implies $\boldsymbol{L}_{sm1} = \boldsymbol{L}'_{sm1}$. Because only one branch leads to category C_{sm}, the expected probability of the b-th RT bin of this category is simply $q_{smb}(\boldsymbol{\theta}, \boldsymbol{L}) = p_{sm}(\boldsymbol{\theta}) L_{sm1b}$. Hence, one obtains

$$p_{sm}(\boldsymbol{\theta}) \boldsymbol{L}_{sm1} = p_{sm}(\boldsymbol{\theta}') \boldsymbol{L}'_{sm1}. \tag{10}$$

From Observation 1 and $\boldsymbol{\theta} \in (0,1)^S$, it follows that $p_{sm}(\boldsymbol{\theta}) = p_{sm}(\boldsymbol{\theta}') \neq 0$ and hence $\boldsymbol{L}_{sm1} = \boldsymbol{L}'_{sm1}$. □

Observation 3 (Recursive Identifiability) Assume that the RT bins of category C_{sm} with $I_{sm} > 1$ branches are modeled by $J \leq I_{sm}$ latent RT distributions. If $J - 1$ of these J latent RT distributions are identifiable from other parts of the extended MPT Model (i.e., from Observation 2 and repeated application of Observation 3), the latency parameters $\boldsymbol{L}_{smJ} = (L_{smJ1}, \ldots, L_{smJB})'$ of the remaining J-th RT distribution are globally identifiable.

Proof We have to show that $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{tkb}(\boldsymbol{\theta}', \boldsymbol{L}')$ for all t, k, and b implies $\boldsymbol{L}_{smJ} = \boldsymbol{L}'_{smJ}$. The category probabilities for the RT bins of category C_{sm} can be split into two parts, the one containing the first $J - 1$ identifiable RT distributions and the remaining one using only the J-th RT distribution:

$$q_{smb}(\boldsymbol{\theta}, \boldsymbol{L}) = \sum_{i \in A} p_{smi}(\boldsymbol{\theta}) L_{smib} + \sum_{i \in B} p_{smi}(\boldsymbol{\theta}) L_{smib}, \tag{11}$$

with the disjunct indexing sets

$$A := \{i : \boldsymbol{L}_{tki} = \boldsymbol{L}_{tkj} \text{ for any } j = 1, \ldots, J-1\} \tag{12}$$

$$B := \{i : \boldsymbol{L}_{tki} = \boldsymbol{L}_{tkJ}\}. \tag{13}$$

Hence, since $\boldsymbol{\theta} = \boldsymbol{\theta}'$ from Observation 1 and $\boldsymbol{L}_{smi} = \boldsymbol{L}'_{smi} \forall i \in A$, the equation $q_{smb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{smb}(\boldsymbol{\theta}', \boldsymbol{L}')$ can be written as

$$\begin{aligned} &\sum_{i \in A} p_{smi}(\boldsymbol{\theta}) L_{smib} + \sum_{i \in B} p_{smi}(\boldsymbol{\theta}) L_{smib} \\ &= \sum_{i \in A} p_{smi}(\boldsymbol{\theta}) L_{smib} + \sum_{i \in B} p_{smi}(\boldsymbol{\theta}) L'_{smib}. \end{aligned} \tag{14}$$

Since $\boldsymbol{L}_{smi} = \boldsymbol{L}_{smJ}$ for all $i \in B$, one obtains

$$L_{smJb} \sum_{i \in B} p_{smi}(\boldsymbol{\theta}) = L'_{smJb} \sum_{i \in B} p_{smi}(\boldsymbol{\theta}). \tag{15}$$

Because of $\boldsymbol{\theta} \in (0,1)^S$, it follows that $p_{smi}(\boldsymbol{\theta}) \neq 0$ for all i and hence that $\boldsymbol{L}_{smJ} = \boldsymbol{L}'_{smJ}$. □

Assume that the recursive application of Observations 2 and 3 did not ensure the identifiability of at least one of the hypothesized latent RT distributions. In such a case, the identifiability of the extended MPT model can be shown as follows. Note that this is the most general observation and that it implies Observations 2 and 3.

Observation 4 (General Identifiability) The extended MPT model is globally identifiable if and only if the rank of the following $(\sum_{t=1}^{T} K_t) \times J$ matrix exceeds J, the total number of hypothesized latent RT distributions parameterized by $L_{1b}, \ldots, L_{Jb}$:

$$P(\boldsymbol{\theta}) = [p_{tkj}(\boldsymbol{\theta})]_{\substack{tk=11,\ldots,1K_1,21,\ldots,TK_T \\ j=1,\ldots,J}}. \tag{16}$$

The rows of this matrix contain all observable categories across all trees. The entries of $P(\boldsymbol{\theta})$ are defined as the branch probabilities of the original MPT model if a category C_{tk} includes a branch associated with the latent RT distribution j and are zero otherwise,

$$p_{tkj}(\boldsymbol{\theta}) = \begin{cases} p_{tki}(\boldsymbol{\theta}), & \text{if } L_{tkib} = L_{jb} \text{ for all } b = 1, \ldots, B \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Proof First, we have to show that $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{tkb}(\boldsymbol{\theta}', \boldsymbol{L}')$ for all t, k, and b implies $\boldsymbol{L} = \boldsymbol{L}'$. Using the definition of the matrix $P(\boldsymbol{\theta})$ and $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = \sum_{i=1}^{I_{tk}} p_{tki}(\boldsymbol{\theta}) L_{tkib}$, the system of equations $q_{tkb}(\boldsymbol{\theta}, \boldsymbol{L}) = q_{tkb}(\boldsymbol{\theta}', \boldsymbol{L}')\ \forall t, k, b$ can be written in matrix notation as

$$P(\boldsymbol{\theta}) \boldsymbol{L}_b = P(\boldsymbol{\theta}') \boldsymbol{L}'_b \quad \forall b = 1, \ldots, B, \tag{18}$$

where $\boldsymbol{L}_b = (L_{1b}, \ldots, L_{Jb})'$ (note that this is a vector for the b-th bin across all distributions in contrast to the previous notation). Observation 1 implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$ and hence $P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}')$. If the rank of $P(\boldsymbol{\theta})$ is equal or larger than J, the null space of $P(\boldsymbol{\theta})$ is trivial (i.e., only contains the null vector $\boldsymbol{0}$). Hence,

$$P(\boldsymbol{\theta})(\boldsymbol{L}_b - \boldsymbol{L}'_b) = \boldsymbol{0} \quad \forall b = 1, \ldots, B \tag{19}$$

implies $\boldsymbol{L}_b - \boldsymbol{L}'_b = \boldsymbol{0}$, or equivalently $\boldsymbol{L}_b = \boldsymbol{L}'_b$ for all bins $b = 1, \ldots, B$.

Second, we have to show that the models' identifiability implies that $P(\boldsymbol{\theta})$ has full rank. We proof this by contradiction and assume that $P(\boldsymbol{\theta})$ does *not* have full rank. As before, the definition of $P(\boldsymbol{\theta})$ and the identifiability of $\boldsymbol{\theta}$ implies

$$P(\boldsymbol{\theta})(\boldsymbol{L}_b - \boldsymbol{L}'_b) = \boldsymbol{0} \quad \forall b = 1, \ldots, B. \tag{20}$$

However, because the rank of $P(\boldsymbol{\theta})$ is smaller than the dimensionality of the vector $\boldsymbol{L}$, its null space is not trivial. Hence, a solution $\boldsymbol{a} \neq \boldsymbol{0} \in \mathbb{R}^J$ exists with $\boldsymbol{L}_b - \boldsymbol{L}'_b = \boldsymbol{a}$. This implies $\boldsymbol{L}_b \neq \boldsymbol{L}'_b$ and contradicts the models' identifiability. □

Note that computing the rank of the matrix $P(\boldsymbol{\theta})$ in Observation 4 is less complex than examining the rank of the whole system of model equations of the extended MPT model. Moreover, computer algebra software is available facilitating the search for a set of restrictions on the latent RT distributions that ensures identifiability (cf. Schmittmann et al., 2010).

As an example for the application of Observation 4, the 2HTM with the assumption of conditional independence of latent RTs for guessing results in the matrix

$$\begin{array}{rcl} & \begin{array}{cccc} L_{do} & L_{g-old} & L_{g-new} & L_{dn} \end{array} & \\ P(\boldsymbol{\theta}) = & \begin{pmatrix} d_o & (1-d_o)g & 0 & 0 \\ 0 & 0 & (1-d_o)(1-g) & 0 \\ 0 & (1-d_n)g & 0 & 0 \\ 0 & 0 & (1-d_n)(1-g) & d_n \end{pmatrix} & \begin{array}{c} Hit \\ Miss \\ FA \\ CR \end{array}, \end{array}$$

which has full rank if $d_o, d_n, g \in (0, 1)$ and thus ensures the identifiability of the model (given that the core parameters d_o, d_n, and g are identified). In contrast, the following (non-theoretical) assignment of four latent RT distributions is not identifiable if $d_o = d_n$ and $g = .5$, because the first and last row are then linearly dependent,

$$\begin{array}{rcl} & \begin{array}{cccc} L_{d} & L_{g-correct} & L_{Miss} & L_{FA} \end{array} & \\ P(\boldsymbol{\theta}) = & \begin{pmatrix} d_o & (1-d_o)g & 0 & 0 \\ 0 & 0 & (1-d_o)(1-g) & 0 \\ 0 & 0 & 0 & (1-d_n)g \\ d_n & (1-d_n)(1-g) & 0 & 0 \end{pmatrix} & \begin{array}{c} Hit \\ Miss \\ FA \\ CR \end{array}. \end{array}$$

Appendix B: Testing stochastic dominance of latent RT distributions

In the following, we illustrate why stochastic dominance cannot be tested directly by linear inequality constraints in a binary MPT representation if more than two RT bins are used. Consider the binary reparameterization for four

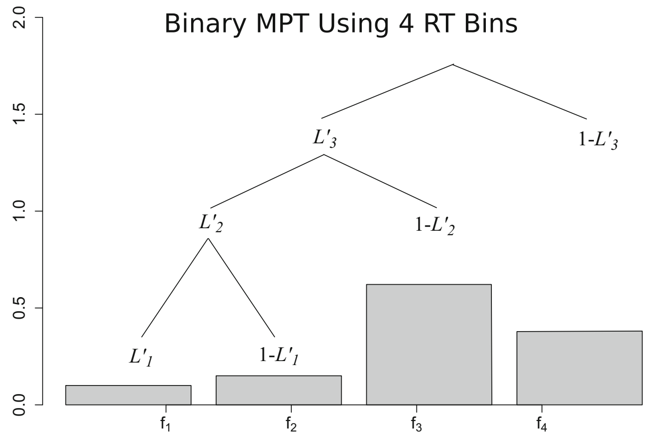

Fig. 11 A possible reparameterization of an RT-extended MPT model with 4 RT bins into a binary MPT model

RT bins in Fig. 11. In this model, the new parameters L'_{ib} of the binary MPT represent the conditional probabilities that a response from process i falls into one of the bins $\{1, \ldots, b\}$, given that it is not in a bin $\{b+2, \ldots, B\}$. Using this reparameterization, the L'_{jb} are mapped to the original parameters L_{ib} by

$$\sum_{k=1}^{b} L_{jk} = \prod_{k=b}^{B-1} L'_{ik} \tag{21}$$

Hence, testing stochastic dominance (Eq. 4) is equivalent to testing

$$\prod_{k=b}^{B-1} L'_{jk} \leq \prod_{k=b}^{B-1} L'_{ik} \;\; \forall b = 1, \ldots, B-1 \tag{22}$$

in the binary MPT model. Note that this is a set of nonlinear order constraints, which cannot directly implemented with the standard MPT methods for linear order constraints (Klauer et al., 2015; Knapp & Batchelder, 2004). Importantly, testing the order of the binary latency parameters directly (i.e., testing $L'_{jb} \leq L'_{ib}$ for all b) will result in an overly conservative test. More precisely, it tests only a sufficient but not a necessary condition for Eq. 22. In other words, the property of stochastic dominance in Eq. 22 can perfectly hold, whereas the order restrictions on the individual binary parameters are violated. For instance, with three RT bins, if $L'_{j2} = .2 \leq L'_{i2} = .8$ (Eq. 22 for $b = 2$), stochastic dominance still holds if $.2L'_{j1} \leq .8L'_{i1}$ (Eq. 22 for $b = 1$), for example, for $L'_{j1} = .8 > L'_{i1} = .4$.

The binary model in Fig. 11 can be reparameterized into statistically equivalent binary MPT models that differ in the definition of the L' parameters. However, it can be expected that the set of nonlinear order restrictions in Eq. 22 will also constrain the parameters of the new binary MPT model in a nonlinear way.

References

Agresti, A. (2013). *Categorical data analysis*. Hoboken: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pp. 267–281.

Ashby, F.G., Prinzmetal, W., Ivry, R., & Maddox, W.T. (1996). A formal theory of feature binding in object perception. *Psychological Review*, *103*, 165–192. doi:10.1037/0033-295X.103.1.165

Bamber, D., & van Santen, J.P.H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. doi:10.1006/jmps.1999.1275

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. doi: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. doi: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. doi:10.3758/BF03210812

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. doi:10.1037/0278-7393.22.1.197

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606. doi:10.1037/a0015279

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

Brown, V. (1998). Recent progress in mathematical psychology: Psychophysics, knowledge, representation, cognition, and measurement. C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), (pp. 253–284). NJ: Mahwah.

Buchner, A., Erdfelder, E., Steffens, M.C., & Martensen, H. (1997). The nature of memory processes underlying recognition judgments in the process dissociation procedure. *Memory & Cognition*, *25*, 508–517. doi:10.3758/BF03201126

Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., & Groom, C.J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*, 469–487. doi:10.1037/0022-3514.89.4.469

Donkin, C., Nosofsky, R.M., Gold, J.M., & Shiffrin, R.M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*, 873–902. doi:10.1037/a0034247

Dube, C., Starns, J.J., Rotello, C.M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language*, *67*, 389–406. doi:10.1016/j.jml.2012.06.002

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

Erdfelder, E., Castela, M., Michalkiewicz, M., & Heck, D.W. (2015). The advantages of model fitting compared to model simulation in research on preference construction. *Frontiers in Psychology*, *6*, 140. doi:10.3389/fpsyg.2015.00140

Erdfelder, E., Küpper-Tetzel, C.E., & Mattern, S.D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*, 7–22.

Fazio, L.K., Brashier, N.M., Payne, B.K., & Marsh, E.J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. doi:10.1037/xge0000098

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1055–1075. doi:10.1037/0278-7393.34.5.1055

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making*, *6*, 23–42.

Goldstein, D.G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.) *Simple heuristics that make us smart*, (pp. 37–58). New York: Oxford University Press.

Grünwald, P. (2007). *The minimum description length principle*. Cambridge: MIT Press.

Heathcote, A., Brown, S., Wagenmakers, E.J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology*, *54*, 454–463. doi:10.1016/j.jmp.2010.06.005

Heck, D.W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, *60*, 29–34. doi:10.1016/j.jmp.2014.06.002

Heck, D.W., Wagenmakers, E.-J., & Morey, R.D. (2015). Testing order constraints: Qualitative differences between Bayes factors and normalized maximum likelihood. *Statistics & Probability Letters*, *105*, 157–162. doi:10.1016/j.spl.2015.06.014

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. doi:10.1037/a0017518

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

Hoijtink, H., Klugkist, I., & Boelen, P.A. (Eds.) (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.

Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, *45*, 603–634. doi:10.1006/jmps.2000.1340

Hu, X., & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/bf02294263

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

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. doi:10.1080/01621459.1995.10476572

Kellen, D., & Klauer, K.C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*, 542–557.

Kellen, D., Klauer, K.C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*, 693–719. doi:10.3758/s13423-013-0407-2

Kellen, D., Singmann, H., Vogt, J., & Klauer, K.C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology*, *62*, 40–53. doi:10.1027/1618-3169/a000272

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

Klauer, K.C., Singmann, H., & Kellen, D. (2015). Parametric order constraints in multinomial processing tree models: An extension of Knapp and Batchelder. *Journal of Mathematical Psychology*, *64*, 1–7. doi:10.1016/j.jmp.2014.11.001

Klauer, K.C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 680–703. doi:10.1037/0278-7393.33.4.680

Klauer, K.C., & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*, 430–450. doi:10.1016/j.jmp.2011.09.002

Klauer, K.C., & Kellen, D. (2015). The flexibility of models of recognition memory: The case of confidence ratings. *Journal of Mathematical Psychology*, *67*, 8–25. doi:10.1016/j.jmp.2015.05.002

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477.

Knapp, B.R., & Batchelder, W.H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*, 215–229. doi:10.1016/j.jmp.2004.03.002

Lahl, O., Göritz, A.S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods*, *41*, 13–19. doi:10.3758/BRM.41.1.13

Link, S.W. (1982). Correcting response measures for guessing and partial information. *Psychological Bulletin*, *92*, 469–486. doi:10.1037/0033-2909.92.2.469

Luce, R.D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324. doi:10.3758/s13428-011-0168-7

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. doi:10.1007/s11336-013-9374-9

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817. doi:10.3758/pbr.16.5.798

Meiser, T. (2005). A hierarchy of multinomial models for multidimensional source monitoring. *Methodology*, *1*, 2–17. doi:10.1027/1614-1881.1.1.2

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 116–137. doi: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. doi:10.1037/a0030734

Meyer, D.E., Yantis, S., Osman, A.M., & Smith, J.K. (1985). Temporal properties of human information processing: Tests of discrete versus continuous models. *Cognitive Psychology*, *17*, 445–518. doi:10.1016/0010-0285(85)90016-7

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

Nadarevic, L., & Erdfelder, E. (2011). Cognitive processes in implicit attitude tasks: An experimental validation of the Trip Model. *European Journal of Social Psychology*, *41*, 254–268. doi:10.1002/ejsp.776

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155–156. doi:10.3758/BF03328004

Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (Vol. 124, p. 125). Vienna, Austria*.

Province, J.M., & Rouder, J.N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*, 14357–14362. doi:10.1073/pnas.1103880109

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471. doi:10.1016/0005-1098(78)90005-5

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47. doi:10.1109/18.481776

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. D. E. Meyer, & S. Kornblum (Eds.) *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, (pp. 611–653). Cambridge: The MIT Press.

Schmittmann, V.D., Dolan, C.V., Raijmakers, M.E., & Batchelder, W.H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. doi:10.3758/BRM.42.3.836

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Schweickert, R., & Chen, S. (2008). Tree inference with factors selectively influencing processes in a processing tree. *Journal of Mathematical Psychology*, *52*, 158–183. doi:10.1016/j.jmp.2008.01.004

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575. doi: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. doi: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. doi: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. doi:10.1037/0096-3445.117.1.34

Steffens, M.C., Buchner, A., Martensen, H., & Erdfelder, E. (2000). Further evidence on the similarity of memory processes in the process dissociation procedure and in source monitoring. *Memory & Cognition*, *28*, 1152–1164. doi:10.3758/BF03211816

Stretch, V., & Wixted, J.T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379–1396. doi:10.1037/0278-7393.24.6.1379

Townsend, J.T., & Ashby, F.G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. doi:10.1016/j.concog.2008.09.006

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424–465. doi:10.3758/BF03214357

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, *2*, 20–54. doi:10.3758/BF03214411

Vandekerckhove, J.S., Matzke, D., & Wagenmakers, E. (2015). Model comparison and the principle of parsimony. In *Oxford Handbook of Computational and Mathematical Psychology (pp. 300–319)*. New York: Oxford University Press.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M.D., Matzke, D., & Morey, R.D. (2015). A power fallacy. *Behavior Research Methods*, *47*, 913–917. doi:10.3758/s13428-014-0517-4

Wu, H., Myung, J.I., & Batchelder, W.H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, *17*, 275–286. doi:10.3758/PBR.17.3.275

Yantis, S., Meyer, D.E., & Smith, J.K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, *110*, 350–374. doi:10.1037/0033-2909.110.2.350

Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes

Additional Model Tests

Daniel W. Heck and Edgar Erdfelder

February 18, 2016

To assess the robustness of our results, we also fitted the standard and RT-extended 2HTM without constraining the detection probabilities d_o and d_n to be identical.

First, we tested whether the core parameters where selectively influenced by the experimental manipulations as expected. This was indeed the case, as indicated by significant model misfit when constraining the detection parameters to be identical across memory strength conditions (all $G^2(2) > 21.3$, $p < .001$) or constraining the guessing parameters to be identical across base-rate conditions (all $G^2(2) > 11.6$, $p < .003$). In contrast, guessing parameters did not differ significantly across memory-strength conditions (all $G^2(2) < 3.8$, $p > .15$).

Table 1 shows that the same models are selected for all participants except for Participant 3. Whereas the more constrained model 'Invariant L_d' was preferred for this participant in the main text including the constraint $d_o = d_n$, the more complex model 'Fast detection' is preferred without the constraint. Thus, the notion of stochastic dominance of detection RTs over guessing RTs is supported, contrary to the hypothesis of identical detection speeds across base rate manipulations.

Figure 1 shows the latency estimates based on the unconstrained version of the model 'CI loss,' which do not substantially differ from those reported in the main text.

		Participant 1			Participant 2			Participant 3			Participant 4		
Model	df	G^2	p	ΔFIA	G^2	p	ΔFIA	G^2	p	ΔFIA	G^2	p	ΔFIA
No mix	0	0.0	–	12.3	0.0	–	13.2	0.0	–	11.8	0.0	–	12.1
CI loss	0	0.2	–	5.9	0.0	–	6.7	0.5	–	5.5	0.0	–	5.5
Fast detection	0	0.3	–	0.4	0.0	–	1.1	0.5	–	**0.0**	0.0	–	**0.0**
Invariant L_{do}	2	2.3	0.32	**0.0**	0.4	0.80	**0.0**	5.4	0.07	1.1	12.7	< 0.01	5.0
Invariant L_d	4	11.6	0.02	2.9	5.3	0.26	0.6	7.5	0.11	0.4	22.2	< 0.01	8.0
Null	15	125.1	< 0.01	45.3	111.1	< 0.01	39.2	87.3	< 0.01	25.9	101.1	< 0.01	33.1

Table 1: Model selection for the RT-extended 2HTM models without constraining d_o and d_n to be identical (for two RT bins). The 1,200 responses per participant exceed the lower-bound $N' = 312$ for the application of FIA (Heck, Moshagen, & Erdfelder, 2014).

	4 RT bins			8 RT bins		
Participant	$\mathrm{B}_{\mathrm{Ldo}}$	$\mathrm{B}_{\mathrm{Ldn}}$	p_{T1}	$\mathrm{B}_{\mathrm{Ldo}}$	$\mathrm{BF}_{\mathrm{Ldn}}$	p_{T1}
1	168.00	29.35	0.40	728.15	0.90	0.15
2	16.86	10.12	0.40	14.65	14.28	0.27
3	199.32	33.54	0.32	1345.20	39.38	0.25
4	172.01	38.49	0.36	239.11	16.36	0.22

Table 2: The Bayes factors $\mathrm{BF}_{\mathrm{Ldo}}$ and $\mathrm{BF}_{\mathrm{Ldn}}$ quantify the evidence that target detection is faster than guessing *old* and that lure detection is faster than guessing *new*, respectively. The model contains separate parameters d_o and d_n for detection of old and new items, respectively.

Table 2 shows the Bayes factors in favor of stochastic dominance of detection RT distributions over the corresponding guessing RT distributions when estimating separate parameters d_o and d_n for target and lure detection, respectively (based on eight MCMC chains with one million samples each).

References

Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, *60*, 29–34. doi:10.1016/j.jmp.2014.06.002

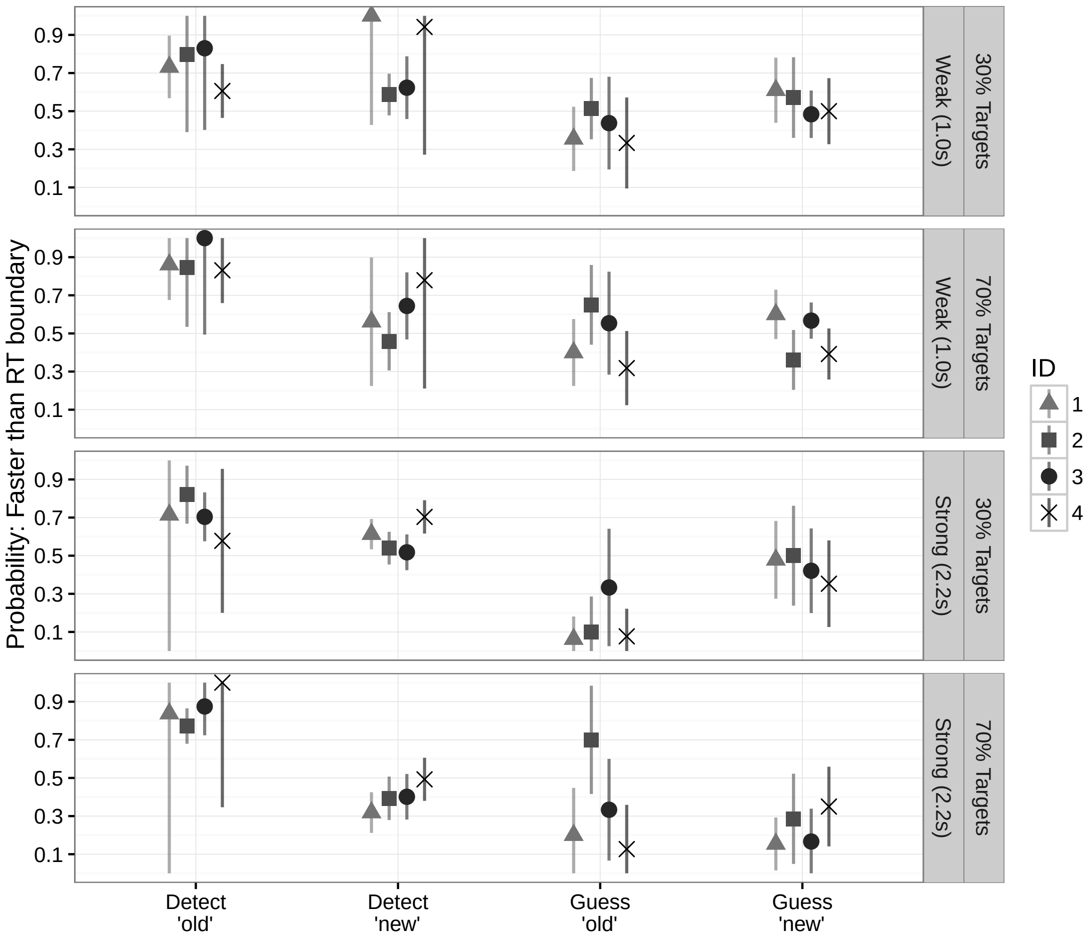

Figure 1: Estimated probabilities $\hat{L}_{j1}$ that responses generated by each of four processes are faster than the individual RT boundaries, based on the model 'CI loss,' without constraining the different detection probabilities d_o and d_n to be identical (including 95% confidence intervals).

Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes

Monte Carlo Simulations

Daniel W. Heck and Edgar Erdfelder

February 18, 2016

In the following, we provide simulation results showing that the data-dependent choice of RT boundaries does not bias the statistical inferences, both with regard to goodness-of-fit tests and standard errors of the latency parameters. Specifically, we compare two strategies to obtain RT boundaries that are both based on the overall RT distribution across all response categories. Whereas the first strategy uses the exact empirical b/B-quantiles (where B is the number of bins and $b = 1, ..., B-1$; e.g., 25%-, 50%-, and 75%-quantiles for $B = 4$), the second strategy uses b/B-quantiles of a log-normal approximation as explained in the main text. Note that the log-normal strategy requires only two statistics of the RT data (mean and variance of log(RT)) and is therefore expected to perform better than the empirical strategy, which requires $B-1$ statistics. For comparison, we also included fixed RT boundaries fixed a-priori for which the asymptotic theory for multinomial models necessarily holds.

As outlined in detail below, RT boundaries based on empirical quantiles resulted in slightly worse results compared to the other two strategies. Given the advantages of data-dependent RT boundaries specifically tailored to each participant, we thus recommend log-normal quantiles as the default for practical applications.

1 Method: Simulation Details

The setting of the simulations is identical to the power simulation in the main text. Data are generated based on the RT-extended 2HTM for a single experimental condition. For the core parameters, we used the true values $d = d_o = d_n = .7$ and $g = .5$. We generated RTs from only two latent RT distributions (detection and guessing), that is, we did not differentiate between target and lure detection and between guessing *old* and *new*. The ex-Gaussian parameters of the guessing latency distribution were chosen as $\mu_g = 1,000$ ms, $\sigma = 100$ ms, and $\lambda = 300$ ms. In the first simulation, the detection and guessing distribution were identical to ensure that the null hypotheses of both goodness-of-fit tests were true, that is, that the model generated the data (tested by the absolute G^2-value) and that the two RT distributions for detection and guessing were identical, H_0: $L_{db} = L_{gb}$ for all bins $b = 1, ..., B - 1$ (tested by the log-likelihood difference ΔG^2). In contrast, in the second simulation, we used a different mean $\mu_d = 800$ ms for the detection RT distribution to test the consistency of the L-parameters in the more interesting case of different underlying RT distributions.

Note that our setting for the simulation was deliberately chosen as a worst-case scenario. First, we included only two MPT trees and only two latent RT distributions. In such a case, the data-dependent choice of RT-boundaries restricts the frequencies within RT-bins to a greater extent compared to a model with many trees and several latent RT distributions. To see this, consider Table 1. If RT boundaries are chosen based on empirical RT quantiles of the overall RT distribution, the column sums are fixed to N/B.[1] If more trees are added to the model, this restriction affects the frequencies in each tree to a smaller extent and hence the relative dependency across trees is reduced. For each set of latency parameters $\{L_{1b}, ..., L_{Jb}\}$, there exists an unknown functional relationship ensuring that the predicted column sums are equal to N/B. This constraint must be fulfilled by design and thus reduces the variability of the L-parameters, because it is not taken into account by the estimated standard errors as shown in the simulations below. However, if the number of latent RT distributions J and thus the number of parameters L_{jb} increases, this constraint will have a smaller impact on the L-parameters compared to the present model with only two RT distributions.

In addition to using only two MPT trees and two latent RT distributions, we included

[1]Note that quantiles based on a log-normal approximation will pose a similar, but more complex constraint on the RT-extended frequencies. However, since the approximation uses only the mean and variance of the overall log(RT) distribution, this constraint will have less impact than that shown in the table. This is also indicated by the simulations below.

Table 1: Comparison of the frequencies used to analyze standard vs. RT-extended MPT models based on empirical quantiles as RT boundaries. Note that the RT bins and the original categories are orthogonal, each with fixed marginal sums.

		RT-extended			**Standard MPT**	
Tree	Category	Fast	...	Slow	Frequencies	N per tree
	C_{11}	n_{111}	...	n_{11B}	n_{11}	
T_1	...	...	...	...	...	N_1
	C_{1I_1}	n_{1I_11}	...	n_{1I_1B}	n_{1I_1}	
	C_{21}	n_{211}	...	n_{21B}	n_{21}	
T_2	...	...	...	...	...	N_2
	C_{2I_2}	n_{2I_21}	...	n_{2I_2B}	n_{2I_2}	
Sum		$\frac{N}{B}$	...	$\frac{N}{B}$		N

small sample sizes in our simulation that are below a reasonable threshold in some cases. For instance, when modeling 100 responses using eight RT bins, these responses must be distributed across $2 \cdot 2 \cdot 8 = 32$ categories in the RT-extended 2HTM. As a consequence, one of the conditions for the asymptotic χ^2 approximation, namely, expected cell frequencies larger than or equal to five, will necessarily be violated. Similarly, such small samples do not ensure that the asymptotic approximation of the standard errors by the observed Fisher information matrix is valid. For all scenarios, the simulation results based on fixed RT boundaries serve as a baseline that differs only in one critical detail, that is, how the RT boundaries where chosen.[2] The results for fixed RT boundaries enable us to assess the differential robustness of multinomial modeling procedures against data-dependent choices of RT boundaries. If the statistical inferences are robust even under such adverse conditions, data-dependent RT boundaries can be assumed to have only negligible effects.

2 Robustness of the Goodness-of-Fit Test

The distributions of simulated G^2-values of the goodness-of-fit tests are compared to the expected χ^2 distribution under the null hypothesis using QQ-plots and Kolmogorov-

[2]The exact RT boundaries were: 1,300 ms using two RT bins; 1,100 ms, 1,300 ms, and 1500 ms using four RT bins; and seven equally-spaced points between 900 ms and 1,700 ms using eight RT bins.

Smirnov tests. Figures 1, 2, and 3 show the QQ-plots of the absolute goodness-of-fit test for fixed boundaries, empirical quantiles, and log-normal quantiles, respectively. In each of these plots, separate QQ-plots are shown for sample sizes of $N = 100$, 200, and 400 (columns) when using $B = 2$, 4, and 8 RT bins (rows) based on 5,000 simulated data sets. For comparison, the diagonal through the origin indicates the optimal result, that is, that the simulated G^2-values follow the asymptotic χ^2 distribution. Moreover, the vertical dashed line indicates the critical χ^2 value for the $\alpha = 5\%$ significance level, and the values $\hat{\alpha}$ show the relative frequency of G^2 values larger than this critical value (i.e., the actual Type I error rate).

As expected, for all three RT-boundary strategies, the simulated G^2 distribution was closer to the asymptotic χ^2 distribution for larger sample sizes. Similarly, the approximation was more adequate when using only two compared to four or even eight RT bins. Note that these results simply emerge from the fact that a sufficiently large expected sample size per cell is required to ensure the asymptotic χ^2 distribution of the G^2-statistic. In all but one case (using eight RT bins and $N = 50$), the actual Type I error rate is slightly too large. Note that this is less critical for the absolute goodness-of-fit test since we aim at not finding a significant p-value. Besides these issues regarding sample size, the distribution of absolute G^2 values does not show a systematic deviation when using data-dependent RT boundaries instead of fixed RT boundaries.

Figures 4, 5, and 6 show the QQ-plots for the simulated ΔG^2 of the nested-model test for the restriction $L_{db} = L_{gb}$ for all bins $b = 1, ..., B - 1$. Resembling the results for the absolute goodness-of-fit test, the approximation of the ΔG^2-distribution by a χ^2-distribution seemed to be sufficiently adequate given large sample sizes. Most importantly, the results are comparable for different strategies of choosing RT boundaries.

To corroborate these visual inspections, we also computed p-values of the Kolgomorov-Smirnov (KS) test, which tests whether the discrepancy between the observed G^2 and the expected χ^2 distribution significantly deviates from chance. We ran this test for a wide range of samples sizes from 300 to 1,000. In each of these conditions, we generated 300 data sets, fitted the RT-extended MPT model, and computed the p-value of an KS test. Figure 7 shows the resulting p-values of the absolute goodness-of-fit test as a function of sample size, separately for the number of bins (rows) and the RT-boundary strategy (columns). The results clearly show that, as predicted by asymptotic theory, less p-values are significant when sample size increases. As shown by the QQ-plots, the χ^2 approximation is more adequate for two RT bins compared to four or eight. Both of these trends emerged for all strategies and number of RT bins. Most importantly, the pattern of p-values does not show a systematic bias when using data-dependent strategies compared to using fixed RT boundaries. The same pattern did again emerge for the nested-model test (Figure 8). Hence, we conclude that the asymptotic χ^2 approximation still holds in our case, provided that the sample size is sufficiently large.

Note that our simulation results are in line with analytical work in statistics concerning χ^2 goodness-of-fit tests based on data-dependent cells (Andrews, 1988a, 1988b; Moore & Spruill, 1975; Pollard, 1979). In many areas, data-dependent cells (also called random cells) are often used for testing whether some continuous data follow a specific distribution, for example, a multivariate Gaussian distribution. Often, such an assumption is tested by partitioning the continuous space into disjoint areas and computing the squared norm of the normalized difference between observed vs. expected frequencies of observations in each bin (a generalized version of Pearson's X^2). This yields a statistic that asymptotically follows the distribution of a sum of χ^2 random variables (Andrews, 1988a, 1988b; Moore & Spruill, 1975). Even though these analytical solutions do not directly apply to the present approach, they hint at the possibility that the asymptotic χ^2 distribution is indeed robust with respect to random cell boundaries.

3 Point Estimates and Standard Errors of Latency Parameters

To check the consistency of the latency parameters, we set the means of the normal component to μ_d = 800ms and μ_g = 1000ms as mentioned above. We derived the expected, 'true' values for the latency parameters L_{jb} by applying the RT-boundary strategies to the full likelihood function of the data generating model.

The results are shown in Figures 9, 10, and 11 when using two, four, and eight bins, respectively. The mean point estimates and standard errors are shown by solid and dashed lines, respectively, whereas the actual distributions of $\hat{L}_{jb}$ are shown by gray areas. A comparison of the point estimates with the true values (i.e., the black points at the right of each panel) indicates that the method is unbiased even with small sample sizes. Given that the empirically determined RT boundaries converge to some fixed values when using a precisely specified strategy, this is to be expected.

The standard errors of these estimates are in line with the standard deviation of the point estimates when using fixed RT boundaries (first row) for sufficiently large samples (e.g., around 100 to 150 when using eight RT bins). When RT boundaries were chosen based on the empirical quantiles of the overall RT distribution (second row), the estimated standard errors of the L-parameters for detection were slightly too large. This result is in line with the fact that the empirical quantiles constrain a whole set of L-parameters as mentioned above, which is not considered by the standard statistical theory for multinomial models. Most importantly, the use of log-normal quantiles (third row) results in valid estimates of the standard errors with a similar precision than those of fixed RT boundaries. Hence, we recommend to rely on this last strategy to achieve an optimal trade-off of statistical accuracy and practical considerations (nonzero frequencies, comparable interpretation of bins across participants).

Note that standard errors can in principle be bootstrapped as in the present simulation (Efron & Tibshirani, 1997). However, when doing so, it is necessary to first sample from a model with continuous RTs and then fit the RT-MPT using the categorization strategy of interest (e.g., based on empirical or log-normal quantiles). If data are generated from the multinomial distribution directly, the constraints of RT-bin frequencies across trees are obviously absent in the data.

References

Andrews, D. W. K. (1988a1). Chi-square diagnostic tests for econometric models. *Journal of Econometrics*, *37*, 135–156. doi:10.1016/0304-4076(88)90079-6

Andrews, D. W. K. (1988b1). Chi-square diagnostic tests for econometric models: Theory. *Econometrica*, *56*, 1419–1453. doi:10.2307/1913105. JSTOR: 1913105

Efron, B. & Tibshirani, R. J. (1997). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Moore, D. S. & Spruill, M. C. (19751). Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics*, *3*, 599–616. JSTOR: 2958431

Pollard, D. (1979). General chi-square goodness-of-fit tests with data-dependent cells. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *50*, 317–331. doi:10.1007/BF00534153

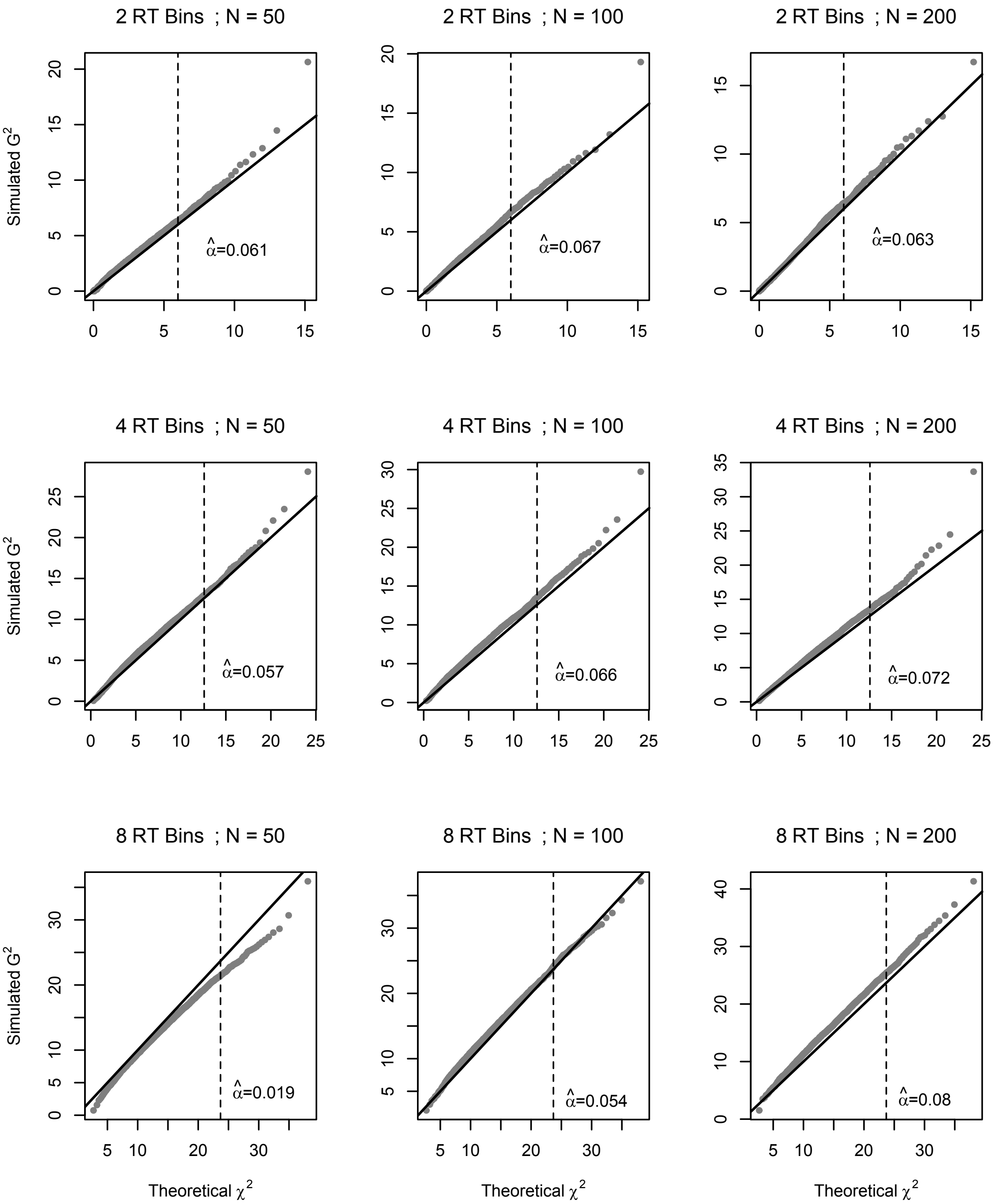

Figure 1: QQ-plot for the G^2-values of the absolute goodness-of-fit test using RT boundaries based on *fixed* RT boundaries. The solid line indicates perfect agreement of the theoretical χ^2 and the observed G^2 distributions. Actual simulation results are indicated by gray dots.

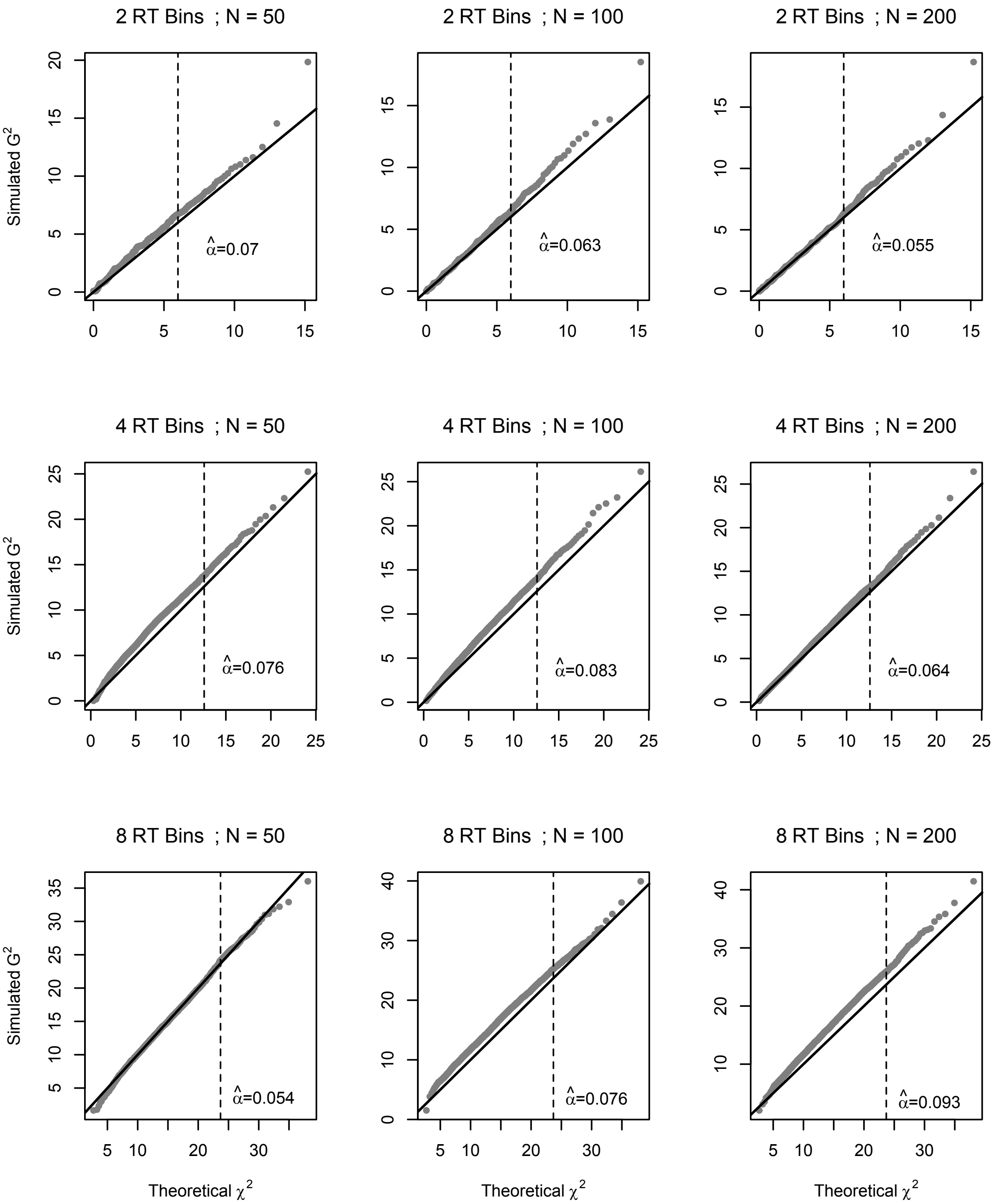

Figure 2: QQ-plot for the G^2-values of the absolute goodness-of-fit test using RT boundaries based on *empirical quantiles*. The solid line indicates perfect agreement of the theoretical χ^2 and the observed G^2 distributions. Actual simulation results are indicated by gray dots.

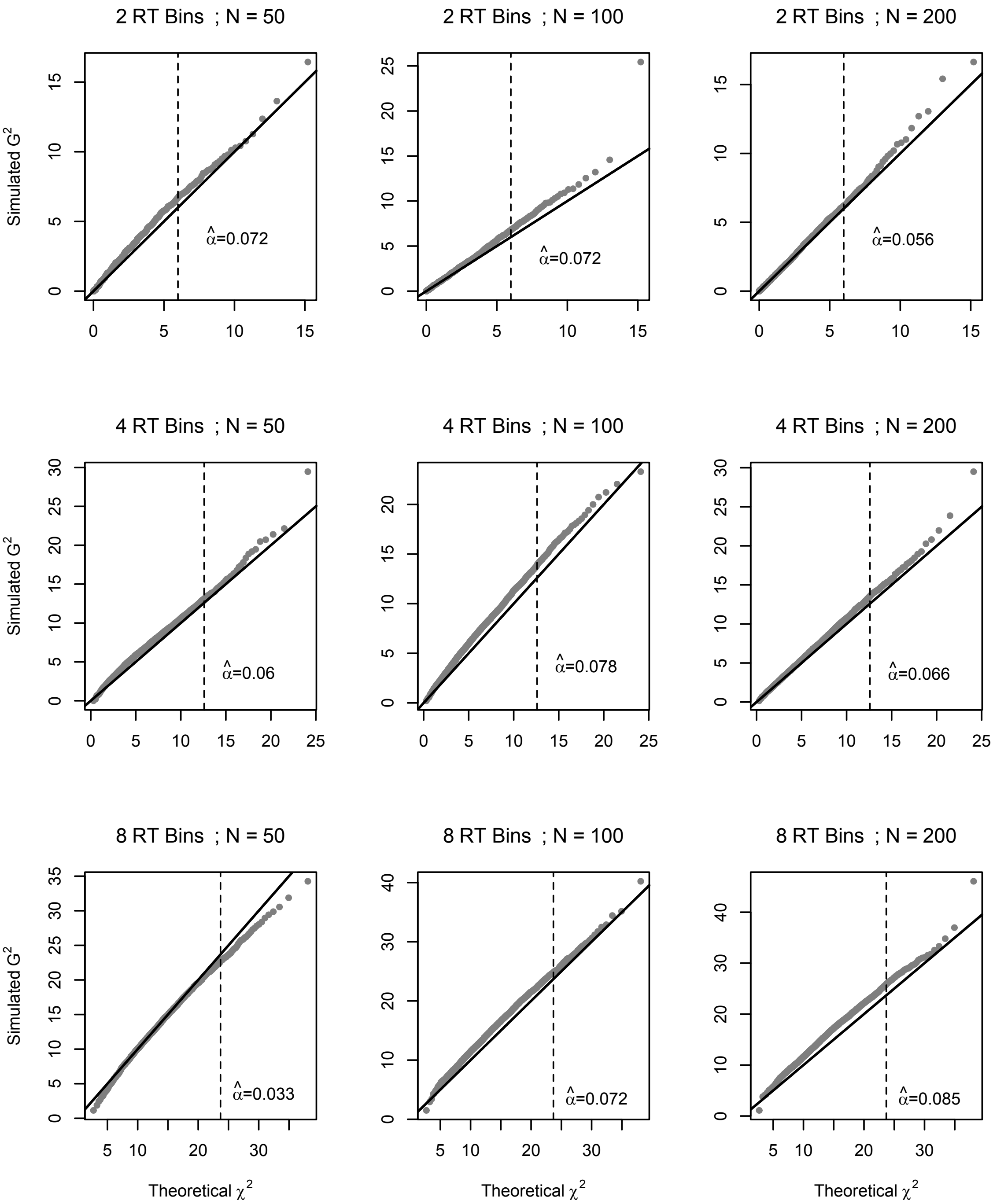

Figure 3: QQ-plot for the G^2-values of the absolute goodness-of-fit test using RT boundaries based on a *log-normal approximation*. The solid line indicates perfect agreement of the theoretical χ^2 and the observed G^2 distributions. Actual simulation results are indicated by gray dots.

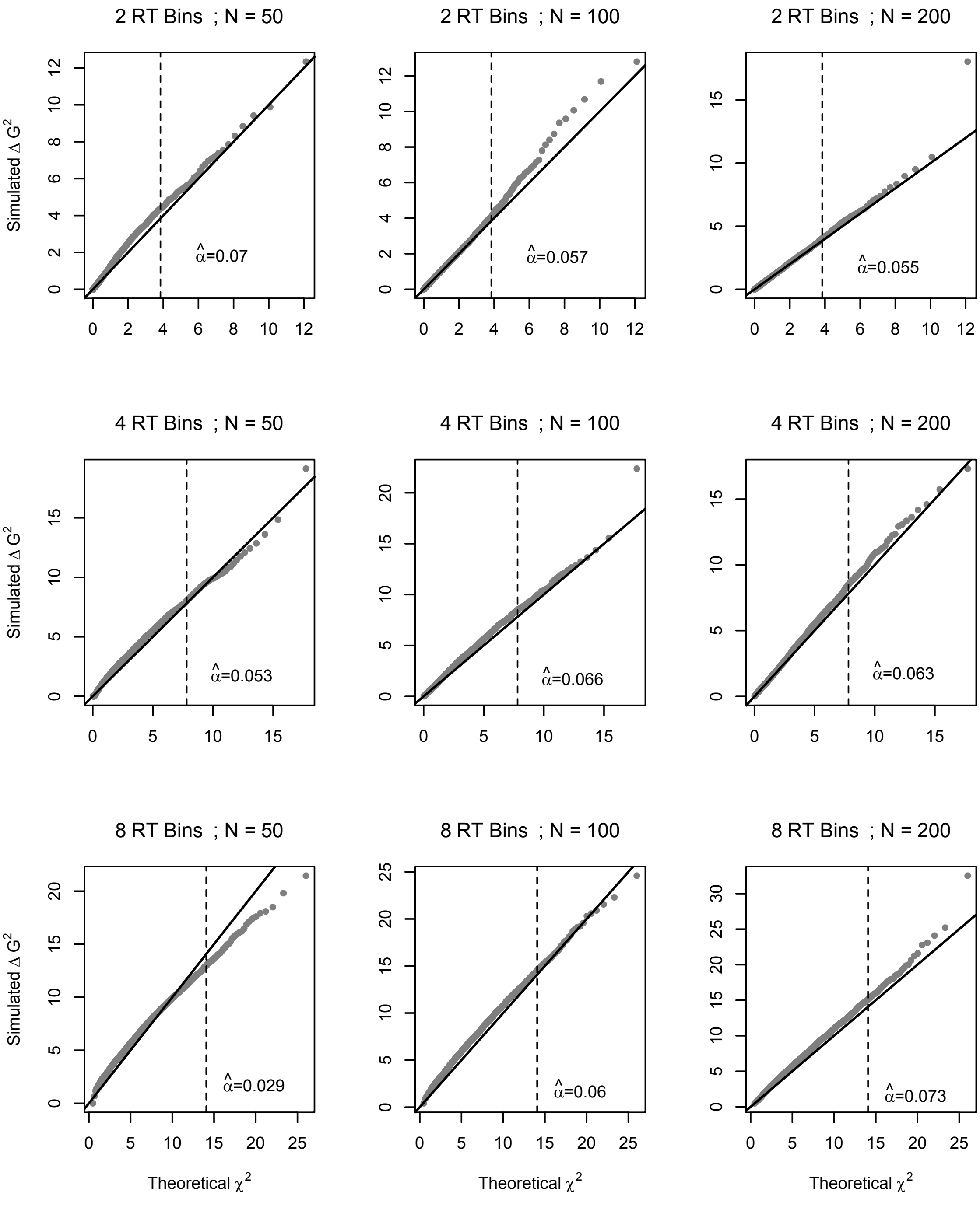

Figure 4: QQ-plot for the ΔG^2-values of the nested goodness-of-fit test using *fixed* RT boundaries. The solid line indicates perfect agreement of the theoretical χ^2 and the observed ΔG^2 distributions. Actual simulation results are indicated by gray dots.

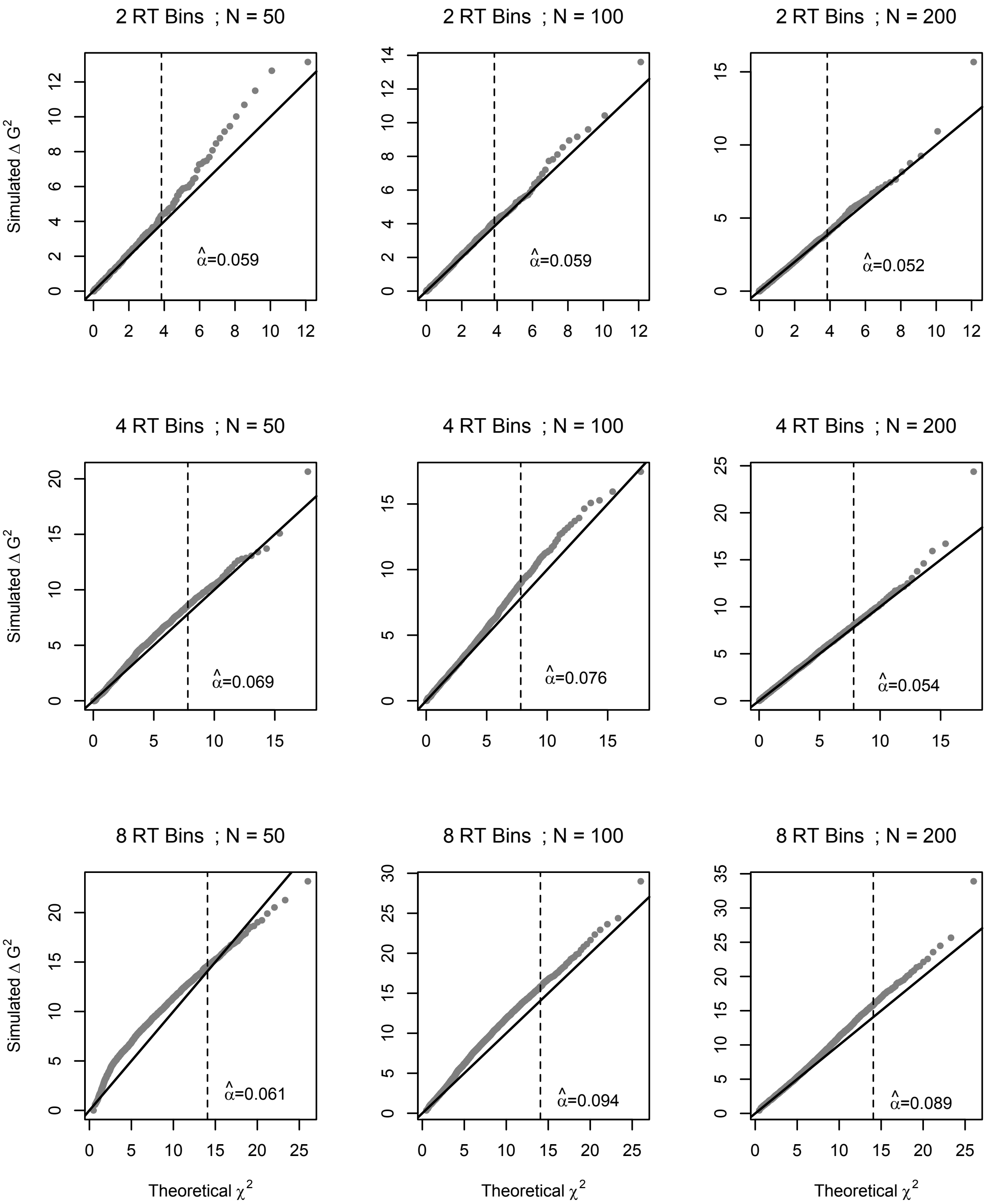

Figure 5: QQ-plot for the ΔG^2-values of the nested goodness-of-fit test using *empirical* RT boundaries. The solid line indicates perfect agreement of the theoretical χ^2 and the observed ΔG^2 distributions. Actual simulation results are indicated by gray dots.

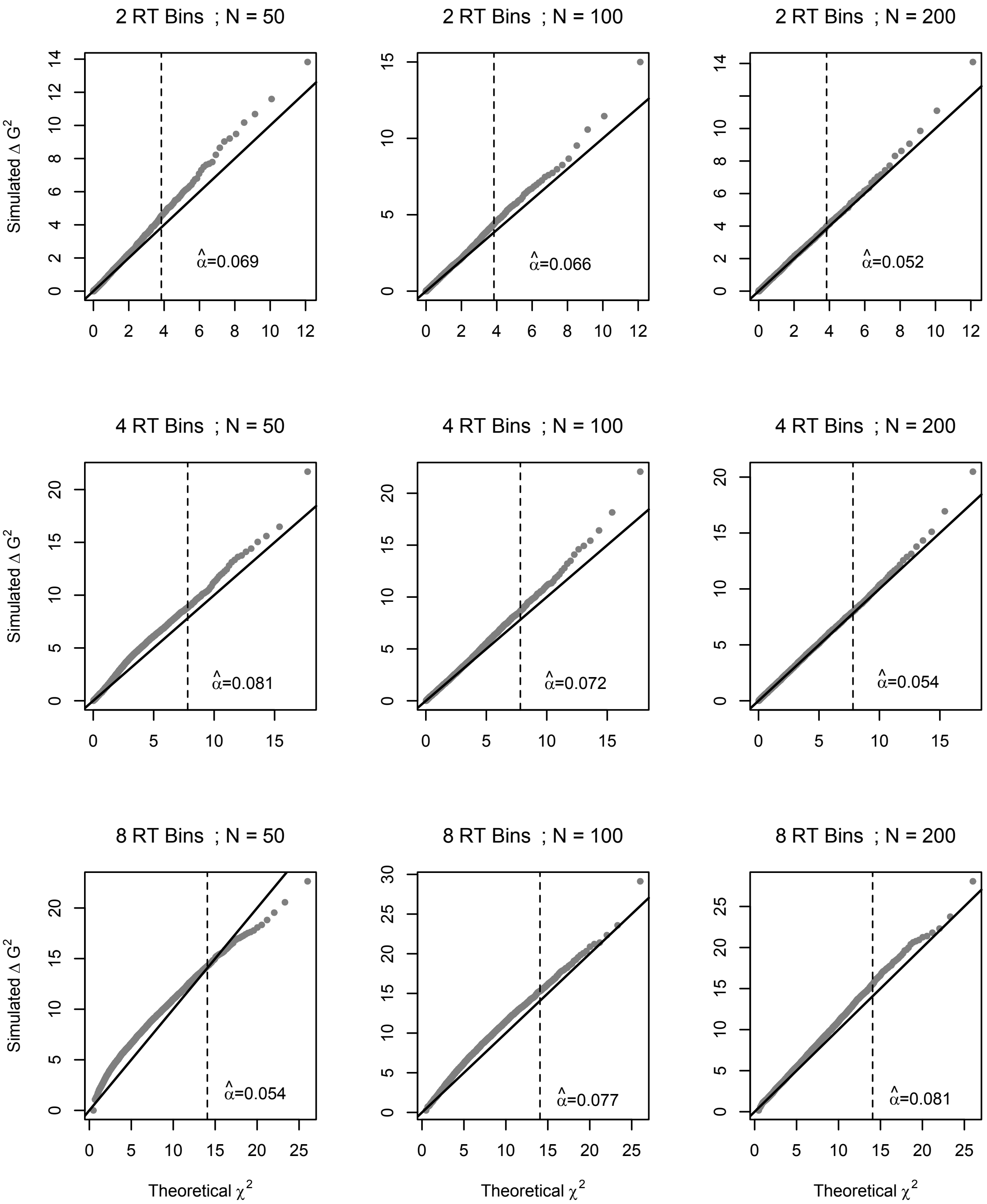

Figure 6: QQ-plot for the ΔG^2-values of the nested goodness-of-fit test using RT boundaries based on a *log-normal approximation*. The solid line indicates perfect agreement of the theoretical χ^2 and the observed ΔG^2 distributions. Actual simulation results are indicated by gray dots.

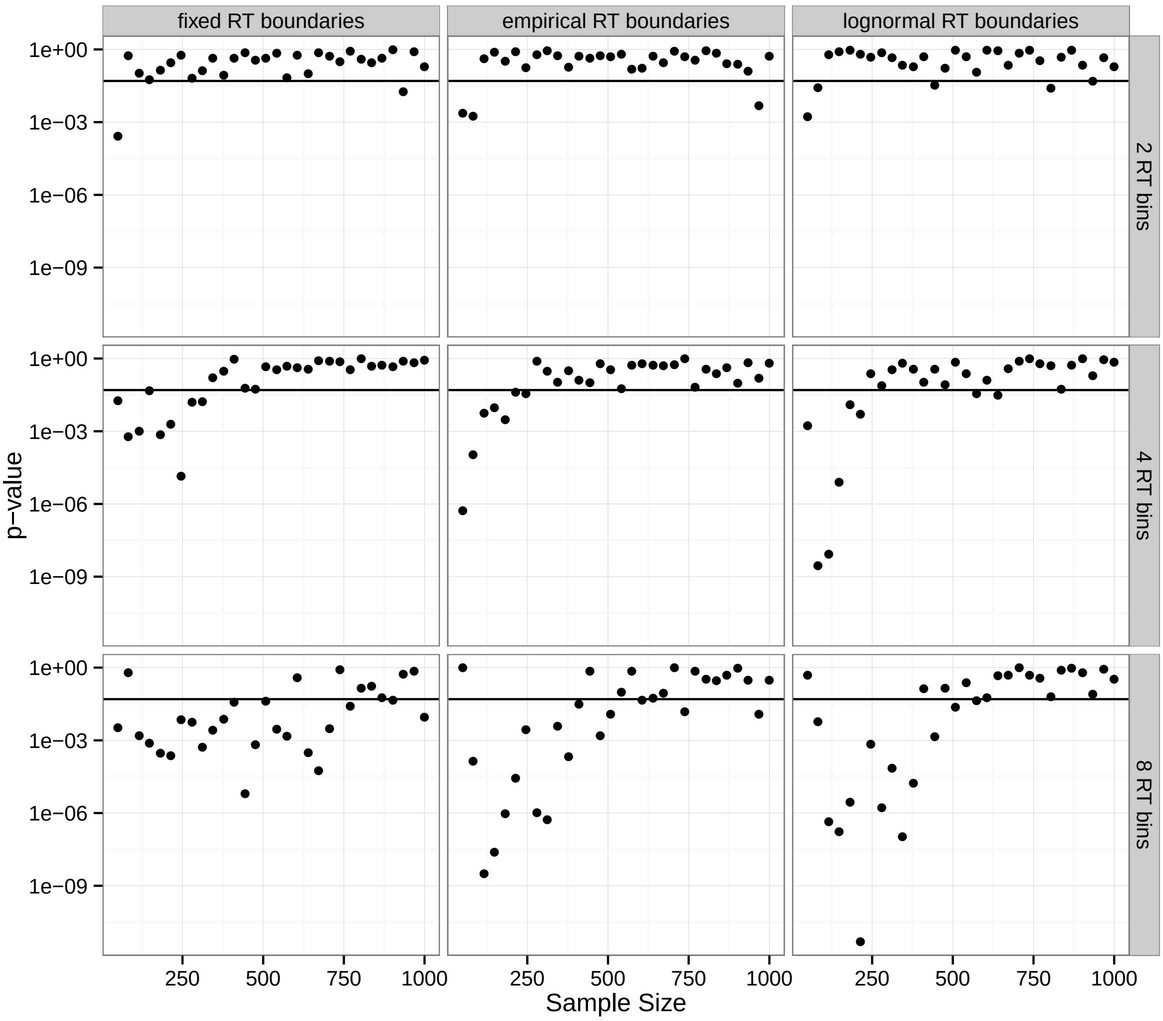

Figure 7: p-values of Kolmogorov-Smirnov tests, each based on the simulated G^2 distribution using 300 simulated data sets. The horizontal, solid line indicates the standard α-level of 5%.

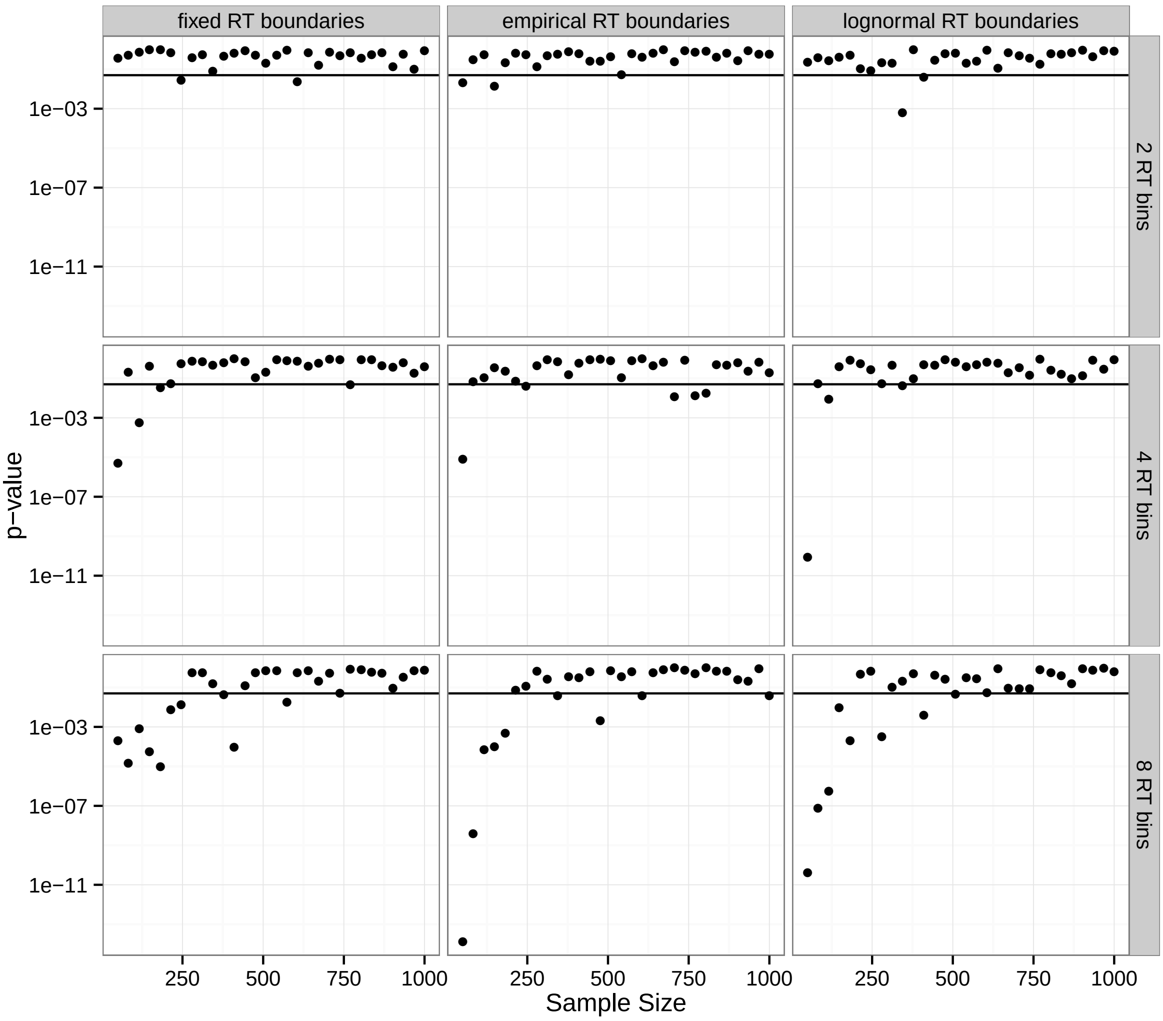

Figure 8: p-values of Kolmogorov-Smirnov tests, each based on the simulated ΔG^2 distribution using 300 simulated data sets. The horizontal, solid line indicates the standard α-level of 5%.

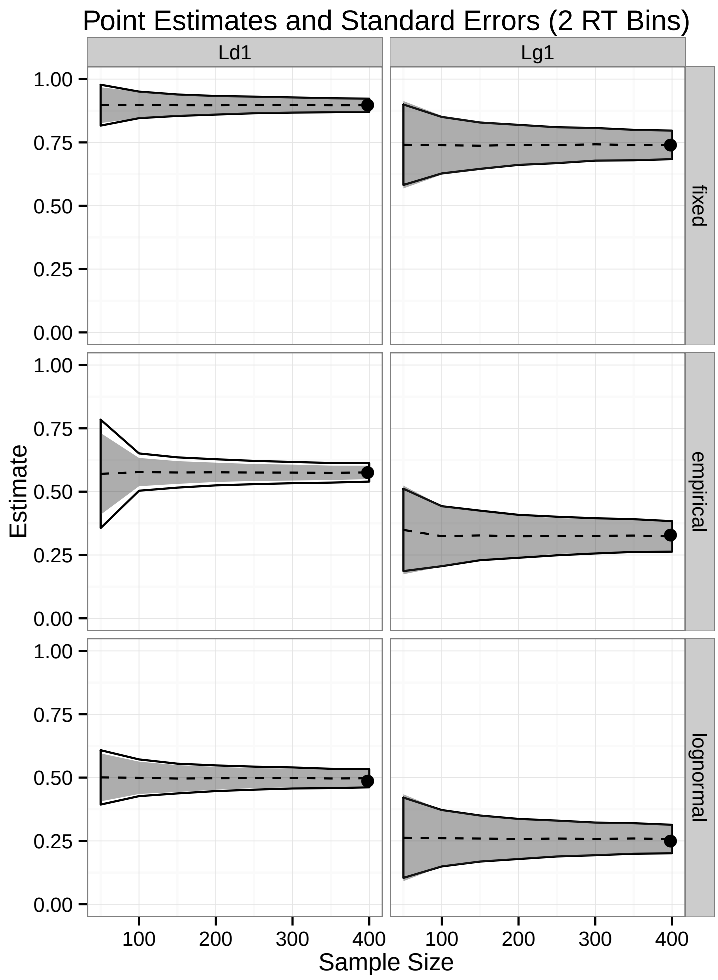

Figure 9: Mean point estimates (dashed line) and mean estimated standard errors (ribbon of solid lines) for the L-parameters using *two RT bins*. The true, data-generating L-values are shown as large points to the right, whereas the empirical standard deviation (i.e., the bootstrapped, actual SE) of the point estimates is shown in gray.

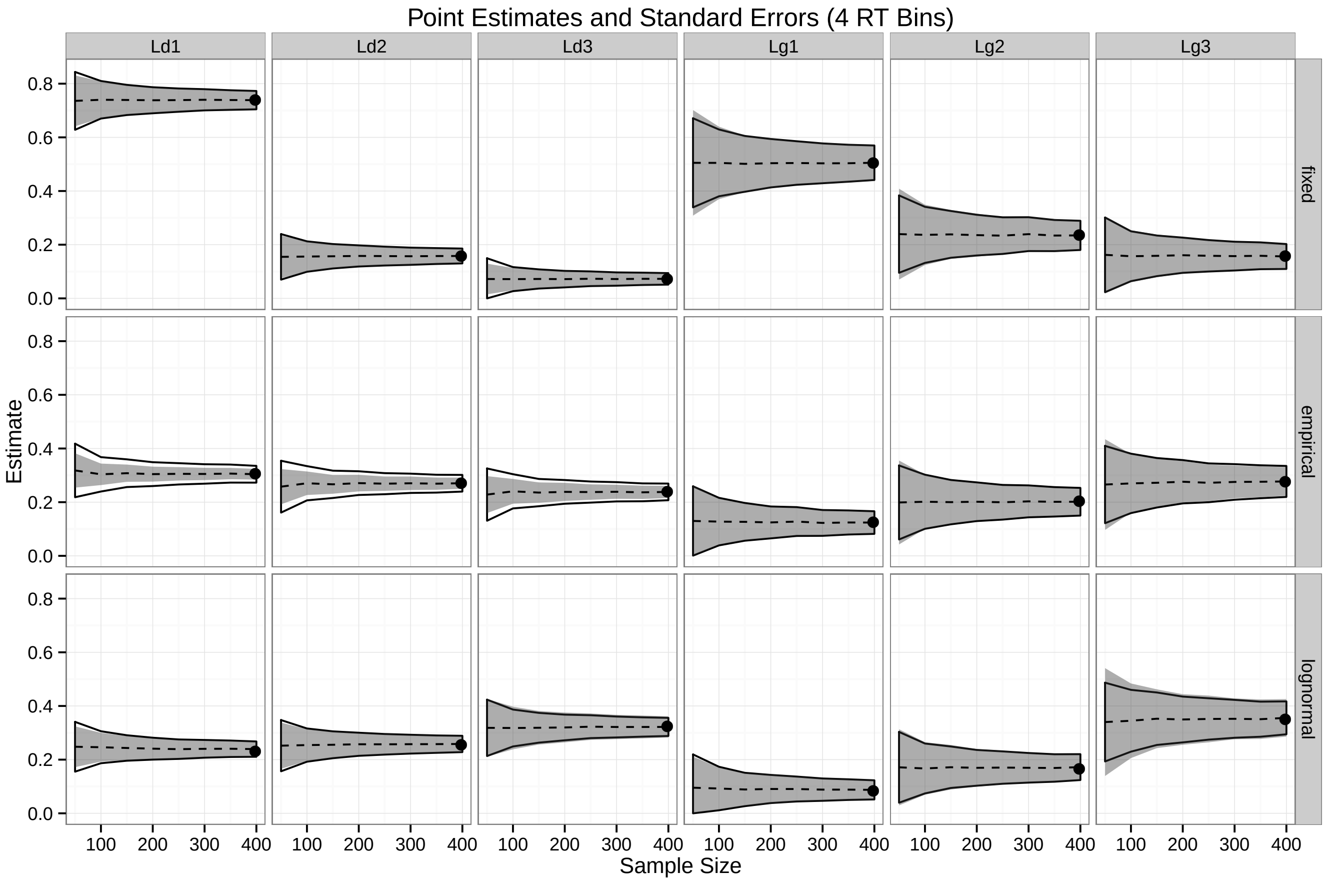

Figure 10: Mean point estimates (dashed line) and mean estimated standard errors (ribbon of solid lines) for the L-parameters using *four RT bins.* The true, data-generating L-values are shown as large points to the right, whereas the empirical standard deviation (i.e., the bootstrapped, actual SE) of the point estimates is shown in gray.

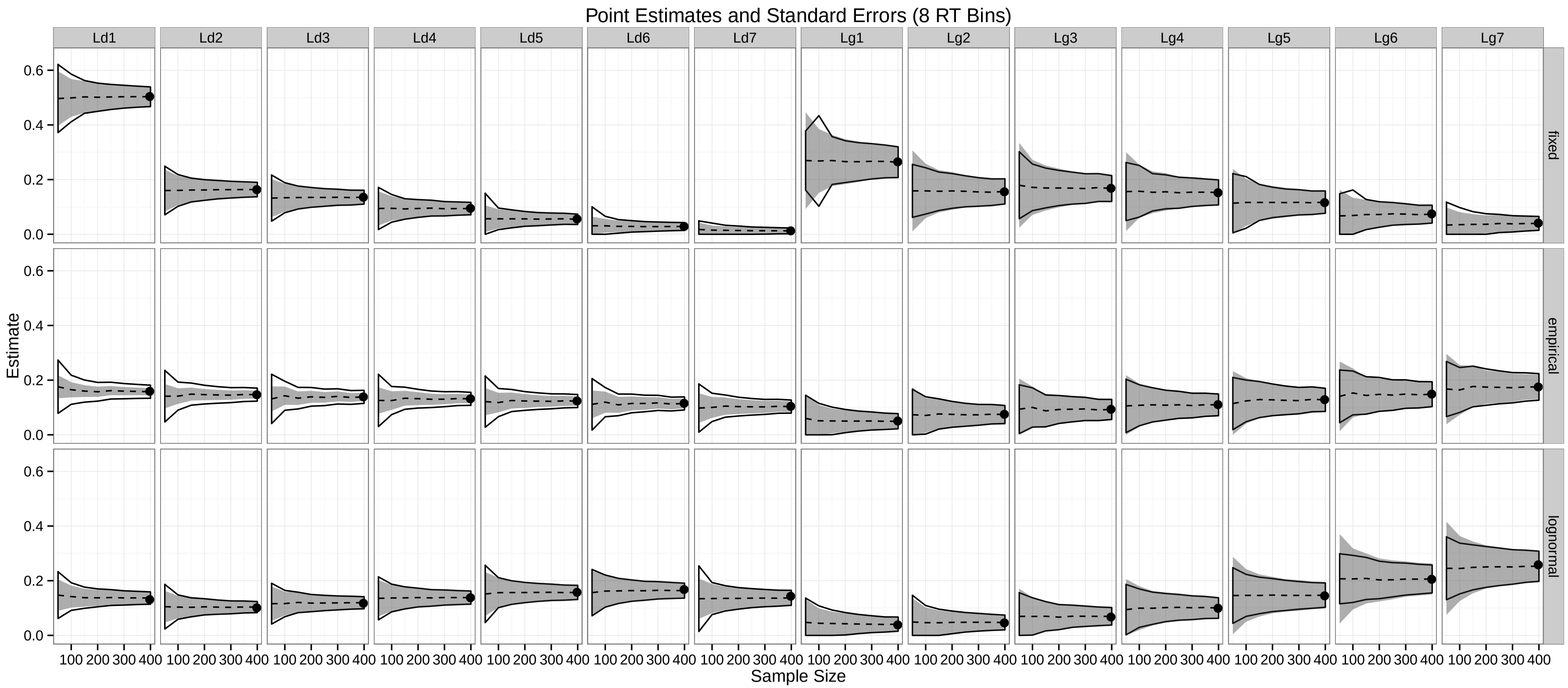

Figure 11: Mean point estimates (dashed line) and mean estimated standard errors (ribbon of solid lines) for the L-parameters using *eight RT bins*. The true, data-generating L-values are shown as large points to the right, whereas the empirical standard deviation (i.e., the bootstrapped, actual SE) of the point estimates is shown in gray.

Psychological Review

0033-295X/17/$12.00 http://dx.doi.org/10.1037/rev0000063

Linking Process and Measurement Models of Recognition-Based Decisions

Daniel W. Heck and Edgar Erdfelder
University of Mannheim

When making inferences about pairs of objects, one of which is recognized and the other is not, the recognition heuristic states that participants choose the recognized object in a noncompensatory way without considering any further knowledge. In contrast, information-integration theories such as parallel constraint satisfaction (PCS) assume that recognition is merely one of many cues that is integrated with further knowledge in a compensatory way. To test both process models against each other without manipulating recognition or further knowledge, we include response times into the r-model, a popular multinomial processing tree model for memory-based decisions. Essentially, this response-time-extended r-model allows to test a crucial prediction of PCS, namely, that the integration of recognition-congruent knowledge leads to faster decisions compared to the consideration of recognition only—even though more information is processed. In contrast, decisions due to recognition-heuristic use are predicted to be faster than decisions affected by any further knowledge. Using the classical German-cities example, simulations show that the novel measurement model discriminates between both process models based on choices, decision times, and recognition judgments only. In a reanalysis of 29 data sets including more than 400,000 individual trials, noncompensatory choices of the recognized option were estimated to be slower than choices due to recognition-congruent knowledge. This corroborates the parallel information-integration account of memory-based decisions, according to which decisions become faster when the coherence of the available information increases.

Keywords: decision making, recognition heuristic, adaptive toolbox, response times, information integration

Supplemental materials: http://dx.doi.org/10.1037/rev0000063.supp

When faced with the decision which of two cities is larger, which river is longer, or which musician is more successful, people often choose the recognized over the unrecognized option (Goldstein & Gigerenzer, 1999). In many natural environments, choosing the recognized option often results in accurate decisions because recognition is highly correlated with the criterion (i.e., larger cities are usually better known; Goldstein & Gigerenzer, 2002). The result that people rely on recognition to make inferences has often been replicated and stimulated a lot of research to shed light on the underlying cognitive processes of memory-based decisions.

Whereas some researchers developed computational process models that allow precise predictions of future behavior (e.g., Goldstein & Gigerenzer, 2002), others used measurement models that decompose observed behavior to obtain process-pure parameter estimates of the latent processes of interest (e.g., Hilbig, Erdfelder, & Pohl, 2010). Even though both lines of research enhanced understanding of how recognition is used in memory-based decisions, they have largely been applied as separate and distinct methods to investigate decision behavior. However, to integrate knowledge from both streams of research, it is necessary to derive strong theoretical links between process and measurement models, as called for by Marewski, Pohl, and Vitouch (2011a, p. 359). "Future research on recognition-based inferences should (a) converge on overcoming past controversies, taking an integrative approach to theory building, and [. . .] (b) test existing models of such strategies competitively [. . .]."

To integrate existing theories and directly test two competing accounts, we develop a unifying approach in which a specific measurement model (an extended version of the r-model; Hilbig, Erdfelder, & Pohl, 2010) serves as an encompassing framework to test opposing predictions of two process models—the serial rec-

Daniel W. Heck and Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim.

This research was supported by the Grant Er 224/2-2 from the Deutsche Forschungsgemeinschaft (DFG) and the University of Mannheim's Graduate School of Economic and Social Sciences, also funded by the DFG. Both authors contributed to the present work in a significant way and have approved the final manuscript. We thank Marta Castela and Benjamin Hilbig for many helpful discussions, and Marta Castela, Ralph Hertwig, Stefan Herzog, Benjamin Hilbig, Guy Hochman, Sebastian Horn, Martha Michalkiewicz, Thorsten Pachur, Rüdiger Pohl, Tobias Richter, and Oliver Schweickart for providing the raw data of their studies.

Preliminary versions of the present work were presented at the Conference of the German Society for Psychology (Leipzig, Germany, 2016), Coherence-Based Approaches to Decision Making, Cognition, and Communication (Berlin, Germany, 2016), the International Meeting of the Psychonomic Society (Granada, Spain, 2016), the Tagung der Fachgruppe Methoden & Evaluation (Jena, Germany, 2015), and the European Mathematical Psychology Group Meeting (Padua, Italy, 2015). The present manuscript, ideas, and results have not been uploaded to any online depository or website previously.

Correspondence concerning this article should be addressed to Daniel W. Heck or Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim, Schloss EO 254, D-68131 Mannheim, Germany. E-mail: heck@uni-mannheim.de or erdfelder@psychologie.uni-mannheim.de

ognition heuristic on the one hand (RH; Goldstein & Gigerenzer, 1999, 2002) against the parallel constraint satisfaction theory of decision making on the other hand (PCS; Glöckner & Betsch, 2008; Glöckner, Hilbig, & Jekel, 2014). Whereas the former is one of the fast and frugal heuristics proposed within the theoretical framework of the adaptive toolbox of decision making (Gigerenzer & Selten, 2001), the latter is a single-process, neural-network model that explains how information from probabilistic cues is holistically and intuitively integrated to form a coherent judgment.

Importantly, the two accounts differ with respect to the effect of additional recognition-congruent information on response times (RT). The serial heuristic account predicts that knowledge-based choices are necessarily slower than choices due to recognition only. In contrast, the parallel model predicts that recognition-congruent information leads to faster choices even though more information is processed (Glöckner & Bröder, 2011). However, these opposing predictions cannot directly be tested in the standard experimental paradigm, since it only includes a recognition phase, in which participants judge cities as "recognized" or "unrecognized," and a decision phase, in which participants have to decide which of two cities is larger. In such a design, the researcher does not know whether and what kind of further knowledge participants have about the presented cities.

As a remedy, previous studies experimentally manipulated recognition, cue knowledge, or both to test the noncompensatory nature of the RH (e.g., Newell & Shanks, 2004; Richter & Späth, 2006). However, evidence based on experimentally induced recognition or knowledge has been challenged by the objection that such artificial scenarios are not within the domain of the recognition heuristic (i.e., the context of natural, memory-based decisions; Goldstein & Gigerenzer, 2011; Pachur, Bröder, & Marewski, 2008). Summarizing the debate, Pohl (2011) already concluded that "if knowledge may not be learned in the lab, how can the nature of additional cue knowledge be controlled?" As a remedy, we propose to link the two process models of interest to a measurement model that allows to disentangle use of knowledge and recognition in the standard paradigm without requiring artificially induced knowledge (the r-model; Hilbig, Erdfelder, & Pohl, 2010).

In the following, we develop such a unifying framework of the existing process and measurement models by (a) formally deriving RT predictions from the serial RH and the PCS information integration account; (b) extending the multinomial r-model to include response times; (c) showing in simulations that the RT-extended measurement model provides an encompassing framework to test between the two process models; and (d) using the RT-extended model to reanalyze 29 data sets from our and other researchers' labs that include approximately 400,000 individual decisions. This reanalysis clearly shows that choices based on recognition-congruent knowledge are faster than those based on recognition only, thus corroborating the parallel information-integration account of memory-based decisions. More generally, our approach might serve as a case study of how measurement and process models can be linked to allow for novel empirical tests of psychological theories.

Process Models of Recognition-Based Decisions

In the following, we derive opposing RT predictions from two prominent process models that aim to explain recognition-based decisions—the serial heuristic (RH) and the parallel information-integration account (PCS). Whereas the former predicts that responses based only on recognition are necessarily faster than responses based on further knowledge, the latter predicts that recognition-congruent knowledge facilitates faster responding.

The Recognition Heuristic: A Serial-Processing Account

As a general framework for judgment and decision making under uncertainty, Gigerenzer and Goldstein (1996) proposed the metaphor of the *adaptive toolbox*—a set of fast and frugal strategies that allow people to make accurate decisions with a minimal amount of effort (Payne, Bettman, & Johnson, 1988; Payne, Bettman, & Johnson, 1993). These heuristics are assumed to have evolved over the history of mankind through a natural selection of strategies optimally adapted to the structure of existing environments (Gigerenzer & Brighton, 2009). Crucially, these heuristics are defined as precise process models by a set of search, stopping, and decision rules. When making decisions, people are assumed to adaptively choose an appropriate heuristic from the adaptive toolbox depending on the context and structure of the environment (Marewski & Schooler, 2011).

In the context of natural, memory-based decisions, one or more of the choice options might not be known and thus decision makers sometimes have to choose between recognized (R) and unrecognized (U) options. In such cases, where not all of the alternatives are recognized, the adaptive toolbox states that people rely on the recognition heuristic (RH; Goldstein & Gigerenzer, 1999), which is defined as a precise, formal model (Goldstein & Gigerenzer, 2002, p. 76) as follows: "If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion." The RH is one of the core heuristics of the adaptive toolbox and has been the subject of three special issues in *Judgment and Decision Making* (Marewski, Pohl, & Vitouch, 2010, Marewski et al., 2011a, 2011b).

Like all heuristics in the adaptive toolbox, the RH is precisely defined as a formal model in terms of a sequence of underlying processing steps (Gigerenzer & Gaissmaier, 2011). Specifically, the RH consists of the following three rules (Goldstein & Gigerenzer, 2002):

1. *Search Rule.* Try to recognize both options.
2. *Stopping Rule.* Stop if exactly one of the two options is recognized.
3. *Decision Rule.* Choose the recognized option.

Note that additional knowledge about the choice options does not enter the process at any stage, which makes this a noncompensatory heuristic.

The noncompensatory nature of the RH has been at the core of many heated discussions (Marewski et al., 2010). Due to its precise definition as a serial process that immediately results in a decision if only one of the two options is recognized, the RH implies that any further knowledge cannot affect the decision process. For instance, in the city-size task, participants might know whether the recognized city has an airport, a university, or a metro. Despite the availability and possible validity of such additional cues, the RH states that further knowledge is not considered and does not affect the decision process. Note that this does not imply that decisions based only on recognition are necessarily less accurate—quite to the contrary, Davis-Stober, Dana, and Budescu (2010) showed that mere reliance on the recognition cue can result in better predictive performance than optimal cue weighting, which is prone to overfitting due to increased model complexity.

Despite its exact definition in terms of processing rules, the formal model of the RH does not make precise predictions about RTs. However, testable predictions about the relative speed of different strategies within the adaptive toolbox can directly be derived by comparing the number of required elementary information processes (EIPs; Glöckner, 2009; Payne et al., 1988). Given a set of different decision strategies, ordinal predictions for their relative speed follow directly by comparing the numbers of required EIPs per strategy. However, before deriving such predictions, it is necessary to consider alternative strategies that people might use when deciding between a recognized and an unrecognized option.

Alternative, Knowledge-Based Heuristics

In most studies, not all participants always choose the recognized option in all trials. Even though the recognized option is chosen very often as indicated by high adherence rates (Goldstein & Gigerenzer, 2002), the unrecognized option is nevertheless systematically chosen at least in some trials by some participants (Hilbig & Richter, 2011). To explain why and when further information influences memory-based decisions, even in environments that are ideally suited for the application of the RH (Hilbig, 2010b), it is necessary to assume that decision makers use an alternative strategy from the adaptive toolbox at least in some trials (Gigerenzer & Goldstein, 2011). The idea that participants adaptively select a strategy in each trial has been advocated, for instance, by Pachur (2011, p. 418): "the decision of whether to use the recognition heuristic or not is made for each individual pair of objects rather than for an entire environment."

What kind of strategy might people use instead of the RH? As a first hypothesis, one could simply attribute all choices of the unrecognized option to guessing or errors in applying the RH. However, both of these strategies can clearly be rejected (Hilbig, Erdfelder, & Pohl, 2010), which calls for an alternative heuristic that integrates further knowledge to systematically decide between the unrecognized or the recognized option. Moreover, within the adaptive toolbox, the existence of such an alternative heuristic is necessary to explain adaptive decision behavior in environments with low recognition validity, in which people often correctly choose the unrecognized option (e.g., in case of infectious diseases; Horn, Pachur, & Mata, 2015).

However, most strategies of the adaptive toolbox that consider cue knowledge (e.g., weighted-additive; Gigerenzer & Goldstein, 1996) are not applicable if only one of the two options is recognized because a full set of cue values is required for both choice options. Given that an option is not recognized in the first place, it is assumed that the decision maker does not have any cue knowledge about that option (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2011). Note that this reasoning is based on the assumption that the stimuli are sufficiently homogeneous and do not contain superficial information (e.g., city names that sound Chinese, which would be a valid cue to judge the city size; Marewski et al., 2011a). In the following, we assume that surface features of the material are reasonably homogeneous, thus resulting in a complete lack of information for the unrecognized option. However, our modeling approach and predictions can directly be generalized to the case where some cue knowledge is available for the unrecognized option.

Marewski, Gaissmaier, Schooler, Goldstein, and Gigerenzer (2010) proposed several decision strategies for recognized-unrecognized (RU) pairs as competitors to the RH. All of these strategies rely on the evaluation and comparison of further cues that are available for the recognized option. Here, we focus on RT predictions of such alternative strategies that might be used instead of the RH in some trials. Given that the RH only requires a single cue to make a decision (recognized vs. unrecognized), any alternative strategy that considers at least one additional cue necessarily requires more processing steps (EIPs) irrespective of how they are compared (Bröder & Gaissmaier, 2007; Glöckner & Bröder, 2011). In addition, Pachur and Hertwig (2006, p. 986) derived the prediction that "recognition is first on the mental stage and ready to enter inferential processes when other probabilistic cues still await retrieval." Even if further cues were already available simultaneously with recognition, the integration of these cues required some additional processing time compared to simply choosing the recognized option.

Figure 1 illustrates the prediction that the integration of any further cues requires at least one additional processing step compared to the RH (cf. Gigerenzer & Goldstein, 1996; Pohl, 2011, and their Figure 2 and 1, respectively). First, attempts are being made to recognize both options. Based on the outcome of this process, an evaluation or strategy selection stage is required to adaptively choose between using the RH or a different strategy (Gigerenzer & Goldstein, 2011). Even though such a stage could already evaluate further information about available cues, this would render any subsequent decisions compensatory by definition—given specific cue configurations, use of the RH could be suspended, thus overruling the recognition cue (Söllner, Bröder, Glöckner, & Betsch, 2014). Hence, such a stage can only consider information about recognition, that is, whether and how fast the options were recognized. Once the appropriate strategy is chosen, either the RH or another strategy is used. Importantly, such an alternative strategy requires the comparison and integration of further cues and thus must be slower. Note that this is in line with the original idea of Goldstein and Gigerenzer (2002) that the RH provides a frugal shortcut to make fast and accurate decisions.

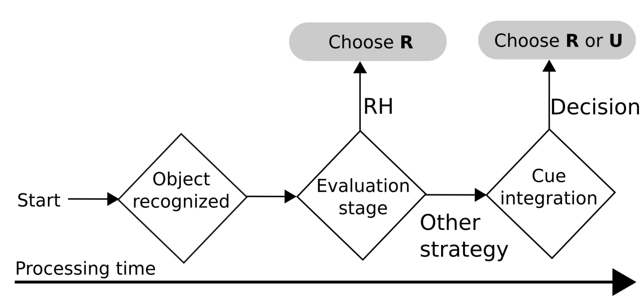

Figure 1. The RH requires less processing stages than alternative strategies that integrate further knowledge (cf. Gigerenzer & Goldstein, 1996).

Stochastic Dominance

To elaborate and formalize the above prediction, we focus on a specific, tallying-like strategy in the following, which is similar to "tallying of positive and negative cues" by Marewski et al. (2010, p. 289). As mentioned above, we assume that no knowledge about the unrecognized option is available but that some additional cues of the recognized option are known due to prior knowledge (Gigerenzer & Goldstein, 1996). Given such partial cue knowledge, the tallying strategy (TALL) simply counts the number of available positive and negative cues for the recognized option:

1. *Search rule.* Count the positive and negative cue values of the recognized option (including the recognition cue).
2. *Stopping Rule.* Stop, if all available cues of the recognized option have been considered.
3. *Decision rule.* Choose the recognized option if the number of positive cue values exceeds the number of negative cue values.

For instance, if city A is recognized and has two positive and three negative cue values, then it is not chosen.

Despite its increased complexity in comparison to the RH, the TALL strategy mainly requires counting positive and negative cues, a process that can be assumed to be relatively frugal (Gigerenzer & Gaissmaier, 2011). However, as discussed above, this strategy requires more EIPs than the RH due to the consideration of further cues. Note that this ordinal prediction is not based on an exact count of the required EIPs. Moreover, due to the serial nature of the elementary processes, the increase in processing time is a monotonic function of the number of available cues considered, irrespective whether cues are positive or negative. For the simulation below, we set the number of required EIPs to the number of available cues that have to be assessed and compared (including the recognition cue).

Note that both strategies, the RH and TALL, may result in the same outcome, that is, choosing the recognized option. However, responses due to RH use (with RTs denoted by the random variable T_{RH}) are necessarily faster than responses due to TALL use (T_{TALL}), assuming that the additional processing steps require some positive, unobservable time $T_{\mathrm{add}} > 0$:

$$T_{\mathrm{TALL}} = T_{\mathrm{RH}} + T_{\mathrm{add}}. \tag{1}$$

This equation of random variables[1] implies a stochastic ordering of the observed RTs, that is, $T_{\mathrm{TALL}} > T_{\mathrm{RH}}$. Such an inequality of random variables is called stochastic dominance in the RT literature (Heathcote, Brown, Wagenmakers, & Eidels, 2010; Townsend, 1990) and constitutes a stronger statement than the corresponding ordinal constraint on the mean RTs. In practice, stochastic dominance can be tested by assessing the following inequality constraint on the cumulative distribution functions $F(t)$:

$$F_{\mathrm{RH}}(t) > F_{\mathrm{TALL}}(t) \qquad \text{for all } t \in \mathbb{R}^{+}. \tag{2}$$

When plotting the cumulative density functions, Equation 2 simply states that the cumulative density of RTs associated with the RH is always larger than that associated with TALL.

The strong prediction of stochastic dominance also follows if some kind of independent strategy-selection stage is assumed that determines whether the RH or TALL is used based on cost-benefit trade-offs, reinforcement learning, or natural environmental constraints (e.g., Goldstein & Gigerenzer, 2011; Marewski & Schooler, 2011; Payne et al., 1993; Rieskamp & Otto, 2006). In a serial-processing model, an independent selection stage adds the same amount of random processing time $T_{\mathrm{SS}} > 0$ irrespective whether the RH or some other strategy such as TALL is used subsequently. Importantly, the prediction of stochastic dominance, as expressed by an ordering of random variables, is not affected by this additional stage:

$$T_{\mathrm{SS}} + T_{\mathrm{RH}} < T_{\mathrm{SS}} + T_{\mathrm{TALL}}. \tag{3}$$

Equation 3 implies that the cumulative density functions of the corresponding RTs are shifted to the right (i.e., RTs are slower) compared to those in Equation 2 but still do not cross.

Note that Equation 3 is even valid without requiring an *independent* evaluation stage. First, the prediction also follows if we assume that the time for strategy selection (T_{SS}) is correlated with the time for strategy execution (T_{RH} and T_{TALL}). This is the case, for instance, if a strategy that has been selected faster can also be applied faster. Second, the prediction is still valid if the time of selecting the RH is systematically faster than that of selecting TALL. In such a case, the single random variable T_{SS} in Equation 3 is replaced by two separate strategy-selection times $T_{\mathrm{SSR}} < T_{\mathrm{SST}}$ (and hence, observed RH responses are still predicted to be faster). In order to violate the prediction that RH-responses are stochastically faster, it would be necessary that selection of TALL is actually *faster* than that of the RH (i.e., $T_{\mathrm{SSR}} > T_{\mathrm{SST}}$). However, we are not aware of a strategy-selection account that predicts faster selection times for more complex strategies. Hence, the prediction of stochastically faster RH responses holds under rather general conditions (and also for other serial strategies as derived by Glöckner & Betsch, 2012). In fact, a process model that predicts slower RH responses than knowledge-based responses would compromise the core assumption that the RH is a "fast and frugal" strategy.

In line with the prediction that RH responses are faster than knowledge-based responses, previous studies showed that people choose the recognized option faster than the unrecognized option (Hilbig & Pohl, 2009). However, without experimentally manipulating further knowledge, it is not possible to directly test whether Equation 2 also holds for all choices of the recognized object, that is, whether choices of the recognized option due to the RH are actually faster than those due to consideration of further knowledge (TALL). Below, we will test this prediction without the need to experimentally manipulate cue or recognition knowledge by using an appropriate measurement model. However, before doing so, we discuss an alternative theoretical account of decision making in the next section.

Parallel Information Integration in Recognition-Based Decisions

The hypothesis that decisions are made in a noncompensatory way as implied by the RH and by other heuristics of the adaptive toolbox such as take-the-best (Gigerenzer & Goldstein, 1996) has often been questioned by evidence showing the impact of further knowledge on various measures such as choices, RTs, confidence

[1] Since T_{TALL}, T_{RH}, and T_{add} are random variables, the equation does only hold P-almost surely, that is, with probability one and not for each random element $\omega \in \Omega$ in the probability space (Ω, A, P).

ratings, and process-tracing measures (e.g., Hilbig, 2010b; Lee & Cummins, 2004; Newell & Shanks, 2004; Richter & Späth, 2006). In contrast to the adaptive toolbox, according to which people adaptively select one of several qualitatively different heuristics depending on the environmental context, single-process mechanisms assume that information about probabilistic cues is integrated in a weighted, compensatory fashion to make a decision (Lee & Cummins, 2004; Söllner et al., 2014). Whereas various implementations of this general idea such as the adjustable spanner (Newell & Bröder, 2008), decision field theory (Busemeyer & Townsend, 1993), MINERVA-DM (Dougherty, Gettys, & Ogden, 1999), magnitude comparison under uncertainty (Schweickart & Brown, 2014), or parallel constraint satisfaction theory (Glöckner & Betsch, 2008) differ with respect to the conceptual and mathematical details of the integration process, they all assume that decisions are driven by a single process in a holistic, compensatory manner.

What do these information-integration accounts predict in case of memory-based decisions? In contrast to the RH, single-process models treat recognition as merely one of many probabilistic cues (Glöckner & Bröder, 2011; Newell & Shanks, 2004; Richter & Späth, 2006). Despite the metacognitive nature of this recognition cue, it is integrated and weighted just as any other cue representing knowledge about the domain of interest. Given an environment with a high subjective recognition validity, recognition will likely be a very valid and thus influential cue, but nevertheless, further available information will affect the probability of choosing the recognized option.

Information-integration theories do not only predict that further information influences choice probabilities; they also specify whether decisions in favor of the recognized option become faster or more confident depending on the type of further knowledge. On the one hand, if further knowledge is congruent with recognition (e.g., one knows that the recognized city has an airport, a metro, and a cathedral), the recognized option will be chosen more often, faster, and with higher confidence (Glöckner & Bröder, 2011; Hilbig & Pohl, 2009). On the other hand, if further knowledge is incongruent with recognition, the recognized option will be chosen less often and decisions become slower and less confident.

Parallel Constraint Satisfaction Theory

In the present article, we rely on the parallel constraint satisfaction theory (PCS; Glöckner & Betsch, 2008; Holyoak & Simon, 1999; Thagard, 1989) to derive predictions and simulate data for the class of single-process information-integration models. PCS assumes that available information is integrated automatically and intuitively by a parallel spreading of activation in a network, where options and cues are represented by nodes as shown in Figure 2A. Essentially, this process is conceptualized as an automatic search for a coherent representation of the available information. Based on a state of maximal coherence, the option with the highest activation is chosen (Glöckner & Betsch, 2008). Importantly, the network is fully determined by the given cue structure and does not account for noise or errors, which makes PCS a deterministic process model (Glöckner et al., 2014). In the following, we explain the conceptual and psychological rationale of the theory; the mathematical details of the model are presented below in the section "Implementation of the Parallel Information-Integration Account."

As shown in Figure 2A, the number of nodes in the network is determined by the number of choice options and probabilistic cues. During the integration process, these nodes are activated differentially, thereby representing their relative importance for the judgment process. The links between cues and options are determined by the available cue structure. A connection is excitatory or inhibitory, if the option has a positive or negative value on the corresponding cue, respectively. Hence, an option will receive a stronger activation when it is supported by many positive cues. In addition, the choice options mutually inhibit each other as shown by the inhibitory link between the option nodes (Glöckner & Betsch, 2008).

Whereas the number of option and cue nodes and their reciprocal links are defined by the context, the "general validity" node is always present and keeps the network activated by constantly spreading activation to the cue nodes. The strength of the corresponding links to the cue nodes is determined by a monotonic function of the cue validities. For instance, a subjectively very valid cue is represented by a strong excitatory link from the general validity to the corresponding cue node (represented by a large weight w_i in Figure 2A) and will thus have a stronger influence compared to less valid cues and their corresponding nodes.

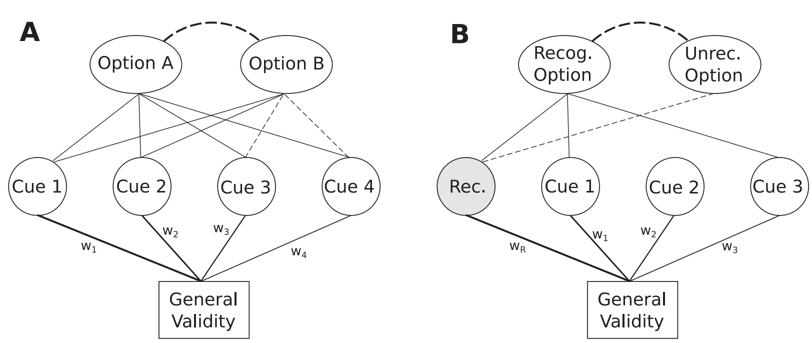

Figure 2. The parallel constraint satisfaction (PCS) model for (A) inferences from given information and (B) memory-based decisions. Information about positive and negative cue values is represented by excitatory (solid) and inhibitory (dashed) links between cue and option nodes, respectively (Glöckner & Betsch, 2008). In memory-based decisions, the lack of further knowledge about the unrecognized option is represented by nonexisting links to the cue nodes (with the exception of the recognition cue).

Once the network structure (i.e., the type and strength of the links) is determined by the task and context, activation spreads between all nodes simultaneously. For instance, a very valid and thus highly activated cue node will increase the activation of all options that have a positive value on this cue. Simultaneously, the option nodes fire back and activate relevant cue nodes. This process of spreading activation is assumed to stop when a state of coherence is reached at which the system stabilizes, that is, when the activation levels of all nodes do not change anymore. The option with the highest final activation is chosen, with a confidence level determined by the relative difference between the choice options. In addition, the number of iterations required to reach stability predict the observed decision times up to an interval scale. Note that the search for a coherent representation of the available information is thought of as an intuitive, automatic process (Glöckner & Betsch, 2008). This is an important feature of the PCS theory, since other frameworks such as the adaptive toolbox conceptualize the compensatory integration of information as a deliberate, cognitively demanding, and time consuming process (Gigerenzer & Goldstein, 1996).

Concerning inferences from given information—that is, when participants are presented with the complete pattern of cue values in each trial—PCS was found to account better for decisions, RTs, and confidence judgments than prominent heuristic strategies such as take-the-best (e.g., Glöckner & Betsch, 2012; Glöckner et al., 2014). In these empirical studies, the researcher usually needs to know the complete structure of positive and negative cue values to derive precise predictions for PCS and the competing strategies (Glöckner, 2009). In the case of recognition-based decisions with natural stimulus material (e.g., German cities), this requirement is usually not met, an issue discussed in the next section.

PCS Predictions for Memory-Based Decisions

To generalize PCS to memory-based decisions, recognition can be included as an additional cue node in the network similar to the nodes representing cue knowledge (Glöckner & Bröder, 2011; Hilbig, Erdfelder, & Pohl, 2010). For each option, this recognition cue (shown as the first node in the cue layer of Figure 2B) is positive for recognized options and negative for unrecognized options, resulting in excitatory and inhibitory links, respectively. Moreover, the weight w_R of this recognition cue is determined by the subjective recognition validity, which decision makers may learn from previous experience or infer from the current context. Computationally, PCS thereby treats the recognition cue as one of many available knowledge cues that are integrated to form a coherent judgment, with the conceptual and psychological difference that the recognition information is of a metacognitive nature.

Even though recognition itself can simply be treated as an additional cue in the PCS framework, memory-based decisions require a special consideration of how to model incomplete cue knowledge (Gigerenzer, Hoffrage, & Goldstein, 2008; Jekel, Glöckner, Bröder, & Maydych, 2014). Given that an option is not recognized, a decision maker will not have further knowledge about it and hence the corresponding cue values for this unrecognized option are necessarily unknown as shown in the first column of Table 1 (Gigerenzer & Goldstein, 2011). Moreover, it is very likely that participants only have partial cue knowledge about the recognized options when making memory-based decisions in natural environments (Gigerenzer & Goldstein, 1996). For instance, one might recall that a city has an airport and a university, but does not know whether it is a state capital. Previous approaches of modeling memory-based decisions using PCS (e.g., Glöckner & Bröder, 2011) have been criticized for ignoring the fact that decision makers have no access to any further information about the unrecognized option because "recognition can be seen as a prior condition for being able to recall further cue values from memory" (Gigerenzer & Goldstein, 2011, p. 107).

Table 1
Possible Cue Structures for Recognized and Unrecognized Choice Options in Memory-Based Decisions

Cue	Unrecognized	R-only	R-congruent	R-incongruent
R	−	+	+	+
1	?	?	+	−
2	?	?	+	−
3	?	?	+	−

Note. The values −, +, and ? represent positive, negative, and missing cue knowledge, respectively.

To address this criticism, we model missing and partial cue knowledge within PCS explicitly by distinguishing positive, negative, and missing cue values (Gigerenzer et al., 2008). More specifically, within the network structure of PCS in Figure 2B, we encode missing cue values by deleting the links between the corresponding cue and option nodes. Together with the assumption that recognition is a precondition for further knowledge, this results in a spreading-activation network without direct links between the unrecognized option node and the knowledge cues. Substantively, this leads to the psychologically plausible prediction that the unrecognized option is chosen only if sufficient evidence is available *against* the recognized option due to the inhibitory link between the two choice options. For instance, if a city is recognized as not being state capital or having an airport, PCS predicts that the unrecognized alternative is more likely to be chosen.

This conceptualization of partial cue knowledge by the nonexistence of excitatory and inhibitory links implies that PCS mimics the RH in making noncompensatory decisions if only the recognition cue is known. To illustrate this, Table 1 shows different cue configurations for the recognized option. If only recognition is available (R-only; Column 2), all other cue values are unknown and hence the network in Figure 2B essentially has a single cue node (i.e., recognition) that solely determines the decision. Obviously, such cases are by definition noncompensatory, since only recognition determines the decision process (Hilbig, Scholl, & Pohl, 2010). To derive opposing predictions for both models, it is necessary to consider recognized options for which further cue knowledge is available to the decision maker (Pohl, 2011).

Whereas the serial heuristic account predicts that processing of further cues requires additional time, PCS predicts the opposite if further knowledge is congruent with recognition (Glöckner & Bröder, 2011). For instance, if the recognized city is known to have an airport, a university, and is a state capital, then all cue values are positive and thus congruent with the recognition cue as shown in the third column of Table 1 (R-congruent). In a neural network, congruent information facilitates and speeds up the search for a coherent representation of the available information and thus leads to faster and more confident choices compared to trials without further knowledge (Glöckner & Betsch, 2012).

In contrast, if additional information is incongruent with recognition as shown in the fourth column of Table 1 (R-incongruent), PCS predicts slower choices because this incoherence needs to be resolved (Glöckner & Bröder, 2011, 2014). For instance, if one knows that the recognized city has neither an airport nor a metro, the recognition cue contradicts the negative cue values representing further knowledge. Hence, more processing time is required to make an intuitive decision, that is, to arrive at a coherent representation of the available information.

Obviously, these three types of recognized options—R-only, R-congruent, and R-incongruent—represent only special cases of possible cue configurations. For some of the recognized options, further knowledge is likely to be partially incomplete or incoherent, as represented by a combination of positive, negative, and missing cue values for the recognized option. However, since PCS is precisely defined as a computational model, it directly predicts choices and RTs for these more ambiguous cases.

Figure 3 shows these predictions for a recognition validity of $\alpha = .80$ and five knowledge cues with validities of .85, .80, .70, .65, and .60. Each point represents a prediction for a specific configuration of the five cue values for the recognized option. Given that each cue can be positive, negative, or missing, there are $3^5 = 243$ possible configurations. Along the *x*-axis, these configurations are ordered by the sum of the cue values for the recognized option (excluding recognition), whereas the *y*-axis shows the required number of iterations as predicted by PCS. Moreover, red (light gray) triangles and blue (dark gray) circles indicate whether PCS predicts the choice of the unrecognized or recognized option, respectively. The three special cases of R-only, R-congruent, and R-incongruent pairs (see Table 1) are highlighted by filled, black dots.

Figure 3 illustrates several implications of a core feature of PCS, namely, that choices and decision times depend more on internal coherence than on the amount of available information (Betsch & Glöckner, 2010). If most of the cues are negative, the unrecognized option is chosen; if most of the cues are positive, the recognized option is chosen. In both directions, decisions become faster as the coherence between cue values increases. For instance, as more and more cues are positive as indicated by larger sums on the *x*-axis, the number of iterations decreases and decisions become faster. Similarly, the unrecognized option is chosen faster if more and more cues are negative. Overall, however, choices of the recognized option are more frequent and faster due to the presence of the positive recognition cue.

Note that our PCS modeling account of memory-based decisions addresses several points that have been criticized previously. First, we present a precise, mathematical implementation that depends on the cue validities and the cue structure only and does not require free parameters. However, since the exact cue values are usually not known in natural, memory-based decisions, we need to test predictions of the model within the framework of the encompassing measurement model introduced below. Second, we directly model partial knowledge in PCS by missing links in the network (cf. Figure 2B) and do not assume that all cues are available when making inferences from memory (Gigerenzer & Goldstein, 2011). Given that the set of valid cues is possibly very large in natural environments, it is very likely that people have only partial cue knowledge leading to sparse cue structures.

Valid Tests of the Noncompensatory Nature of Recognition

To test compensatory versus noncompensatory accounts of recognition-based decisions, previous studies often manipulated recognition or cue knowledge. In one of the earliest studies, Oppenheimer (2003) found that fictional (and thus unrecognized) cities were often chosen over existing, recognized cities. By inducing both recognition and cue knowledge, Newell and Shanks (2004) showed that participants relied on further knowledge in a stock market prediction game. Similarly, Richter and Späth (2006) manipulated cue knowledge using existing material from natural environments to show that the influence of further knowledge in

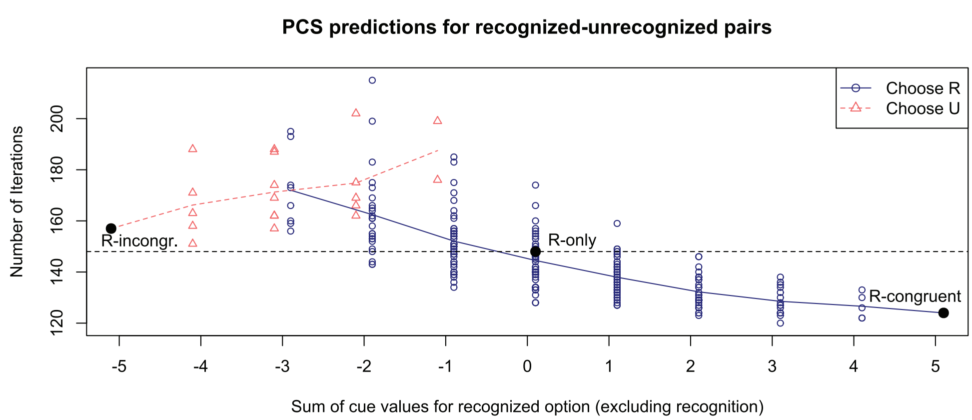

Figure 3. Choice and RT predictions of PCS for recognized-unrecognized (RU) pairs based on a recognition validity of $\alpha = .80$ and cue validities of .85, .80, .70, .65, and .60. Each point represents one of the $3^5 = 243$ possible configurations of cue values for the recognized option (i.e., a combination of positive, negative, and missing cue values). Dashed red and solid blue lines show the mean number of iterations when choosing the unrecognized and the recognized option, respectively. The dashed horizontal line serves as a reference for the number of iterations if only the recognition cue is available. See the online article for the color version of this figure.

decisions about the population of animal species, the safeness of airlines, or the size of cities.

More recently, Glöckner and Bröder (2011) manipulated cue knowledge for existing cities by presenting three cue values to the participants in the decision phase. Based on the completely known cue structure, they used choices, confidence ratings, and RTs to classify participants as users of the recognition heuristic or users of PCS. Many participants responded faster if further knowledge was congruent with recognition, which corroborated the compensatory PCS account. In a second study, Glöckner and Bröder (2014) replicated their previous finding using memory-based instead of screen-based cue knowledge in the decision phase and found only slightly higher rates of RH users. In addition to behavioral measures, Hochman, Ayal, and Glöckner (2010) found that an additional, recognition-congruent cue reduced physiological arousal as measured by a change in blood volume in the fingertips. This result is in line with PCS and opposes the special status attributed to recognition by the RH.

A potentially severe limitation of these studies concerns the experimental manipulation of either recognition, cue knowledge, or both. Originally, the RH was proposed to be an adaptive strategy for decision making in natural environments, where any information needs to be retrieved from memory (Goldstein & Gigerenzer, 2002). Moreover, recognition itself is conceptualized as knowledge that is obtained naturally prior to an experiment as opposed to knowledge that is experimentally induced (Gigerenzer & Goldstein, 2011). Hence Gigerenzer, and Goldstein (2011) questioned whether studies relying on experimental manipulations of knowledge represent a fair test of the RH. Specifically, participants might overly rely on further knowledge due to demand characteristics and an increased accessibility of further cue knowledge (Pachur et al., 2008).

Given that some scholars argue that the manipulation of cue or recognition knowledge jeopardizes a fair test of the RH, the question arises whether it is possible at all to test the effect of further knowledge in the standard RH paradigm, which only consists of a decision and a recognition phase. Indeed, some studies tested the noncompensatory nature of the RH under such conditions. For instance, Pachur et al. (2008) simply asked participants about their knowledge with respect to cues that were judged to be relevant and valid by other participants. Compared to experimentally induced knowledge, the natural knowledge had a smaller but still reliable effect on decisions. Similarly, Pohl (2006) proposed to ask participants whether a city is merely recognized (mR) or recognized with further knowledge (R+). Whereas this distinction should be irrelevant if recognition is used in a noncompensatory way as stated by the RH (but see Marewski & Schooler, 2011), compensatory theories predict differences in choice probabilities and response latencies for both types of stimuli. For instance, information-integration theories predict that mean RTs are faster in pairs including additional knowledge (R+/U) compared to pairs including merely recognition (mR/U; Hilbig & Pohl, 2009).

However, these alternative methods come with their own drawbacks. First, these approaches are not directly connected to the two process models discussed above. Second, the introspective method of asking participants directly about further knowledge can be questioned. If asked for specific cues as in Pachur et al. (2008), important information that is available to the participants might be missed; if asked more generally about the distinction of "merely recognized" and "recognized with further knowledge," it is still not clear whether the additional knowledge is congruent or incongruent with recognition (i.e., whether the cue values are mostly positive or negative). To overcome these problems of both approaches—experimental manipulations of knowledge on the one hand and introspective judgments on the other hand—we rely on a new measurement model as an encompassing framework to test whether and how further knowledge affects recognition-based decisions.

Toward an Encompassing Measurement Model

In this section, we first introduce the r-model, a multinomial processing tree model for recognition-based decisions. Second, we show how to extend this model to account for choices and response times simultaneously. Third, we show that the extended model can serve as an encompassing framework to test the two process models of interest against each other.

The r-Model: Measuring Noncompensatory Use of Recognition

Hilbig, Erdfelder, and Pohl (2010) proposed a multinomial measurement model to disentangle the noncompensatory use of recognition from the use of further knowledge. The r-model belongs to the family of multinomial processing tree (MPT) models that explain observed response frequencies by a finite number of latent cognitive states (Batchelder & Riefer, 1999; Erdfelder et al., 2009). The r-model is tailored to the standard RH paradigm, which includes a recognition and a decision phase, and does not require an experimental manipulation of recognition or cue knowledge.

The r-model as shown in Figure 4 (Hilbig, Erdfelder, & Pohl, 2010) accounts for responses in all three types of trials in the decision phase. Given the binary recognition judgments for all options, each trial in the decision phase is categorized either as knowledge case (both objects recognized; RR pair), recognition case (exactly one object recognized; RU pair), or guessing case (neither object recognized; UU pair).

If both alternatives are recognized (RR pairs), participants are assumed to rely on prior knowledge, which leads to a correct decision with probability b, an estimate of knowledge validity. If exactly one of the two alternatives is recognized (RU pairs), several underlying cognitive processes can lead to identical responses. With probability r, the recognition cue is used in a noncompensatory manner, thus always resulting in the choice of the recognized option irrespective of further knowledge. Note that this core parameter r has often been used as a process-pure measure for the probability of RH use (e.g., Horn et al., 2015; Michalkiewicz & Erdfelder, 2016). Depending on the recognition validity a, reliance on recognition either results in a correct or incorrect choice.

With probability $1 - r$, further knowledge determines the decision and either leads to a correct or incorrect response with probability b and $1 - b$, respectively. In the first case, the conditional probability of choosing the recognized option is given by the recognition validity a, in the latter case, it is given by its complement $1 - a$. Note that the r-model explicitly assumes that two types of processes can lead to a decision in favor of the recognized option: Either only the recognition cue determines the decision in

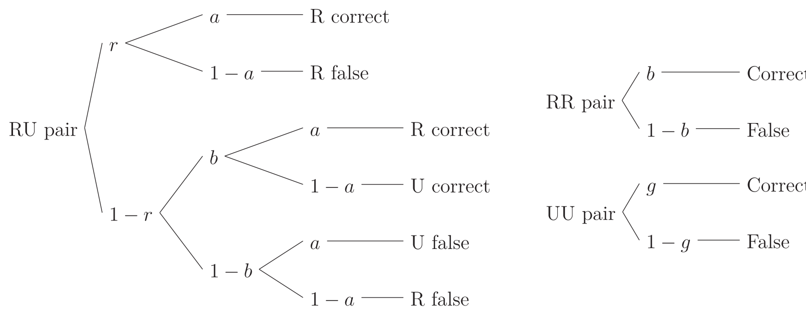

Figure 4. The r-model disentangles observed responses to pairs of recognized (R) and/or unrecognized (U) options to obtain estimates for *r*, the probability of noncompensatory use of recognition, the recognition validity *a*, and the knowledge validity *b* (Hilbig, Erdfelder, & Pohl, 2010).

a noncompensatory fashion (RH use) or further knowledge results in choosing the recognized option. Finally, if both of the alternatives are not recognized (UU pair), participants are assumed to merely guess and give a correct response with probability *g*. The *g*-parameter can be used to test whether participants performed better than chance ($g > .5$) and hence may serve as a validity check whether people have knowledge despite categorizing the options as unrecognized.

As a necessary condition for the validity of a model, Hilbig, Erdfelder, and Pohl (2010) showed that the parameters of the r-model are selectively influenced by experimental manipulations as predicted by their theoretical interpretation. In contrast to measures such as the adherence rate (i.e., the observed proportion of R-choices; Goldstein & Gigerenzer, 2002) or the discrimination index (the difference in the proportions of correct vs. incorrect R-choices; Hilbig & Pohl, 2008), the r-model results in an unbiased estimate of the probability to rely solely on the recognition cue (Hilbig, 2010a). Moreover, the r-model has been used in many studies to test the influence of experimental manipulations on the probability of RH use. For instance, the probability of RH use decreased when raising the availability of further information (Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015). Concerning individual differences, Michalkiewicz and Erdfelder (2016) demonstrated the stability of the parameter *r* within participants across time and different contexts. Moreover, RH use increases in old compared to young adults (Horn et al., 2015).

Overall, previous studies clearly showed that the parameter *r* measures the adaptive, noncompensatory use of recognition. However, given that information-integration models such as PCS can mimic noncompensatory decision strategies, the *r* parameter cannot directly be interpreted as the probability to rely on a qualitatively distinct heuristic. Instead, the parameter *r* might merely represent a measure of the dominance of the recognition cue compared to other cues within a compensatory framework (Hilbig, Scholl, & Pohl, 2010). Importantly, the standard r-model does not allow to discriminate these two opposing explanations, since it relies on choice frequencies only, where both models make similar predictions. However, by incorporating information about RTs into the model, it is possible to disentangle the two process models and shed light on the speed of noncompensatory use of recognition.

Modeling RTs Within the MPT Framework

To account for the speed of responding, we rely on a general method for including RTs in MPT models (Heck & Erdfelder, 2016). This method is based on the fundamental assumption of MPT models that a finite number of cognitive processes determines the observed responses. This assumption has a direct implication for the observed RT distribution, which must necessarily be a mixture of a finite number of latent RT distributions associated with the different cognitive processing paths of the MPT model. Figure 5 illustrates this idea of assigning a separate latent RT distribution to each of the six hypothesized processing paths of the r-model. Whereas some of these distributions are directly observable (e.g., the RT distribution of correct choices of the unrecognized option), others are not (e.g., the RT distribution of correct choices of the recognized option due to RH use). However, the model directly predicts a mixture structure for the observed RT distributions with the mixing probabilities given by the MPT path probabilities.

To illustrate this, consider correct choices of the recognized option, which either emerge from RH use with probability $r \cdot a$ or from further knowledge with probability $(1 - r) \cdot b \cdot a$. The corresponding RTs also emerge from one of two latent RT distributions, correct RH use (with density f_{RH^c}) or correct use of further knowledge (f_{K^c}). It follows that the density of the observed RTs of correctly choosing the recognized option is given by the mixture

$$f_{\mathrm{R}^c}(t) = \frac{r \cdot a}{r \cdot a + (1 - r) \cdot b \cdot a} f_{\mathrm{RH}^c}(t) + \frac{(1 - r) \cdot b \cdot a}{r \cdot a + (1 - r) \cdot b \cdot a} f_{\mathrm{K}^c}(t). \tag{4}$$

The objective of the RT-extended measurement model is to estimate the latent RT distributions (e.g., f_{RH^c}) based on the choice proportions and the observed RT distributions (e.g., f_{R^c}).

Whereas the assumption that observed RT distributions are mixtures follows directly from the basic structure of any MPT model, it is less obvious how to model the latent RT distributions. To avoid any auxiliary parametric assumptions about the exact shape of the underlying, unobservable RT distributions, we instead rely on a distribution-free approach (Heck & Erdfelder, 2016). Essentially, we replace the continuous distribution in Figure 5 by histograms and estimate the resulting bin probabilities, which does

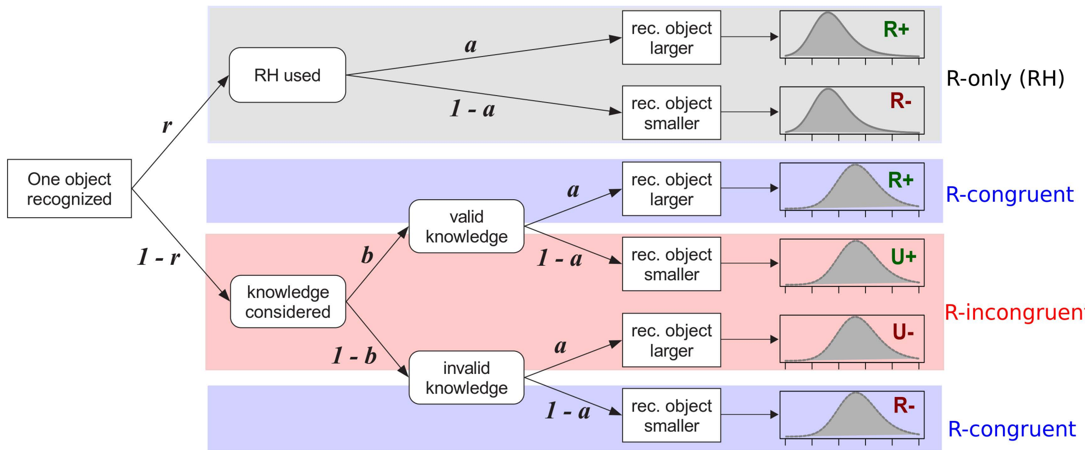

Figure 5. The RT-extended r-model. Observed RTs are assumed to follow a mixture distribution of the three latent RT distributions of interest (R-only, R-congruent, and R-incongruent). In this illustration, the indicated latent RT distributions correspond to predictions of the serial RH theory. See the online article for the color version of this figure.

not restrict the estimated shape of the latent RT distributions in any way.

Even though this approach induces a loss of information (ratio-scaled RTs are categorized into discrete bins), it provides some crucial advantages besides not requiring parametric assumptions (Heck & Erdfelder, 2016; Yantis, Meyer, & Smith, 1991). First, the RT-extended model belongs to the class of MPT models itself and merely has more observed categories (i.e., each response category is split into B more fine-grained categories, where B is the number of RT bins). Thus, existing methods and software for this model class can readily be used (e.g., Heck, Arnold, & Arnold, in press; Moshagen, 2010; Singmann & Kellen, 2013). Second, the approach allows for a direct estimation of the latent cumulative density functions and thus for a test of stochastic dominance. Moreover, we use a principled strategy to choose individual boundaries for the categorization of RTs into bins, thereby ensuring that the more fine-grained RT-frequencies are comparable across participants. For instance, when using only two RT bins, we rely on the mean of the log RTs across all responses per participant (i.e., the geometric mean: $T_{\text{bnd}} = \exp(\overline{\log T})$) to ensure that fast and slow responses are approximately equally likely a priori. In contrast to fixed RT boundaries, this strategy allows to compare and aggregate the individual RT-frequencies because the RT bins are precisely and identically defined per participant (Heck & Erdfelder, 2016).

Linking Latent RT Distributions to Both Process Models

Before applying the RT-extended r-model, it is important to ensure that the new model parameters that measure the relative speed of responses can be uniquely estimated, or in other words, that the model is identifiable. Without further constraints, it is not possible to uniquely estimate the six latent RT distributions shown on the right side of Figure 5 in a distribution-free way based on only four observed RT distributions (for details, see Heck & Erdfelder, 2016). As a remedy, we derive suitable theoretical constraints on the latent RT distributions of the RT-extended r-model that allow for a test between the two competing process models. Even though we constrain the latent RT distributions in a way that is compatible with both process models, the two theoretical accounts predict distinct orderings of the latent RT distributions. Hence, a single encompassing measurement model will allow for an empirical test between the serial heuristic and the parallel information-integration account.

First, we assume that the speed of noncompensatory use of recognition is invariant with respect to the outcome, that is, whether the choice is correct or incorrect. In other words, we assume a single latent RT distribution called “R-only” that determines the speed of responses emerging from the first two processing paths in Figure 5. This assumption is compatible with both process accounts: The serial heuristic account directly predicts that responses due to recognition are always similarly fast, since the same search, stopping, and decision rules are used irrespective whether the resulting response is correct or not. The same is true for PCS—if only the recognition cue is available to make a decision, the choice probabilities and the speed of responding do not depend on the correctness of the recognition cue in a specific trial. In sum, the latent RT distribution “R-only” measures the relative speed of relying on recognition only in both theoretical accounts.

Concerning the RT distributions of relying on additional available information, we differentiate between R-congruent and R-incongruent further knowledge, which results in choosing the recognized or unrecognized option, respectively. Similarly, as for noncompensatory responses, we assume that the speed of correct and incorrect responses is identical conditional on the type of further knowledge. In other words, the speed of responding depends on the coherence of further cues with recognition but not on the validity of the cues in a specific trial. This distinction between R-congruent and R-incongruent further knowledge is shown graphically using different colors/shades in the lower four processing paths in Figure 5.

Importantly, the distinction between R-congruent and R-incongruent further knowledge allows us to derive distinct predictions for both process models. The serial heuristic account predicts that use of further information requires additional processing time (Equation 1), which results in slower responses compared to noncompensatory RH responses—irrespective of the type of further knowledge (Pachur & Hertwig, 2006). Within the encompassing framework of the RT-extended r-model, this stochastic dominance hypothesis is formalized by two inequalities on latent RT distribution functions:

$$\begin{aligned}\mathcal{H}_{\text{serial}}:\ & F_{\text{R-only}}(t) > F_{\text{R-congruent}}(t)\\ & F_{\text{R-only}}(t) > F_{\text{R-incongruent}}(t) \quad \forall\, t \in \mathbb{R}^{+}.\end{aligned} \tag{5}$$

Hence, the RT-extended r-model allows to test the predicted ordering of RT distributions as derived from the serial-processing account in Equation 2.

In contrast to the serial heuristic account, PCS makes distinct predictions depending on the coherence of further cues with the recognition cue. If further knowledge is R-congruent, the decision maker perceives a state of increased coherence and hence makes faster decisions compared to the case that no further cues are available (Glöckner & Bröder, 2011, 2014). If further knowledge is R-incongruent, the incoherence of the available information must be resolved and hence decisions require more time. The PCS hypothesis of stochastic dominance is given by the inequalities

$$\mathcal{H}_{\text{parallel}}: F_{\text{R-congruent}}(t) > F_{\text{R-only}}(t) > F_{\text{R-incongruent}}(t) \qquad \forall\, t \in \mathbb{R}^{+}. \tag{6}$$

Again, the RT-extended r-model serves as an encompassing framework to test whether congruent information actually leads to faster decisions as predicted in Figure 3. Even though the new measurement model does not allow to differentiate between different underlying configurations of cue values (i.e., which cue values are positive, negative, or missing), it allows to estimate the overall effect of R-congruent and R-incongruent information on RTs.

The latent RT distributions of the remaining two multinomial processing trees (either both options recognized or both unrecognized) are not informative to test between the serial and parallel accounts of recognition-based decisions. In these cases, the recognition cue is either positive or negative for both options and hence it cannot be tested whether recognition enters the decision process in a noncompensatory fashion or not. These cases are also not informative to test how several knowledge cues are integrated—to answer such questions, paradigms with known cue matrices are much better suited (e.g., Glöckner et al., 2014). Hence, the following constraints are only of minor interest.

If both options are unrecognized (UU pairs), we assume that participants merely guess, assuming that the two underlying RT distributions are independent of the outcome and thus identical. In contrast, if both options are recognized (RR pairs), both theoretical accounts do not allow clear-cut predictions; hence, we did not impose constraints on the corresponding RT distributions. Appendix A formally shows that this set of constraints on the latency parameters renders the RT-extended r-model identifiable if the core parameters *r, a, b,* and *g* are in the interior of the parameter space (i.e., not equal to zero or one) and if $b \neq .5$. Note that the latter constraint substantively means that the model is unidentifiable if further knowledge is invalid (i.e., at chance level). Given that virtually all natural and experimental environments of interest have recognition and knowledge validities above chance level ($a > .5$, $b > .5$), this special case does not restrict the applicability of the RT-extended r-model.

Overall, the RT-extended r-model allows to test between the two competing process models by estimating the relative speed of R-only, R-congruent, and R-incongruent responses. Note that the predictions of both theoretical accounts refer to *latent* RT distributions that need to be estimated within an encompassing measurement model, since they cannot be observed directly (because of the ambiguity whether R-choices were due to recognition-congruent knowledge or due to recognition only).

Simulation Study

Before analyzing empirical data with the RT-extended r-model, we show in a simulation study that the proposed measurement model can indeed differentiate between data generated by the serial heuristic and the parallel information-integration account. To simulate a natural environment, we used a classical data set that includes information about the largest German cities in the form of nine probabilistic cues with validities ranging from .51 (East Germany) to 1.00 (national capital; Gigerenzer & Goldstein, 1996; Schooler & Hertwig, 2005). Note that the supplementary material provides more technical details (including R code) and a second simulation based on an updated database of cue and criterion values that lead to identical conclusions.

Data Generation

Both cognitive process models, the serial RH and PCS, rely on the same input. Besides the recognition and cue validities, two vectors of cue values are required for the available options:

$$\begin{aligned}C_1 &= (R_1, C_{11}, \ldots, C_{1K})\\ C_2 &= (R_2, C_{21}, \ldots, C_{2K}).\end{aligned}$$

Whereas the recognition status of both alternatives is necessarily known ($R_i \in \{-1, +1\}$), the other cue values may encode missing information by $C_{ic} = 0$. Note that we do not simulate responses and RTs for the case that both options are unrecognized, since these cases cannot differentiate whether or how knowledge is integrated. Moreover, it is sufficient to generate RTs up to an interval scale because the RT-extended r-model relies on the *relative* speed of the latent processes only. Hence, to generate RTs, we simulated the number of required elementary information processes (EIPs; Payne et al., 1988) in case of the serial account and the number of required iterations in case of PCS (plus continuous random noise in both cases).

Recognition Process

The probabilities to recognize a city were modeled by the sigmoid function of the criterion values shown in Figure 6A (a probit link; cf. Hilbig, 2010a; Schooler & Hertwig, 2005). Substantively, this function induces the ecological correlation that people are more likely to recognize larger cities. In our simulation, this resulted in a mean recognition validity of $\alpha \approx .80$ and a mean discrimination rate of $d \approx .50$.

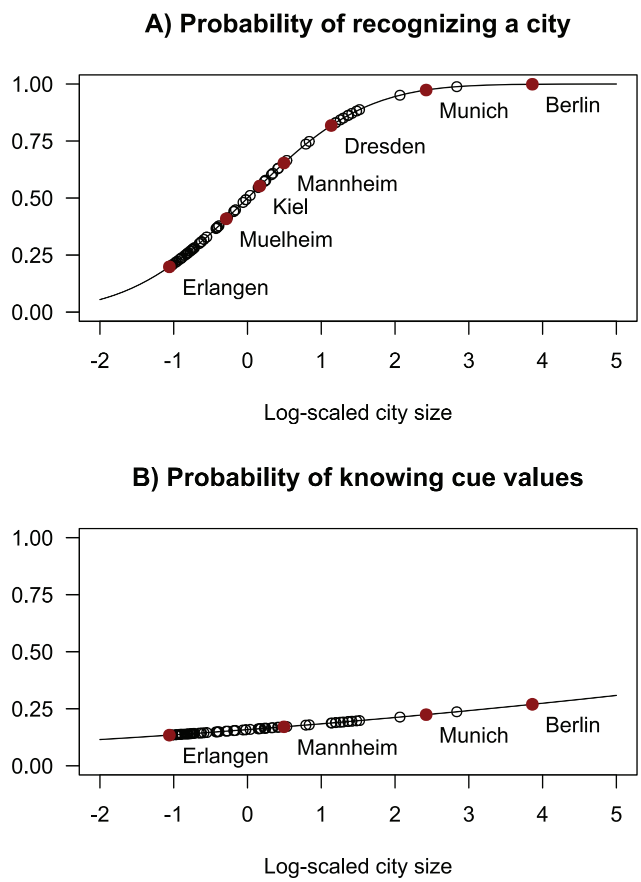

Figure 6. Probit-link functions used to model (A) recognition probabilities and (B) partial cue knowledge. City sizes were first log-transformed and then *z*-standardized. For illustration, the named cities are indicated by red (gray) points. See the online article for the color version of this figure.

For each participant, a random vector of recognized ($R_i = +1$) and unrecognized ($R_i = -1$) cities was drawn. For the unrecognized cities, all cue values were set to zero, thereby modeling the absence of knowledge ($C_{ic} = 0$). Within the subset of recognized cities, we assumed that cue values were more likely known for larger than for smaller cities. For this purpose, we generated probabilities of knowing a cue value ($C_{ic} = 1$ or $C_{ic} = -1$) based on the sigmoid function in Figure 6B. This procedure resulted in individual cue structures with partial knowledge about some cue values of the recognized cities.

Since our focus is on the processing time required for the integration of recognition with further available information (i.e., in which way and how fast the cue values R_i and C_{ic} are combined), we did not model the time required to recognize an option. However, we address this possible issue in the Discussion below and also show in Appendix B that partialing out the recognition times from the decision times does not change the empirical results qualitatively.

Implementation of the Serial Heuristic Account

According to the serial heuristic account, the take-the-best strategy is applied if both alternatives are recognized (Gigerenzer & Goldstein, 1996). This heuristic evaluates the cues by decreasing validity and stops if a cue favors one option ($C_{1c} = +1$) but not the other one ($C_{2c} \in \{-1, 0\}$; Gigerenzer & Goldstein, 1996, Figure 3) and results in random guessing if no discriminating cue is found. The number M of EIPs of this strategy is given by the number of cues compared (Bröder & Gaissmaier, 2007).

If one of the options is recognized and the other not, the RH is used with probability r, whereas the TALL strategy is used with probability $1 - r$. Note that one of the two strategies is selected randomly in each trial without considering any cost-benefit trade-offs, reinforcement learning, or natural environmental constraints (e.g., Marewski & Schooler, 2011; Payne et al., 1993; Rieskamp & Otto, 2006). We did not model such mechanisms of strategy selection to allow for a direct link of the two process accounts of information integration to the parameters of the new measurement model. Moreover, a preceding strategy-selection stage would add some positive processing time to both strategies, RH and TALL, but would not change the prediction of stochastic dominance (cf. "Stochastic Dominance," above).

If the RH is used, the R-option is always chosen and a single processing step is required ($M = 1$), since the RH relies on the recognition cue only. If TALL is used, the unrecognized option is chosen if the number of negative cue values is larger than or equal to the number of available cues. This TALL strategy requires as many EIPs as there are informative cues available for the recognized option (i.e., the number of $C_{ic} \in \{-1, +1\}$ plus one for R_i).

Implementation of the Parallel Information-Integration Account

PCS uses the same input as the serial-processing model, but relies on maximizing the coherence in the connectionist network shown in Figure 2B. We obtained choice and RT predictions for PCS based on the iterative updating algorithm described by Glöckner and Betsch (2008) using the same set of fixed parameters as in other applications of PCS (e.g., Glöckner et al., 2014). In the following, we show how the available cue values determine the network structure (i.e., the weights of the links) and how the spreading of activation through the network is modeled mathematically (cf. Glöckner & Betsch, 2008).

For each paired comparison, the available cue values C_1 and C_2 determine whether the links between cue and option nodes are inhibitory or excitatory. Specifically, the cue-option weight between cue node c and option node i is fixed to $w_{ci} = 0.01 \cdot C_{ic}$ based on the cue value C_{ic} (and similarly for R_i). For instance, a negative cue value manifests itself in the network as a bidirectional, inhibitory link with weight $w_{ci} = -0.01$ between the corresponding cue and option node. Moreover, missing cue information is simply modeled by nonexistent links, or equivalently, $w_{ci} = 0$.

Regarding the environmental structure, the cue validities v determine the weights of the links between the general-validity and the cue nodes. Mathematically, the validities v are transformed by the weighting function $w_i = (v_i - 0.5)^P$, where $P \geq 0$ is a sensitivity parameter that determines how strongly large validities are overweighted compared to small ones. The sensitivity parameter P directly determines whether decisions are rather compensatory ($P < 1$) or noncompensatory ($P > 1$) compared to a model that weighs the cue validities linearly ($P = 1$). Based on previous empirical and theoretical work (Glöckner et al., 2014), we used the default $P = 1.9$, which results in choice predictions of PCS that are similar (but not identical) to those of a rational Bayesian choice rule (Jekel, Glöckner, Fiedler, & Bröder, 2012; Lee & Cummins, 2004). Finally, the weight of the links between both options is set to the default $w = -0.2$, thus modeling strong inhibition between the choice options. Psychologically, this implies that a strong preference for an option decreases the preference for alternative options (and vice versa).

Once the weights of the network are fixed, activation is assumed to spread in parallel through the network. Mathematically, this is modeled by an iterative algorithm that updates the activations of all cues and option nodes in each iteration simultaneously (Glöckner & Betsch, 2008). The network is initialized with the activations of all cue and option nodes set to zero. Only the general validity node has a constant activation of one from the beginning and supplies the network with energy. To update the activation $a_i(t)$ of a node i at time t, the weighted sum of the current activation values of all nodes serves as the input value $i_i(t)$,

$$i_i(t) = \sum_{j=1}^{J} w_{ij} \cdot a_j(t). \tag{7}$$

This input is passed to a piecewise-linear activation function to update the current activation of node i,

$$a_i(t+1) = (1-d) \cdot a_i(t) + \left[1 + \mathrm{sign}(i_i(t)) \cdot a_i(t)\right] \cdot i_i(t), \tag{8}$$

where sign(x) is -1 if $x < 0$ and $+1$ otherwise, and d a constant decay factor (set to $d = .1$; Glöckner et al., 2014).[2] As mentioned above, all nodes are updated simultaneously in parallel until a state of coherence is reached. This is the case if the overall change in activation falls below some threshold, or more precisely, if the energy $e(t)$ of the network,

$$e(t) = -\sum_i \sum_j w_{ij} \cdot a_i(t) \cdot a_j(t), \tag{9}$$

differs less than 10^{-6} across 10 consecutive iterations. Based on the final state of the network, the option with the highest activation is chosen. Moreover, interval-scaled RT predictions are provided by the number of iterations until convergence is reached.

Results

We generated separate configurations of known cities and available cue values for $N = 500$ participants. For each of these individual cue patterns, we applied both process models to all $\binom{83}{2} = 3{,}403$ binary decisions. These simulated choices, RTs, and recognition judgments served as input for the analysis using the RT-extended r-model. Importantly, the measurement model does not rely on any information regarding the simulated cue values, which were hidden similar as in an empirical setting. In the following, the simulated data are used to fit two versions of the RT-extended r-model that differ in the number of bins used to categorize RTs.

The Relative Speed of Processes

First, we fitted the RT-extended r-model using only two RT bins. This allows to summarize each latent RT distribution i by a single latency parameter L_i defined as the probability of responding faster than the RT boundary. Hence, larger estimates of L_i indicate a faster cognitive process. To categorize responses as "fast" or "slow," we used geometric mean RTs per participant. This allows to aggregate the frequencies across individuals, which is not possible with arbitrary or fixed RT boundaries (Heck & Erdfelder, 2016). Since the simulation did not induce heterogeneity between participants, we fitted the RT-extended r-model to the summed frequencies across all 500 individual data sets.

Figure 7 shows the resulting maximum-likelihood estimates for the core parameters of the r-model (*r, a, b,* and *g*) and the latency parameters that measure the relative speed of the latent processes (most importantly, $L_{\text{R-only}}$, $L_{\text{R-con}}$, and $L_{\text{R-inc}}$). Both types of process models, the serial heuristic and the parallel information-integration account, resulted in similar estimates for *r,* the probability of relying on the recognition cue only, for the recognition validity *a,* and for the knowledge validity *b.* This shows that the standard r-model without RTs does in general not allow to discriminate between both types of data-generating models. Even though the parameter *r* has usually been interpreted as "the probability of RH use" (Hilbig, Erdfelder, & Pohl, 2010), it might as well represent the probability of relying only on the recognition cue within the PCS framework, without assuming a distinct, qualitatively different heuristic.

However, the latency parameters of the RT-extended r-model allow to differentiate between both process accounts as shown in the right part of Figure 7. Note that the latency parameters L_i can only be interpreted conditional on the RT boundaries, which are differently defined for both data-generating models. Hence, the latency estimates cannot be compared across process models. Importantly, however, the relative speeds of latent processes can be compared *within* each process model to infer which one generated the data. For the serial heuristic, RH responses (R-only) were estimated to be faster than those based on further knowledge regardless of coherence (cf. Equation 5). In contrast, for PCS, R-congruent responses were clearly faster than those due to noncompensatory use of recognition, whereas R-incongruent responses were even slower. This ordering reflects the crucial prediction of PCS that the coherence of the available information determines the speed of responding and not the amount of cues entering the decision process (cf. Equation 6).

Testing Stochastic Dominance

For a more fine-grained test of stochastic dominance, it is important to estimate the cumulative density functions across the entire range of observable RTs (Heathcote et al., 2010). A sufficiently high precision is important to detect intersections of the cumulative densities at the tails (e.g., for very fast or slow responses). In RT-extended MPT models, such more fine-grained tests can be implemented by using larger numbers of RT bins (Heck & Erdfelder, 2016).

In general, the use of B RT bins requires the estimation of $B - 1$ latency parameters, each resembling the height of a bin in an RT histogram. More precisely, each latent RT distribution is described by a vector $L = (L_1 \ldots, L_{B\text{-}1})$, where L_b is the probability that RTs from the corresponding process fall into the b-th RT bin. We fitted the RT-extended r-model in a Bayesian framework using the software JAGS (Plummer, 2003), which provides posterior sam-

[2] In Equation 8, we assume that the minimum and maximum activation of a node is -1 and $+1$, respectively.

ples using Markov chain Monte Carlo (MCMC) methods. Based on the posterior samples, the cumulative density function is estimated by the partial sums of the latency parameters, that is, $L_b^{cum} = \sum_{k=1}^{b} L_k$. Note that this is done for each MCMC replication separately and thus provides posterior credibility intervals for the cumulative densities.[3] We used uniform priors on the substantive core parameters of the r-model, assuming that all parameter values are a priori equally likely. Moreover, we adopted a noninformative Dirichlet prior for the latency vectors, $\boldsymbol{L} \sim \text{Dir}\,(1/B, \ldots, 1/B)$. This prior assigns equal probability to all RT bins and has an influence similar to a single observation (Heathcote et al., 2010).

Based on 30 RT bins, Figure 8 shows posterior means and 95% credibility intervals for the cumulative densities of the latent RT distributions. Depending on the data-generating model, different patterns emerged for the RT distributions of noncompensatory use of recognition (R-only) compared to R-congruent and R-incongruent use of further information. For the serial heuristic account, the RT-extended r-model recovered the theoretically predicted order of latent RT distributions: responses due to use of the RH were stochastically faster than responses due to reliance on any further knowledge (Equation 5). In contrast, for PCS, the latent RT distributions were stochastically faster for R-congruent responses compared to noncompensatory R-responses, whereas R-incongruent responses were even slower (Equation 6).

Generalizability of Simulation Results

The previous simulation showed that the RT-extended r-model can in principle differentiate between data generated by both process models. To ensure that this result generalizes to other situations, for instance, involving different numbers of cues or different recognition or knowledge validities, we replicated the simulation using randomly chosen parameters to generate data. This approach is similar to parameter space partitioning (Pitt, Myung, & Zhang, 2002), which allows one to investigate qualitative model predictions across the parameter space. Essentially, we changed the cue structure by including different numbers of cues and applying different recognition functions, thereby manipulating the amount of knowledge available (see Supplementary Material for details). As in the previous simulation, these cue structures served as input for the two process models to generate responses and RTs.

Based on 10,000 replications, each including $N = 50$ hypothetical participants, we fitted the RT-extended r-model using two RT bins. Before summarizing the results, we excluded replications that were not of theoretical interest. First, we excluded data sets with parameter estimates $\hat{r}$ at the boundary ($\hat{r} < .05$ or $\hat{r} > .95$) for which latency parameters of R-only and R-congruent responses were not empirically identified and could thus not be compared. Substantively, these replications resembled cases in which further knowledge was never or always used. Second, we excluded replications with recognition or knowledge validities smaller or equal than chance ($\hat{a} \leq .50$ or $\hat{b} \leq .50$).

For the 6,617 remaining, substantively interesting replications, Table 2 shows summary statistics of the estimated relative speed of R-congruent and R-only responses. As expected, the RT-extended

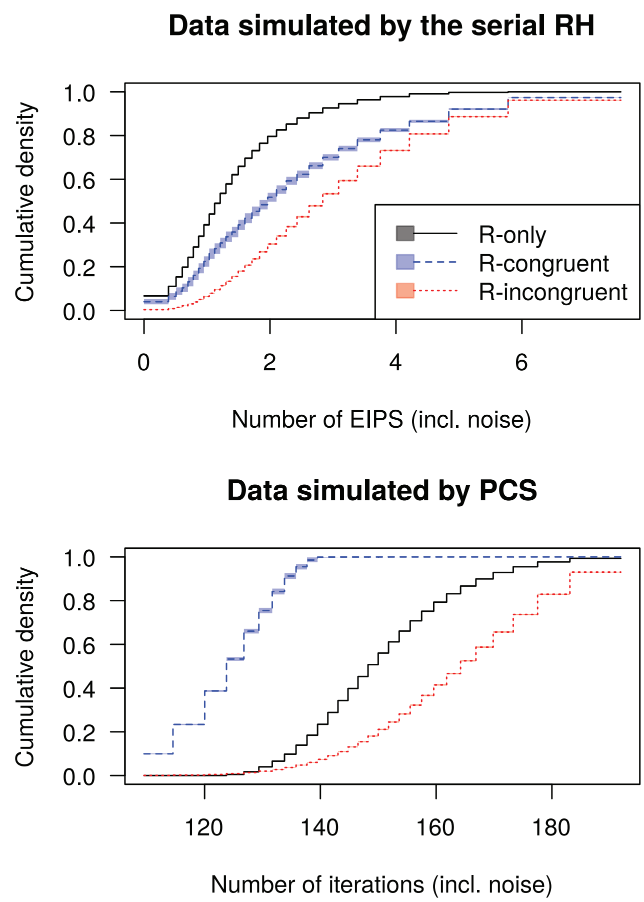

Figure 8. Estimated cumulative density functions of the latent RT distributions based on data simulated by the serial-heuristic account and PCS. See the online article for the color version of this figure.

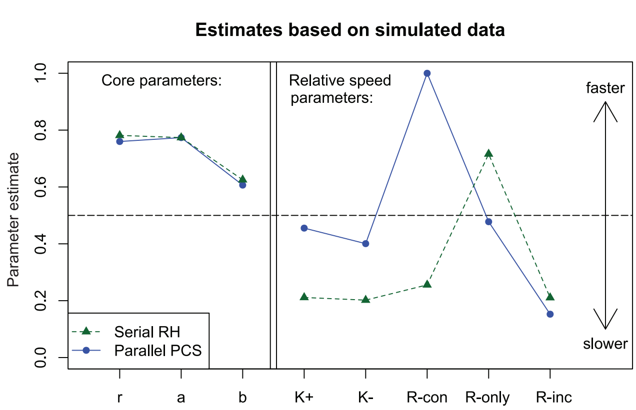

Figure 7. Parameter estimates for the RT-extended r-model with two RT bins based on data generated by both process models, the serial heuristic account and PCS. See the online article for the color version of this figure.

[3] Maximum-likelihood software for MPT models such as multiTree (Moshagen, 2010) requires a *binary* MPT model and hence a reparameterization of each set of 30 latency parameters. To obtain confidence intervals for the sums of the original latency parameters, which describe the estimated cumulative density, a transformation of the standard errors is required (Rao, 1973, Ch. 6).

Table 2
Relative Speed of R-Congruent and R-Only Responses Based on the RT-Extended R-Model in the Robustness Simulation

Data-generating model	R-congruent	R-only	Difference Δ	*SD* (Δ)	*P* (Δ > 0)
Serial RH	.401	.786	−.385	.099	0.0%
Parallel PCS	.882	.174	.708	.186	99.3%

r-model reliably discriminated between the two data-generating process models. If data were generated by the serial RH process model, responses due to R-congruent knowledge were estimated to be slower than those due to RH use in all replications with a mean difference of −.39 (*SD* = .10). In contrast, the opposite pattern emerged for 99.3% of the replications if data were generated by the PCS account of information integration (mean difference of .71, *SD* = .19).

Overall, our simulations show that the RT-extended r-model allows to test between the two process models without knowing the actual underlying cue structures for a variety of data-generation scenarios. As derived theoretically, the crucial difference between the data-generating process models emerged in the latency parameters that measure the relative speed of the underlying processes. When the serial-heuristic model generated the data, R-only choices due to RH use were estimated to be faster than choices due to further knowledge. This order reversed for PCS, according to which increased coherence of the available information facilitates faster responding (Betsch & Glöckner, 2010). Therefore, the RT-extended model can be used as an encompassing measurement model to differentiate between the serial RH and the PCS account of information integration.

Empirical Reanalysis With the RT-Extended r-Model

In the following, we reanalyze 19 data sets by researchers associated with the University of Mannheim. Table 3 lists these studies, which include 1,074 individual data sets with 274,296 memory-based decisions in total.[4] All of these studies used the standard RH paradigm with two phases; a decision phase, in which participants had to choose one of the two presented options, and a recognition phase, in which the stimuli had to be judged as recognized or unrecognized. Moreover, natural, realistic domains such as city size, river length, or celebrity success were used, in which both recognition and further knowledge provided valid information about the criterion (i.e., $\alpha > .5$, $\beta > .5$). We included only studies in which neither recognition nor cue knowledge was experimentally manipulated and in which participants were not asked to respond within a fixed time limit.

In each data set, we removed individual data sets of participants who recognized either all or none of the options. Note that RT outliers were not removed, since our distribution-free approach uses only relative information about the speed of responding and is thus robust against outliers. Within-subject manipulations of the stimulus material were included as separate data sets (e.g., celebrity and movie success in the "Related" condition of Exp. 3 by Michalkiewicz & Erdfelder, 2016). Moreover, many studies additionally asked whether an object was merely recognized (mR) or recognized with further knowledge (R+; Pohl, 2006). In our main analyses, we pooled these different types of items and used only the distinction between recognized and unrecognized options. In Appendix C, we show that the results were not affected when differentiating between the recognized options based on the subjective judgments concerning further knowledge.

To test the PCS against the RH account, we relied on three different versions of the RT-extended r-model. First, we used only two RT bins ("fast" and "slow"), leading to $8 \times 2 = 16$ observed categories and 10 model parameters. Due to the moderate size of this model, we accounted for the nested data structure (responses nested in participants nested in data sets) in a hierarchical MPT model (Klauer, 2010). Second, for a more fine-grained test of stochastic dominance (Heck & Erdfelder, 2016), we used 40 RT bins to summarize the latent RT distributions. With $8 \times 40 = 320$ observed categories, this model version requires much more observed responses to obtain reliable estimates. Hence, we refrained from modeling the hierarchical structure and aggregated individual frequencies. Third, we adapted the outcome-based strategy classification approach by Bröder and Schiffer (2003) to analyze each participant separately. As shown in the following sections, all three approaches provide converging evidence that choices due to recognition-congruent knowledge are faster than those due to recognition only.

A Hierarchical Bayesian Model

In line with our simulation study, we used geometric mean RTs to categorize responses as "fast" and "slow" on the individual level, thereby ensuring comparable frequencies across participants. To account for heterogeneity between participants and data sets, we fitted a hierarchical latent-trait MPT model (Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015). Specifically, each of the MPT parameters θ in the interval [0, 1] was modeled additively on a latent probit-scale,

$$\Phi^{-1}(\theta_{sji}) = \mu_s + \beta_{sj} + \delta_{si}, \tag{10}$$

where Φ^{-1} is the inverse of the cumulative density function of the standard normal distribution (the probit-link).

The transformed parameters were modeled by additive effects similar to those in generalized linear mixed models (Bates, Mächler, Bolker, & Walker, 2015). First, an overall group mean μ_s with a standard-normal prior was included for each parameter *s*. Second, the fixed effect β_{sj} accounted for mean differences across data sets (e.g., different recognition validities). As a prior, we adopted the weakly informative Cauchy-priors proposed by Rouder, Morey, Speckman, and Province (2012) for ANOVAs of directly measurable dependent variables. Note that the use of fixed effects for the different data sets differs from fitting a hierarchical model to each data set separately, since a single covariance matrix Σ is assumed for the random effects $\boldsymbol{\delta}_s$. In other words, our model accounts for mean differences of the MPT parameters across data sets but assumes that the variances and correlations of the parameters are identical across data sets (Heck et al., in press). Third, the random effect terms δ_{si} with mean zero and covariance matrix Σ accounted for variation between participants. Following Klauer

[4] The Supplementary Material provides data and R code for our reanalysis.

Table 3
Data Sets Included in the First Reanalysis

#	Source	Domain	N	Responses
1	Castela and Erdfelder (2017b) Exp. 1[#]	Celebrity success	74	14,060
2	Castela and Erdfelder (2017b) Exp. 1[#]	City size	73	13,870
3	Castela and Erdfelder (2017b) Exp. 1[#]	River length	72	13,680
4	Castela and Erdfelder (2017a) Exp. 1 (Session 1)	City size	44	10,560
5	Hilbig and Pohl (2009) Exp. 1	City size	24	4,560
6	Hilbig and Pohl (2009) Exp. 3	City size	62	5,642
7	Hilbig, Erdfelder, and Pohl (2010) Exp. 6a	City size	16	2,176
8	Hilbig, Erdfelder, and Pohl (2010) Exp. 6b	City size	18	2,448
9	Michalkiewicz and Erdfelder (2016) Exp. 1 (Session 1)	City size	64	19,200
10	Michalkiewicz and Erdfelder (2016) Exp. 2 (Session 1)	City size	83	24,900
11	Michalkiewicz and Erdfelder (2016) Exp. 3 (Different)[†]	Island size	64	19,200
12	Michalkiewicz and Erdfelder (2016) Exp. 3 (Different)[†]	Musician success	64	19,200
13	Michalkiewicz and Erdfelder (2016) Exp. 3 (Related)[*]	Celebrity success	68	20,400
14	Michalkiewicz and Erdfelder (2016) Exp. 3 (Related)[*]	Movie success	68	20,400
15	Michalkiewicz and Erdfelder (2016) Exp. 4 (Names)	Celebrity success	43	12,900
16	Michalkiewicz and Erdfelder (2016) Exp. 4 (Pictures)	Celebrity success	43	12,900
17	Michalkiewicz, Minich, and Erdfelder (2017) Neutral[‡]	River length	76	22,800
18	Michalkiewicz, Minich, and Erdfelder (2017) Standard[‡]	Musician success	74	22,200
19	Michalkiewicz, Arden, and Erdfelder (2017) Equal	Celebrity success	44	13,200

Note. N refers to the number of participants after exclusion of those who recognized either all or none of the items. Within-subject manipulations with different material across blocks are indicated by the symbols #, *, †, ‡ and hence refer to overlapping sets of participants.

(2010), we used a scaled inverse-Wishart prior on the covariance matrix (with a diagonal matrix and $S + 1$ degrees of freedom, where S is the number of MPT parameters).

The model was fitted with the R package TreeBUGS (Heck et al., in press), which extends the hierarchical MPT implementation by Matzke et al. (2015) and obtains posterior samples from JAGS (Plummer, 2003). We used four MCMC chains with 100,000 iterations each, of which the 20,000 first samples were excluded as burn-in to reduce the dependency on the starting values. The remaining 80,000 posterior samples where thinned by a factor of 50. Convergence was checked graphically, and by ensuring small Gelman-Rubin statistics ($R^2 < 1.1$) and sufficiently large estimated effective sample sizes.

Due to the large sample size of 274,296 observations, statistical tests of the model fit have very high power to reject the model even for minor discrepancies between observed and predicted frequencies. Therefore, we assess model fit qualitatively as shown in Figure 9. For each of the 8×2 categories, mean observed frequencies are compared to box plots of the mean frequencies sampled from the posterior-predictive distribution. As indicated by the tiny ranges of the box plots (whiskers indicate the 1.5 interquartile range), the large sample size resulted in a posterior-predictive distribution with very small variance. Importantly, however, the predicted frequencies were very close to the observed frequencies thus indicating a satisfactory model fit.

The left part of Table 4 shows the posterior estimates for the overall mean parameters and the parameter heterogeneity between participants. Across all data sets, recognition and knowledge validity were sufficiently high to render both sources of information useful for decision makers. Recognition was frequently used in a noncompensatory way as indicated by a large r parameter. Nevertheless, r was clearly smaller than one, thus showing that further knowledge affected memory-based decisions.

Note that participant heterogeneity was much larger for the r parameter compared to the validity and guessing parameters. This is in line with the conceptualization of the r-model, according to which the latter parameters refer to structural properties of the environment whereas the r parameter refers to an individual style of decision making (Hilbig, Erdfelder, & Pohl, 2010; Michalkiewicz & Erdfelder, 2016).

To assess the speed of the latent processes of interest, Figure 10 shows the overall, data-set, and individual estimates for the latency parameters $L_{\text{R-only}}$, $L_{\text{R-con}}$, and $L_{\text{R-inc}}$. Estimates close to one resemble latent cognitive processes that resulted in fast choices with a high probability. On all three levels, a clear ordering of the speed of latent processes emerged. Whereas R-incongruent responses were estimated to be slightly slower than noncompensatory R-responses, R-congruent responses were clearly estimated to be faster than noncompensatory R-responses. This ordering supports the PCS account according to which R-congruent knowledge fosters fast responding.

Testing Stochastic Dominance

To test stochastic dominance of latent RT distributions, we used 40 RT bins, thus increasing the resolution with respect to the decision speed. Since this results in a model with $8 \times 40 = 320$ categories, we refrained from hierarchical modeling and treated responses as being independently and identically distributed across participants.[5] Note that the analysis of averaged data has well-known drawbacks (e.g., Gigerenzer & Goldstein, 2011; Klauer, 2006), but combined with hierarchical and individual analyses, the present model allows for an additional, more fine-grained test of stochastic dominance. To categorize choices into bins from "fast" to "slow," we used the same principled strategy as in the simulation reported above. First, the individual RTs were approximated

[5] In a Bayesian framework, the weaker assumption of exchangeability suffices, that is, permutations of the observations do not change the joint distribution of the data (Bernardo, 1996).

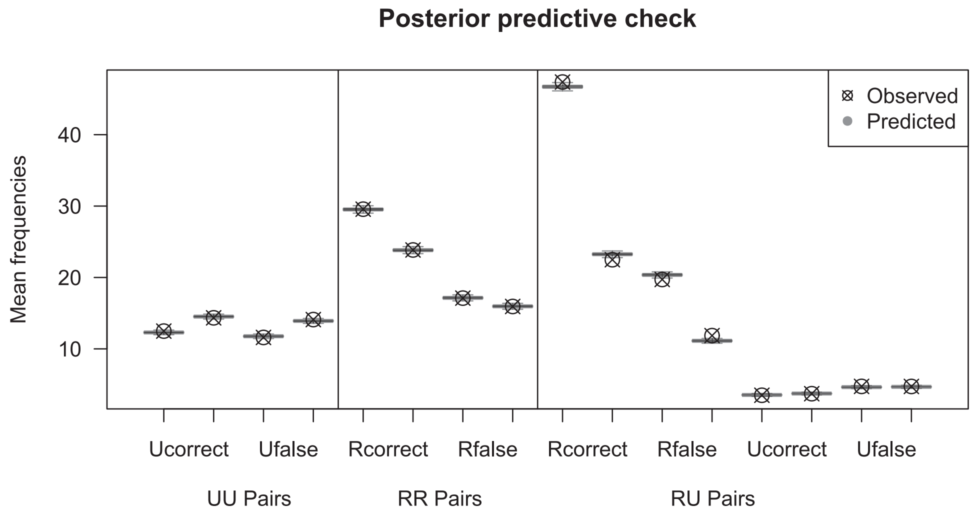

Figure 9. Model fit of the hierarchical RT-extended r-model to the 8 × 2 data categories based on the 19 studies in Table 3. For each response category of the r-model, the observed and predicted frequencies of the "fast" category are presented first, and those of the "slow" category second.

by a log-normal distribution, which was then used to obtain equally spaced quantiles as RT boundaries (e.g., the 25%, 50%, 75% quantiles for four RT bins). As mentioned above, these principled boundaries yield comparable frequencies across participants and result in approximately equally likely RT bins. Importantly, it is not necessary that the log-normal approximation fits the data well, since it only serves as a standardization to define suitable RT boundaries.

The RT-extended MPT model was implemented in the same way as in the simulation above (i.e., with uniform priors for the core parameters and noninformative Dirichlet priors for the latency parameters). Figure 11 shows that the posterior predictive distributions (blue/gray box plots) matched the pattern of observed frequencies (black circles) well, thus indicating a satisfactory model fit. Moreover, the core parameter estimates were comparable to the overall-mean estimates of the hierarchical model, $\hat{r}$ = .669 [.692–.701], $\hat{a}$ = .653 [.651–656], and $\hat{b}$ = .637 [.634–640]. Even though the 95% credibility intervals had a tendency to be slightly too small (a well-known effect of aggregating nested data; Klauer, 2010), the point estimates did not indicate a severe bias in parameter estimation.

Figure 12 shows the estimated cumulative densities for the speed of noncompensatory R-responses as well as R-congruent and R-incongruent responses (i.e., posterior means and 95% credibility intervals). To obtain a meaningful scaling for the *x*-axis, we computed the means of the individual RT boundaries initially used for categorization. The results show that choices due to R-congruent knowledge were stochastically faster than choices due to other processes, as indicated by a clear order of the estimated cumulative density functions. The credibility intervals were sufficiently small to ensure that this difference was not due to chance.[6] Moreover, choices due to recognition only were estimated to be slightly faster than those due to R-incongruent knowledge (with a minor reversal for very fast decisions). Overall, the estimated decision times of integrating recognition with further knowledge closely matched the PCS prediction in the simulation but clearly differed from the prediction of the serial-heuristic account (cf. Figure 8).

Outcome-Based Strategy Classification

The two previous analyses assumed that the average performance of the two competing process models across individuals is of core interest. To address individual differences in decision-strategy preferences, we adapted the method of outcome-based strategy classification (Bröder & Schiffer, 2003; Glöckner, 2009; Hilbig & Moshagen, 2014), which classifies each participant independently as a user of one of a set of decision strategies. Since the competing strategies are precisely defined as statistical models, model-selection methods can be used on the individual level. Compared to the previous analyses, this approach has the advantage that each participant is analyzed and classified independently, which allows detecting qualitative differences in decision behavior (Gigerenzer & Goldstein, 2011).

Based on the RT-extended r-model with two RT bins, we classify participants as users of the RH or PCS. For this purpose, we adapted the model in Figure 5 and imposed order constraints on the latency parameters according to the predictions derived from PCS and the serial-heuristic account (Equation 6 and 5, respectively). Whereas the RH predicts noncompensatory use of recognition to be faster than any knowledge-based strategy ($L_{\text{R-only}} > L_{\text{R-con}}$ and $L_{\text{R-only}} > L_{\text{R-inc}}$), PCS predicts a latency order according to decreasing coherence of the available information ($L_{\text{R-con}} > L_{\text{R-only}} > L_{\text{R-inc}}$). Moreover, we included a second model for the serial-heuristic account called "RH (strong)" that additionally as-

[6] One can compute a Bayes factor as the ratio of posterior to prior probability in line with the hypothesized order constraint (Hoijtink, Klugkist, & Boelen, 2008). However, since the posterior distributions do not overlap much in the present case, this will result in a large, but also imprecisely estimated Bayes factor, thus not adding any substantive value to the conclusions.

Table 4
Posterior Mean Estimates for the Hierarchical, RT-Extended r-Model (Including 95% Credibility Intervals)

	19 Data sets in Table 3				10 Data sets in Table 5			
	Φ ($\hat{\mu}$)	[95% CI]	$\hat{\sigma}$	[95% CI]	Φ ($\hat{\mu}$)	[95% CI]	$\hat{\sigma}$	[95% CI]
Core parameters								
r	.74	[.73–.76]	0.69	[0.66–0.73]	.39	[.36–.43]	0.94	[0.88–1.01]
a	.70	[.69–.71]	0.24	[0.23–0.25]	.65	[.64–.66]	0.21	[0.19–0.23]
b	.63	[.63–.64]	0.18	[0.17–0.19]	.67	[.66–.68]	0.21	[0.19–0.23]
g	.52	[.51–.53]	0.15	[0.13–0.16]	.50	[.49–.51]	0.17	[0.14–0.19]
Latency parameters								
L_{guess}	.43	[.43–.44]	0.29	[0.27–0.31]	.43	[.41–.44]	0.23	[0.20–0.26]
$L_{\text{RR correct}}$	.55	[.54–.55]	0.22	[0.20–0.24]	.57	[.56–.58]	0.12	[0.10–0.14]
$L_{\text{RR false}}$	.50	[.49–.51]	0.25	[0.23–0.27]	.48	[.47–.49]	0.16	[0.13–0.20]
$L_{\text{R-congruent}}$	.97	[.95–.99]	0.91	[0.77–1.08]	.77	[.73–.81]	0.32	[0.25–0.39]
$L_{\text{R-only}}$	.59	[.58–.60]	0.38	[0.35–0.41]	.54	[.50–.57]	0.28	[0.22–0.33]
$L_{\text{R-incongruent}}$	.44	[.42–.46]	0.37	[0.33–0.40]	.45	[.44–.47]	0.20	[0.17–0.24]

Note. Estimated group means Φ ($\hat{\mu}$) are reported on the probability scale, whereas group SDs $\hat{\sigma}$ are reported on the latent probit-scale (and can thus be larger than one).

sumes that R-congruent choices are faster than those due to R-incongruent knowledge ($L_{\text{R-only}} > L_{\text{R-con}} > L_{\text{R-inc}}$). This additional prediction only follows for some decision strategies people might use instead of the RH. For instance, R-incongruent choices were slower than R-congruent choices in our simulation using the TALL model (cf. Figure 8), but different patterns might be predicted by other strategies. However, since the additional assumption of slower R-incongruent choices is psychologically plausible and in line with at least one strategy (i.e., TALL), we included the "RH (strong)" model in the comparison, even though it gives the serial-heuristic account an advantage because it is instantiated by two model versions. Besides these three substantive models, we included a baseline model that does not pose any order constraints on the latency parameters $L_{\text{R-only}}$, $L_{\text{R-con}}$, and $L_{\text{R-inc}}$ and can therefore account for data not predicted by any of the two process models (Hilbig & Moshagen, 2014).

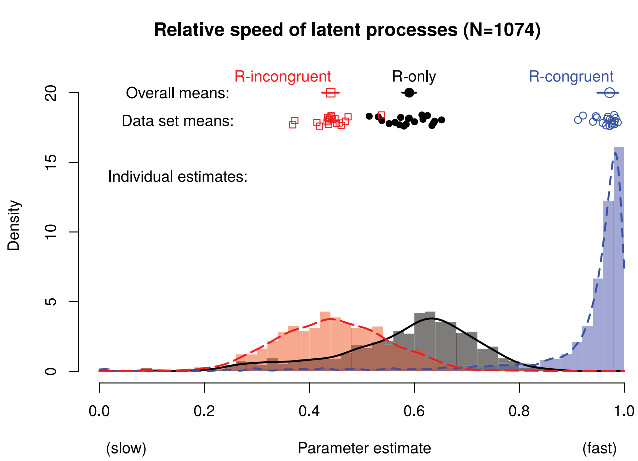

Figure 10. Estimated relative speed of latent processes based on the RT-extended r-model across the 19 data sets in Table 3. Density functions were fitted to the individual estimates to enhance the visual distinction between the three histograms. See the online article for the color version of this figure.

Since the three competing models have the same number of free parameters, we relied on the Fisher information approximation (FIA) for model selection (Rissanen, 1996; Wu, Myung, & Batchelder, 2010). FIA is based on the minimum description length principle (Grünwald, 2007) and trades off goodness-of-fit against complexity of a model (Myung, 2000). In contrast to information criteria such as AIC or BIC, FIA takes the functional complexity of a model into account, which provides a crucial advantage for outcome-based classification of order-constrained strategies (Heck, Hilbig, & Moshagen, 2017; Hilbig & Moshagen, 2014). In the present case, the identical number of parameters per model implies that FIA can be applied even in small samples (which is not necessarily the case otherwise; Heck, Moshagen, & Erdfelder, 2014).

When applying this method, 87.5% of the 1,074 participants were classified as PCS users, compared to 6.8% for RH and 4.5% for RH (strong), whereas the baseline model was only selected for 1.2% of the participants. Figure 13 shows that the proportion of PCS users was high across the 19 data sets listed in Table 3. The lowest percentage of 65.8% PCS users was observed in a study by Michalkiewicz, Minich, and Erdfelder (2017), in which recognition and knowledge validities regarding the length of rivers were rather low ($\hat{a}$ = .58, $\hat{b}$ = .58), which might have decreased the coherence effect predicted by PCS. In sum, however, these results clearly show that PCS was supported for a majority of participants across data sets.

Replication Based on Data From Other Labs

All data sets reanalyzed above (cf. Table 3) were collected by researchers associated with the University of Mannheim. To replicate and test the robustness of our results, we also reanalyzed data by other researchers and labs. For this purpose, we contacted authors of articles with studies on the RH that met the necessary conditions for reanalysis with the RT-extended r-model. As mentioned in the derivation of the model, these criteria require that (a) both recognition and knowledge have a sufficiently high validity, (b) neither cue knowledge nor recognition was manipulated, and (c) choice RTs were collected and not manipulated via time

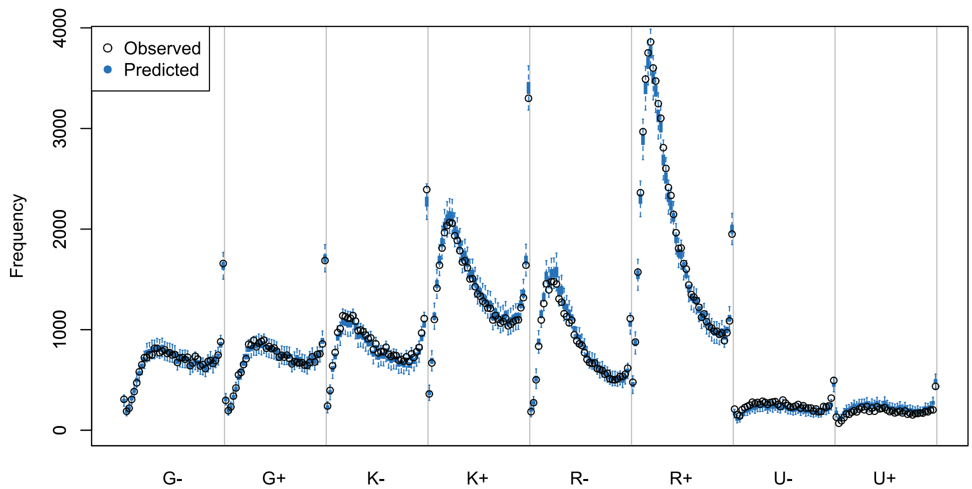

Figure 11. Model fit of the RT-extended r-model with 40 RT bins to the 40 × 8 data categories based on the 19 studies in Table 3. Within each type of response, frequencies from fast to slow RT bins are shown from left to right. See the online article for the color version of this figure.

pressure. Table 5 lists the 10 reanalyzed studies, overall including 121,324 decisions by 671 participants.[7]

We reanalyzed the studies in Table 5 independently of the data from Mannheim (see Table 3) for two reasons. First, the data sets by other labs were obtained after knowing the results of the first reanalysis, which renders the second one an independent replication that allows testing the robustness of the findings by applying exactly the same methods to novel data. Second, the two sets of studies differed in the total number of trials (274,296 vs. 121,324 trials for the data collected by researchers associated with Mannheim and other researchers, respectively). Hence, any results from a pooled reanalysis would have likely been dominated by the first set of studies, thereby covering possible inconsistencies with the new, independent data.

We performed the same three statistical analyses as above. First, based on two RT bins, we fitted the hierarchical MPT version of the RT-extended r-model, which accounted for the observed frequencies well (the supplementary material provides a goodness-of-fit plot closely resembling Figure 9). The estimated latency parameters are shown in the right part of Table 4 and plotted in

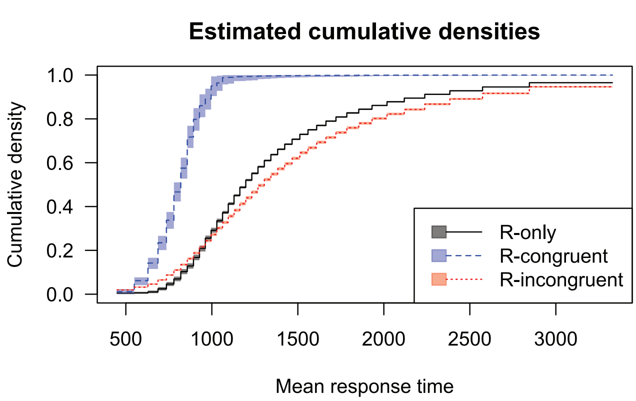

Figure 12. Estimated latent RT distributions for the 19 studies in Table 3. See the online article for the color version of this figure.

Figure 14. As in the reanalysis above, R-congruent choices were faster than choices due to recognition only, which in turn were slightly faster than R-incongruent choices. This result again supports PCS, which predicts that memory-based decisions become slower as the coherence of further knowledge with recognition decreases.

Next, we aggregated frequencies across participants to fit the RT-extended r-model with 30 RT bins.[8] The estimates for the core parameters were comparable to the group-level estimates in the hierarchical model (but with smaller, possibly underestimated 95% credibility intervals), $\hat{r}$ = .408 [.400–.416], $\hat{a}$ = .646 [.642–.650], and $\hat{b}$ = .659 [.655–.662]. Figure 15 shows that R-congruent choices were again stochastically faster than R-only and R-incongruent choices (i.e., the cumulative density was larger across the full RT range). However, the relative speed of the latter two processes was estimated to be very similar, with a slight tendency of R-incongruent choices being faster than choices due to recognition only, a pattern neither predicted by the RH nor by PCS. Given that aggregated analyses tend to underestimate credibility intervals, we doubt the reliability of this result. However, the integration of R-congruent knowledge clearly lead to the fastest choices, thereby supporting the PCS account.

Finally, the results of the outcome-based strategy classification on the individual level are shown in Figure 16. Across data sets, a majority of 83.0% of the 671 participants was best described by PCS, compared to a minority of 15% best described by the serial-heuristic account (4.6% and 10.9% for RH and RH-strong, respectively), and only 1.5% were classified by the baseline model.

[7] Unfortunately, despite making their raw data available, we could not include Study 3 by Hertwig, Herzog, Schooler, and Reimer (2008) because of the small number of individual trials for each of the three domains (52 trials per participant, whereas the other studies in Table 5 had between 120 and 276 trials). The three domains differed considerably with respect to choice frequencies and RTs, thereby preventing an aggregation across domains.

[8] We only used 30 instead of 40 RT bins in the second reanalysis due to the smaller number of observations available.

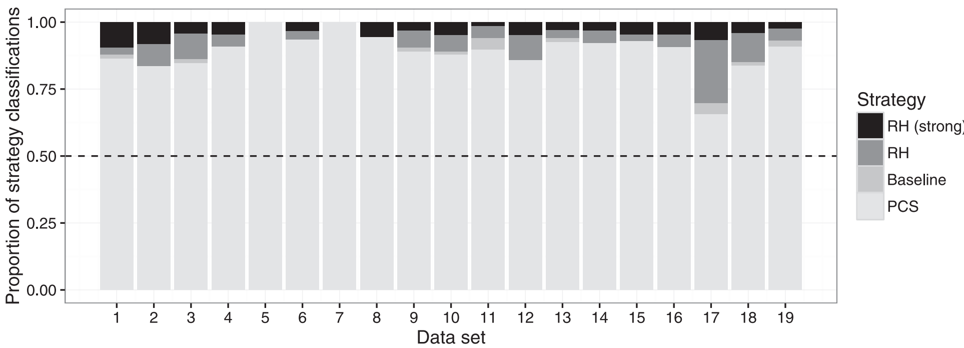

Figure 13. Estimated proportion of PCS and RH users for the 19 data sets in Table 3.

Similar as above, these proportions of strategy users were robust across data sets.

Overall, the reanalysis of an independent set of studies by researchers from other nine labs provided converging evidence in favor of the PCS account. All analyses revealed that memory-based decisions that take additional, R-congruent knowledge into account were faster than those due to recognition only, in contrast to the prediction of the serial RH. This result holds not just on average, but also for the vast majority of participants investigated.

Discussion

Since its proposal, the RH theory (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002) stimulated much research investigating whether and how people rely on recognition in memory-based decisions. Whereas researchers agree that recognition is used as a valid and influential metacognitive cue in the decision process (e.g., Gigerenzer & Goldstein, 2011; Glöckner & Bröder, 2014; Hilbig, Erdfelder, & Pohl, 2010; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011), a controversy has evolved concerning the noncompensatory nature of recognition use as stated by the original RH theory (Hilbig & Richter, 2011; Marewski et al., 2010, 2011a).

To test whether recognition is used in a noncompensatory way, we derived opposing predictions from two major theoretical accounts, both implemented as precise process models. On the one hand, the serial heuristic account assumes that in each trial, either the fast RH or a slower, knowledge-based strategy is used (Goldstein & Gigerenzer, 2002, 2011). Based on the formal definition of this model, we derived the strong prediction of stochastic dominance stating that RH responses are stochastically faster than responses due to the use of further knowledge (Equation 2). On the other hand, the information-integration account assumes that recognition is merely one of many cues that are integrated jointly (e.g., Newell & Shanks, 2004). We implemented this account as a process model based on the PCS theory, which assumes that choices and RTs depend on the coherence of the available information (Glöckner & Betsch, 2008; Glöckner et al., 2014). In contrast to earlier PCS accounts of memory-based decisions (e.g., Glöckner & Bröder, 2011, 2014), we explicitly modeled nonexistent cue knowledge for the unrecognized option and partial cue knowledge for the recognized option by deleting links between cue and option nodes in the PCS network. This compensatory account predicts that choices due to recognition-congruent knowledge are actually faster than choices due to recognition only (even though more information is integrated; Betsch & Glöckner, 2010).

Importantly, the opposing predictions of both process models refer to latent, unobservable RT distributions and can thus not be tested directly. To allow for an empirical test between the two

Table 5
Data Sets Included in the Second Reanalysis

Data Set	Citation	Material	*N*	Responses
1	Hochman, Ayal, and Glöckner (2010) (Trials without additional cue)	City size	24	3,354
2	Horn, Ruggeri, and Pachur (2016) (Cities)*	City size	56	8,568
3	Horn, Ruggeri, and Pachur (2016) (Diseases, old adolescents)†	Disease frequency	21	3,213
4	Pachur and Hertwig (2006) Study 1	Disease frequency	39	10,764
5	Pachur, Mata, and Schooler (2009) Study 1 (Cities)*	City size	39	10,752
6	Pachur, Mata, and Schooler (2009) Study 1 (Diseases)*	Disease frequency	39	10,742
7	Pachur, Mata, and Schooler (2009) Study 2 (Old adults)	Disease frequency	59	16,284
8	Pachur, Mata, and Schooler (2009) Study 2 (Young adults)	Disease frequency	60	16,560
9	Richter and Späth (2006) Exp. 1#	Animal population	42	6,047
10	Schweickart and Brown (2014) Study 2	Country GDP	292	35,040

Note. *N* refers to the number of participants after exclusion of those who recognized either all or none of the items.
† We did not include data of children and young adolescents in the "Disease" condition, since the knowledge validity was at chance level (leading to an empirically nonidentifiable model). # Since individual recognition judgments were not available, animals were categorized as "unrecognized" and "recognized" for all participants identically, which is warranted by unambiguous average recognition rates of .97 and .05, respectively. * Includes data by different age groups.

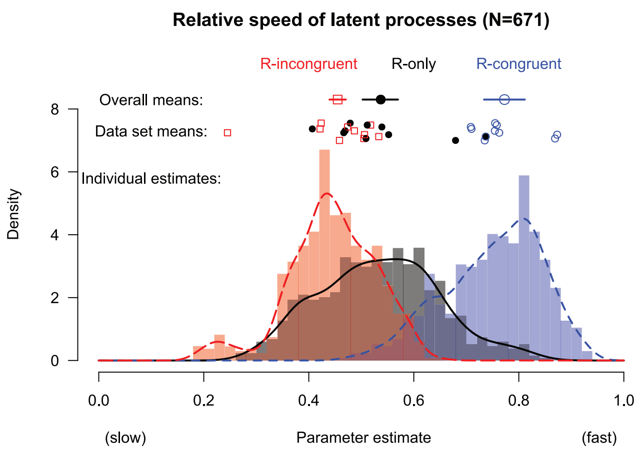

Figure 14. Estimated relative speed of latent processes based on the RT-extended r-model based on the 10 data sets in Table 5. Density functions were fitted to the individual estimates to enhance the visual distinction between the three histograms. See the online article for the color version of this figure.

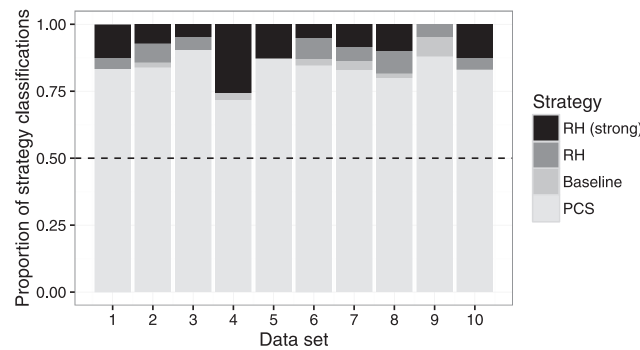

Figure 16. Estimated proportion of PCS and RH users for the 10 data sets in Table 5.

competing process models, we developed an encompassing measurement model. This RT-extended version of the r-model (Hilbig, Erdfelder, & Pohl, 2010) can be fit to data collected in the standard RH paradigm (which includes only a decision and a recognition phase) and does not require an experimental manipulation of cue or recognition knowledge. Without parametric assumptions on the latent RT distributions (Heck & Erdfelder, 2016), this measurement model allows to estimate the relative decision speed when integrating R-congruent further knowledge and when relying on recognition only.

Following the methodology of previous works regarding the RH (e.g., Gigerenzer & Goldstein, 1996; Marewski & Mehlhorn, 2011; Schooler & Hertwig, 2005), simulations showed that the encompassing measurement model indeed allows to differentiate between the two process models. Next, we used the RT-extended r-model to reanalyze 29 data sets including around 400,000 memory-based decisions. The Bayesian hierarchical version with two RT bins explicitly modeled the dependencies of parameters across participants and data sets, showing that choices due to R-congruent knowledge were substantially faster than choices due to recognition only. This main result was replicated in a more fine-grained test of stochastic dominance using 40 RT bins under the more restrictive assumption of independently and identically distributed observations across participants. It is well-known that an isolated analysis of aggregated data has severe limitations (e.g., Gigerenzer & Goldstein, 2011; Klauer, 2006). However, these problems are mitigated in the present context by the cross-validation with the hierarchical and individual analyses. Using the RT-extended r-model with individual data, an outcome-based strategy classification showed that choices and RTs of 85% of the participants were better described by PCS than by the RH. Importantly, the reanalysis of data by researchers from the University of Mannheim led to substantively identical conclusions as that of data by other labs.

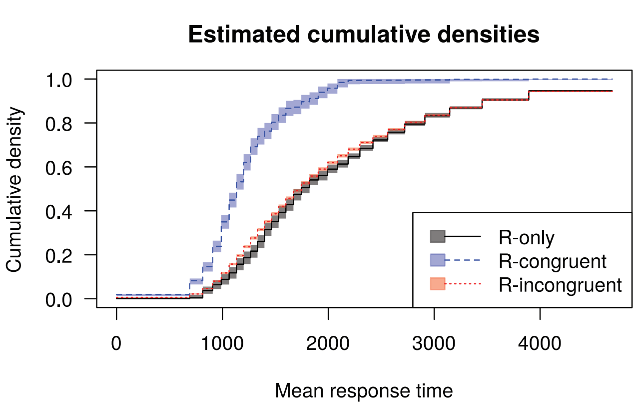

Figure 15. Estimated cumulative density functions of the latent RT distributions for the 10 data sets in Table 5. See the online article for the color version of this figure.

Overall, our analyses provide converging evidence independent of the source of the data (cf. Table 3 and 5) and independent of the different statistical versions of the RT-extended r-model used (hierarchical Bayesian, stochastic dominance on a finer RT scale, and model selection on the individual level). All analyses corroborated the compensatory information-integration account represented by the PCS model, which predicts that the coherence of the integrated knowledge determines decision speed of memory-based decisions. Note that our results thereby replicate previous studies that demonstrated coherence effects by experimentally manipulating recognition or cue knowledge (e.g., Glöckner & Bröder, 2011, 2014; Hochman et al., 2010).

From a methodological perspective, our approach outlines a novel approach of linking precise process models to empirically testable measurement models. Moreover, RTs are only one of many process-tracing measures that allow more powerful tests of decision theories compared to relying on choice frequencies only. Other popular methods include eye-tracking, mouse-tracking, active information search, and thinking aloud (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011), as well as confidence ratings, neurophysiological data, and physiological arousal (e.g., Hochman et al., 2010). In principle, some of these variables could be included in our encompassing measurement approach similar to RTs. For instance, PCS predicts that confidence in a choice increases monotonically with the coherence of the available information, operationalized by a larger difference in the final activation levels of the option nodes (e.g., Glöckner & Bröder, 2011; Hochman et al., 2010). Assuming that each of the cognitive states of the r-model results in a distinct distribution of confidence ratings, a mixture emerges similar as in the case of RTs (cf. Figure 5). In principle, such a "confidence-extended r-model" might allow to obtain process-pure confidence estimates conditional on the latent cognitive states. However, future work is required to test whether other process-tracing measures can actually be modeled within the proposed framework.

The Recognition Process

In all theoretical derivations and empirical tests so far, we have focused on information integration and neglected the recognition process itself. In the following, we discuss the generalizability and implications of our results in light of a closer look at the recognition process (Erdfelder, Küpper-Tetzel, & Mattern, 2011; Pleskac, 2007).

Recognition Speed and Further Knowledge

Often, studies investigating memory-based decisions have also examined the time required to recognize items as a proxy for processing fluency (e.g., Hilbig & Pohl, 2009; Schooler & Hertwig, 2005). The fluency heuristic states that the faster recognized item is chosen in trials involving two recognized objects (RR pairs; Hertwig et al., 2008; Schooler & Hertwig, 2005). Despite empirical evidence against such an account (Hilbig, Erdfelder, & Pohl, 2011; Pohl, Erdfelder, Michalkiewicz, Castela, & Hilbig, 2016), the recognition speed could still confound any analysis of choice RTs. Specifically, memory-based decisions in trials with further knowledge might not be faster because of the increased coherence of further information with recognition (as stated by PCS) but rather due to faster recognition itself (Gigerenzer & Goldstein, 2011). Indeed, it is very plausible to assume that objects for which a lot of knowledge is available are also retrieved faster.

Concerning the parallel information-integration account, the assumption that items with additional knowledge are recognized faster strengthens the prediction that R-congruent choices are faster than R-only choices. However, it also weakens the prediction that R-only choices are faster than R-incongruent choices. Importantly, this matches our results in Figure 12 and 15 that the discrepancies are larger between the estimated latent RT distributions of R-congruent and R-only choices than between those of R-only and R-incongruent choices.

According to the serial heuristic account, recognition speed determines the time of entering the evaluation or strategy-selection stage (see Figure 1).[9] To derive decision-time predictions, we subsume recognition and strategy selection under a single preceding stage. Then, as derived above, the prediction of stochastically faster RH responses (Equation 3) holds for a variety of assumptions regarding this preceding stage (for instance, if recognition-speed is correlated with the speed of executing the selected strategy). However, as an important exception, stochastic dominance does not follow anymore if the preceding stage requires more time for selecting the RH than a knowledge-based strategy. If this assumption is made, the serial-heuristic account would be compatible with our results. However, such an assumption would imply that knowledge-based strategies are used for faster-recognized items (i.e., items with further knowledge) whereas the RH is used for slower-recognized items (i.e., items without or with less further knowledge). Put differently, this implies that the RH is only used in cases in which less or no further knowledge is available, that is, in cases in which decisions are noncompensatory by definition regardless of the theoretical framework. This reasoning narrows the scope of the RH to a trivial case and renders the information-integration account the better theory due to its larger scope and explanatory power (Glöckner & Betsch, 2011).

Taken together, these theoretical arguments show that the RH prediction of fast R-only choices cannot be reversed by considering recognition speed without giving up a core feature of the theory. Moreover, one can empirically test whether recognition speed can account for our results. In Appendix B, we used the approach by Hilbig and Pohl (2009) in which RT residuals from a regression of decision times on recognition times are analyzed (i.e., choice RTs are predicted in each trial by the recognition RTs of the two presented items). It follows that the RT residuals are statistically controlled for recognition speed and can be used as the input for the RT-extended r-model. The analyses in Appendix B clearly show that the R-congruent RT residuals are still faster than R-only responses. Hence, our results are robust regarding possible confounds by recognition speed.

Recognition as a Binary Variable

When proposing the RH, Gigerenzer and Goldstein (1996) conceptualized recognition as an all-or-none, binary variable. Recently, this simplifying assumption has been questioned (Gigerenzer & Goldstein, 2011; Hilbig & Pohl, 2009) and recognition models from the memory literature have been combined with the RH (Erdfelder et al., 2011; Pleskac, 2007). On the one hand, Pleskac (2007) adapted signal detection theory, assuming a continuous memory-strength signal. On the other hand, based on the two-high threshold model (Bröder & Schütz, 2009), Erdfelder et al. (2011) proposed the memory-state heuristic (MSH) assuming that the discrepancy between three latent recognition states determines choices (i.e., recognition certainty, uncertainty, and rejection certainty; see also Castela & Erdfelder, 2017a; Castela, Kellen, Erdfelder, & Hilbig, 2014). For instance, the MSH predicts that the recognized object is more likely to be chosen if it is recognized with certainty compared to being in an uncertainty state, even if the recognition judgment is positive in either case.

Similarly as the RH, the parallel information-integration account can be modified to take more fine-grained recognition information into account. Since the PCS network is set up in each trial anew depending on the available information, it is natural to assume that the subjective recognition validity depends on the strength of the recognition signal (cf. Figure 2B). For instance, recognition validity might subjectively be larger in trials involving a certainly recognized and a certainly rejected item compared to trials involving an uncertain item. Similarly, subjective recognition validity can be modeled as a function of a continuous memory strength signal based on signal detection theory.

[9] Note that in each trial, the second choice option also needs to be judged as unrecognized, so the effect of memory strength on the time of entering the evaluation stage will be diminished.

Across trials, the assumption of varying subjective recognition validities results in a mixture of PCS predictions. In other words, the weight of the recognition cue (i.e., the link w_R in Figure 2B) is assumed to be smaller in some trials but larger in others. Crucially, however, the prediction that R-congruent choices are faster than R-only choices still holds within each of these cases for a given subjective validity. If we assume that the recognition validities α are identically distributed for R-only and R-congruent RT distributions according to some density $\pi(\alpha)$ (which depends on the distribution of memory-strength signals), the PCS prediction in Equation 6 generalizes to the emerging mixture distributions of RTs across trials. When marginalizing over the distribution of subjective validities π (α), the prediction of stochastic dominance of R-congruent choices still holds:

$$\int_0^1 F_{\text{R-only}}(t \mid \alpha)\pi(\alpha)\mathrm{d}\alpha < \int_0^1 F_{\text{R-congruent}}(t \mid \alpha)\pi(\alpha)\mathrm{d}\alpha \quad \text{for all } t \in \mathbb{R}^+. \quad (11)$$

Note that in case of discrete memory states as assumed by the MSH, $\pi(\alpha)$ becomes a point measure that assigns all probability mass to a finite number of possible values for α, thereby reducing the integral to a finite sum. Irrespective whether the recognition signal is assumed to be discrete or continuous, the predictions of PCS with varying subjective recognition probabilities across trials are hence perfectly in line with the observed ordering of RT distributions.

Besides these theoretical considerations, the effect of memory states on RTs can be empirically tested by adapting the r*-model that was developed by Castela et al. (2014) to test the MSH. This extended version of the r-model additionally differentiates between items that are subjectively judged as merely recognized (mR) or recognized with further knowledge (R+). According to the MSH, R+ items are more likely to be in the recognition certainty state. Based on this indicator, one can empirically test whether R-congruent responses are faster irrespective of the underlying memory states, that is, for both mR/U (lower discrepancy in memory states) and R+/U pairs (higher discrepancy). In Appendix C we show that this is indeed the case: even though recognition is more influential in R+/U pairs as indicated by a larger estimate for *r,* R-congruent responses are faster than R-only responses in both conditions.

Reinterpretation of the r-Model and Related Models

Given that our results clearly favor information-integration theories, it appears unwarranted to interpret the r-model as measuring the probability of using a fast and frugal heuristic, namely, the RH (Hilbig, Erdfelder, & Pohl, 2010). By definition, the *r* parameter represents the probability of relying solely on the recognition cue such that this cue determines the judgment, irrespective of further knowledge. However, although RH use implies reliance on recognition only, the reverse does not necessarily hold. Noncompensatory reliance on the recognition cue can occur as a special case of a single, compensatory information-integration process, simply by giving the recognition cue a very large weight (cf. Hilbig, Scholl, & Pohl, 2010). Thus, although the definition of the *r* parameter is unique, its theoretical interpretation is not. This difference in theoretical interpretations becomes important when considering the effect of experimental manipulations on *r*: Given our results, instead of assuming an adaptive qualitative shift in decision strategies (e.g., "temporarily suspending the recognition heuristic"; Pachur & Hertwig, 2006, p. 992), an observed decrease in *r* rather appears to indicate that recognition determines decisions less often in a noncompensatory way because the relative weight of cues based on further knowledge increases—or the relative weight of the recognition cue decreases—during information integration (Glöckner et al., 2014; Hilbig, Scholl, & Pohl, 2010).

This reinterpretation of the *r* parameter might change some but not all conclusions of previous research that relied on the r-model. On the one hand, validation studies that tested whether recognition is used more often in a noncompensatory way under suitable experimental manipulations remain valid, since the theoretical assumption of a distinct heuristic is not required. In contrast, interpretations in terms of a shift in the underlying strategies need to be adjusted. For instance, the increase in *r* due to time pressure (Hilbig, Erdfelder, & Pohl, 2012) can be explained by further knowledge being available too late (this is in line with the decreased knowledge validity in the time-pressure condition). Accordingly, further knowledge can only overrule recognition if knowledge cues enter the integration process sufficiently fast. A different example concerns individual differences in *r,* which were found to be stable across several weeks and domains (Michalkiewicz & Erdfelder, 2016). Instead of assuming a trait-like preference for or against using the RH, these results might simply be due to stable individual tendencies in the relative weight of the recognition cue in the global information integration process. Similarly, younger and older participants might assign different subjective weights to recognition that results in different estimates of *r* for different age groups (Horn et al., 2015).

Overall, these examples show the far-reaching consequences of interpreting *r* either as a discrete switch in using the RH versus a continuous shift of the relative influence of the recognition cue (Hilbig, Scholl, & Pohl, 2010). Importantly, only the RT-extended and not the standard r-model allows for differentiation between both interpretations of the *r* parameter. Given the strong evidence for the information-integration account, further applications of the (standard) r-model should interpret *r* as the subjective weight assigned to recognition such that it effectively becomes a noncompensatory cue. Similar conclusions hold with respect to extensions of the r-model based on the memory-state heuristic, that is, the r*-model (Castela et al., 2014) or the latent-states MSH model (Castela & Erdfelder, 2017a).

Conclusion

Given that memory-based decisions are very common in our everyday life, it is important to know whether recognition is used in a noncompensatory fashion as stated by the RH. To answer this question, we developed an encompassing measurement model that allows to test precise RT predictions of two competing process models. A reanalysis of 29 data sets including approximately 400,000 memory-based decisions clearly showed that responses due to R-congruent knowledge were actually faster than responses due to recognition only. These results support the parallel constraint satisfaction account according to which different sources of information are intuitively integrated in a compensatory way. Accordingly, the coherence and not the amount of further knowledge with recognition determines the speed of memory-based decisions.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bernardo, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences, 4,* 111–122.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry, 21,* 279–294. http://dx.doi.org/10.1080/1047840X.2010.517737

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14,* 895–900. http://dx.doi.org/10.3758/BF03194118

Bröder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multiattribute decision making. *Journal of Behavioral Decision Making, 16,* 193–213. http://dx.doi.org/10.1002/bdm.442

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 587–606. http://dx.doi.org/10.1037/a0015279

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459. http://dx.doi.org/10.1037/0033-295X.100.3.432

Castela, M., & Erdfelder, E. (2017a). The memory state heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 205–225. http://dx.doi.org/10.1037/xlm0000299

Castela, M., & Erdfelder, E. (2017b). Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences. Manuscript submitted for publication.

Castela, M., Kellen, D., Erdfelder, E., & Hilbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review, 21,* 1131–1138. http://dx.doi.org/10.3758/s13423-014-0587-4

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision Making, 5,* 216–229.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106,* 180–209. http://dx.doi.org/10.1037/0033-295X.106.1.180

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

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making, 6,* 7–22.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–143. http://dx.doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62,* 451–482. http://dx.doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making, 6,* 100–121.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological Review, 115,* 230–239. http://dx.doi.org/10.1037/0033-295X.115.1.230

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox*. Bounded rationality: The adaptive toolbox. Cambridge, MA: The MIT Press.

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making, 4,* 186–199.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making, 3,* 215–228.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making, 6,* 711–721.

Glöckner, A., & Betsch, T. (2012). Decisions beyond boundaries: When more information is processed faster than less. *Acta Psychologica, 139,* 532–542. http://dx.doi.org/10.1016/j.actpsy.2012.01.009

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making, 6,* 23–42.

Glöckner, A., & Bröder, A. (2014). Cognitive integration of recognition information and additional cues in memory-based decisions. *Judgment and Decision Making, 9,* 35–50.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition, 133,* 641–666. http://dx.doi.org/10.1016/j.cognition.2014.08.017

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & A. B. C. Research Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York, NY: New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90. http://dx.doi.org/10.1037/0033-295X.109.1.75

Goldstein, D. G., & Gigerenzer, G. (2011). The beauty of simple models: Themes in recognition heuristic research. *Judgment and Decision Making, 6,* 392–395.

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology, 54,* 454–463. http://dx.doi.org/10.1016/j.jmp.2010.06.005

Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*.

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review, 23,* 1440–1465. http://dx.doi.org/10.3758/s13423-016-1025-6

Heck, D. W., Hilbig, B. E., & Moshagen, M. (2017). *From information processing to decisions: Formalizing and comparing probabilistic choice models*. Manuscript submitted for publication.

Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology, 60,* 29–34. http://dx.doi.org/10.1016/j.jmp.2014.06.002

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1191–1206. http://dx.doi.org/10.1037/a0013025

Hilbig, B. E. (2010a). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making, 5,* 272–284.

Hilbig, B. E. (2010b). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review, 17,* 923–930. http://dx.doi.org/10.3758/PBR.17.6.923

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

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 827–839. http://dx.doi.org/10.1037/a0022638

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

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

Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review, 21,* 1431–1443. http://dx.doi.org/10.3758/s13423-014-0643-0

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

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

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

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

Hochman, G., Ayal, S., & Glöckner, A. (2010). Physiological arousal in processing recognition information: Ignoring or integrating cognitive cues? *Judgment and Decision Making, 5,* 285–299.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128,* 3–31. http://dx.doi.org/10.1037/0096-3445.128.1.3

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

Jekel, M., Glöckner, A., Bröder, A., & Maydych, V. (2014). Approximating rationality under incomplete information: Adaptive inferences for missing cue values based on cue-discrimination. *Judgment and Decision Making, 9,* 129–147.

Jekel, M., Glöckner, A., Fiedler, S., & Bröder, A. (2012). 24). The rationality of different kinds of intuitive decision processes. *Synthese, 189,* 147–160. http://dx.doi.org/10.1007/s11229-012-0126-7

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika, 71,* 7–31. 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

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review, 11,* 343–352. http://dx.doi.org/10.3758/BF03196581

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2010). From recognition to decisions: Extending and testing recognition-based models for multialternative inference. *Psychonomic Bulletin & Review, 17,* 287–309. http://dx.doi.org/10.3758/PBR.17.3.287

Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making, 6,* 439–519.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2010). Recognition-based judgments and decisions: Introduction to the special issue (Vol. 1). *Judgment and Decision Making, 5,* 207–215.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011a). Recognition-based judgments and decisions: Introduction to the special issue (Vol. 2). *Judgment and Decision Making, 6,* 1–6.

Marewski, J. N., Pohl, R. F., & Vitouch, O. (2011b). Recognition-based judgments and decisions: What we have learned (so far). *Judgment and Decision Making, 6,* 359–380.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118,* 393–437. http://dx.doi.org/10.1037/a0024143

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

Michalkiewicz, M., Arden, K., & Erdfelder, E. (2017). Do smarter people make better decisions? The influence of intelligence on adaptive use of the recognition heuristic. Manuscript submitted for publication.

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

Michalkiewicz, M., Minich, B., & Erdfelder, E. (2017). Explaining individual differences in fast-and-frugal decision making: Need for cognition, faith in intuition, and the recognition heuristic. Manuscript submitted for publication.

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

Myung, J. I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. http://dx.doi.org/10.1006/jmps.1999.1283

Newell, B. R., & Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. *Judgment and Decision Making, 3,* 195–204.

Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 923–935. http://dx.doi.org/10.1037/0278-7393.30.4.923

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90,* B1–B9. http://dx.doi.org/10.1016/S0010-0277(03)00141-0

Pachur, T. (2011). The limited value of precise tests of the recognition heuristic. *Judgment and Decision Making, 6,* 413–422.

Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making, 21,* 183–210. http://dx.doi.org/10.1002/bdm.581

Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 983–1002. http://dx.doi.org/10.1037/0278-7393.32.5.983

Pachur, T., Mata, R., & Schooler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging, 24,* 901–915. http://dx.doi.org/10.1037/a0017211

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Cognitive Science, 2,* 147. http://dx.doi.org/10.3389/fpsyg.2011.00147

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552. http://dx.doi.org/10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY: Cambridge University Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. http://dx.doi.org/10.1037/0033-295x.109.3.472

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review, 14,* 379–391. http://dx.doi.org/10.3758/BF03194081

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, p. 125). Vienna, Austria.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19,* 251–271. http://dx.doi.org/10.1002/bdm.522

Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making, 6,* 423–438.

Pohl, R. F., Erdfelder, E., Michalkiewicz, M., Castela, M., & Hilbig, B. E. (2016). The limited use of the fluency heuristic: Converging evidence across different procedures. *Memory & Cognition, 44,* 1114–1126. http://dx.doi.org/10.3758/s13421-016-0622-y

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: Wiley.

Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 150–162. http://dx.doi.org/10.1037/0278-7393.32.1.150

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236. http://dx.doi.org/10.1037/0096-3445.135.2.207

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42,* 40–47. http://dx.doi.org/10.1109/18.481776

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–628. http://dx.doi.org/10.1037/0033-295X.112.3.610

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making, 6,* 733–739.

Schweickart, O., & Brown, N. R. (2014). Magnitude comparison extended: How lack of knowledge informs comparative judgments under uncertainty. *Journal of Experimental Psychology: General, 143,* 273–294. http://dx.doi.org/10.1037/a0031451

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

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica, 146,* 84–96. http://dx.doi.org/10.1016/j.actpsy.2013.12.007

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12,* 435–467. http://dx.doi.org/10.1017/S0140525X00057046

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567. http://dx.doi.org/10.1037/0033-2909.108.3.551

Wu, H., Myung, J. I., & Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review, 17,* 275–286. http://dx.doi.org/10.3758/PBR.17.3.275

Yantis, S., Meyer, D. E., & Smith, J. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin, 110,* 350–374. http://dx.doi.org/10.1037/0033-2909.110.2.350

(*Appendices follow*)

Appendix A

Identifiability of the RT-Extended r-Model

Heck and Erdfelder (2016) proved that the substantive core parameters of an RT-extended MPT model are identifiable if the underlying basic MPT model is identifiable. For the RT-extended r-model, this implies the identifiability of *r, a, b,* and *g*. Moreover, the identifiability of the latency parameters is guaranteed if the following matrix has full rank (for details, see Heck & Erdfelder, 2016):

$$P(\theta) = \begin{array}{c} \\ \\ \end{array} \begin{array}{cccccc} L_{\text{guess}} & L_{\text{RR correct}} & L_{\text{RR false}} & L_{\text{R-only}} & L_{\text{R-congruent}} & L_{\text{R-incongruent}} \end{array}$$

$$P(\theta) = \begin{pmatrix} g & 0 & 0 & 0 & 0 & 0 \\ 1-g & 0 & 0 & 0 & 0 & 0 \\ 0 & b & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-b & 0 & 0 & 0 \\ 0 & 0 & 0 & ra & (1-r)ba & 0 \\ 0 & 0 & 0 & r(1-a) & (1-r)(1-b)(1-a) & 0 \\ 0 & 0 & 0 & 0 & 0 & (1-r)b(1-a) \\ 0 & 0 & 0 & 0 & 0 & (1-r)(1-b)a \end{pmatrix} \begin{array}{c} G+ \\ G- \\ K+ \\ K- \\ R+ \\ R- \\ U+ \\ U- \end{array}$$

For parameters $\theta = (r, a, b, g) \in (0, 1)^4$, $P(\theta)$ has full rank if and only if $b \neq 0.5$. Note that this special case is not of substantive interest, since it refers to scenarios where further knowledge is invalid and noninformative. Under such conditions, it is intuitively clear that we cannot learn how further knowledge is integrated with the recognition cue. Hence, for scenarios of interest (i.e., $a > .5$, $b > .5$), the RT-extended r-model is identifiable.

Appendix B

Analyzing RT Residuals

In the main text, we used the observed RTs of the decision phase to test the relative speed of integrating the recognition cue with further knowledge. However, this analysis might have been affected by the recognition speed itself, which often serves as a proxy for processing fluency (Hilbig et al., 2011; Schooler & Hertwig, 2005). In other words, the analysis of choice RTs might have confounded the speed of information integration with speed of recognizing the objects. Hence, the main result that R-congruent responses were estimated to be slower than R-only responses might be explained by those items being recognized faster.

To test this alternative explanation, we repeated our analysis using RT residuals, thereby accounting for recognition speed. Specifically, we followed the method used by Hilbig and Pohl (2009) and predicted the choice RTs in the decision phase by the RTs of the recognition phase for both objects (which are often used as a proxy for processing fluency; Schooler & Hertwig, 2005) separately per person. The RT residuals of this regression are then categorized and used as input for the RT-extended r-model as explained in the main text. The following analysis only differed with respect to the strategy of categorizing continuous RT residuals into discrete bins. Instead of a lognormal approximation, we used a normal distribution to approximate the RT residuals across all response categories per person and obtain individual boundaries for categorizing responses from "fast" to "slow." Specifically, we computed the mean and SD of all RT residuals per person to compute equally-spaced quantiles from a corresponding normal distribution. The reason for using a normal instead of a lognormal approximation lies in the nature of RT residuals, which assume both positive and negative values.

(*Appendices continue*)

Table B1
Posterior Mean Estimates for the Hierarchical, RT-Extended r-Model Based on RT Residuals (Including 95% Credibility Intervals)

	Φ ($\hat{\mu}$)	95% CI	$\hat{\sigma}$	95% CI
Core parameters				
r	.74	[.72–.75]	0.69	[0.66–0.73]
a	.67	[.66–.68]	0.24	[0.22–0.25]
b	.62	[.61–.62]	0.18	[0.17–0.19]
g	.52	[.51–.52]	0.15	[0.13–0.16]
Latency parameters				
L_g	.56	[.56–.57]	0.18	[0.16–0.19]
$L_{\text{RR,false}}$	.64	[.63–.65]	0.15	[0.13–0.17]
$L_{\text{RR,correct}}$	.61	[.60–.61]	0.17	[0.15–0.19]
$L_{\text{R-congruent}}$	.93	[.89–.96]	0.64	[0.52–0.77]
$L_{\text{R-only}}$	.71	[.70–.72]	0.22	[0.20–0.25]
$L_{\text{R-incongruent}}$	.57	[.56–.58]	0.27	[0.24–0.30]

Note. Estimated group means Φ ($\hat{\mu}$) are reported on the probability scale, whereas group SDs $\hat{\sigma}$ are reported on the latent probit-scale (and can thus be larger than one).

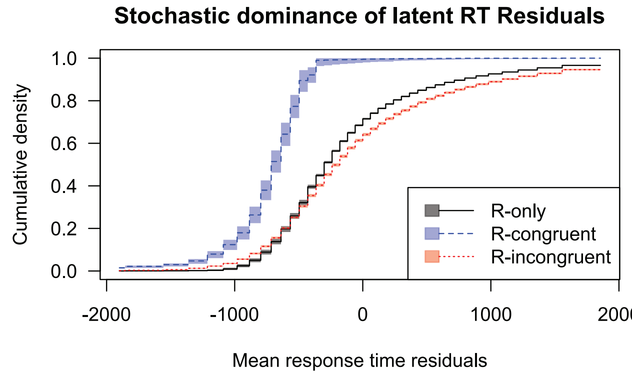

Figure B1. Estimated latent distributions of RT residuals. See the online article for the color version of this figure.

Using this approach, we reanalyzed a subset of 15 data sets from Mannheim (cf. Table 3) including 259,470 individual choices for which recognition RTs were available. The results of this analysis were very similar to those from the main text as shown by the parameter estimates for the hierarchical model in Table B1. In line with the prediction of PCS, R-congruent responses were estimated to be faster than those based on recognition only. The same result emerged in a test of stochastic dominance when using 40 RT bins as shown in Figure B1 (the core estimates were $\hat{r}$ = 0.69, $\hat{a}$ = .65, and $\hat{b}$ = .63 with standard deviations of the posterior distributions ≤ .003).

Overall, this analysis shows that the main results cannot simply be explained by retrieval fluency as operationalized by the recognition speed of the objects. In other words, when taking into account that some of the objects were recognized faster, R-congruent responses were still estimated to be faster than responses based on recognition only, thus corroborating the parallel information-integration account.

(*Appendices continue*)

Appendix C

Subjective Knowledge Ratings (mR vs. R+)

Besides binary recognition judgments (R/U), several studies also asked participants whether cities were either merely recognized (mR) or recognized with further knowledge (R+) as proposed by Pohl (2006). Under the assumption that these subjective, introspective judgments are valid indicators for the availability of further knowledge, they allow to test hypotheses concerning the effect of cue knowledge (e.g., Hilbig & Pohl, 2009) and memory states (Castela et al., 2014) on the probability and speed of recognition use. In the following, we show that our main result—R-congruent responses are faster than responses due to recognition only—also emerged when estimating separate latent RT distributions for R+ and mR items.

If the subjective knowledge ratings were perfectly valid, people could only rely on the recognition cue and not on further knowledge when deciding between a merely recognized and an unrecognized option (mR/U pair). Accordingly, without any explicit or implicit knowledge, the knowledge validity for trials involving only merely recognized items (mR/mR pairs) must be at chance level, that is, $b = .50$. However, previous studies showed that this was not the case and that participants still had some kind of implicit knowledge despite reporting to merely recognize an object (e.g., Castela et al., 2014). Hence, the subjective knowledge ratings cannot be interpreted as perfectly valid indicators of further knowledge.

In deriving predictions from the PCS account, we did not differentiate between explicit and implicit knowledge. Hence, we assume that any type of relevant knowledge enters the information integration process (Glöckner & Betsch, 2008). This implies that the model predicts the same ordering of R-congruent, R-only, and R-incongruent RT distributions within both mR/U and R+/U pairs. However, assuming that the subjective ratings reflect the amount of both explicit and implicit knowledge to some degree, we expect a larger effect of knowledge on the RT distributions for R+/U pairs compared to mR/U pairs (namely, larger discrepancies between the three latent RT distributions).

To test our predictions, we fitted the r*-model by Castela et al. (2014) that extends the standard r-model to all paired combinations of R+, mR, and U item types, resulting in six MPT trees and 18 observed categories. We adapted the restrictions of the b parameters by Castela et al. (2014), namely, that knowledge validity may differ for R+ and mR item types but is identical for RR and RU pairs (i.e., $b_{R+,U} = b_{R+,R+}$ and $b_{mR,U} = b_{mR,mR}$). Moreover, we modeled the latent RT distribution similarly as in the main text, that is, we fitted three latent RT distributions (R-congruent, R-only, and R-incongruent) separately for the R+/U and the mR/U pairs.

Table C1
Core Parameter Estimates for the r-Model (Castela, Kellen, Erdfelder, & Hilbig, 2014)*

	Mean	*SD*	95% CI
$r_{\mathrm{R+}}$	.802	.003	[.797–.808]
r_{mR}	.596	.005	[.585–.607]
$a_{\mathrm{R+}}$	.695	.002	[.691–.699]
a_{mR}	.588	.003	[.581–.594]
$b_{\mathrm{R+}}$	.650	.003	[.645–.655]
b_{mR}	.598	.004	[.590–.606]
$b_{\mathrm{R+,mR}}$	.647	.003	[.640–.653]
g	.519	.003	[.513–.524]

Due to the large number of observed categories of the RT-extended r*-model ($18 \cdot B$ instead of $8 \cdot B$ when using B RT bins) and the smaller sample size (only 13 data sets in Table 3 included mR/R+ judgments, resulting in a total of 165,786 responses), we did not fit a hierarchical model. Instead, we relied on the assumption of identically and independently distributed responses and used 20 RT bins to test stochastic dominance.

Table C1 shows the core parameter estimates of the r*-model. Both knowledge and recognition validity were higher for R+ compared to mR items. Note that participants' accuracy was substantially above chance for merely recognized items ($b_{\mathrm{mR}} = .598$), which indicates that introspective knowledge ratings were imperfect indicators for the availability of further knowledge. Moreover, participants relied on the recognition cue more often for trials for which further knowledge was available ($r_{\mathrm{R+}} = .802$ vs. $r_{\mathrm{mR}} = .596$). At first glance, this result seems to contradict the PCS prediction, according to which the recognition cue should have less impact when more knowledge becomes available. However, this effect can in principle be overruled if the weight associated with the recognition cue changes, that is, if participants assign a higher recognition validity to the R cue for R+ compared to mR items. In other words, decision makers simply adjust the subjective weight of the recognition cue (Hilbig, Scholl, & Pohl, 2010). Note that this theoretical conceptualization of 'adaptivity' refers to the adaptive use of a cue within a single process instead of a qualitative change of the processing strategy as assumed by the serial RH account (Glöckner et al., 2014).

(*Appendices continue*)

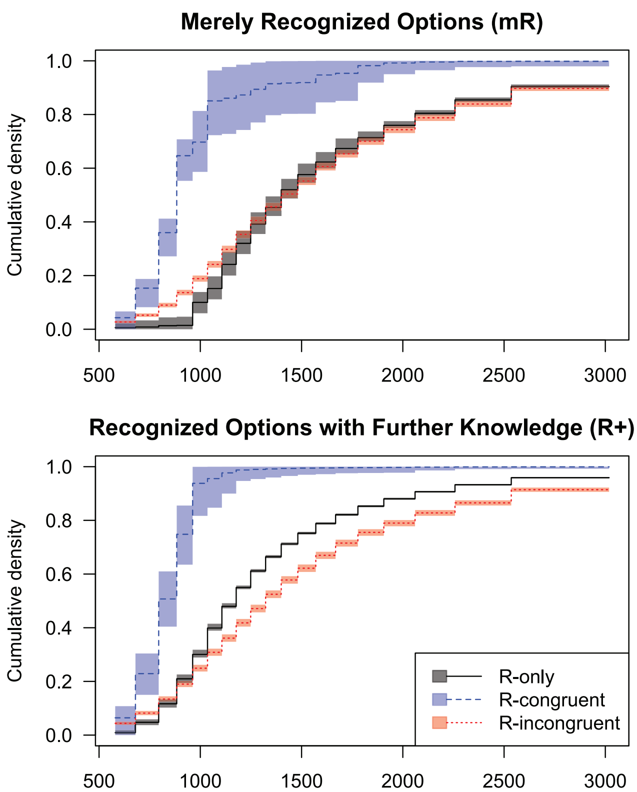

Figure C1. Estimated latent RT distributions for pairs including an unrecognized and a merely recognized option (mR/U) or an unrecognized and a recognized option with further knowledge (R+/U). See the online article for the color version of this figure.

Figure C1 shows the cumulative density estimates for the six latent RT distributions of interest, that is, the R-congruent, R-only, and R-incongruent distributions separately for R+/U and mR/U trials. For both types of trials, the R-congruent responses were estimated to be substantially faster than responses due to recognition only across the whole range of RTs. For merely recognized objects, R-incongruent responses were similarly fast as those due to recognition only, but slower for items that were recognized with further knowledge.

Note that the speed of noncompensatory use of recognition (R-only) was estimated to be slightly faster for R+/U pairs compared to mR/U pairs. At first sight, this seems to contradict the PCS prediction, according to which responses should be similarly fast if the integration process is determined by the recognition cue only. However, this prediction rests on the additional assumption that the weight assigned to the recognition cue is identical in both cases. However, given that the recognition validity is higher for R+ compared to mR objects, adaptive decision makers should assign a higher weight to the recognition cue accordingly. For the RT-extended r-model, this implies that both the *r* parameter and the speed of R-only responses will increase for R+/U pairs compared to mR/U pairs, exactly as observed.

Overall, this analysis of subjective knowledge ratings corroborates our conclusion from the main text, that is, responses due to R-congruent knowledge were faster than those due to recognition only. Moreover, our analysis showed that some type of implicit knowledge must be available even when participants judge items as being "merely recognized without further knowledge." This strengthens the importance of using a measurement model such as the r-model to decompose observed responses that are due to further knowledge and responses that are due to recognition only instead of relying on introspective judgments.

Received July 12, 2016
Revision received January 16, 2017
Accepted January 16, 2017 ■

Linking Process and Measurement Models of Recognition-Based Decisions
(Supplementary Material)

Daniel W. Heck & Edgar Erfelder
University of Mannheim

January 16, 2017

Contents

1 Supplementary Material

In the following, we provide technical details regarding the simulation study based on the classic German city data set by Gigerenzer and Goldstein (1996) and present an updated data set of cue and criterion values for the German cities with population greater than 100,000. We repeated the simulation for this data set, which lead to identical conclusions regarding the ability of the RT-extended r-model to differentiate between both process models. Moreover, the supplementary material includes goodness-of-fit plots for the second empirical reanalysis.

2 Technical Details Regarding the Simulation Study

2.1 The 83 Largest German Cities

The data set by Gigerenzer and Goldstein (1996) contains information about the largest German cities in form of the following cues, with validities provided in parentheses: national capital (1.00), exposition site (0.91), soccer team (0.87), intercity train line (0.78), state capital (0.77), license plate (0.75), university (0.71), industrial belt (0.56), and East Germany (0.51). The correct cue values for option i and cue c are coded by $C_{ic} \in \{-1, +1\}$. Below, we encode unavailable cue values due to failed recognition or partial knowledge by $C_{ic} = 0$.

2.2 Recognition Process and Incomplete Knowledge

The recognition probabilities were modeled by a sigmoid function of the criterion values (Hilbig, 2010; Schooler & Hertwig, 2005). First, we log-transformed and z-standardized the city populations reported by Gigerenzer and Goldstein (1996, p. 668). Second, these scaled city sizes z_i served as input for a probit-link (i.e., the inverse of the cumulative density function Φ of the standard normal distribution) to determine the probability that city i is recognized,

$$\mathrm{P}(R_i = +1) = \Phi(\alpha' + \alpha^* z_i), \tag{1}$$

where the intercept α' and the slope α^* jointly determine the recognition validity and discrimination rate. We used values of $\alpha' = 0$ and $\alpha^* = 0.8$ to obtain the recognition probabilities shown in Figure 1A that result in a mean recognition validity of $\alpha \approx .80$ (cf. Gigerenzer & Goldstein, 1996) and a mean discrimination rate of $d \approx .50$.

Based on the recognition probabilities in Figure 1A, a random vector of recognized ($R_i = +1$) and unrecognized ($R_i = -1$) cities was drawn separately for each participant. Different participants were treated as random replications of the same stochastic process. In other words, no systematic heterogeneity between participants was assumed. To model incomplete knowledge represented by missing information regarding the unrecognized options, all corresponding cue values were set to zero. In natural environments, cue values are more likely to be known for larger cities. Hence, to account for partial cue knowledge concerning the recognized cities, we generated probabilities of knowing a cue value using the sigmoid function shown in Figure 1B (with intercept $\alpha' = -1$ and slope $\alpha^* = .1$). Based on these probabilities, cue values were removed randomly for each participant (i.e., setting $C_{ic} = 0$). This procedure resulted in individual cue structures with partial knowledge about some cue values of some options.

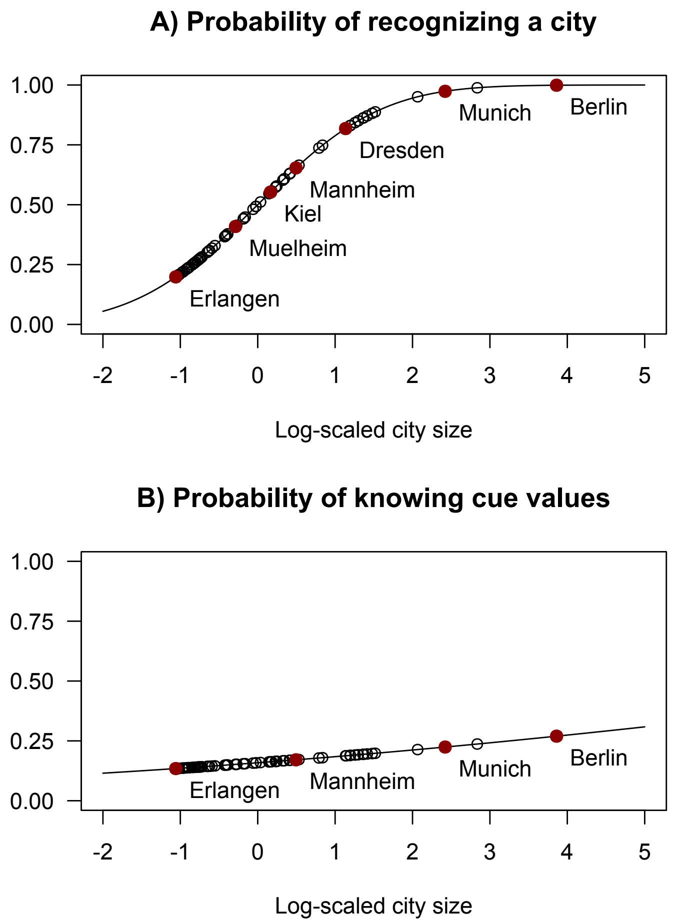

Figure 1: Probit-link function used to model recognition and cue knowledge probabilities.

2.3 Process Model Implementations

For the serial heuristic account, the predicted number of EIPs in each trial is described in the main text. Based on this number M of required EIPs, the observed RTs are modeled by the sum of a uniform nondecision time T_0 (with values in the interval $[0, 1]$) and M independent random variables E_i that represent the time required for each EIP,

$$T = T_0 + \sum_{i=1}^{M} E_i, \tag{2}$$

where each E_i is exponentially distributed with rate $\lambda = 1$.

As described in the main text, PCS chooses the option with the highest activation.

Moreover, interval-scaled RT predictions are provided by the number of iterations until convergence is reached. To simulate continuous RT values, we added independent, normally distributed nondecision times to the number of iterations. We chose a standard deviation (SD = 10) for the nondecision times, which is in the range of the SD of the raw iterations produced by PCS. However, values of SD = 1 and SD = 25 did not affect the pattern of results qualitatively.

2.4 Details Regarding the Generalizability Simulation

Each replication was generated under the following conditions. A subset between two and nine of the available cues was sampled with uniform probability (without replacement). Regarding the probabilities to recognize a city, we used the probit function in Equation 1 with randomly chosen parameters. We sampled the intercept α' and the slope α^* from truncated normal distributions with a standard deviation of .25 and means equal to the previous simulation ($\mu_{\alpha'} = 0$, $\mu_{\alpha^*} = 0.8$). Similarly, the probability of a cue value being known to a participant was determined by a probit-function with intercept and slope drawn from truncated normal distributions with a standard deviation of .25 and means $\mu_{\beta'} = -1$ and $\mu_{\beta^*} = 0.1$, respectively.

Based on 10,000 generated cue structures, the process models were applied as above, that is, with the true cue validities and a subjective recognition validity of $\alpha = .80$ in case of PCS and a probability $r = .80$ of RH use in case of the serial RH. We simulated responses for $N = 50$ hypothetical participants and we fitted the RT-extended r-model using two RT bins for each process model.

3 Simulation Based on Updated Data Base

To check the robustness of our simulation, we updated the data base by Gigerenzer and Goldstein (1996), which lists city sizes and cue values for German cities with a population larger than 100,000. Cue validities were similar compared to the classic data set (national capital = 1.00, exposition site = 0.92, license plate = 0.87, intercity train line = 0.83, state capital = 0.81, university = 0.76, soccer team = 0.76, East Germany = 0.58, industrial belt = 0.55). All parameters were identical to those reported in the simulation in the main text. The resulting recognition probabilities for cities and cue values are shown in Figure 3.

Figure 3 show the results of the simulation. Overall, the same conclusions can be drawn. First, the core parameter estimates of the RT-extended r-model are very similar between data-generating models and do not allow to distinguish both theoretical

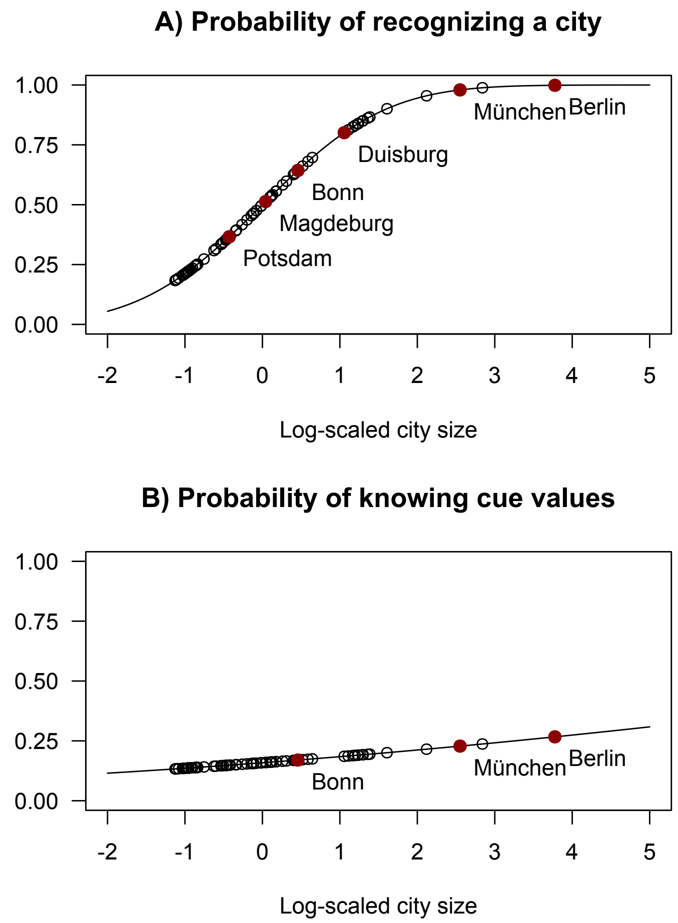

Figure 2: Recognition probabilities for the simulation based on the updated data base of German cities.

accounts. In contrast, for PCS, the relative speed of R-congruent knowledge use is estimated to be faster than reliance on recognition only (and vice versa for the RH). Hence, we conclude that our simulation results also hold in a different, up-to-date ecological environment.

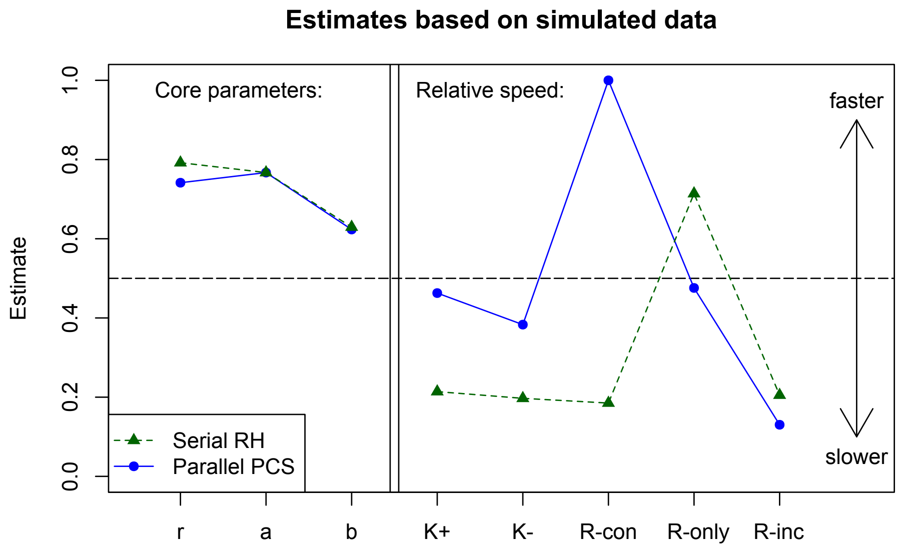

Figure 3: Estimated core parameter and relative speed of processes in based on the updated data base.

4 Goodness-of-Fit in the Second Reanalysis

Concerning the second reanalysis of 10 studies by researchers not associated with Mannheim, Figure 4 shows that the hierarchical MPT model fitted the observed frequencies well. Similarly, Figure 5 indicates a good model fit of the RT-extended r-model with 30 RT bins to the observed data.

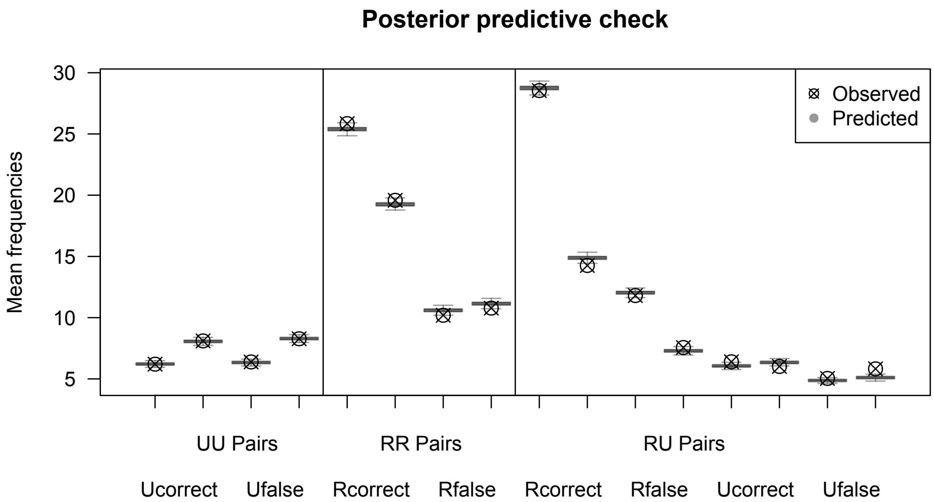

Figure 4: Model fit of the hierarchical RT-extended r-model to the 8×2 data categories. For each response category of the r-model, the observed and predicted frequencies of the 'fast' category are presented first, and those of the 'slow' category second.

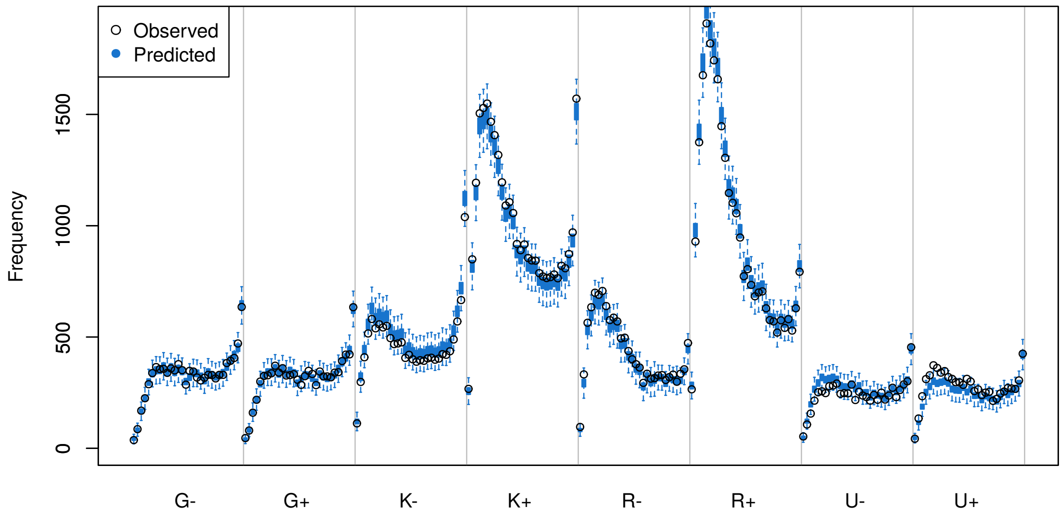

Figure 5: Model fit of the RT-extended r-model with 30 RT bins to the 30 × 8 data categories. Within each type of response, frequencies from fast to slow RT bins are shown from left to right.

References

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. doi:10.1037/0033-295X.103.4.650

Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making*, *5*, 272–284.

Schooler, L. J. & Hertwig, R. (2005). How Forgetting Aids Heuristic Inference. *Psychological Review*, *112*, 610–628. doi:10.1037/0033-295X.112.3.610

GENERALIZED PROCESSING TREE MODELS: JOINTLY MODELING DISCRETE AND CONTINUOUS VARIABLES

DANIEL W. HECK, EDGAR ERDFELDER, & PASCAL J. KIESLICH

UNIVERSITY OF MANNHEIM

Author Note

This research was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the DFG. All data and R scripts for the simulation and empirical analysis are available in the supplementary material at https://osf.io/fyeum.

Correspondence concerning this article should be addressed to Daniel W. Heck, Department of Psychology, School of Social Sciences, University of Mannheim, Schloss EO 254, D-68131 Mannheim, Germany. E-mail: heck@uni-mannheim.de; Phone: (+49) 621 181 2145.

GENERALIZED PROCESSING TREE MODELS: JOINTLY MODELING DISCRETE AND CONTINUOUS VARIABLES

Abstract

Multinomial processing tree models assume that discrete cognitive states determine observed response frequencies. Generalized processing tree (GPT) models extend this conceptual framework to continuous variables such as response times, process-tracing measures, or neurophysiological variables. Essentially, GPT models assume a finite mixture distribution, with weights determined by a processing-tree structure, and continuous components modeled by parameterized distributions such as Gaussians with separate or shared parameters across states. We discuss identifiability issues, parameter estimation, and propose a modeling syntax. A simulation shows the higher precision of GPT estimates, and an empirical example tests a GPT model for computer-mouse trajectories of semantic categorization.

Key words: multinomial processing tree model; discrete states; mixture model; cognitive modeling; response times; mouse-tracking

1. Introduction

In psychology, many theories comprise the assumption that identical observed responses can emerge from qualitatively distinct cognitive processes. For instance, in recognition memory, correct responses might not only be due to actual memory performance, but can also be the result of lucky guessing (Batchelder & Riefer, 1990). As a remedy, measurement models assume an explicit structure of latent processes and thereby provide process-pure measures for psychological constructs. A particularly successful class of such models are multinomial processing tree (MPT) models which account for choice frequencies by assuming discrete cognitive states (Batchelder & Riefer, 1990, 1999; Erdfelder et al., 2009). Essentially, MPT models assume a finite number of cognitive processes where the parameters are defined as the (unconditional or conditional) probabilities of entering different latent states.

A prominent example is the two-high threshold model (2HTM) for old-new recognition judgments shown in Figure 1 (Snodgrass & Corwin, 1988). Given a learned item, correct *old* responses (hits) emerge either from recognition certainty with probability d_o or from recognition uncertainty with the complementary probability $(1 - d_o)$. Whereas recognition certainty determines an *old* response with probability 1, recognition uncertainty results in guessing an *old* response with probability g. The two processing branches are disjoint and thus the probability of hits is $d_o + (1 - d_o)g$. Similarly, the probability of correct rejections (*new* responses to lures) is $d_n + (1 - d_n)(1 - g)$. Besides a good quantitative fit to data, the psychological validity of the parameters needs to be established by showing selective influence of experimental factors as predicted, for example, memory-strength manipulations only affect d_o and d_n whereas response bias manipulations only affect g (e.g., Bröder & Schütz, 2009). Once the validity of an MPT model such as the 2HT model has been established, it can be used to test effects of experimental manipulations or population characteristics on specific cognitive processes as represented by the parameters (an approach termed *cognitive psychometrics*; Batchelder, 1998).

Despite their wide applicability, the scope of MPT models is inherently limited by

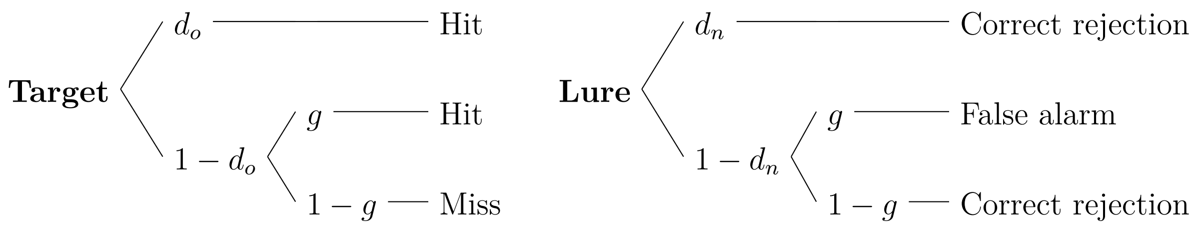

Figure 1. The two-high threshold model (2HTM) of recognition memory. The parameters d_o and d_n are the probabilities to detect old and new items (targets and lures), respectively, whereas g is the probability to guess *old* conditional on non-detection.

the fact that only discrete data such as response frequencies can be analyzed (Batchelder & Riefer, 1990). However, due to the availability of computerized testing, eye-tracking, and neurophysiological methods, psychologists increasingly rely on continuous variables to investigate cognitive functioning. Given the success of MPT models, we aim at generalizing their underlying theoretical and statistical structure to continuous variables by modeling observed behavior as an outcome of discrete cognitive states. For instance, theories of memory retrieval might not only predict the probability of recognizing items with certainty, but also that responses due to detection states are faster and given with higher confidence relative to guessing responses (Heck & Erdfelder, 2016; Province & Rouder, 2012).

Conceptually, the proposed generalized processing tree (GPT) framework formalizes this line of reasoning and assumes that latent cognitive processes affect discrete responses and continuous variables jointly. The GPT approach focuses on data in which one or more discrete and one or more continuous variables are observed in each trial. In case of binary recognition judgments, *old* and *new* responses could be recorded jointly with variables such as response times (Dube, Starns, Rotello, & Ratcliff, 2012), confidence ratings (Bröder, Kellen, Schütz, & Rohrmeier, 2013), or process-tracing measures (e.g., computer-mouse trajectories as modeled in the empirical example in Section 4; Koop & Criss, 2016). Assuming that different latent processes may result in the same vector of observed variables (e.g., Province & Rouder, 2012), an indeterminacy problem arises because observations cannot directly be assigned to the underlying latent processes. In other words, each observed vector with one discrete and one or more continuous observations can in principle emerge from more than one of the cognitive states. GPT models solve this indeterminacy problem by explicitly assuming a finite mixture distribution for discrete and continuous variables jointly.

In psychology, mixture models have often been used to account for response times (e.g., Dixon, 2012; Heck & Erdfelder, in press; Miller, 2006; Ollman, 1966; Province & Rouder, 2012; Yantis, Meyer, & Smith, 1991). The proposed GPT framework generalizes both mixture models, which are usually applied to response times, and MPT models, which are usually applied to discrete data, and provides some crucial advantages: (a) the mixture weights are modeled by a theoretically motivated and psychologically interpretable processing tree (similar to MPT modeling, see Section 2.1), (b) the simple model structure facilitates the development of GPT models for specific theories and experimental paradigms (following the philosophy of the MPT approach, an easy-to-use modeling syntax is developed in Section 2.4), (c) the MPT parameters, which are defined as probabilities of entering different cognitive states, can be estimated with higher precision than in equivalent MPT models (as shown in the simulation in Section 3), and (d) mixture assumptions can be tested for continuous variables other than response times (e.g., curvature of mouse-movement trajectories, as in Section 4; Dale, Kehoe, & Spivey, 2007). Before highlighting these advantages in a simulation and an empirical example, we first define GPT models formally, derive an expectation-maximization algorithm for parameter estimation, discuss issues of identifiability and model testing, and propose an easy-to-use modeling language implemented in the R package `gpt`.

2. Generalized Processing Tree Models

2.1. Jointly Modeling Discrete and Continuous Variables

The following notation for GPT models extends that of Hu and Batchelder (1994) for MPT models in order to include one or more continuous variables. In each trial, a vector $(x, \boldsymbol{y})$ is observed, where $x \in \{C_1, \ldots, C_J\}$ denotes the discrete choice of one of the response categories $C_1, \ldots, C_J$, and $\boldsymbol{y} \in \mathbb{R}^D$ is a vector of D continuous observations.

The discrete and continuous data across $k = 1, \ldots, K$ independent trials are collected in a vector $\boldsymbol{x} \in \{C_1, \ldots, C_J\}^K$ and a matrix $\boldsymbol{Y} \in \mathbb{R}^{K \times D}$, respectively. Discrete observations are summarized by the observed response frequencies $n_1, ..., n_J$ that sum to K and are defined as $n_j = \sum_k \delta_{C_j}(\{x_k\})$, where $\delta_{C_j}(A)$ is the Dirac measure, which is one if C_j is in the set $A \subset \{C_1, \ldots, C_J\}$ and zero otherwise.

The hypothesized cognitive states are modeled by I processing branches, which are indexed by $i = 1, \ldots, I_j$ within category j (with $I = \sum_j I_j$). Similar to MPT models, the category probabilities $p_j(\boldsymbol{\theta})$ are modeled by a binary tree structure, conditional on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{S_1})^\mathsf{T} \in \Theta = [0,1]^{S_1}$. The category probabilities are thus given by summing the corresponding branch probabilities $p_{ij}(\boldsymbol{\theta})$ (Hu & Batchelder, 1994), which are polynomial functions of $\boldsymbol{\theta}$,

$$p_{ij}(\boldsymbol{\theta}) = c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}} (1 - \theta_s)^{b_{ijs}}, \tag{1}$$

where the variables $a_{ijs} \in \mathbb{N}$ and $b_{ijs} \in \mathbb{N}$ indicate how often θ_s and $(1 - \theta_s)$ occur in the j-th branch leading to category C_i, respectively (with $\mathbb{N}$ including zero). Moreover, all fixed parameter values in this branch are collected in a single constant $c_{ij} \geq 0$.

The continuous observations $\boldsymbol{y}$ are modeled conditional on the cognitive states, that is, separately for each of the processing branches. More specifically, GPT models assign the continuous probability density function $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta})$ to branch i leading to category C_j, which is parameterized by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{S_2})^\mathsf{T} \in \Lambda \subset \mathbb{R}^{S_2}$. For the univariate case $D = 1$, Gaussian, gamma, beta, Wald, ex-Gaussian (Luce, 1986), or other commonly used continuous distributions can be defined for the component densities $g_{ij}(y)$ in order to match theoretical assumptions and the type of variable (e.g., response times or continuous confidence judgments). For the multivariate case $D > 1$, the continuous variables $\boldsymbol{y}$ may be assumed to be independent conditional on the latent states. To this end, univariate distributions $g_{ij}^{(d)}$ are defined separately for each of the continuous variables $y^{(d)}$,

$$g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}) = \prod_{d=1}^{D} g_{ij}^{(d)}(y^{(d)} \mid \boldsymbol{\eta}). \tag{2}$$

Alternatively, multivariate continuous distributions can be defined for g_{ij} to model dependencies conditional on the latent states.

Depending on the choice of the component densities g_{ij}, the parameters in $\boldsymbol{\eta}$ refer to means, standard deviations, shapes, rates, or other parameters of the these continuous distributions. Theoretical predictions are implemented by constraining some of these parameters to be identical across cognitive states (e.g., by assuming a constant variance for all components g_{ij}). Note that some dimensions of the parameter space Λ may be constrained to be positive or within a specific range (e.g., variance and rate parameters must be nonnegative). Such constraints render Λ a proper subset of $\mathbb{R}^{S_2}$ and need to be considered in parameter estimation (Section 2.3).

Based on the core assumption that discrete and continuous responses emerge from a finite number of latent cognitive states, a mixture distribution is predicted for the vector of observations $(x, \boldsymbol{y})$. However, in contrast to standard mixture models, the GPT framework uses the branch probabilities $p_{ij}(\boldsymbol{\theta})$ to assume an explicit structure for the mixture weights. More specifically, a GPT model accounts for the data by the parameterized probability density function

$$f(x, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \delta_{C_j}\left(\{x\}\right) \sum_{i=1}^{I_j} p_{ij}(\boldsymbol{\theta}) g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}). \tag{3}$$

Note that only one term of the sum across categories $C_1, \ldots, C_J$ differs from zero (i.e., the term for which $x = C_j$), and thus the inner sum across branches $i = 1, \ldots, I_j$ gives the defective density function for continuous observations $\boldsymbol{y}$ in category C_j. Taken together, a generalized processing tree (GPT) model is defined as a set of parameterized probability density functions,

$$\mathcal{M}^{\text{GPT}}(\Theta = [0,1]^{S_1}, \Lambda \subset \mathbb{R}^{S_2}) = \{f(x, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) \mid \boldsymbol{\theta} \in \Theta, \boldsymbol{\eta} \in \Lambda\} \tag{4}$$

with the mixture density f defined by Equations (1) and (3). Assuming independently and identically distributed responses across trials, the likelihood of the model is given as a function of the parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$ by the product

$$L(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{Y}) = \prod_{k=1}^{K} f(x_k, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{5}$$

Often, MPT models include $T > 1$ independent category systems, usually called ‘trees’, that refer to different item types or experimental manipulations and are modeled by separate processing tree structures. Based on the independence assumption, the resulting product-multinomial model can be analyzed similarly as a multinomial model (Hu & Batchelder, 1994, p. 31). For GPT models, it is also desirable to define different tree structures for disjoint sets of trials. For instance, in recognition memory, learned and new items could be represented by $T = 2$ disjoint sets $M_1 \subset \{1, \ldots, K\}$ and $M_2 = \{1, \ldots, K\} \setminus M_1$, respectively. For these different types of trials, separate GPT structures f_1 and f_2 can be assumed that have a distribution as defined in Equation (3). In general, there are T subsets of trials $M_1, \ldots, M_t, \ldots, M_T$ corresponding to tree structures $f_1, \ldots, f_t, \ldots, f_T$ with J_t categories, parameter counts a_{tijs}, b_{tijs}, c_{tij}, and components $g_{tij}(\boldsymbol{y} \mid \boldsymbol{\eta})$, but shared parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Since the conditions are independent, the overall likelihood becomes

$$L(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{Y}) = \prod_{t=1}^{T} \prod_{k \in M_t} f_t(x_k, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{6}$$

This more general form is required for the analysis of factorial designs and implemented in the R package `gpt`. However, for readability, the index $t = 1, \ldots, T$ is omitted in the following without loss of generality.

Figure 2 illustrates the GPT approach when including one continuous variable (i.e., $D = 1$) in the 2HTM (Heck & Erdfelder, 2016; Province & Rouder, 2012). Hits can either emerge from recognition certainty with probability d_o or from guessing *old* with probability $(1 - d_o)g$. Accordingly, values on the continuous variable y follow either the component distribution g_{do} or g_{go}, respectively. Then, the joint density of a hit and an observation y is given by the mixture

$$f(\text{hit}, y \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = d_o \cdot g_{do}(y \mid \boldsymbol{\eta}) + (1 - d_o) \cdot g \cdot g_{go}(y \mid \boldsymbol{\eta}). \tag{7}$$

Note that this is a defective density function with respect to the argument y because integrating over y gives the probability of hits, $d_o + (1 - d_o)g \leq 1$. The joint density for a false alarm and an observation y of the continuous variable can be derived accordingly.

What is gained by making parametric assumptions about the g_{ij} instead of adopting a distribution-free approach for the component densities that avoids restrictive assumptions about their shapes (see, e.g., Yantis et al., 1991)? In fact, based on the mixture assumption, Heck and Erdfelder (2016) recently estimated the component distributions g_{ij} by categorizing response times from fast to slow into several bins. Thereby, component distributions could be estimated by histograms without requiring specific parametric assumptions for $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta})$ as in GPT modeling. However, often researchers have a-priori knowledge about the approximate shape of these component distributions conditional on the cognitive states. For instance, response-time distributions are known to be nonnegative and right-skewed (Luce, 1986), and confidence judgments are usually bounded and unimodal. Incorporating these assumptions explicitly within the GPT framework constrains the statistical model and reduces model complexity. Moreover, compared to parametric approaches, categorization of continuous variables is not feasible for multivariate outcomes (i.e., $D > 1$), often requires larger sample sizes (Van Zandt, 2000), and might not provide unique parameter estimates if the core MPT structure is not identifiable (Heck & Erdfelder, 2016).

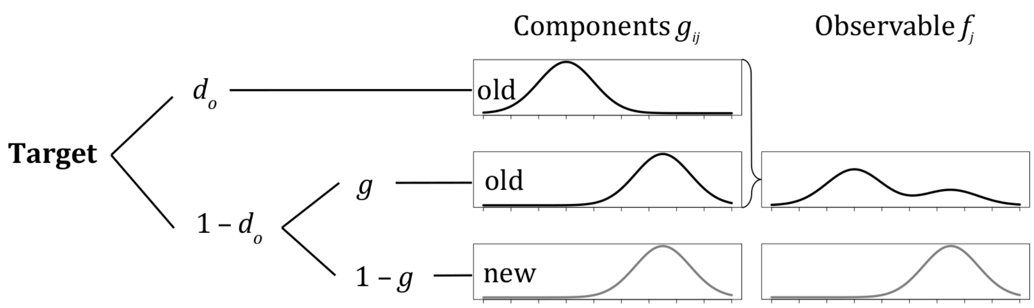

Figure 2. The two-high threshold model (2HTM) of recognition memory extended to one continuous variable. The component densities g_{ij} (Gaussian distributions in our example) are defined conditional on the processing branches, such that their observable distributions for separate response categories are, in general, finite mixture distributions.

2.2. Identifiability

To ensure unique parameter estimates, a statistical model must be identifiable, that is, the prediction function

$$h : \Theta \times \Lambda \longrightarrow P_{\boldsymbol{\theta},\boldsymbol{\eta}}, \tag{8}$$

which assigns the probability measure $P_{\boldsymbol{\theta},\boldsymbol{\eta}}$ to the parameters $(\boldsymbol{\theta},\boldsymbol{\eta})$, must be one-to-one (Lehmann & Casella, 1998, Definition 5.2). For GPT models, the measure $P_{\boldsymbol{\theta},\boldsymbol{\eta}}$ is defined as

$$\begin{aligned} P_{\boldsymbol{\theta},\boldsymbol{\eta}} : \mathcal{P}\left(\{C_1,\ldots,C_J\}\right) \times \mathcal{B}\left(\mathbb{R}^D\right) &\longrightarrow [0,1] \\ (A,B) &\longmapsto \sum_{j=1}^{J} \delta_{C_j}(A) \int_{\boldsymbol{y}\in B} f(C_j,\boldsymbol{y} \mid \boldsymbol{\theta},\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{y}, \end{aligned} \tag{9}$$

where $\mathcal{P}$ denotes the power set and $\mathcal{B}$ the Borel σ-algebra. If T independent conditions are modeled, Equation (9) needs to be expanded to the Cartesian product space

$$\Omega = \bigtimes_{t=1}^{T} \left(\{C_1,\ldots,C_{J_t}\} \times \mathbb{R}^D\right). \tag{10}$$

To prove identifiability of the parameters, it is necessary to show that $P_{\boldsymbol{\theta},\boldsymbol{\eta}} = P_{\boldsymbol{\theta}',\boldsymbol{\eta}'}$ implies $(\boldsymbol{\theta},\boldsymbol{\eta}) = (\boldsymbol{\theta}',\boldsymbol{\eta}')$. However, proving the identifiability for such complex models analytically is often infeasible, even for the more simple special case of MPT models (but see Batchelder & Riefer, 1990; Meiser & Bröder, 2002).

As a remedy, a heuristic approach facilitates assessing the identifiability of a given GPT model. In the following, it is assumed that each component distribution g_{ij} is identifiable with respect to its 'relevant' parameters $\boldsymbol{\eta}^*_{ij}$ that affect the predicted probability. Formally, the relevant parameters are defined as the shortest subvector $\boldsymbol{\eta}^*_{ij} = (\eta_{s_1},\ldots,\eta_{s_S}) \in \mathbb{R}^S$, $S \leq S_2$, for which $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}) = g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{ij})$ almost surely.

In a first step, the continuous variables $\boldsymbol{y}$ can be ignored to assess the identifiability of $\boldsymbol{\theta} \in (0,1)^{S_1}$ in isolation by using standard methods for MPT models. A necessary (but not sufficient) condition for the identifiability of an MPT model is that the number of free parameters must not be larger than (a) the number of free categories and (b) the rank of the Jacobian matrix for any observed frequencies $n_1,\ldots,n_J$ (Bamber & van Santen, 2000). Moreover, computer algebra software (Schmittmann, Dolan, Raijmakers, & Batchelder, 2010) and simulations (Moshagen, 2010; Rouder & Batchelder, 1998) are useful to assess the identifiability of the MPT structure.

If the MPT part of a GPT model is identifiable, it follows that the GPT parameters $\boldsymbol{\theta}$ are identifiable, because unique parameter estimates for $\boldsymbol{\theta}$ are obtained even when the continuous variables are ignored (Heck & Erdfelder, 2016, Observation 1). Therefore, only the identifiability of the parameters $\boldsymbol{\eta}$ needs to be shown, which can be done using the iterative approach proposed in Heck and Erdfelder (2016, Appendix A). Originally, this heuristic was developed for the distribution-free MPT extension, but it generalizes to GPT models. First, categories C_j are considered that are reached by a single processing branch i. For such categories, the corresponding component distributions g_{ij} are directly observable because

$$f(C_j, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = p_j(\boldsymbol{\theta}) g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{ij}), \tag{11}$$

and hence, the relevant parameters $\boldsymbol{\eta}^*_{ij}$, which determine the component density g_{ij}, are also identifiable (because p_j and g_{ij} are one-to-one). Next, categories are considered that are reached by two processing branches i and i', where the parameters $\boldsymbol{\eta}^*_{ij}$ are identifiable due to the first step. Since

$$f(C_j, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = p_{ij}(\boldsymbol{\theta}) g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{ij}) + p_{i'j}(\boldsymbol{\theta}) g_{i'j}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{i'j}), \tag{12}$$

the parameters $\boldsymbol{\eta}^*_{i'j}$ relevant for the component i' become identifiable.

This iterative scheme can be applied until the identifiability of the full vector $\boldsymbol{\eta}$ has been shown. Note that for some GPT models, this recursive approach might not be sufficient to determine the identifiability of all parameters $\boldsymbol{\eta}$. In such cases, identifiability can be assessed with the more general method of computing the rank of a matrix that summarizes the assignment of component distributions to processing branches (Heck & Erdfelder, 2016, Observation 4).

If the core MPT structure of a GPT model is *not* identifiable, it is in principle still possible to construct a GPT model with identifiable parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$ by appropriately constraining the parameters $\boldsymbol{\eta}$. As an example, consider a simple GPT model with two processing branches ($I = 2$) with two Gaussian component distributions $\mathcal{N}(\mu_i, \sigma_i^2)$, where both branches collapse within a single response category ($J = 1$). This special case is a standard two-component mixture model, which is faced with an issue called *label switching* (Frühwirth-Schnatter, 2001): within the vector $\boldsymbol{\eta} = (\mu_1, \sigma_1, \mu_2, \sigma_2)$, the parameters $\boldsymbol{\eta}^*_{11} = (\mu_1, \sigma_1)$ and $\boldsymbol{\eta}^*_{21} = (\mu_2, \sigma_2)$ can be switched (i.e., the indices in $\boldsymbol{\eta}$ can be permuted) without changing the predicted model density

$$f(C_1, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = p_{11}(\boldsymbol{\theta})\, g_{11}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{11}) + [1 - p_{11}(\boldsymbol{\theta})]\; g_{21}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{21}), \tag{13}$$

because the branch probabilities $p_{11}(\boldsymbol{\theta})$ and $1 - p_{11}(\boldsymbol{\theta})$ can be switched as well. Intuitively, the GPT model does not 'know' which of the two processing branches belongs to which of the two component distributions. As a remedy, the model can be rendered identifiable by constraining the component densities $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta})$ (Frühwirth-Schnatter, 2001). For instance, in Section 4, an identifiable GPT model is obtained by constraining the order of the mean parameters $\mu_1 \leq \mu_2$ of the Gaussian component distributions. Importantly, such strategies for constraining the components $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta})$ of general mixture models ensure the identifiability of the branch probabilities $p_{ij}(\boldsymbol{\theta})$. Thereby, GPT parameters $\boldsymbol{\theta}$ can be rendered identifiable that would not be identifiable in an equivalent MPT model.

2.3. Parameter Estimation

In the following, an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is developed to obtain maximum-likelihood estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. For this purpose, the latent state indicators z_{ijk} represent which of the processing branches ij caused the observation $(x_k, \boldsymbol{y}_k)$ in trial k (i.e., $z_{ijk} = 1$ if branch ij is the data-generating cognitive state and $z_{ijk} = 0$ otherwise). In the EM algorithm, two estimation steps (computing the expectation of z_{ijk} and maximizing the complete data likelihood) are repeated until the discrepancy of two successive parameter estimates falls below a desired precision $\epsilon > 0$.

2.3.1. Complete-Data Likelihood. For each trial k, the vector $\boldsymbol{z}_k = (z_{11k}, \ldots, z_{ijk}, \ldots, z_{I_J Jk})$ indicates the unobservable cognitive state and is filled with zeros except for the data-generating process $z_{ijk} = 1$. The latent state indicators $\boldsymbol{z}_k$ are missing by design, and are modeled by the complete-data likelihood

$$f(x_k, \boldsymbol{y}_k, \boldsymbol{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\boldsymbol{y}_k, \boldsymbol{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{14}$$

Note that the observable discrete responses x_k are a deterministic function of the latent states $\boldsymbol{z}_k$ (since $z_{ijk} = 1$ implies $x_k = C_j$) and can therefore be omitted from the complete-data likelihood. To simplify this function, the probability density function in (14) is factorized as

$$f(\boldsymbol{y}_k, \boldsymbol{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\boldsymbol{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{y}_k | \boldsymbol{z}_k, \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{15}$$

This expression can further be simplified because the GPT framework assumes separate sets of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ for modeling the discrete and continuous variables, respectively. Accordingly, the distribution of latent states $\boldsymbol{z}_k$ in (15) is independent of $\boldsymbol{\eta}$ and reduces to the branch probabilities,

$$f(\boldsymbol{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\boldsymbol{z}_k \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{i=1}^{I_j} \left[p_{ij}(\boldsymbol{\theta})\right]^{z_{ijk}}. \tag{16}$$

Essentially, this function simply selects the branch probability $p_{ij}(\boldsymbol{\theta})$ for which $z_{ijk} = 1$. Similarly, conditional on being in the latent state z_{ijk}, the second term in Equation (15) simplifies to the data-generating component density

$$f(\boldsymbol{y}_k | \boldsymbol{z}_k, \boldsymbol{\theta}, \boldsymbol{\eta}) = f(\boldsymbol{y}_k | \boldsymbol{z}_k, \boldsymbol{\eta}) = \prod_{j=1}^{J} \prod_{i=1}^{I_j} \left[g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta})\right]^{z_{ijk}}. \tag{17}$$

2.3.2. E-Step. In the first step, the EM algorithm computes the expectation of the latent state indicators $\boldsymbol{z}_k$ (Dempster et al., 1977) conditional on the current parameter values $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and on the observed data $(x_k, \boldsymbol{y}_k)$,

$$\hat{z}_{ijk} = \mathbb{E}[z_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k, \boldsymbol{y}_k] = P(z_{ijk} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k, \boldsymbol{y}_k). \tag{18}$$

In standard MPT models (Hu & Batchelder, 1994), the expected value of z_{ijk} is

$$P(z_{ijk} = 1 \mid \boldsymbol{\theta}, x_k) = \begin{cases} p_{ij}(\boldsymbol{\theta}) / \sum_i p_{ij}(\boldsymbol{\theta}) & \text{if } x_k = C_j \\ 0 & \text{else.} \end{cases} \tag{19}$$

However, in GPT models, additional information from the continuous variables is used to compute the expected value of the state indicator z_{ijk}. More precisely, Bayes' theorem and the functional independence of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ results in

$$P(z_{ijk} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k, \boldsymbol{y}_k) = \frac{P(z_{ijk} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) f(\boldsymbol{y}_k \mid z_{ijk} = 1, \boldsymbol{\theta}, \boldsymbol{\eta}, x_k)}{f(z_{ijk} = 1, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) + f(z_{ijk} = 0, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k)} \tag{20}$$

$$= \frac{P(z_{ijk} = 1 \mid \boldsymbol{\theta}, x_k) g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta})}{\sum_i P(z_{ijk} = 1 \mid \boldsymbol{\theta}, x_k) g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta})}, \tag{21}$$

$$= \frac{p_{ij}(\boldsymbol{\theta}) g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta})}{\sum_i p_{ij}(\boldsymbol{\theta}) g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta})}, \tag{22}$$

if $x_k = C_j$, and $P(z_{ijk} = 1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k, \boldsymbol{y}_k) = 0$ otherwise. Note that the normalizing constant in the denominator in Equation 21 is derived as

$$f(z_{ijk} = 1, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) + f(z_{ijk} = 0, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) \tag{23}$$

$$= f(z_{ijk} = 1, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) + \sum_{i' \neq i} f(z_{i'jk} = 1, \boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}, x_k) \tag{24}$$

$$= \sum_{i=1}^{I} P(z_{ijk} = 1 \mid \boldsymbol{\theta}, x_k) g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta}) \tag{25}$$

because exactly one of the indicators $z_{1jk}, \ldots, z_{I_j jk}$ is equal to one if $x_k = C_j$. Intuitively, Equation (22) provides refined estimates $\hat{z}_{ijk}$ for the latent states that weigh the branch probabilities p_{ij} by the corresponding component densities g_{ij} (in contrast to Equation (19) for standard MPT models which is obtained from (22) by assuming unit weights $g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta}) = 1$ for all i, j.).

2.3.3. M-Step. In the second step, the EM algorithm maximizes the complete-data likelihood in (15) conditional on the current estimates $\hat{\boldsymbol{z}}_k$ in (22). Due to the decomposition assumption of GPT models, this function can be maximized separately for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. First, the parameters $\boldsymbol{\theta}$ are estimated essentially in the same way as in standard MPT models (Hu & Batchelder, 1994),

$$\hat{\theta}_s = \frac{\sum_j \sum_i a_{ijs} \hat{m}_{ij}}{\sum_j \sum_i (a_{ijs} + b_{ijs}) \hat{m}_{ij}}, \tag{26}$$

where $\hat{m}_{ij} = \sum_{k=1}^{K} \hat{z}_{ijk}$ is the expected frequency for the processing branch ij. To obtain estimates $\hat{\boldsymbol{\eta}}$, the log likelihood in (17) needs to be maximized, with the estimated latent states $\hat{\boldsymbol{z}}_k$ serving as weights,

$$\log f(\boldsymbol{Y} \mid \hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_K, \boldsymbol{\eta}) = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I_j} \hat{z}_{ijk} \log g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta}) \tag{27}$$

An analytical solution for $\hat{\boldsymbol{\eta}}$ can be obtained for some component distributions (e.g., for Gaussian distributions). In general, however, it will be necessary to use numerical optimization methods such as gradient descent to maximize (27).

2.3.4. Implementation and Standard Errors. A few iterations of the standard EM algorithm for MPT models can be used to obtain sensible starting values for $\hat{\boldsymbol{\theta}}$ and $\hat{z}_{ijk}$ before estimating all parameters $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$. This is simply achieved by using Equation (19) instead of (22) when computing the expectation of the latent state indicators.

To estimate the standard errors of $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$, the observed Fisher information matrix

$$\left[\boldsymbol{I}(\hat{\boldsymbol{\xi}})\right]_{sr} = - \frac{\partial^2 \log L(\boldsymbol{\xi} \mid \boldsymbol{x}, \boldsymbol{Y})}{\partial \xi_s \, \partial \xi_r} \bigg|_{\xi = \hat{\xi}} \tag{28}$$

is computed numerically using the observed-data likelihood in (5). Then, the inverse of the observed Fisher information $\boldsymbol{V} = \boldsymbol{I}(\hat{\boldsymbol{\xi}})^{-1}$ provides the asymptotic covariance of the normally distributed estimate $\hat{\boldsymbol{\xi}}$. Accordingly, confidence intervals are obtained as $\hat{\xi}_s \pm z_{1-\alpha/2} \, (V_{ss})^{1/2}$.

2.3.5. Model Testing and Model Selection. To test the goodness of fit of MPT models, fitted category probabilities can be compared against the saturated model by means of power-divergence statistics such as the likelihood-ratio statistic G^2 or Pearson's X^2, which are asymptotically χ^2 distributed (Read & Cressie, 1988). However, it is more difficult to derive tests of absolute goodness of fit for continuous data for which a saturated model does not exist. If a GPT model includes only one continuous variable, its goodness of fit can be assessed by categorizing the continuous observations into a finite number of intervals (Klauer, 2001). Thereby, the predicted probabilities of observations falling into these intervals can be compared against the corresponding observed proportions using any power-divergence statistic. However, the resulting test statistic follows a standard χ^2 distribution only if the likelihood of the parameters $\boldsymbol{\xi}$ is maximized based on the binned data, but not if the original data are used for optimization as in the EM algorithm in Section 2.3 (Chernoff & Lehmann, 1954). As a remedy, analytical solutions such as the Rao-Robson statistic (Rao & Robson, 1974) or computational strategies such as the parametric bootstrap (Efron & Tibshirani, 1997) can be used. Moreover, goodness of fit can be assessed graphically by comparing the empirical against the fitted probability or cumulative density function as in Section 4.

Often, it will be necessary to test between several nested or nonnested GPT models. For instance, psychologists might be interested in testing whether experimental manipulations affect the parameters $\boldsymbol{\theta}$ of entering different cognitive states or the parameters $\boldsymbol{\eta}$ which summarize the continuous component distributions. If a GPT model

$\mathcal{M}_0$ is nested in $\mathcal{M}_1$, the likelihood-ratio statistic

$$G^2(\nu) = -2\left[\log L_0(\hat{\boldsymbol{\xi}}_0 \mid \boldsymbol{x}, \boldsymbol{Y}) - \log L_1(\hat{\boldsymbol{\xi}}_1 \mid \boldsymbol{x}, \boldsymbol{Y})\right] \tag{29}$$

is asymptotically χ^2 distributed with degrees of freedom ν given by the difference in the dimensionality of the parameter vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_0$ (Casella & Berger, 2002, Chapter 10.3). In contrast to statistical tests, model-selection criteria such as AIC or BIC trade-off model fit and complexity to select the model that best generalizes to new data (Myung, Pitt, & Kim, 2005). Note that these methods can be applied to both nested and nonnested GPT models.

2.4. A Modeling Syntax for GPT Models

MPT models are more specific than general-purpose methods such as ANOVA and need to be adapted to different psychological theories and experimental paradigms (Erdfelder et al., 2009). However, the availability of software for MPT modeling (e.g., Heck, Arnold, & Arnold, in press; Moshagen, 2010; Singmann & Kellen, 2013) greatly facilitates standard tasks such as model development, fitting, and evaluation, which has contributed to the wide popularity and success of MPT models. Similarly, the GPT framework requires the development and testing of specific models for each application. Therefore, it is important to design a modeling syntax for GPT models that uniquely defines a specific model structure (i.e., the MPT part and the component densities).

In the following, a syntax for GPT models is developed based on the .eqn-standard for MPT models (Hu, 1999; Moshagen, 2010) that simply lists the equations for the processing-branch probabilities $p_{ij}(\boldsymbol{\theta})$ in a text file. In each row, a processing branch is defined by (a) the label (or number) of the corresponding tree, (b) the label (or number) of the response category C_j, and (c) the MPT equation. GPT models can be specified in a similar way by defining each branch probability $p_{ij}(\boldsymbol{\theta})$. In addition to the standard .eqn-syntax for MPT models, it is necessary to define the type and parameterization of the component distributions g_{ij}. For this purpose, an additional column lists the parameters $\boldsymbol{\eta}^*_{ij}$ that are relevant for the component density $g_{ij}(\boldsymbol{y} \mid \boldsymbol{\eta}^*_{ij})$ (e.g., mean and SD of a normal distribution). Note that the distributional family is not defined within the model file but needs to be provided when fitting a GPT model. For instance, the R package `gpt` requires the argument `latent = "normal"` to define Gaussian component distributions for one continuous variable, or the argument `latent = c("normal", "exgauss")` to define Gaussian and ex-Gaussian component distributions for the first and second of two continuous variables, respectively.

As an example, a GPT extension of the 2HTM to both confidence ratings and response times could be defined as follows (model equations for lures are omitted):

```
# tree  ; category  ; MPT equation    ; confidence  ; response time
Target  ; hit       ; do              ; CR_d, sd    ; RT_d, sig, nu
Target  ; hit       ; (1-do)*g        ; CR_g, sd    ; RT_g, sig, nu
Target  ; miss      ; (1-do)*(1-g)    ; CR_g, sd    ; RT_g, sig, nu
```

Each line first lists the tree, category, and MPT equation, separated by a semicolon. Next, the latent continuous distributions are defined by distinct parameter labels for (a) mean and SD parameters of the Gaussian component distributions for confidence ratings and (b) three parameters of the ex-Gaussian component distributions for response times (i.e., mean and SD of a Gaussian random variable, and mean of an independent exponential random variable). Note that this GPT model assumes that confidence ratings and response times are independent conditional on the latent states (see Section 2.1).

The proposed syntax has the advantage that a-priori constraints on the parameters $\boldsymbol{\eta}$ can easily be specified across component distributions. For instance, the Gaussian components for modeling confidence ratings are assumed to have the same standard deviation (because a single parameter `sd` is used). Such constraints are important to obtain a parsimonious model (Myung, 2000) and to test theoretical predictions about the effect of cognitive processes on continuous variables (e.g., by comparing a GPT model with a nested special case).

In addition to using identical parameter labels in the model file, the package `gpt` also allows to put equality constraints on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ within R when fitting a model. Moreover, the package uses sensible parameter boundaries for the estimation of $\boldsymbol{\eta}$ (which is important for parameters such as variances that must be nonnegative). However, the range of admissible values can also be defined for each parameter η_s separately. The latter feature is useful to obtain an identifiable GPT model, for instance, by constraining the mean parameters to be on separate sides of a constant C, that is, $\mu_1 \leq C$ and $\mu_2 \geq C$ (see Section 4). Currently, the `gpt` package features various continuous distributions (e.g., Gaussian, lognormal, Wald, gamma, beta; Heathcote, 2004) and implements functions for model fitting evaluation, plotting, and generating simulated data. Note that details about the implementation and use of the `gpt` package will be provided in a separate article.

3. Simulation: Increased Precision of GPT Estimates

By modeling discrete and continuous variables jointly, GPT models can provide more precise parameter estimates $\hat{\boldsymbol{\theta}}$ (as indicated by smaller standard errors) relative to equivalent MPT models. Essentially, the additional information due to inclusion of continuous variables results in better estimates for the latent state indicators $\hat{z}_{ijk}$ (as shown in the E-step of the EM algorithm in (22)). For instance, Figure 2 shows an example in which detection responses in the 2HTM are associated with smaller values on the continuous variable y, whereas guessing-responses are associated with larger values. In such a case, an *old* responses x_k with a small value y_k did more likely emerge from being in the detection state, whereas an *old* response $x_{k'}$ with a large value $y_{k'}$ did more likely emerge from being in the uncertainty state. GPT models utilize this information to estimate the latent state in each trial, thereby increasing the precision of the estimates $\hat{\boldsymbol{\theta}}$.

The following simulation shows that the increase in precision depends on the discrepancy between the latent component distributions. More precisely, the continuous variables provide most information when the component distributions differ strongly across cognitive states (e.g., in terms of mean, variance, or higher moments). In contrast, the continuous variables provide no additional information if the component distributions are identical across cognitive states. In this case, GPT parameter estimates $\hat{\boldsymbol{\theta}}$ are as precise as the equivalent MPT estimates.

3.1. Methods

The GPT version of the 2HTM in Figure 2 was used to generate 3,000 data sets with $K = 100$ responses each (i.e., 50 targets and 50 lures). For the data-generating parameters $\boldsymbol{\theta}$, the detection probabilities were set to $d_o = d_n = .60$ with a guessing probability of $g = .50$. To model the effect of item detection and guessing on the continuous variable, two normal distributions with a standard deviation of $\sigma = 1$ were assumed. The discrepancy between these two component distributions was manipulated by varying the mean μ_g of the guessing distribution between zero and five, whereas the mean μ_d of the detection distribution was held constant at zero (Figure 2).

In each replication, the GPT version of the 2HTM was fitted to the simulated data. The detection probabilities were constrained to be identical, $d_o = d_n$, to ensure the identifiability of the MPT part of the model. Moreover, three free parameters were fitted for the component distributions: A common standard deviation σ, and two means μ_d and μ_g for detection and guessing, respectively. Overall, this resulted in a model with five free parameters, $\boldsymbol{\theta} = (d, g)$ and $\boldsymbol{\eta} = (\mu_d, \mu_g, \sigma)$. For comparison, the corresponding MPT model with the parameters d and g (i.e., the standard 2HTM) was fitted to the response frequencies only.

3.2. Results

Figure 3 shows the means of the estimated standard errors across replications for varying levels of the data-generating mean μ_d. Whereas the precision of the MPT estimates remains constant, the GPT estimates become more precise as the difference in the means μ_d and μ_g increases. Note that the precision of the GPT estimates is identical to that of the MPT estimates if the component distributions are identical (i.e., if $\mu_d = \mu_g$). Hence, the standard error of the MPT estimates provides an upper bound for the standard error of the GPT estimates.

Moreover, Figure 3 also shows the theoretical boundary for an increase in precision by GPT modeling that is obtained by treating the latent state indicators z_{ijk} as observable variables. If the data-generating states z_{ijk} were known, the MPT parameters $\boldsymbol{\theta}$ can directly be estimated using the M-step of the EM algorithm in (26), with the corresponding standard error for binomial-rate parameters,

$$\mathrm{SE}(\hat{\theta}_s) = \left(\frac{\hat{\theta}_s \, (1 - \hat{\theta}_s)}{M_s} \right)^{1/2}, \tag{30}$$

where M_s is the denominator of the estimator in (26) based on the observable states z_{ijk}

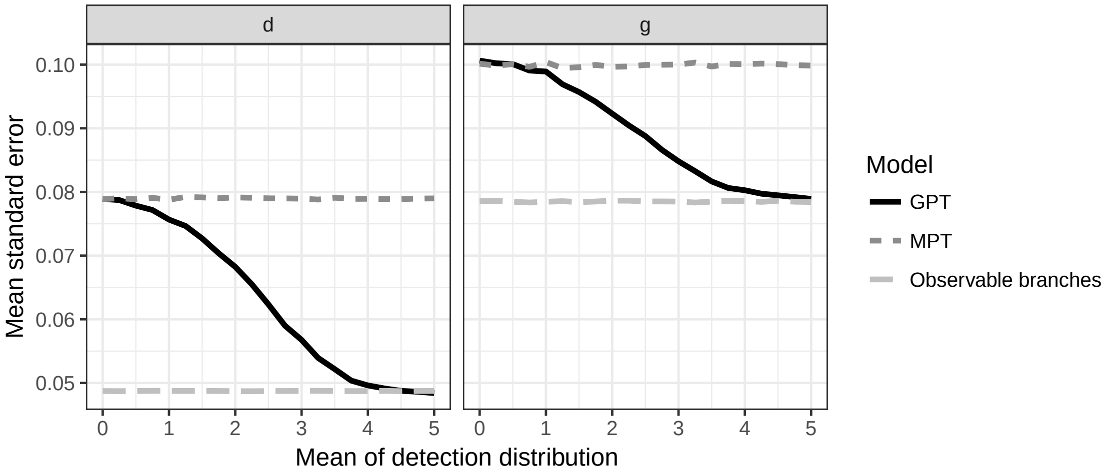

Figure 3. Mean standard error for the estimates $\hat{\boldsymbol{\theta}} = (\hat{d}, \hat{g})$ of the 2HTM obtained by the GPT model, the MPT model, and conditional on observable processing branches.

instead of the latent-state estimates $\hat{z}_{ijk}$. This standard error only depends on the frequencies $\sum_k z_{ijk}$ of being in state ij and on the specific processing tree structure which is represented by a_{ijs} and b_{ijs}. As shown in Figure 3, Equation (30) serves as a lower bound for the standard error of the GPT estimates $\hat{\boldsymbol{\theta}}$.

In sum, the simulation shows that GPT models utilize information provided by the continuous variables to obtain more precise estimates of the latent states z_{ijk}. If the hypothesized cognitive processes affect the component distributions, this information results in more precise estimates for $\boldsymbol{\theta}$. Note that the increase in precision will result in higher statistical power to detect differences in the parameters $\boldsymbol{\theta}$.

4. Application: GPT Modeling of Mouse-Tracking Trajectories

In the following, a GPT model for semantic categorization is developed and empirically tested that follows the spirit of the feature comparison model (Smith, Shoben, & Rips, 1974). In the experimental paradigm of interest (Dale et al., 2007), participants assign animals to one of two categories (e.g., categorize *bat* as a *mammal* or *bird*). Importantly, the trials differ in how typical the presented animal is for its category: whereas some animals are typical members of their class (e.g., *hawk* for *bird*) and do not share many features with the distractor category (e.g., *reptile*), some are atypical members (e.g., *bat*) and share features with both the correct (*mammal*) and the distractor category (*bird*).

Often, participants are very good at classifying animals, and thus the proportion of correct responses is usually high for both typical and atypical animals. Therefore, Dale et al. (2007) also recorded the computer-mouse movements of participants during each trial. For this purpose, the two response options (i.e., the categories) are presented at the upper left and the upper right of the screen, while the mouse cursor is placed at the lower center in the beginning of each trial as illustrated in Figure 4. Over the course of each trial, the coordinates of the mouse cursor are recorded with a constant sampling rate. Using mouse-tracking, Dale et al. (2007) showed that, relative to typical items, responses to atypical items resulted in less direct mouse trajectories that were curved toward the alternative option. They concluded that atypical items triggered a conflict between the two response options as reflected by the mouse movements.

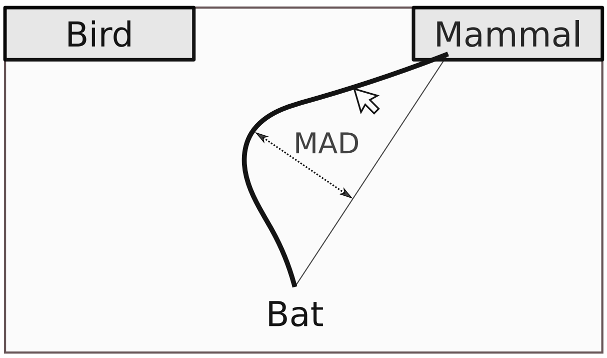

Figure 4. In the mouse-tracking paradigm, movements of the mouse cursor are recorded over the course of each trial. The maximum absolute deviation (MAD) is defined as the maximum perpendicular distance of the trajectory from a straight line connecting its start and end points.

The mouse-tracking methodology provides one set of time series data for each trial (i.e., a sequence of x-y coordinates of the mouse cursor from start to end point). Usually, these data are preprocessed for further analysis by computing summary statistics per trial (Freeman & Ambady, 2010). In the following, we focus on the maximum absolute deviation (MAD) that assesses the curvature of mouse-tracking trajectories and is defined as the maximum perpendicular discrepancy between the observed trajectory and a direct trajectory as shown in Figure 4 (i.e., a straight line from start to end position; see Kieslich, Wulff, Henninger, & Haslbeck, 2017, for alternative statistics). This preprocessing results in a single continuous observation per trial that is jointly observed with the discrete response (*correct* versus *incorrect*).

In most mouse-tracking studies, the MAD (and similar) summary statistics are analyzed only for correct responses, whereas incorrect trials are simply discarded. However, when using such an approach, subsequent statistical tests will not take classification uncertainty into account, even though the same responses can be due to different cognitive processes. As a remedy, a GPT model is proposed that accounts for this ambiguity by explicitly modeling how different hypothesized processes affect discrete responses and process-tracing measures jointly. This theoretical conceptualization contrasts with the use of observed eye- or mouse-tracking data as one-to-one indicators of hypothesized latent processes.

4.1. A GPT Version of the Feature Comparison Model

The feature comparison model (FCM; Smith et al., 1974) provides a theoretical account of the cognitive processes underlying semantic categorization. Essentially, the model assumes that features of objects and categories fall into two disjoint sets. Whereas defining features determine whether an object belongs to a category (e.g., birds breed eggs), characteristic features are shared among many members of a category, but are not strictly necessary (e.g., many, but not all birds can fly). On the basis of these two sets of features, the FCM assumes that semantic categorization proceeds in two steps. First, an overall-comparison process determines the similarity of an object to a category on the basis of all features, without discriminating between defining and characteristic ones. If the outcome of this first step provides sufficient evidence in favor for or against category membership, a corresponding response is made. In contrast, if the outcome is ambiguous, only the defining features are compared in detail in a second step.

The first process is assumed to be relatively fast and to have a higher accuracy for categorizing typical relative to atypical animals, because the overall comparison will more often be valid for typical items. In contrast, the second process is assumed to be relatively slow and to have identical accuracy for categorizing typical and atypical animals, because characteristic features become irrelevant. Note that, according to dual-process theories (Sloman, 1996), the second comparison step can also be conceptualized as a rule-based process, in which lay theories about animal classes are used for categorization. In the present context, however, this rule-based mechanism for the second-stage process results in identical predictions as the defining-feature comparison process of the FCM.

Originally, the FCM was tested in response-time experiments (Smith et al., 1974). However, the model is very general and directly suggests a corresponding GPT model for the mouse-tracking version of the semantic categorization task (Dale et al., 2007). Figure 5 illustrates the model predictions for discrete responses and the mouse-tracking measure MAD for typical items. With probability f_t, the overall-comparison process provides clear evidence for judging category membership, which in turn results in a correct response with probability $c_{1,t}$. In contrast, if the outcome of the first-stage comparison is ambiguous with probability $(1 - f_t)$, the defining-feature comparison provides a correct response with probability $c_{2,t}$. The same processing structure is assumed for atypical items with a separate set of parameters f_a, $c_{1,a}$, and $c_{2,a}$ (not shown in Figure 5).

Besides defining the processing tree for discrete responses, the proposed GPT model in Figure 5 also assumes that the mouse-tracking measure MAD is normally distributed conditional on the type of comparison process. If the overall comparison of all features provides unambiguous evidence for responding, MADs are assumed to follow a normal distribution with mean μ_1 and SD σ_1 irrespective of the correctness of responses. If this is not the case, the defining-feature process determines response behavior, and MADs are assumed to follow a normal distribution with mean μ_2 and SD σ_2. Importantly, the FCM entails that responses are relatively fast when the overall-comparison process provides unambiguous evidence for category membership (Smith et al., 1974). In this case, the comparison leads to less conflict between the two response options, and thus to rather direct mouse trajectories. In contrast, the defining-feature comparison process is assumed

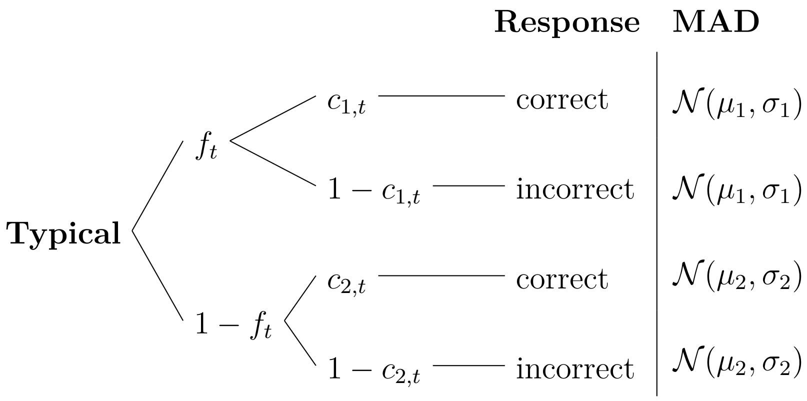

Figure 5. A GPT version of the feature comparison model (FCM; Smith, Shoben, & Rips, 1974) for the mouse-tracking version of a semantic classification task. Given a typical item, the overall-comparison processing provides clear evidence for category membership with probability f_t, which in turn results in a correct response with probability $c_{1,t}$. This first stage does not result in a response with probability $1 - f_t$, and in a second stage, a comparison of the defining features leads to a correct response with probability $c_{2,t}$. Moreover, the maximum absolute deviation (MAD) of a mouse trajectory in each trial follows a normal distribution, with mean and SD determined by the type of comparison process. The same processing structure is assumed for atypical items with a different set of probability parameters f_a, $c_{1,a}$, and $c_{2,a}$ (not shown).

to be slower and accompanied by more indirect trajectories due to ambiguity and conflict between the two choice options. Hence, the FCM implies the order constraint $\mu_1 \leq \mu_2$ which also ensures that all model parameters are identifiable. Note that the package `gpt` currently does not support such order constraints between parameters. As a technical trick, the mean parameters were instead restricted to fall into different ranges (i.e., $\mu_1 \leq 500$ and $\mu_2 \geq 500$). Thereby, the resulting model becomes equivalent to the theoretical model with the constraint $\mu_1 \leq \mu_2$ as long as the parameter estimates are not at the boundaries of the order constraints (which was not the case in the following analysis). In principle, this technical limitation can also be overcome by a reparameterization $\mu_2 = \mu_1 + \alpha$ using an additive parameter $\alpha \in \mathbb{R}^+$.

If the FCM provides a valid account of semantic categorization, the most general GPT version in Figure 5 with 10 parameters (i.e., three probability parameters for typical and atypical animals each, and four parameters for the two component distributions) should describe the observed distribution of responses and MADs well. In contrast, a nested model version with a single component distribution (i.e., $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$) should result in considerable misfit, because the two feature comparison processes are assumed to influence mouse trajectories (i.e., the distribution of MADs). Moreover, the FCM implies three hypotheses concerning the probability parameters $\boldsymbol{\theta}$ (Smith et al., 1974). First, the overall-comparison process is more likely to determine responses for typical than for atypical items (i.e., $f_t > f_a$), because characteristic features of atypical animals conflict with defining features resulting in increased discrepancy (and lower similarity). Second, the overall-comparison process is expected to be more accurate for typical than for atypical items (i.e., $c_{1,t} > c_{1,a}$), because characteristic features of typical animals are more valid than those of atypical ones. Finally, the defining-feature comparison process is expected to have the same accuracy for typical and atypical animals (i.e., $c_{2,t} = c_{2,a}$), because both classes mainly differ with respect to the characteristic features. These substantive hypotheses can be tested using likelihood-ratio tests of nested models with the corresponding parameter constraints.

4.2. Methods

The proposed model is tested by reanalyzing data of Kieslich and Henninger (in press), in which $N = 60$ participants had to categorize 13 typical and 6 atypical animals each (taken from Dale et al., 2007, Experiment 1) that were presented in random order. In each trial, the two response categories were presented first (in the top left and top right corners of the screen), after which participants had to click on a start button in lower middle of the screen. Following the button click, the name of the animal was presented, and participants had to click on the button of the response category to which the animal belonged. Mouse trajectories were recorded using the `mousetrap` plugin (Kieslich & Henninger, in press) for OpenSesame (Mathôt, Schreij, & Theeuwes, 2012), and the data were processed and analyzed using the mousetrap R package (Kieslich et al., 2017). MADs were calculated as the maximum perpendicular deviation between the start

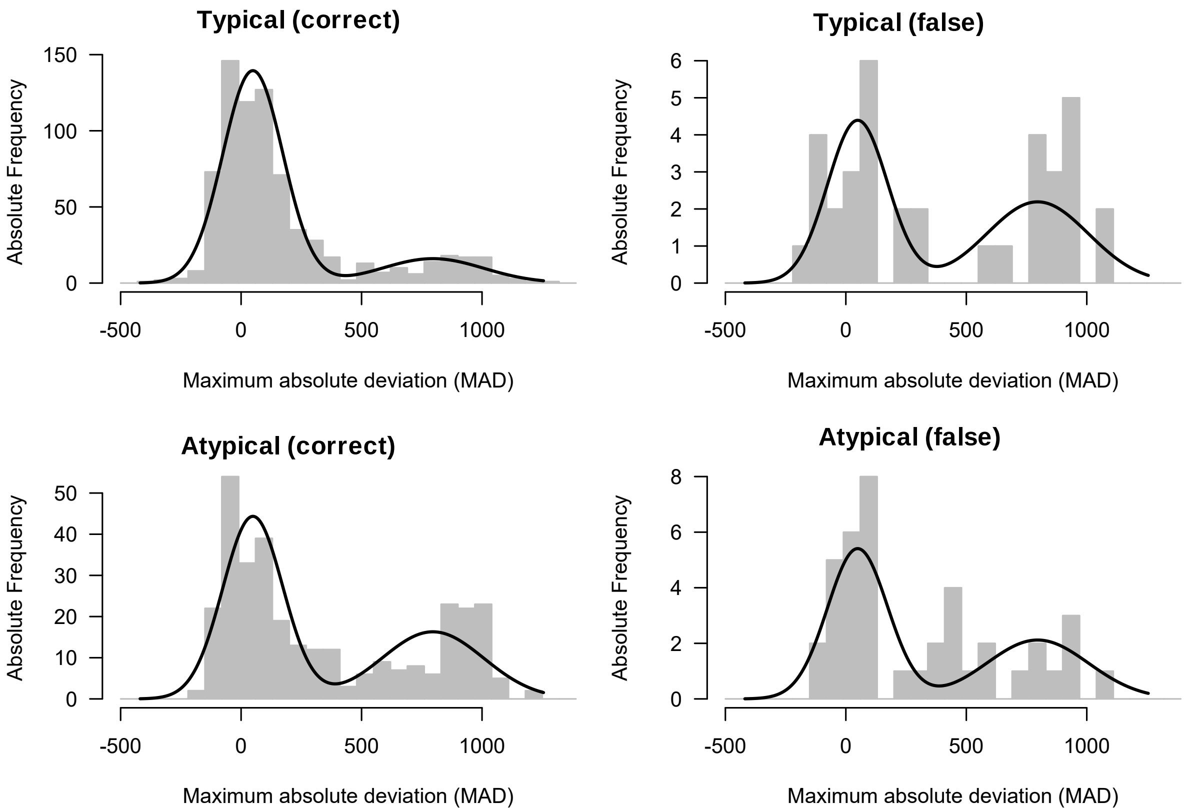

Figure 6. Observed (gray histograms) and fitted (black curves) MAD frequencies based on the general GPT version of the FCM with 10 parameters. Note that the scaling of the ordinate differs across panels.

and end point of each trajectory, with units referring to the absolute distance in pixels. Note that MADs were positive (negative) when the maximum absolute deviation from a direct trajectory was in the direction of the non-chosen (chosen) category and that, by taking the absolute value, MADs close to zero were rarely observed.

4.3. Results

Figure 6 shows separate histograms of the observed MADs for correct and incorrect categorizations of typical and atypical animals. Overall, participants had a high accuracy to select the correct category for typical (95.4%) and atypical (88.9%) animals. Note, however, that performance clearly was not error-free which underscores the necessity to analyze both correct and incorrect responses. Moreover, mouse trajectories exhibited considerable variability and the MAD distribution appeared to be bimodal.

As outlined in Section 2.3.5, when modeling continuous data, the definition and derivation of absolute goodness-of-fit tests is not as straightforward as for MPT models, where a natural saturated model exists. As a remedy, Figure 6 provides a graphical assessment of model fit by plotting gray histograms of observed MADs against black curves for the distributions predicted by the GPT model with 10 parameters. Overall, the model captured all salient features of the observed distributions well. Most importantly, the two Gaussian component distributions reflected the bimodality of MADs, one describing the overall-comparison process with direct trajectories (i.e., small MADs; $\hat{\mu}_1 = 48.9$, SE = 5.1; $\hat{\sigma}_1 = 125.1$, SE = 4.2), and the second one describing the defining-feature comparison process with less direct and more variable trajectories (i.e., large MADs; $\hat{\mu}_2 = 796.0$, SE = 19.0; $\hat{\sigma}_2 = 211.3$, SE = 15.6). Note that constraining both distributions to be identical increased misfit significantly ($G^2(2) = 976.1$, $p < .001$).

The model also described the relative contribution of the two mixture components in each response category very well, and provided estimates for the probability parameters $\boldsymbol{\theta}$ that were in line with the psychologically motivated hypotheses. First, the overall-comparison process determined responses more often for typical than for atypical animals ($\hat{f}_t = .824$, SE = .015; $\hat{f}_a = .615$, SE = .028; $G^2(1) = 81.8$, $p < .001$). This result was predicted by the FCM, because defining and characteristic features are less consistent for atypical than for typical items, resulting in more ambiguous similarity signals. Second, the overall-comparison process resulted in a higher accuracy for typical than for atypical animals ($\hat{c}_{1,t} = .970$, SE = .007; $\hat{c}_{1,a} = .891$, SE = .021; $G^2(1) = 26.8$, $p < .001$) in line with the expectation that characteristic features of typical stimuli provide more valid information. Third, the defining-feature comparison had the same accuracy for both typical and atypical animals ($\hat{c}_{2,t} = .880$, SE = .028; $\hat{c}_{2,a} = .885$, SE = .028; $G^2(1) < 0.1$, $p = .864$), supporting the hypothesis that characteristic features do not affect the accuracy of the second-stage comparison. Figure 7 shows that this constrained model (i.e., $c_{2,t} = c_{2,a}$) had an excellent fit, as indicated by the empirical (gray) and the fitted (black) cumulative density functions (note that minor discrepancies close to zero are due to the fact that MADs are determined based on absolute deviations). This model was

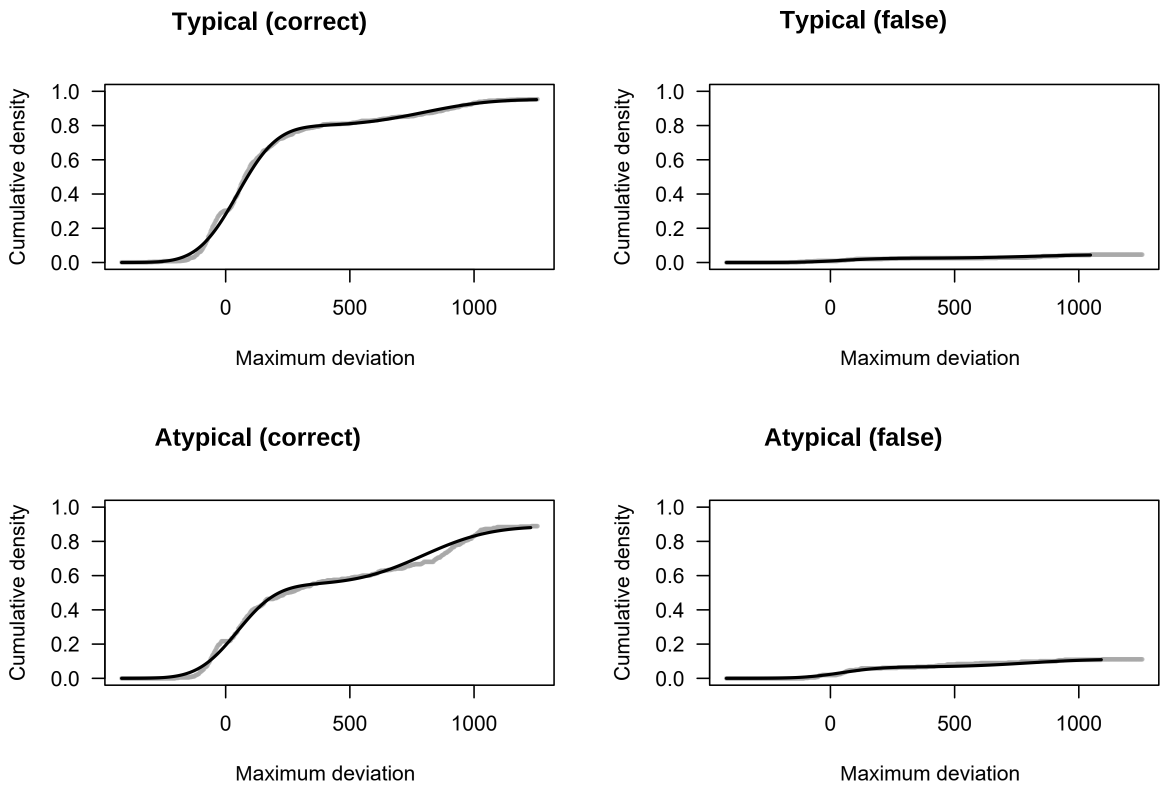

Figure 7. Empirical (gray) and fitted (black) cumulative density functions per response category, on the basis of the GPT version of the FCM with 9 parameters (i.e., with the constraint $c_{2,t} = c_{2,a}$).

also selected by the model-selection index BIC, with a posterior probability of .97 (and an AIC weight of .73) when assuming equal prior probabilities for the five competing GPT models (i.e., the general model, a single-component model, and the three constrained models $f_t = f_a$, $c_{1,t} = c_{1,a}$, and $c_{2,t} = c_{2,a}$).

In sum, the proposed GPT model provided a good account of the distribution of discrete responses and MADs. On the one hand, the model fitted the observed distributions of mouse trajectories very well, and on the other hand, the pattern of parameter estimates was in line with substantive hypotheses implied by the FCM (Smith et al., 1974).

5. Discussion

Generalized processing tree (GPT) models assume that discrete cognitive states affect discrete and continuous variables jointly, and thus predict a mixture distribution with weights determined by a processing tree structure. Similar to MPT models, the new model class is motivated by psychological theories that assume qualitatively distinct cognitive processes (Erdfelder et al., 2009). Moreover, GPT models are mathematically tractable, easy to apply using a flexible modeling syntax, and can increase the precision of parameter estimates relative to MPT models. In an empirical example, a theoretically motivated GPT model provided a good account of responses and the curvature of mouse trajectories in a semantic categorization task. Importantly, the hypothesized processing structure on the mixture weights allowed for a test of psychologically motivated predictions (e.g., concerning the accuracy of the two comparison processes for categorizing typical and atypical animals; Smith et al., 1974). Note that standard analyses of mouse-tracking data often discard faulty trials and do not account for the inherent ambiguity of responses (e.g., Dale et al., 2007), thereby incorrectly assuming a one-to-one mapping of cognitive processes to observed responses. Moreover, GPT models provide a conceptual and statistical framework to account for bimodality, a phenomenon that has been addressed in many mouse-tracking studies (Freeman & Dale, 2013).

Conceptually, the GPT framework differs from previous approaches to including continuous covariates in MPT modeling (Coolin, Erdfelder, Bernstein, Thornton, & Thornton, 2015; Heck et al., in press; Klauer, 2010; Michalkiewicz & Erdfelder, 2016). These alternative methods include external covariates to predict MPT parameters in a (logistic or probit) regression model. For instance, memory performance in the 2HTM could be modeled as a function of general intelligence. In contrast, GPT models assume that continuous variables are affected by the hypothesized cognitive states across trials, and thereby reverse the underlying causal reasoning for the psychological constructs. Overall, regression approaches are more often of interest when modeling inter-individual differences, whereas the proposed GPT framework primarily aims at testing effects of experimental manipulations or population characteristics on continuous and discrete variables. In principle, however, both approaches can be merged to investigate the two types of research questions jointly.

The proposed model class $\mathcal{M}^{\text{GPT}}$ makes a strong decomposition assumption, that is, category probabilities and continuous variables are modeled by separate parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, respectively. More precisely, the component distributions are assumed to be conditionally independent of the probabilities $\boldsymbol{\theta}$ of entering any of the cognitive states,

$$g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = g_{ij}(\boldsymbol{y}_k \mid \boldsymbol{\eta}). \tag{31}$$

In principle, the assumption of conditional independence can be relaxed, for instance, by assuming an evidence accumulation process that determines response probabilities and component distributions jointly (e.g., Donkin, Nosofsky, Gold, & Shiffrin, 2013). However, such generalizations of GPT models are not useful for the present purpose of developing a flexible, but mathematically tractable model class. First, the conditional independence assumption is at the core of threshold models that assume discrete processing states for recognition memory (Province & Rouder, 2012), source memory, (Batchelder & Riefer, 1990), visual working memory (Donkin et al., 2013), and word perception (Swagman, Province, & Rouder, 2015). Second, the substantive development and interpretation of complex functional relationships will in general be more difficult conceptually and mathematically. In contrast, when defining the GPT class as in the present paper, model development is separated into (a) theorizing about latent cognitive processes (i.e., developing an MPT-like structure based on the parameters $\boldsymbol{\theta}$), and (b) theorizing how these processes might affect continuous variables (i.e., choosing the distributions g_{ij} and constraining the parameters $\boldsymbol{\eta}$).

Despite its advantages, the GPT framework also has theoretical and practical limitations. Concerning the former, GPT models make the strong assumption that discrete states determine behavior. This assumption fits well with psychological theories that assume qualitatively distinct states such as dual-process theories (Sloman, 1996) or the FCM (Smith et al., 1974). However, the adaptation of GPT models for theories assuming graded signals within the cognitive system (e.g., familiarity, preference, or similarity) requires the assumption (a) that this fine-grained information is mediated by discrete states (Rouder & Morey, 2009), or alternatively, (b) that the GPT approach provides a sufficiently good approximation as a measurement model, even though the underlying process is assumed to be continuous (Batchelder & Alexander, 2013). However, the success of MPT models indicates that these are not very severe limitations.

A rather practical limitation concerns the assumption that trials are identically and independently distributed (Equations (5) and (6)), which is violated if participants or items are heterogeneous. As a remedy, Bayesian hierarchical modeling (Lee, 2011) could be adopted for GPT models, similar as for MPT models (e.g., Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Smith & Batchelder, 2010). According to such approaches, the core model structure in (3) is assumed to hold within each participant p conditional on the individual parameters $(\boldsymbol{\theta}_p, \boldsymbol{\eta}_p)$, which in turn follow a hierarchical distribution on the group level. Whereas this model structure is conceptually simple, it requires a sensible choice of priors and the development of an estimation algorithm (e.g., a Markov chain Monte Carlo sampler). Moreover, in the Bayesian analysis of mixture models, special consideration has to be given to the issue of label switching when sampling from the posterior if the component distributions of continuous variables are not sufficiently constrained (Frühwirth-Schnatter, 2001).

From a larger perspective, GPT models provide a conceptual framework to formalize how discrete psychological states affect observed behavior, irrespective of whether behavior is measured in terms of categorical or continuous variables or both. In the present paper, GPT models were applied successfully to a process-tracing measure that summarizes mouse trajectories. Similarly, GPT models can be applied to neurophysiological variables that have gained increasing interest in cognitive modeling recently (e.g., Forstmann & Wagenmakers, 2015). In principle, GPT models provide a framework for formalizing the idea that discrete cognitive states are associated with specific patterns of neural activation. On the basis of such an assumption, GPT models can be used to formalize this theoretical link, and to model behavioral and neuropsychological outcomes jointly (Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017).

In sum, the GPT framework provides a tractable approach for fitting psychologically motivated models to discrete and continuous data jointly. The model class follows the MPT philosophy according to which measurement models should provide a trade-off between flexibility and ease-of-use. Most importantly, however, GPT models show new directions for testing psychological theories that assume qualitatively distinct cognitive processes.

References

Bamber, D. & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. doi:10.1006/jmps.1999.1275

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H. & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin*, *139*, 1204–1212. doi:10.1037/a0033894

Batchelder, W. H. & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564. doi: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. doi:10.3758/BF03210812

Bröder, A. & Schütz, J. (2009). Recognition ROCs are curvilinear - or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606. doi:10.1037/a0015279

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 1–29. doi:10.1080/09658211.2013.767348

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Chernoff, H. & Lehmann, E. L. (1954). The use of maximum likelihood estimates in chi-square tests for goodness of fit. *The Annals of Mathematical Statistics*, *25*, 579–586. doi:10.1214/aoms/1177728726

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. doi:10.3758/s13423-014-0691-5

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*, 15–28. doi:10.3758/BF03195938

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Dixon, P. (2012). Assessing the evidence for response time mixture distributions. *Behavior Research Methods*, *44*, 706–724. doi:10.3758/s13428-012-0244-7

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*, 873–902. doi:10.1037/a0034247

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language*, *67*, 389–406. doi:10.1016/j.jml.2012.06.002

Efron, B. & Tibshirani, R. J. (1997). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

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

Forstmann, B. U. & Wagenmakers, E. (2015). *An introduction to model-based cognitive neuroscience*. New York, NY: Springer.

Freeman, J. B. & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*, 226–241. doi:10.3758/BRM.42.1.226

Freeman, J. B. & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*, 83–97. doi:10.3758/s13428-012-0225-x

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, *96*, 194–209. doi:10.1198/016214501750333063

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*, 678–694. doi:10.3758/BF03206550

Heck, D. W., Arnold, N. R., & Arnold, D. (in press). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*. doi:10.3758/s13428-017-0869-7

Heck, D. W. & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Heck, D. W. & Erdfelder, E. (in press). Linking process and measurement models of recognition-based decisions. *Psychological Review*. doi:10.1037/rev0000063

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, & Computers*, *31*, 689–695. doi:10.3758/BF03200747

Hu, X. & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/bf02294263

Kieslich, P. J. & Henninger, F. (in press). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*. doi:10.3758/s13428-017-0900-z

Kieslich, P. J., Wulff, D. U., Henninger, F., & Haslbeck, J. M. B. (2017). Mousetrap: An R package for processing and analyzing mouse-tracking data. *Version 3.1.0*. doi:10.5281/zenodo.581591

Klauer, K. C. (2001). Model Testing and Selection, Theory of. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 9927–9931). Oxford: Pergamon. doi:10.1016/B0-08-043076-7/00599-4

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

Koop, G. J. & Criss, A. H. (2016). The response dynamics of recognition memory: Sensitivity and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 671–685. doi:10.1037/xlm0000202

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

Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation*. New York, NY: Springer.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324. doi:10.3758/s13428-011-0168-7

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. doi: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. doi:10.1037/0278-7393.28.1.116

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. doi:10.3758/s13421-015-0567-6

Miller, J. (2006). A likelihood ratio test for mixture effects. *Behavior Research Methods*, *38*, 92–106. doi:10.3758/BF03192754

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

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In K. Lamberts & R. Goldstone (Eds.), *Handbook of Cognition* (pp. 422–436). Thousand Oaks, CA: Sage.

Myung, J. I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204. doi:10.1006/jmps.1999.1283

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, *6*, 155–156. doi:10.3758/BF03328004

Province, J. M. & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*, 14357–14362. doi:10.1073/pnas.1103880109

Rao, K. C. & Robson, B. S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics*, *3*, 1139–1153. doi:10.1080/03610927408827216

Read, T. R. C. & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer.

Rouder, J. N. & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology: Psychophysics, knowledge, representation, cognition, and measurement* (pp. 195–225). Mahwah, NJ: Erlbaum.

Rouder, J. N. & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*, 655–660. doi:10.1037/a0016413

Schmittmann, V. D., Dolan, C. V., Raijmakers, M. E., & Batchelder, W. H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. doi:10.3758/BRM.42.3.836

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. doi:10.1037/0033-2909.119.1.3

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241. doi:10.1037/h0036351

Smith, J. B. & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi: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. doi:10.1037/0096-3445.117.1.34

Swagman, A. R., Province, J. M., & Rouder, J. N. (2015). Performance on perceptual word identification is mediated by discrete states. *Psychonomic Bulletin & Review*, *22*, 265–273. doi:10.3758/s13423-014-0670-x

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65–79. doi:10.1016/j.jmp.2016.01.001

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424–465. doi:10.3758/BF03214357

Yantis, S., Meyer, D. E., & Smith, J. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, *110*, 350–374. doi:10.1037/0033-2909.110.2.350