# The Memory State Heuristic

## A tale of two cities
## (and their underlying memory states)


**Marta Castela**
**M.Sc.**

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 at the University of Mannheim.

| | |
|---|---|
| Dean of the School of Social Sciences: | Prof. Dr. Michael Diehl |
| Academic Director of the CDSS: | Prof. Dr. Edgar Erdfelder |
| Thesis Advisors: | Prof. Dr. Edgar Erdfelder |
| | Prof. Dr. Rüdiger Pohl |
| Thesis Reviewers: | Prof. Dr. Rüdiger Pohl |
| | Prof. Dr. Arndt Bröder |
| Date of Thesis Defense: | October $14^{th}$, 2016 |

# Contents

# Abstract

The recognition heuristic (RH) is one of the most prominent models of inferential decision making, but also one of the simplest. Its basic premise is straightforward: Whenever a decision maker is evaluating two objects according to a given criterion (e.g., population size), one object being recognized and the other not, recognition by itself is used to make an inference, ignoring all further knowledge one might have about the recognized object (Goldstein & Gigerenzer, 1999). Surprisingly, this simple strategy can be quite accurate. This accuracy stems from an exploitation of the environmental structure. In fact, for objects in many domains (e.g., world cities) there is a correlation between recognition and the corresponding criterion value (e.g., population size). For example, if a city name is recognized and another one is not, the former city is likely to be more populous than the latter one. Goldstein and Gigerenzer assumed this to be the case because recognition judgments are positively correlated with the criterion. However, Erdfelder, Küpper-Tetzel, and Mattern (2011) questioned whether it makes sense to rely on the recognition cue regardless of the *memory strength* associated with a certain recognition judgment. Specifically, they proposed that memory strength, and not recognition judgments per se, should influence reliance on recognition. Erdfelder et al. therefore proposed to extend the RH to the memory state heuristic (MSH) by incorporating the notions of a well-supported recognition memory model, the two-high-threshold model (Snodgrass & Corwin, 1988), into the theory. The MSH assumes that three orderly defined memory states can underlie recognition judgments - recognition certainty, uncertainty, and rejection certainty - and that, when comparing two objects, people should infer that the one in a higher memory state scores higher on the given criterion. Moreover, it predicts that the higher the distance between memory states of the objects under comparison, the higher should be the preference for objects in a higher state. This implies that the MSH should be used more often when the objects under comparison are in recognition certainty and rejection certainty, respectively,

than when they are in recognition certainty and uncertainty, or uncertainty and rejection certainty. It follows that, unlike the RH, the MSH's spectrum of predictions goes beyond so-called recognition pairs (pairs where one object is judged as recognized and the other as unrecognized), and cover any combination of objects in different memory-states.

Erdfelder et al. (2011) tested qualitative predictions of the MSH for recognition pairs, but some questions were not addressed by them. The present thesis describes a research program developed to overcome that gap and test some of the predictions that stem from the core assumptions of the MSH. The first manuscript tested qualitative predictions of the MSH which complement the work developed by Erdfelder et al. (2011). By relying on a simple assumption - that the uncertainty memory state is associated with longer recognition or rejection latencies than the certainty memory states - we tested the MSH predictions for three types of pairs: recognition pairs (one object is recognized and the other is not), knowledge pairs (both objects are recognized) and guessing pairs (none of the objects is recognized). In a second manuscript, we relied on a formal model to test the MSH against the RH and knowledge integration accounts. We found evidence in favor of the MSH across 16 published data sets. Finally, in a third manuscript, we developed and successfully tested a new paradigm and formal model of the MSH, which incorporates its predictions for all possible combinations of memory states.

In sum, the present thesis describes converging support for the MSH. From qualitative predictions to tests of a formal implementation of the MSH, I show how memory states predict reliance on recognition and are correlated with the criterion value. Therefore, I conclude that the inspiring but nevertheless simplistic RH theory as originally proposed by Goldstein and Gigerenzer (1999) must be abandoned in favor of an account that gives a more fine-grained characterization of the underlying mnemonic processes involved in inferential decision making.

# Manuscripts

This thesis is based on three manuscripts which have been published or are currently submitted for publication in peer-reviewed journals. The manuscripts are listed below and appended to this thesis in the order in which they will be discussed.

1. Castela, M., & Erdfelder, E. (2016). *Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences.* Manuscript submitted for publication.

2. 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.

3. Castela, M., Erdfelder, E. (in press). The Memory State Heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

# Introduction & Theoretical Background

It is hard to deny that there is something remarkable about the feeling of recognition. Despite the frequency and easiness with which it occurs, recognition seems to carry along a considerable amount of information with it. We recognize people that we have seen or met before (even if only once), brands or products that we have encountered or were exposed to somehow, names of books we may or not have read before, etc. Along with all those feelings of recognition, some information can be inferred. Anyone who ever watched a program like "Who wants to be a Millionaire" will probably relate to the feeling of thinking you know the correct answer, while at the same time not being able to retrieve any argument for it other than the fact that the option sounds familiar. We find ourselves compelled to produce a response on the basis of familiarity alone. But can those inferences be accurate?

Goldstein and Gigerenzer (1999; see also Gigerenzer & Goldstein, 1996) thought so. Based on the remarkable human ability to distinguish between what we have or have not experienced before, they proposed a decision strategy that exploits that simple distinction - the recognition heuristic (RH). The heuristic is defined within the context of inferring which of two objects scores higher on a given criterion as follows: "If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value" (Goldstein & Gigerenzer, 1999, p. 41). Importantly, they clearly stated that, in this context, what they mean by recognition involves a division between the previously experienced and the novel. While this distinction is sensible, it ignores important aspects, such as the degree of familiarity of recognized items. Goldstein and Gigerenzer assumed the RH would ignore such information, since once one object is recognized and another one is not, an inference

can be made on the grounds of recognition alone. However, differences in the degree of familiarity between recognized objects can be very marked, which brings up the question: Do different degrees of familiarity impact reliance on recognition as a single cue? This thesis explores this important question by testing an extension of the RH which assumes that the memory states underlying recognition judgments, and not the judgments per se, correlate with the criterion value and can be used to make inferences. This extension has been introduced by Erdfelder et al. (2011) and named the memory state heuristic (MSH). However, they only tested a few qualitative predictions derived from a verbal model of the MSH. In this thesis I extend these initial tests to cover the full spectrum of predictions of the MSH, test it against competing models, and ultimately develop and test a formal implementation of it. Before introducing the MSH, in the next section I will address the RH in more detail by bringing it into context and discussing the major findings around it.

## 1.1 The Recognition Heuristic

According to a classical view of human reasoning, inferences are rational by virtue of the use of the laws of probability and statistics. In turn, deviations from those laws are perceived as errors (e.g., Gigerenzer & Goldstein, 1996; Tversky & Kahneman, 1974). However, given that most decisions are made under limited time and cognitive resources, is it reasonable to expect judgments under uncertainty to perfectly follow those laws? Gigerenzer and Goldstein revived Herbert Simon's notions of *bounded rationality* and *satisficing* (a combination of sufficing and satisfying) arguing that accurate inferential strategies can deviate from classical norms of rational inference. Briefly, these two concepts can be summarized by the idea that organisms are bounded by limited external and internal resources (e.g., time, available information, processing speed, memory, etc) and therefore will accept "good enough" solutions rather than always try to optimize (e.g., Simon, 1956). It is in fact one of Simon's quotes that best describes Gigerenzer and Goldstein's understanding of human reasoning: "Human rational behavior is shaped by a scissors whose two blades are the structure of task environments and the computation capabilities of the actor" (Simon, 1990, p., 7).

Gigerenzer and Goldstein's (1996; Goldstein & Gigerenzer, 1999, 2002) approach to human reasoning aimed at demonstrating how the use of heuristics can lead to accurate inferences through the exploitation of the structures of the environment. In other words, they intended to show that heuristics are *ecologically rational*, such that they can adaptively explore the structure of the environment, leading to good inferences with minimal effort. Hence, the fast and frugal heuristics program was born. This program describes a metaphorical *adaptive toolbox*, that is, a set of adaptive heuristics which can be used to make accurate inferences. One of the most studied examples within the adaptive toolbox is the RH. As described above, the RH relies on a single cue, recognition, to make inferences. It is therefore proposed to function as a one-reason decision making process, since it bases judgments on recognition alone, ignoring all further cues.

### 1.1.1 Ecological Rationality of the RH

The ecological rationality of the RH depends on three concepts: *ecological correlation*, *surrogate correlation* and *recognition validity* (Goldstein & Gigerenzer, 1999). If one were asked to estimate which of two cities is larger, the fact that one recognizes one city but not the other would be a good indicator that the recognized city is larger. This occurs because in that domain recognition positively correlates with the size of cities, meaning that *recognition validity* (proportion of recognition pairs for which the recognized option is the correct one) is high. That correlation occurs through the interaction with mediators. For example, the fact that larger cities are more likely to be mentioned in TV, which corresponds to the *ecological correlation*. This, in turn, creates a *surrogate correlation*: the number of times the city is mentioned in TV will positively correlate with the probability that that city is recognized. Some preconditions are necessary for the RH to be a good strategy (Gigerenzer & Goldstein, 2011). First, it is only a good strategy for domains where recognition is a valid cue. Second, recognition should be natural, learned from the interaction with the environment, and not manipulated in the laboratory. Finally, the RH is a model for inferences from memory, not from givens, meaning that other cues should not be made readily available. These three preconditions determine what should be the ideal paradigm for studying the RH. In what follows, I describe

that paradigm, which I adhered to (sometimes with necessary extensions) in all studies reported in this thesis.

### 1.1.2 Paradigm and Measurement of RH-use

The typical paradigm for investigating the RH involves a comparison task (usually a two-alternative forced choice task) where items from a certain domain are compared regarding a given criterion. Additionally, a recognition test is performed where all items must be judged as recognized or not. This paradigm has been applied in many domains from the length of rivers or size of islands (e.g., Hilbig & Pohl, 2008) to the success of celebrities or musicians (Michalkiewicz & Erdfelder, 2015), among many others. However, the most vastly used domain is the population size of cities, and the paradigm is in these cases referred to as the city-size task. For example, in the domain of World cities, a typical trial in the comparison task would be "Delhi - Foshan". Following the RH, if a given participant recognized *Delhi* but not *Foshan*, he or she should infer that *Delhi* has a larger population. In this case, "Delhi - Foshan" would be a so-called *recognition pair*, a pair where one object is recognized and the other is not. Additionally, there will be pairs where both objects are recognized, so-called *knowledge pairs*, and pairs where both objects are not recognized, so-called *guessing pairs*. Importantly, only recognition pairs are appropriate for use of the RH, since the heuristic cannot be applied to the other cases.

The measurement of RH-use was initially done by relying on adherence rates, that is, the proportion of times the recognized option is chosen in recognition pairs. However, this is an inherently biased measure, since by simply looking into choice patterns it is not possible to discriminate between use of the RH or reliance on other strategies, like knowledge-use or even guessing (e.g., Hilbig, 2010). In other words, this means that the fact that the recognized option is chosen in a recognition pair does not imply that recognition alone motivated that choice (e.g., Hilbig, Erdfelder, & Pohl, 2010; Hilbig & Pohl, 2008).

Different measures have been proposed to improve the estimation of the RH (see Hilbig & Pohl, 2008; Pachur, Mata, & Schooler, 2009), but one of them stands out (Hilbig, 2010) and has proven very fruitful in the last years (e.g., Castela, Kellen,

Erdfelder, & Hilbig, 2014; Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015; Horn, Pachur, & Mata, 2015; Michalkiewicz & Erdfelder, 2015; Schwikert & Curran, 2014), the r-model. The r-model was proposed by Hilbig et al. (2010) to estimate RH-use. It belongs to the class of multinomial processing tree models (Batchelder & Riefer, 1999; Erdfelder et al., 2009), which postulate a set of latent discrete states as the basis for observable categorical responses. Each latent state is associated with a parameter, which represents the probability of its occurrence. Multinomial processing tree models can be represented by a tree structure, with each branch representing the sequence of presupposed processes that should lead to a specific response category. In the last decades, they have become an increasingly attractive tool for psychologists, and have been successfully applied in a variety of domains, including recognition memory (Snodgrass & Corwin, 1988), preference construction (Erdfelder, Castela, Michalkiewicz, & Heck, 2015), consensus analysis (Romney, Weller, & Batchelder, 1986), and attitude measurement (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005), among others (see Erdfelder et al., 2009, for a comprehensive review of many applications).

The r-model has three trees and four parameters. Each tree corresponds to a type of pair, namely knowledge, recognition and guessing pairs (see Figure 1.1).
In both the knowledge and guessing trees, a single parameter accounts for accuracy, the $b$ parameter and the $g$ parameter, respectively. In the recognition tree, the probability of using the RH-use is estimated through parameter $r$. Additionally, parameter $a$ accounts for recognition validity (the proportion of times that choosing the recognized object in recognition pairs leads to a correct inference). Therefore, when using the RH, if the recognized object is the one scoring higher on the criterion, the inference will be correct with probability $r \cdot a$, if it is not, it will be wrong with probability $r \cdot (1 - a)$. Whenever the RH is not used, accuracy will depend on the validity of knowledge $b$ (or other judgment strategy taking place).

The RH has inspired a lot of research in the last decades. Along with demonstrations of its impressive ability to make fast and frugal yet accurate inferences in a vast diversity of domains (e.g., Goldstein & Gigerenzer, 1999, 2002; Pachur & Hertwig, 2006; Richter & Späth, 2006; Scheibehenne & Bröder, 2007; Serwe & Frings, 2006), it also led to a wave of criticism. Certainly, the most challenged aspect of the

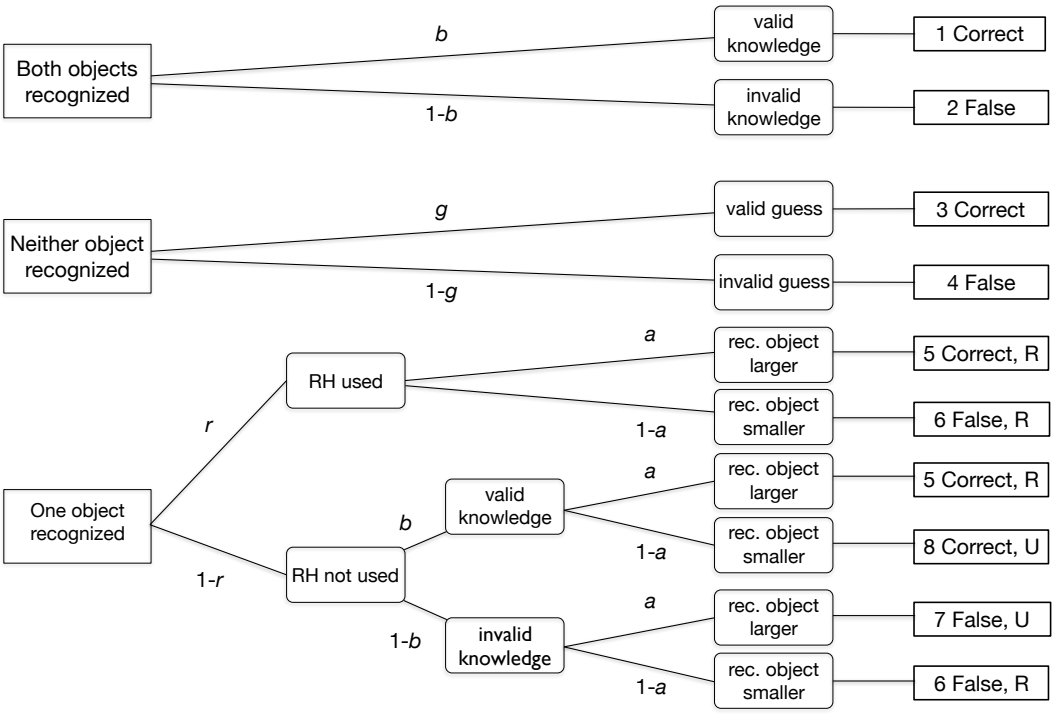



Figure 1.1: Graphical representation of the r-model: Parameter $r$ denotes the probability of applying the recognition heuristic as originally proposed, that is, by ignoring any knowledge beyond recognition. $a$ = recognition validity (probability of the recognized object representing the correct choice in a recognition case); $b$ = probability of valid knowledge; $g$ = probability of a correct guess; rec. = recognized; R = recognized; U = unrecognized.

RH was its noncompensatory nature (e.g., Bröder & Eichler, 2006; Hilbig & Pohl, 2008, 2009; Newell & Fernandez, 2006; Richter & Späth, 2006). In different ways, the assumption that further knowledge is ignored whenever recognition is diagnostic for inferences has been questioned time and time again. At times, the criticism has been so harsh as to question the RH altogether due to an inability to find evidence for its main premises (Newell & Fernandez, 2006). But what if a simple extension of the RH that replaces recognition judgments with memory strength could (1) accommodate all those challenging findings, (2) and extend the spectrum of predictions? Erdfelder et al.'s (2011) memory state heuristic (MSH) offers a promising start.

In the next section, I will address the role of recognition memory in the RH literature and describe the MSH in detail. Then, I will demonstrate how the MSH

redefines the literature by reviewing several findings which can be reinterpreted through its lenses. Finally, in the *Summary of Manuscripts* section I will put forward the building blocks of my thesis by describing a complete research program to test the MSH which addresses the questions left unanswered by Erdfelder et al. (2011).

## 1.2 The Memory State Heuristic

### 1.2.1 Recognition Memory in the RH literature

One important yet largely overlooked aspect regarding the RH is the nature of the recognition process underlying the recognition judgments on which the RH operates. While the basis of the RH is a memory process, Goldstein and Gigerenzer (1999, 2002) assumed that the heuristic operates only on the output of that process, i.e., the recognition judgments, and that the process itself may be ignored. As outlined above, this assumption is rather questionable, especially in as much as it implies that differences in familiarity between recognized objects are inconsequential for reliance on the RH. This simplification has been often questioned (e.g., Dougherty, Franco-Watkins, & Thomas, 2008; Erdfelder et al., 2011; Hilbig & Pohl, 2008) and explicit calls for theory integration have been made (e.g., Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011; Pohl, 2011; Tomlinson, Marewski, & Dougherty, 2011). The few attempts done so far to link theories of recognition memory with the RH have demonstrated that this exercise helps not only to understand RH-use better, but also to draw new predictions (Pachur et al., 2011).

Schooler and Hertwig (2005), for instance, have integrated the RH within the ACT-R cognitive architecture (Anderson, Bothell, Lebiere, & Matessa, 1998), which involves a model of memory. The advantage of implementing the RH within the ACT-R is that it enables a direct assessment of how differences in memory can affect it. Specifically, they have shown that a moderate level of forgetting is beneficial for the heuristic, as it creates partial ignorance. However, they have relied on an all-or-none notion of recognition, aligned with Goldstein and Gigerenzer's (2002) definition. Therefore, while they take a step towards theory integration, they do not improve on the simplified understanding of the recognition memory process associated with the RH.

Pachur and Hertwig (2006) also linked recognition memory theories with the RH. They investigated whether the well documented distinction between familiarity and recollection (e.g., Kelley & Jacoby, 2000) is relevant for the RH. Specifically, they assumed that recognition processes relevant for use of the RH only regard familiarity, while retrieval of other cues implies recollection. Since familiarity is known to enter the mental stage earlier than information which needs to be recollected, they aimed at demonstrating the retrieval primacy of recognition within the context of the RH. In fact, they found evidence that recognition-based inferences are faster than choices inconsistent with recognition. Moreover, that RH-use increases with time pressure, which they interpreted as support for the retrieval primacy of recognition.

Schwikert and Curran (2014) also used the distinction between familiarity and recollection to investigate the RH further. By using event-related potentials, they found evidence suggesting that mostly familiarity processes are involved in RH-use. These two related approaches represent another attempt to link theories of recognition memory with the RH, but again, they do not tackle the question of whether different levels of familiarity (or memory strength), will be associated with a differential use of recognition as a single cue.

Furthermore, Pleskac (2007) has relied on signal detection theory in order to demonstrate that the accuracy of memory affects the validity of the RH. Specifically, he has shown that with the increase of false alarms (recognizing an item that has not been experienced before) and misses (failing to recognize an item that has been experienced before) the accuracy of the RH decreases. Pleskac's work points out the important fact that memory is not perfect and, by implication, if recognition judgments per se are the base for inferences, the ecological rationality of the RH can be compromised by memory errors (see also Erdfelder et al., 2011).

Dougherty et al. (2008) have also pointed out that the all-or-none treatment of recognition adopted by Gigerenzer and Goldstein (1996) violates known aspects of recognition memory, namely the fact that recognition is based on a continuous underlying memory variable. While Gigerenzer, Hoffrage, and Goldstein (2008) countered that argument by clarifying that the RH is not a model of memory processes but of how inferences are made from the output of those processes, and that the notion of binary recognition judgments is well integrated with the recognition

memory literature, the point remains of whether something can be learned from a deeper consideration of the memory processes. Dougherty et al. (2008) implemented both the RH and a familiarity-based model in a simulation. The familiarity-based model compares items regarding their echo intensity, which corresponds to the sum of activation levels of all traces present in memory for that item (see Hintzman, 1988, for more details). Whenever two objects differed in their echo intensity, the model chose the one with a higher one. This means that not only recognition pairs can be compared in terms of recognition, but all pairs for which echo intensity differs. Through this simulation, they demonstrated that a familiarity-based model can explain results observed for the RH.

All the studies described above have contributed to decreasing the unfortunate distance between the recognition memory literature and the RH. Certainly, understanding the role of forgetting in RH-use (Schooler & Hertwig, 2005) and how familiarity (versus recollection) seems to be the driving memory process behind the heuristic (Schwikert & Curran, 2014), allowing it to drive fast inferences (Pachur & Hertwig, 2006), helps us realize the relevance of the memory processes involved. Moreover, the added value of considering the memory processes more carefully is well demonstrated by the fact that the accuracy of the heuristic is affected by memory errors, which are themselves a function of mnemonic (e.g., sensitivity) and decision making processes (e.g., response bias; Pleskac, 2007). Finally, the work of Dougherty et al. (2008) demonstrates that the spectrum of application of an heuristic relying on recognition does not need to be limited to comparisons between one object judged as recognized and another judged as unrecognized. Nevertheless, an important step is missing. The RH theory and its notion that binary judgments determine reliance on recognition should be replaced by a framework that considers the memory processes as the relevant information on which inferences are based. In the next section I describe such an approach.

### 1.2.2 Threshold Models of Recognition Memory and the RH

Erdfelder et al. (2011) developed a framework with provides "(...) a formal link between (1) the memory strengths of choice option names - a latent variable which is

affected by environmental frequency and previous processing - and (2) binary recognition judgments for choice option names - an empirical variable which is assumed to affect decision behavior" (Erdfelder et al., 2011, p. 8). While there are different ways to link memory processes with recognition-based inferences, Erdfelder et al. focused on a rather straightforward extension of the RH. Specifically, they proposed to extend the RH to the MSH, a framework which assumes that three memory states can underlie the binary recognition judgments. The central idea of this framework is that those memory states, and not recognition judgments per se, will influence reliance on recognition.

The MSH is based on the two-high-threshold model (Snodgrass & Corwin, 1988), a well-supported model of recognition memory (e.g., Bröder, Kellen, Schütz, & Rohrmeier, 2013; Kellen & Klauer, 2015). The two-high-threshold model assumes that recognition judgments are determined by three underlying memory states, namely recognition certainty, uncertainty and rejection certainty. The two-high threshold model is a multinomial processing tree model (Batchelder & Riefer, 1999; Erdfelder et al., 2009) with two trees, one for items experienced before, and another for non-experienced items (see Figure 1.2). An object that has been experienced before will enter the recognition certainty state with probability $r$ if the memory strength associated with that object exceeds the high threshold. Whenever that happens, a *yes* recognition judgment will be given. If the memory strength associated with that object lies below the high threshold $(1 - r)$, the object will be in uncertainty, and a second process of guessing will occur. In this case, with probability $g$ a correct *yes* judgment will be given, and with probability $1 - g$ an incorrect *no* judgment will be given. In the tree for new items, the logic is analogous. Whenever the memory strength of a new object lies below the rejection threshold (with probability $d$), the object will be in rejection certainty and a correct *no* judgment will be given. If the memory strength lies above the rejection threshold (with probability $1 - d$), the object will be in uncertainty and, again, a guessing process will determine whether a correct *no* judgment (with probability $1 - g$) or an incorrect *yes* judgment (with probability $g$) will be given.

Building up on the two-high-threshold model, the MSH operates under two assumptions: Whenever comparing two objects in different memory states, (1) there

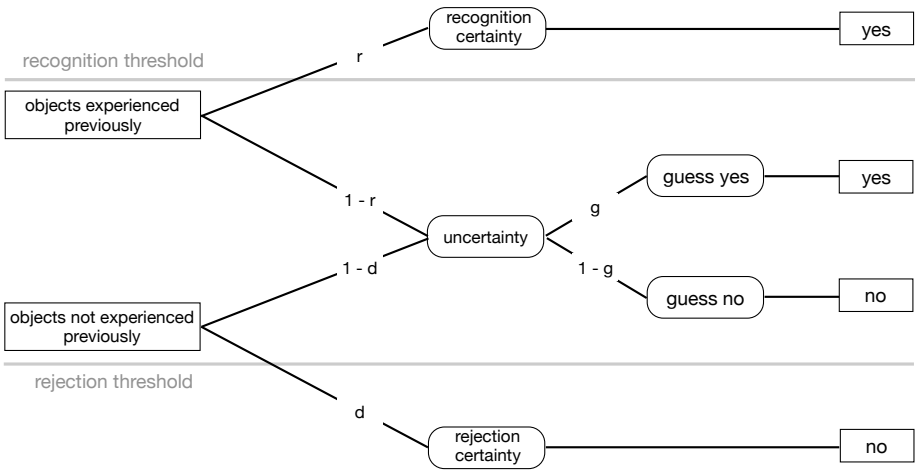



Figure 1.2: Graphical representation of the two-high-threshold model. Parameter $r$ denotes the probability of old objects exceeding the recognition thresholds. Parameter $d$ denotes the probability of new objects exceeding the rejection threshold. Parameter $g$ denotes the conditional probability of guessing *yes* in the uncertainty state.

is a preference for the object in a higher state; (2) the larger the distance between the memory states of the objects under comparison, the larger is the probability of following the MSH. From these two assumptions it is possible to derive predictions for any combination of memory-states that involves two different ones. Whenever the objects in a pair are in the same state, the MSH cannot be applied. For those cases, inferences may rely on knowledge (if available), guessing processes, or other inferential strategies.

Erdfelder et al. (2011) tested some of the predictions of the MSH by relying on a serial processing interpretation of the two-high-threshold model that assumes that each cognitive stage involved in a given branch is processed sequentially (Batchelder & Riefer, 1999; Heck & Erdfelder, in press). It follows that the number of cognitive processing stages on each branch influences its total processing time. In the case

of the two-high-threshold model, this leads to the prediction that the response time distributions associated with the uncertainty state are stochastically larger than the ones associated with the certainty states, given the additional processing stage of guessing in the former (Heck & Erdfelder, in press). Following from this, predictions can be drawn regarding recognition latencies and decision times. Specifically, Erdfelder et al. tested the following predictions:

1. RH accordance rates increase with decreasing recognition and rejection latencies, and these effects are additive.

2. Decision latencies in recognition pairs increase with both the recognition latency of the recognized object and the rejection latency of the unrecognized object, and these effects are additive.

3. Response bias manipulations (aimed at selectively affecting the guessing probability) in the recognition test affect recognition judgments but not performance in the comparison task.

Besides finding support for all of these hypotheses, Erdfelder et al. demonstrated how the MSH framework allows a new interpretation of previous findings which had challenged the RH. In the next session, I will describe those findings under the light of the MSH.

### 1.2.3 Old findings, New explanations

Stating that recognition judgments alone determine inferences for recognition pairs in domains where recognition is valid is certainly a bold assumption underlying the RH. Unsurprisingly, it has been shown that other factors also play a role. Specifically, it has been shown that (1) the preference for the recognized object is stronger for objects recognized faster (Hertwig, Herzog, Schooler, & Reimer, 2008; Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2010; Newell & Fernandez, 2006); (2) recognition pairs for which recognition leads to a correct inference are chosen more often than recognition pairs for which recognition leads to an incorrect inference (Hilbig & Pohl, 2008); (3) recognition pairs for which there is further knowledge about the recognized object are preferred over recognition pairs for which the recognized object is merely recognized (Hilbig & Pohl, 2008; Hilbig, Pohl, & Bröder,

2009; Marewski et al., 2010; Newell & Fernandez, 2006; Pohl, 2006). These findings seem to suggest that fluency and further cue knowledge are also taken into account when making inferences about recognition pairs. But what if the MSH would be sufficient to explain all of these? Erdfelder et al. (2011) convincingly demonstrated that this is the case, since the MSH clearly predicts the findings described. First, recognition accordance rates should decrease with recognition latencies (as stated in the first prediction listed in section 1.2.2), because objects recognized faster are more likely to be in recognition certainty than objects recognized slower.

Second, accordance rates should be larger for recognition pairs when RH-consistent decisions are correct because the MSH should be more valid and followed more often whenever the recognition pair originates from certainty memory states. In other words, different combinations of memory states can underlie a recognition pair, but according to the MSH, these different combinations should be treated differently. Specifically, recognition pairs that originate from recognition and rejection certainty states should lead to higher accordance rates and, in turn, to more correct inferences. This is in line with Hilbig and Pohl (2008).

Finally, the higher accordance rates for recognition pairs when there is further knowledge about the recognized object versus when the recognized object is merely recognized invite the explanation that further knowledge is being used. But this too can be explained by the MSH under a very reasonable assumption. Specifically, this requires only the assumption that recognized objects for which further knowledge is present are more likely to originate from a recognition certainty memory-state than objects which are merely recognized. From this assumption it follows directly that the MSH predicts the observed result, since accordance rates should be higher when the recognized object is in the certainty state than when it is in the uncertainty state.

Besides these findings regarding the differential treatment of different recognition pairs, Erdfelder et al. (2011) also showed that the MSH can explain results involving knowledge pairs. The RH, on the other hand, only makes predictions for recognition pairs. However, another heuristic has been put forward to describe how fast and frugal inferences can be made in knowledge pairs - the *fluency heuristic* (Hertwig et al., 2008; Schooler & Hertwig, 2005; see also Pohl, Erdfelder, Michalkiewicz, Castela,

& Hilbig, in press). The fluency heuristic is defined as follows: "If two objects, *a* and *b*, are recognized, and one of two objects is more fluently retrieved, then infer that this object has the higher value with respect to the criterion" (Hertwig et al., 2008, p. 1192). In fact, Hertwig et al. showed that, in line with the fluency heuristic, there is a preference for the object recognized faster in knowledge pairs. However, the MSH also makes this prediction, since the object recognized faster in a knowledge pair is more likely to be in a recognition certainty memory-state. It follows that, while the RH and the fluency heuristic could be invoked together to explain observations in recognition and knowledge pairs, the MSH alone predicts all those observations.

In sum, the MSH can explain several intriguing results regarding recognition cases and also knowledge cases under a single decision heuristic. Furthermore, it can also make predictions for guessing cases. As long as objects are in different memory states, the MSH predicts a preference for the one in a higher state. Moreover, it predicts that this preference should be higher whenever the distance between the memory state of the objects under comparison is maximal (one object in the recognition certainty state and the other in rejection certainty). If the objects are in adjacent memory states (recognition certainty and uncertainty or uncertainty and rejection certainty) the preference for the one in a higher state should be less marked.

Although Erdfelder et al. (2011) already provided evidence for some of the MSH predictions and showed how it allows a new interpretation of previously intriguing results, important predictions were left untested. Moreover, the MSH was only developed as a verbal model. The present thesis describes the execution of a research program aimed at addressing those questions and ultimately testing the MSH by implementing it in a formal model. This research program is developed in three manuscripts, which will be summarized in the next section.

# Summary of Manuscripts

In this section I summarize the three manuscripts on which this thesis is based. The focus will be on the main research question and the contribution to the literature. For the sake of brevity, I will not address most details, including the method and an exhaustive description of all results, since these can be found in the manuscripts appended. After these summaries, I will draw some overall conclusions clarifying the connection between all manuscripts and the general contribution of this thesis. Finally, I will discuss limitations of my work and possible future directions.

## 2.1 On the relation between recognition latencies and inference strategies

Castela, M., & Erdfelder, E. (2016). *Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences.* Manuscript submitted for publication.

As discussed above, Erdfelder et al. (2011) tested core predictions of the MSH by assessing how RH accordance rates are affected by recognition latencies. The rationale behind this lies on the assumption (derived from a serial interpretation of the two-high-threshold model) that certainty memory-states are associated with shorter recognition and rejection latencies than the uncertainty memory-state (Erdfelder et al., 2011; Heck & Erdfelder, in press). Specifically, they showed that accordance rates decrease with increasing recognition and rejection latencies, and that those effects are additive. Also, that decision latencies for recognition pairs increase with the recognition and rejection latencies of the recognized and the unrecognized objects, respectively, and that, again, those effects are additive. However, they did not

address the predictions for knowledge and guessing cases. Moreover, they relied on accordance rates to test the predictions for recognition cases, but better methods for estimating RH-use are available.

We conducted two studies that complemented Erdfelder et al. (2011). In our first study, we tested the MSH predictions regarding the effect of recognition latencies on inferences for knowledge and guessing pairs. These predictions follow the same logic as the predictions for recognition cases, as they too stem from the two core assumptions of the MSH: 1) If objects are in different memory states, there should be a preference for the one in the higher state; 2) this preference should increase with the distance between the states. Therefore, assuming the association between longer recognition or rejection latencies and the uncertainty state, in knowledge pairs there should be a preference for the faster recognized object, and in guessing pairs there should be a preference for the object rejected more slowly.

Regarding the size of the effect, it is important to note that for recognition pairs there are four possible memory-state combinations underlying the $yes - no$ recognition judgements (recognition certainty and rejection certainty, recognition certainty and uncertainty, uncertainty and rejection certainty, uncertainty and uncertainty). These cover all possible distances between memory-states, from maximal distance to same state cases. However, for both knowledge and guessing pairs, only three memory-state combinations are possible (certainty and certainty, certainty and uncertainty, uncertainty and uncertainty). Therefore, the highest distance that can underlie a knowledge or guessing pair is the case of adjacent memory-states. This leads to the prediction that the preference for the object in a higher state in knowledge and guessing cases should be weaker than what was observed for recognition cases, where a maximal memory-state distance can occur. Accordingly, we observed a consistent, although not large, preference for objects that are likely to be in a higher memory state, both for knowledge and guessing cases (see Figure 2.1).

In a second study, we addressed recognition cases, but using a superior method for estimating RH-use than Erdfelder et al. (2011). When investigating the association between latencies and RH-use, Erdfelder et al. relied on RH accordance rates. This necessarily leads to a bias in the estimation of RH-use, since accordance rates cannot disentangle between the choice of the recognized object due to use of the

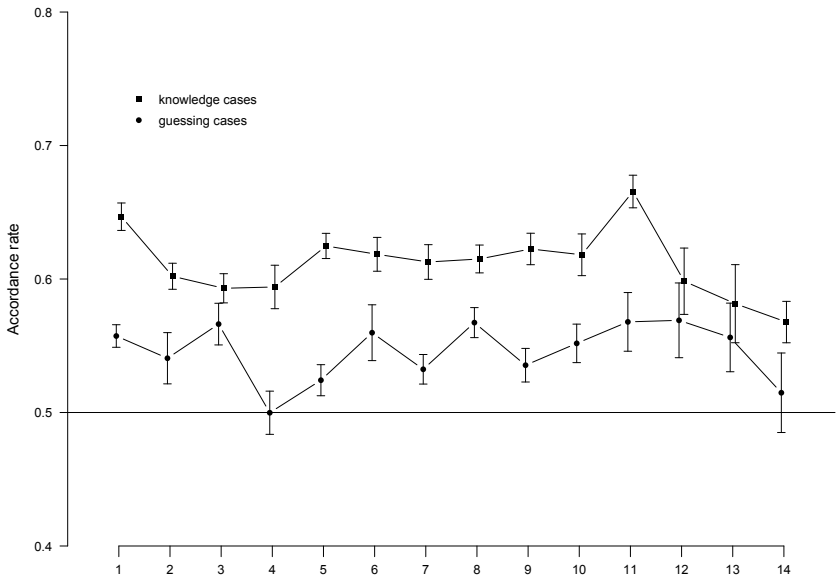



Figure 2.1: Proportion of choices of the fastest or slowest recognized or unrecognized object for knowelde and guessing cases, respectively, for all 14 reanalyzed datasets. Error bars represent standard error of the mean.

RH or due to reliance on further knowledge or any other strategy. Therefore, we wanted to replicate their finding using the r-model (Hilbig et al., 2010). Specifically, we aimed at showing that shorter recognition and rejection latencies are associated with higher $r$ estimates. To test this, we fitted the r-model to four subsets of our data sets. Each subset contained only objects in one of the four quartiles of the individual recognition and rejection latency distributions. When fitting the r-model to the data split into the four subsets, we obtained four different $r$ parameters, and could then test our prediction that $r$ should be higher in the subsets which contained objects with faster recognition and rejection latencies. This could be described as an order restriction such that the $r$ parameters decrease from $r_1$ to $r_4$, with 1 corresponding to the first quartile of the distributions (only the fastest recognized and unrecognized objects are included) and 4 the last quartile (only the slowest recognized and unrecognized objects are included). In short, we found support for our hypothesis (see Figure 2.2).

In sum, this first manuscript consolidated the support Erdfelder et al. (2011) found for the MSH by extending the tests of the MSH's latency predictions to

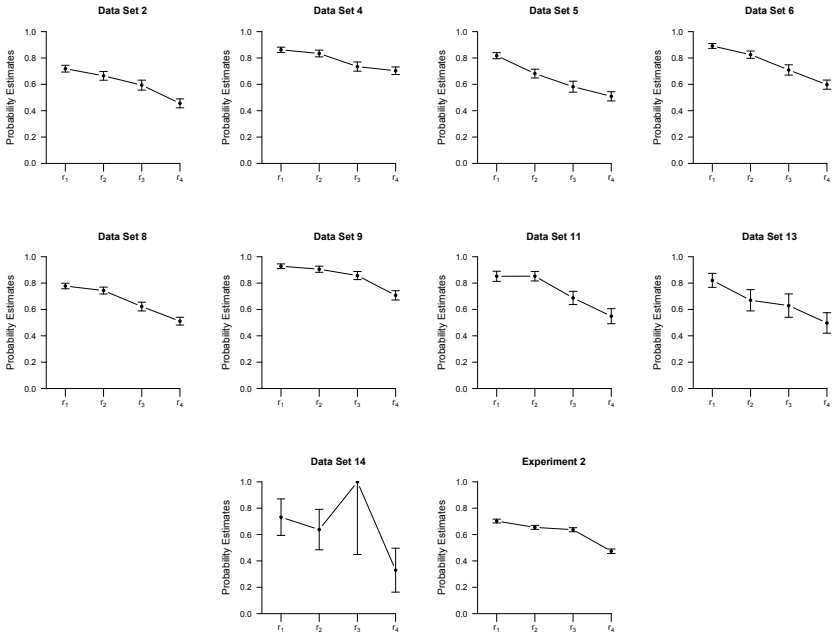



Figure 2.2: $r$ probability estimates in all four quartiles of recognition and rejection latency distributions for all reanalyzed datasets and for Experiment 2. Error bars represent standard errors.

knowledge and guessing cases, and finding converging support for its core prediction regarding recognition cases with a superior method for assessing RH-use.

## 2.2 Competitive testing of the MSH

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.

The goal of the second manuscript was to test the MSH against both the RH and an account which assumes knowledge integration. This was possible by comparing RH-use for two types of recognition pairs, which differ regarding the subjective experience associated with the recognized object. Specifically, we compared recognition pairs for which the recognized object was said to be merely recognized (mR) or recognized along with further knowledge ($R^+$). Comparing RH-use for these two types of pairs was ideal because the RH, the knowledge integration account and the

MSH all make different predictions. According to the RH, RH-use should not differ between the two types of pairs, because once an object is recognized, the availability of further knowledge is inconsequential. In contrast, knowledge integration accounts predict that RH-use should be lower for pairs involving a $R^+$ object than pairs involving a mR object, because if knowledge is available it should be integrated into the decision, leading to a decrease in reliance on the RH. Finally, the MSH predicts the opposite pattern, that is, it predicts that RH-use will be larger in pairs involving a $R^+$ object than pairs involving a mR object. This prediction follows from the reasonable assumption that objects for which further knowledge is available ($R^+$) are more likely to originate from a recognition certainty memory-state than objects that are merely recognized (Erdfelder et al., 2011).

In order to test these three predictions, we relied on a simple extension of the r-model, the r*-model. This extension is straightforward, and essentially involves duplicating the tree for recognition pairs in the original r-model, such that there is one tree for recognition pairs involving a $R^+$ object and another for recognition pairs involving a mR object[1]. In this way, we can separately estimate RH-use for both types of pairs. Importantly, this allows us to represent the three different accounts through different parameter restrictions. These can be summarized as follows:

**RH** $r_1 = r_2$

**knowledge-integration** $r_1 < r_2$

**MSH** $r_1 > r_2$

where $r_1$ is the estimate of RH-use for recognition pairs involving a $R^+$ object and $r_2$ the estimate of RH-use for recognition pairs involving a mR object.

In sum, through the reanalysis of 16 published data sets, we consistently found that RH-use was higher for recognition pairs involving a $R^+$ object compared to pairs involving a mR object. These results strongly supported the MSH, and could not be accommodated by the other two accounts.

[1]For the sake of simplicity, I omit here the extension regarding the knowledge tree. However, all details can be found in Castela et al. (2014)

## 2.3 Developing and testing a formal model of the MSH

Castela, M., & Erdfelder, E. (in press). The memory state heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

The first two manuscripts accumulated significant amount of evidence in favor of the MSH. First, the predictions of the MSH appear to hold not only for recognition pairs, but also for knowledge and guessing pairs. Second, when critically tested against the RH and knowledge integration accounts, the MSH came out as the best model. However, both manuscripts are limited by the fact that they do not measure MSH-use directly, but must rely on assumptions (association between latencies and memory-states in the first manuscript, and association between availability of further knowledge and memory-states in the second manuscript) to test it. As stated in Castela et al. (2014, p. 1137), "Despite the plausibility of [these] assumption[s], future efforts should be placed on implementing a complete model that associates choice predictions to latent memory states that are themselves estimated from the data". This is precisely what we aimed at in the last manuscript.

The biggest challenge associated with developing a formal model of the MSH derives from the fact that the heuristic acts on natural recognition, that is, it exploits the memory strength associated with experiencing different objects through mediators like newspapers, TV or the internet. Therefore, a crucial variable is missing: We do not know which objects were experienced before or not (Erdfelder et al., 2011; Pleskac, 2007). It follows that there is no obvious way to categorize an item as a hit or false alarm (or a correct rejection or a miss). Without the experience variable, the estimation of the memory states is far beyond straightforward. To solve this problem we relied on a proxy which allows for a good (although not perfect) estimation of memory-states - the consistency of recognition judgments. This proxy is associated with a simple extension of the r-model, only involving the addition of two extra recognition tests. Furthermore, its association with memory-states can be

derived from the two-high-threshold model. If an object is in a certainty memory state, the recognition judgment can only be *yes* in the case of recognition and *no* in the case of rejection. Consequentially, any inconsistent judgment should be associated with the uncertainty memory state. On the other hand, consistent judgments are likely to be associated with certainty memory-states, although it is possible that they occur through consistent guesses.

We developed a MSH model (called latent-states MSH model) by using the consistency of recognition judgments to model the probability of an object being in a certain memory state, and relied on the r-model to model the adequate decision strategy for each combination of memory states. As can be seen in Figure 2.3, for inconsistent judgments we directly assume that the objects are in the uncertainty state. For consistent judgments we take into account the possibility of consistent guesses, and therefore directly model the probability that consistent recognition judgments are associated with certainty ($h$ and $l$) or uncertainty ($1-h$ and $1-l$) memory states. This is done for the two objects in a pair, and once the memory-state combination is established, the appropriate decision strategy is modeled. Here, a distinction between pairs of objects in the same memory state versus different memory states is useful. If objects are in the same memory state, the decision strategy modeled corresponds to either the knowledge tree of the r-model (when both objects are in recognition certainty) or the guessing tree (when both objects are in uncertainty or both objects are in rejection certainty). If objects are in different memory states, the recognition tree of the r-model is used, with different parameters for different combinations of memory states. Specifically, there is a distinction between pairs of objects in (1) recognition certainty and rejection certainty (REC - REJ), (2) recognition certainty and uncertainty (REC-UNC) and (3) uncertainty and rejection certainty (UNC-REJ). This permits the estimation of MSH-use for those different types of pairs, which allows the test of the MSH core predictions: There is a preference for objects in a higher state, and this preference is stronger the larger the distance between the memory states of the objects under comparison. Within our model, this can be tested with a set of parameter restrictions involving the $r$ parameter for the three types of pairs. Specifically, the core prediction of the MSH corresponds to the following restrictions, $r_{REC-REJ} > r_{REC-UNC}$ and

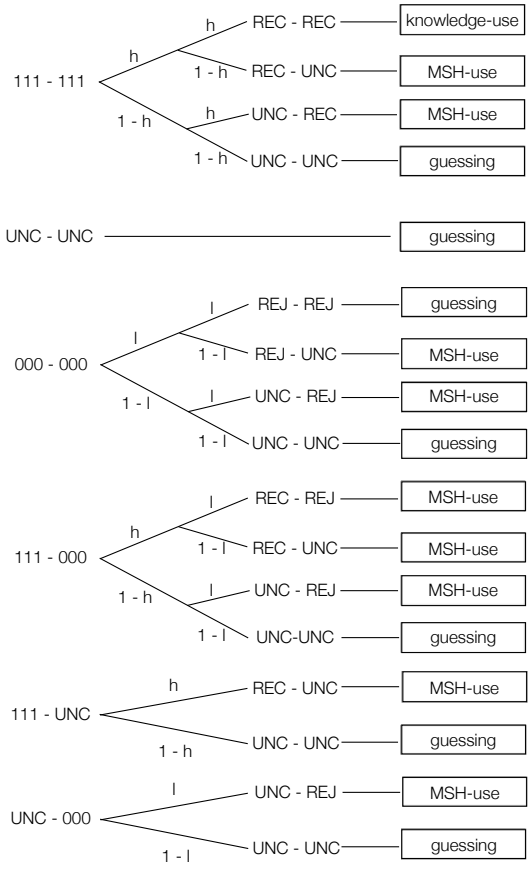



Figure 2.3: Abstract representation of the latent-states MSH model, denoting how the filter parameters determine the memory-state combination under comparison and, consequently, the appropriate decision process. $h$, probability that consistent recognition judgments originate from recognition certainty; $l$, probability that consistent rejection judgments originate from rejection certainty; 111, consistently recognized objects; 000, consistently rejected objects; REC, recognition certainty; UNC, uncertainty; REJ, rejection certainty. The full model can be found in Appendix A of the corresonding manuscript.

$r_{REC-REJ} > r_{UNC-REJ}$. Additionally, the same pattern should be observed for memory-state validity. We found support for this in two Experiments (see Figure 2.4).

Additionally, we tested an approximate MSH model which ignores the possibility

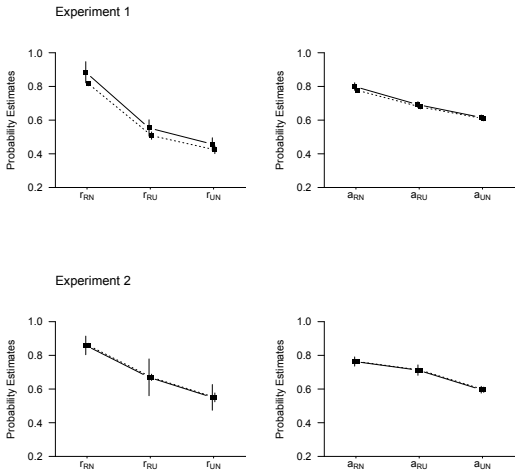



Figure 2.4: Estimates of the three $r$ and $a$ parameters for Experiment 1 and 2. Solid lines represent the estimates from the latent-states MSH model while dashed lines represent estimates from the approximate MSH model. Error bars represent bootstrapped standard errors of the parameter estimates.

of consistent guesses. In other words, it assumes that our proxy is a perfect indicator of the memory states, and so the same way that inconsistent judgments indicate a memory state of uncertainty, consistent judgments indicate a memory sate of certainty. Despite the inherent error of the approximate model, it performed quite well, and led to the same conclusions obtained with the latent-states MSH model. Because this model has less parameters ($h$ and $l$ are fixed to 1) it is not so vulnerable to sampling error and therefore more adequate for hypotheses testing. The fact that both models fit the data nicely and that we find convergent results is ideal since it allows us to not blindly rely on the assumption of the approximate model but, at the same time, be able to assert that our hypotheses hold with a version of the model not so vulnerable to sampling error.

In addition to using our models to test the core predictions of the MSH in Experiment 1 and 2, in Experiment 2 we further tested whether choice consistency is in line with the predictions of the MSH. In order to do so, we repeated through the three sessions not only the recognition test but also the comparison task. If, for simplicity, we assume that participants always resort to one of three strategies, namely, MSH-use, knowledge-use and guessing, we can predict that (1) consistency

should be highest when the distance between the memory-states of the objects under comparison is maximal, (2) consistency should be lower when objects are in adjacent memory-states, and (3) lowest when they are in the same state. However, within the adjacent and same-state cases, (4) consistency should be higher when knowledge is likely to be available, and lower when plain guessing is involved. We found support for all these hypotheses.

Finally, we also validated critical parameters of our model, specifically, the filter parameters $h$ and $l$, and the $r$ parameters. The former have to be validated since they are newly developed. The latter are "borrowed" from the r-model and have been validated before (see Hilbig et al., 2010). However, due to their critical role, and because there are three instead of a single $r$ parameter, we validated them too. With a cross-validation study we showed that the filter parameters, consistent with what they are meant to measure, are larger when we consider two repetitions of the recognition test versus a single one. Moreover, in a third experiment we compared the $r$ parameters between two conditions: One where memory-states were valid and another where memory-states validity was very low. Since MSH-use should decrease with memory-state validity, we predicted that all $r$ parameter estimates should be smaller in the latter condition. Accordingly, this is what we observed.

In sum, in this last paper we addressed the ultimate goal of this research program - developing and testing a formal model of the MSH. This model has the main advantage of incorporating all the possible memory state combinations and formalizing the MSH predictions for all of them. It therefore allows us to test the MSH without requiring further assumptions.

# General Discussion & Outlook

In the three manuscripts summarized above I report the results from my research program aimed at testing the MSH. The work described in this thesis succeed in testing the MSH in three different ways. First, in Castela and Erdfelder (2016) we started where Erdfelder et al. (2011) left off by showing how the recognition and rejection latency predictions of the MSH find support beyond recognition cases. Specifically, we have shown that in knowledge cases there is a preference for choosing the object recognized faster, while in guessing cases there is a preference for choosing the object rejected slower. Both these results are in line with the MSH hypothesis that there should be a preference for the object in a higher memory-state. Additionally, we also tested this prediction for recognition cases, but instead of using accordance rates like Erdfelder et al., we relied on the r-model (Hilbig et al., 2010). In this way, we added support for the latency prediction regarding recognition cases with a measure of RH-use that does not suffer from the biases of accordance rates.

Second, in Castela et al. (2014) we pitted the MSH against the RH and accounts assuming knowledge integration by using a formal model which consists of a simple extension of the r-model. To do so, we relied on information beyond binary recognition judgments, namely the distinction between two subjectively different experiences of recognition, recognition with further knowledge and mere recognition. Consistently only with the MSH and not with the other two accounts, we found that RH-use is higher for recognition pairs involving a recognized object with further knowledge than for recognition pairs involving a merely recognized object. This further hinted at the superiority of the MSH as an account able to explain and predict many findings in the literature.

Finally, in Castela and Erdfelder (in press) we extended the RH paradigm in order to test a formal model of the MSH. This model captures all possible combi-

nations of memory-states and holds the MSH predictions for all of them. In two experiments, the model fit the data well, and allowed us to test the MSH core predictions, namely, that MSH-use differs between the three possible combinations of different memory-states. Consistently with the MSH, we observed that MSH-use is higher when the difference in memory-states is maximal, and that it decreases for adjacent memory-states combinations. In Experiment 2 we additionally tested choice consistency predictions of the MSH, thereby accumulating more support for it. Finally, we validated core parameters of the latent-states MSH model through a cross-validation method and a third experiment, finding support for the role of the filter parameters and the psychological meaning of all three MSH-use parameters. In sum, in this third manuscript, Erdfelder et al.'s (2011) call for the development of a formal model of the MSH incorporating all of its predictions has been answered, with successful results.

Taken together, the three manuscripts gather a considerable amount of converging evidence for the MSH. By using different proxies, including recognition and rejection latencies, subjective recognition experiences, and consistency of recognition judgments, and by relying on increasingly sophisticated methods, evidence for the advantages of considering the underlying recognition process when investigating recognition-based inferences has been put forward. Therefore, I believe the three manuscripts nicely complement each other and together consist of a well-founded research program. In the following, I will discuss a few points which have so far not been addressed, and possible limitations and future directions of my work.

The goal of this thesis is to demonstrate that the RH should be extended to a framework which considers the recognition process itself, and not just the output of that process, i.e., recognition judgments. However, it is important to note that the MSH is not only a rather straightforward extension of the RH, but also that it reduces to it under ideal conditions (Erdfelder et al., 2011). Put simply, as the threshold parameters in the two-high-threshold model approach 1, the uncertainty state does not occur, and therefore the predictions of the MSH are the same as that of the RH. Importantly, though, I want to stress that this does not imply that for those cases the MSH should be used at all times and by all participants. In this thesis, I support a probabilistic version of the RH and the MSH. This implies that

when two objects are in different certainty states (recognition certainty and rejection certainty) reliance on recognition should be highest, but it does not imply that it is the only strategy used at all times. In all three manuscripts involved in this thesis that probabilistic interpretation has been tested by checking whether for those recognition cases more likely to contain only objects in certainty states reliance on recognition occurs every time, or whether other strategies (like knowledge-use) still take place. Accordingly, in all three manuscripts there was evidence for this probabilistic version, suggesting that people rely a lot on recognition-based inferences under ideal conditions, but other strategies can also take place.

The choice to test the MSH, such a simple extension of the RH, is justified as it allows one to draw simple predictions and is more testable than other options. However, this is one possibility among several, and therefore worth questioning. Ultimately, the MSH is based on the two-high-threshold model, which, despite being a very prominent model in the recognition memory literature, is by no means the only possible model on which to develop such an approach. One of its fiercest competitors, signal detection models (Kellen & Klauer, in press; Macmillan & Creelman, 2004), would be another option. Briefly, signal detection models assume there is a continuous memory strength variable described by two normal distributions, one for old items and one for new items. The degree of overlap between those distributions corresponds to the ability to discriminate between old and new items. Recognition judgments are in turn determined by the placement of a criterion. If memory strength surpasses it, an item will be judged as "old", otherwise it is judged as "new" (Kellen & Klauer, in press). I wish at this point to clarify that the current work does not dismiss the possibility that signal detection models are more appropriate than the two-high-threshold model. Additionally, while predictions were drawn from the latter, the findings accumulated do not rule out the former.

The essential distinction between signal detection and the two-high-threshold as models of recognition memory is that the former assumes that recognition judgments reflect a direct mapping of graded memory representations while the latter proposes that recognition judgments are mediated by a discrete-state representation (Kellen & Klauer, 2015). How would that impact the predictions for use of recognition as a cue in inferences? Perhaps the most evident implication is that, within a

signal detection framework, any difference in memory strength could be explored, regardless of the memory-state. That is, while the MSH predicts that recognition can only be exploited when objects are in different memory-states, a signal-detection based framework would predict that recognition can always be used since there should be a preference for the object associated with a higher memory strength.

It would certainly be interesting to pit these two models against each other in the context of recognition-based inferences. To do so, one could design inventive scenarios where the predictions of each model collapse and the two can thereby be critically tested. If memory-states and the underlying memory strength were observable variables, a simple way to approach this would be to test whether within the same memory-state there is a preference for objects with a higher memory strength. Unfortunately, accessing those latent variables is rather challenging. Another, slightly more straightforward approach, would only require knowledge about the experience variable. If it were known which objects were experienced before and which were not, a simple way to test signal detection-based inference models against the MSH would be to assess whether there is a preference for false alarms over misses. While a signal detection account would predict such preference (because false alarms will be necessarily associated with a higher memory strength than misses), the MSH predicts that there should be no preference for one over the other, since both objects necessarily originate from the uncertainty state.

However, as repeated throughout this thesis, one does not have access to the experience variable, complicating any attempt to model the recognition process. Moreover, while with the two-high-threshold model we used consistency as a proxy and could make predictions in a relatively straightforward way, it is hard to think either of what would be a signal detection model prediction for consistency, or which other proxy could be used. For these reasons, while I do not at all dismiss the possibility that signal detection models are more appropriate than the two-high-threshold as a model of recognition memory in this context, I believe that the two-high-threshold model does a better job in terms of eliciting testable predictions, therefore being the more useful model to advance knowledge in this area and allow a proof of concept regarding the possibility of incorporating recognition memory models in recognition-based inferences.

Besides the arguable option discussed above, and specific limitations of each manuscript which are respectively discussed in each of them, one important limitation of this thesis has so far not been addressed. In all formal modeling analysis present in the three manuscripts we relied on aggregation across participants. However, for aggregation in multinomial processing tree models to be unproblematic, there must be homogeneity between participants. In turn, if this assumption is violated, parameter estimates might be biased, and standard errors and confidence intervals underestimated (see Michalkiewicz, 2016, for a detailed account of why this might be a problem for the r-model). For these reasons, it would be an important next step to try to extend the current MSH model to a hierarchical version, in order to ensure that individual differences are not distorting our analysis. While I find this to be a crucial development, let me clarify why it has not been done so far. It is important to note that, as it is, the latent-states MSH is already a rather complex model which, as discussed in detail in Castela and Erdfelder (in press), is vulnerable to sampling error and not always ideal for hypotheses-testing. Given the challenges surrounding the modeling of latent memory-states, it appeared essential to first establish this possibility, before turning to another big challenge. However, now that the model has been established and tested, an important future direction would be to see how individual differences can be adequately taken into account.

Other potentially interesting routes to take in the future regard the way to estimate the underlying memory strength or memory states. One possibility would be to control the experience variable. To do this, we could have participants learn the material in the laboratory and thereby control which objects were effectively experienced before or not (see Bröder & Eichler, 2006; Newell & Shanks, 2004, for a similar approach). An obvious criticism to such approach is that one would no longer be dealing with natural recognition, and therefore the domain may not be adequate for the application of strategies like the RH or the MSH (Gigerenzer & Goldstein, 2011; Pachur, Bröder, & Marewski, 2008). Nevertheless, there might be some merit to this approach. If it would work, it would at least allow one to establish that the memory-strength-based inferences extend to experimentally induced recognition settings. However, if it does not work, that finding cannot be extrapolated to natural recognition settings. Another possibility would be to

look for a middle ground by designing an experiment which holds a learning period spanning a reasonable amount of time. In such an experiment, participants could be exposed to the material in a somewhat natural way, by, for example, reading news pieces involving the target objects at different points in time. While this would certainly be an effortful procedure, it might be worthwhile to consider, since it has the benefit of simulating a natural learning experience while at the same time giving the experimenter control over which objects are experienced and which are not.

# Conclusion

The research program developed in my thesis has addressed a largely overlooked aspect of the RH literature: The influence of recognition memory processes underlying recognition judgments on recognition-based inferences. Through testing and ultimately developing a formal model of the MSH, I have shown that, indeed, recognition judgments are a poor approximation of what determines recognition-based inferences. In turn, considering the memory-states underlying those judgments is worthwhile, as it presents a new and interesting pattern of results: Different levels of memory strength are associated with different validity and use of recognition as a cue. Ignoring those differences is ignoring a large chunk of the story. While that story may never be finished, I hope my work will inspire further developments, and that new and exciting chapters will follow.

# Bibliography

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

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

Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica*, *121*(3), 275–284.

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

Castela, M., & Erdfelder, E. (2016). *Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences.* (Manuscript submitted for publication)

Castela, M., & Erdfelder, E. (in press). The memory state heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory & Cognition*.

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*(5), 1131–1138.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of personality and social psychology*, *89*(4), 469.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological review*, *115*(1), 199.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic,

L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108–124.

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*.

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*(1), 7–22.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 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). , *11*(1), 230–239.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that makes us smart.* Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90.

Heck, D., & Erdfelder, E. (in press). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*.

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*(5), 1191–1206.

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, *17*(6), 923–930.

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

Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition*, *43*(4), 659–671.

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology*, *55*(6), 394–401.

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

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review*, *95*(4), 528–551.

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.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating rocs: A critical test with minimal assumptions. *Psychological review*, *122*(3), 542–557.

Kellen, D., & Klauer, K. C. (in press). Elementary signal detection and threshold theory. In E.-J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience, fourth edition (vol. V).* New York: John Wiley & Sons, Inc.

Kelley, C. M., & Jacoby, L. L. (2000). Recollection and familiarity. In E.Tulving & F. I. M. Craik (Eds.), *The oxford handbook of memory.* Oxford University Press.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide.* Psychology Press.

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*(3), 287–309.

Michalkiewicz, M. (2016). *Assessing and explaining individual differences within the adaptive toolbox framework: New methodological and empirical approaches to the recognition heuristic*. Unpublished doctoral dissertation, University of Mannheim.

Michalkiewicz, M., & Erdfelder, E. (2015). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, 1–15.

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(4), 333–346.

Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 923.

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*(2), 183–210.

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*(5), 983–1002.

Pachur, T., Mata, R., & Schooler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging*, *24*(4), 901.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology*, *2*(147), 1–14.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review*, *14*(3), 379–391.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(3), 251–271.

Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making*, *6*(5), 423.

Pohl, R. F., Erdfelder, E., Michalkiewicz, M., Castela, M., & Hilbig, B. (in press).

The limited use of the fluency heuristic: Converging evidence across different procedures. *Memory & Cognition*.

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*(1), 150.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

Scheibehenne, B., & Bröder, A. (2007). Can lay people be as accurate as experts in predicting the results of wimbledon 2005. *International Journal of Forecasting*, *23*, 415–426.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628.

Schwikert, S. R., & Curran, T. (2014). Familiarity and recollection in heuristic decision making. *Journal of Experimental Psychology: General*, *143*(6), 2341–2365.

Serwe, S., & Frings, C. (2006). Who will win wimbledon? the recognition heuristic in predicting sports events. *Journal of Behavioral Decision Making*, *19*(4), 321–332.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Tomlinson, T., Marewski, J. N., & Dougherty, M. (2011). Four challenges for cognitive research on the recognition heuristic and a call for a research strategy shift. *Judgment and Decision Making*, *6*(1), 89.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

# Statement of Originality

I hereby declare that I am the sole author of this thesis and have made use of no other sources that those cited in this work.

June 2016, Marta Castela

# Co-Author's Statements

# Co-author statement

I hereby confirm that the following article was primarily conceived and written by M.Sc. Marta Castela, School of Social Sciences, University of Mannheim.

Castela, M., Kellen, D., Erdfelder, E., & Hillbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review, 21*(5), 1131-1138.

June 2016, Dr. David Kellen

# Co-author statement

I hereby confirm that the following article was primarily conceived and written by M.Sc. Marta Castela, School of Social Sciences, University of Mannheim.

Castela, M., Kellen, D., Erdfelder, E., & Hillbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review, 21*(5), 1131-1138.

June 2016, Dr. Benjamin E. Hilbig

# Co-author statement

I hereby confirm that the following articles were primarily conceived and written by M.Sc. Marta Castela, School of Social Sciences, University of Mannheim.

Castela, M., & Erdfelder, E. (2016). *Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences.* Manuscript submitted for publication.

Castela, M., Kellen, D., Erdfelder, E., & Hillbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review, 21*(5), 1131-1138.

Castela, M., &, Erdfelder, E. (in press). The Memory State Heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

June 2016, Prof. Dr. Edgar Erdfelder

# APPENDIX: Copies of Manuscripts

Running head: Further evidence for the memory state heuristic

# Further evidence for the memory state heuristic: Recognition latency predictions for binary inferences

Marta Castela and Edgar Erdfelder
University of Mannheim

## Author Note

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 (GESS), also funded by the DFG.
Correspondence concerning this article should be addressed to Marta Castela or Edgar Erdfelder at the Department of Psychology, School of Social Sciences, University of Mannheim, D-68161 Mannheim, Germany.
Electronic mail may be sent to castela@psychologie.uni-mannheim.de or erdfelder@psychologie.uni-mannheim.de

Word count: 7309

## Abstract

According to the recognition heuristic (RH) theory, for decision domains where recognition is a valid predictor of a choice criterion, recognition can be used to make inferences whenever one object is recognized and the other is not, irrespective of further knowledge. Erdfelder, Küpper-Tetzel, and Mattern (2011) questioned whether the recognition judgment itself affects decisions or rather the memory strength underlying it. Specifically, they proposed to extend the RH to the memory state heuristic (MSH), which assumes a third memory state of uncertainty. They tested several qualitative predictions of the MSH, but left some questions unanswered that we address in two studies. First, we show that in knowledge pairs (both objects recognized) and guessing pairs (none of the objects recognized), the object that is more likely to be in a higher memory state is chosen. Second, we used a better measure of RH-use to show that reliance on recognition increases with the proportion of objects in certainty states. In sum, our two studies nicely complement the work of Erdfelder et al. by lending additional evidence to the MSH.


**Keywords:** recognition heuristic; memory-state heuristic; threshold models; multinomial processing tree models

The recognition heuristic (RH) is a fast and frugal decision strategy proposing that, for binary decisions, if one object is recognized and the other is not, one should infer that the recognized object scores higher on the criterion under consideration (Goldstein & Gigerenzer, 2002). This simple decision rule has gained a lot of attention, and there is a large body of research dedicated to it (see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011, for reviews). However, one key concept of the RH seems to be often neglected: *recognition*. While literally at the core of the heuristic, not so much research has been dedicated to understanding the role of recognition in use of the RH. However, there are some exceptions (e.g., Erdfelder et al., 2011; Pachur & Hertwig, 2006; Pleskac, 2007). Notably, Erdfelder et al. proposed a framework that extends the RH by accommodating the role of recognition memory. In this paper, we aim at generalizing their work by providing further tests of their framework. First, we will describe the RH theory in more detail. Then, we will discuss how recognition memory has been understood in the RH theory so far. Afterwards we will describe the contribution of Erdfelder et al. to linking recognition memory and the RH, along with the evidence they accumulated. Finally, we will introduce two new studies that complement the evidence presented by Erdfelder et al. (2011).

## The RH theory

To better understand how recognition memory has been (or can be) integrated in the RH, it is first essential to describe more precisely how the heuristic has been proposed. To simplify that process, we will refer to the most prominent paradigm associated with the RH as an illustrative example. This is the city size paradigm, which involves a pairwise comparison task where people must infer which of two cities has a larger population, and a recognition task, where for all cities involved people must judge whether they have heard of them before or not. With the data from the

recognition task, all pairs in the comparison task can be categorized into three types: knowledge pairs (both objects are recognized), recognition pairs (one object is recognized and the other is not), and guessing pairs (none of the objects is recognized). The RH applies only to recognition pairs, for the obvious reason that it cannot discriminate between objects in the other two types of pairs. Importantly, Gigerenzer and Goldstein (2011) specified additional preconditions for use of the RH. First, there should be a strong correlation between recognition and the decision criterion. In our example, recognition should be strongly correlated to the size of a city (which, indeed, it is). Additionally, further cues should not be readily available. This means that, for example, when comparing the sizes of Berlin and Mannheim, the information that Berlin is the capital of Germany, or that it has an international airport, should not be presented to the participant simultaneously (whereas, of course, it could be retrieved from memory). Finally, they asserted that the RH applies only to natural recognition, that is, artificially inducing recognition in the laboratory (by, for example, presenting the objects several times) should not necessarily lead to use of the RH.

This relates to the notion of how recognition comes to be a valid cue, that is, its ecological rationality. Here, three concepts are important: *recognition validity*, *ecological correlation* and *surrogate correlation* (Goldstein & Gigerenzer, 1999). Going back to our cities example, the ecological correlation – the correlation between the city size and the frequency of occurence of the city in natural encounters – is exemplified by the fact that larger cities are more likely to be mentioned in the Internet, TV, newspapers, or other type of mediator. This, in turn, affects the surrogate correlation, which is the correlation between the number of times a city is mentioned and the recognition of the name of that city. Naturally, cities that are mentioned more often will have a higher probability of being remembered. Finally, recognition validity is defined as the strength of the relationship between recognition and the criterion (Goldstein & Gigerenzer, 1999).

## Recognition memory in the RH theory

Now with a clearer understanding of how and under which conditions the RH was proposed to apply, we can address the question of how recognition memory plays a role in the heuristic. In its original definition, the RH was not related to recognition memory, but only to recognition judgments. Goldstein and Gigerenzer (2002) assumed that the RH works on the output of the recognition process, and that the process itself can be disregarded. In other words, they claimed the RH operates on *yes* or *no* recognition judgments, and whatever underlies that judgment can be ignored for the purpose of investigating the heuristic. This also implies that the frequency with which an object has been encountered does not affect use of the RH, but merely the final all-or-none process of remembering any encounter or not. As stated by Pachur et al. (2011, p. 4), "the recognition heuristic does not distinguish between objects one has encountered 10 times and those encountered 60 times (as long as both are recognized or unrecognized)". Erdfelder et al. (2011, p. 8) challenged this view by stating that "Showing that the RH is an ecologically rational and well-adapted choice strategy obviously requires a formal theoretical link between (1) the memory strengths of choice option names - a latent variable which is affected by environmental frequency and previous processing - and (2) binary recognition judgments for choice option names - an empirical variable which is assumed to affect decision behavior".

Following from this understanding of a necessary link between memory strength and recognition judgments, Erdfelder et al. (2011) proposed to integrate a model of recognition memory with the RH theory. To do so, they relied on one of the most well-supported models of recognition memory available - the two-high-threshold (2HT) model (Snodgrass & Corwin, 1988). Besides being one of the most successful models of recognition memory, the 2HT model has the added advantage of being easily combinable with the RH (Erdfelder et al., 2011).

The 2HT model belongs to the class of multinomial processing tree models

(Batchelder & Riefer, 1999; Erdfelder et al., 2009). Like other multinomial processing tree models, the 2HT model is based on the assumption that observed categorical responses are a result of a defined set of discrete states and that the probability of such states being entered depends on the probability of certain cognitive processes occurring or not. The basic premise of the 2HT model is that there are three possible memory states underlying recognition judgments - recognition certainty, uncertainty, and rejection certainty. The probability of those states being entered depends on the probability of two thresholds being exceeded (see Figure 1). Specifically, for objects experienced before, if the memory strength exceeds the first threshold with probability $r$, the object will be in the recognition certainty state and a *yes* recognition judgment will be given. If, with complementary probability $1 - r$, the memory strength lies below this threshold, the object will be in the uncertainty state, and the recognition judgment will depend on a second process of guessing, resulting in a *yes* judgment with probability $g$ and a *no* judgment with probability $1 - g$. For objects not experienced before, if the memory strength lies below the second threshold with probability $d$, the object will be in the rejection certainty state and a *no* recognition judgment will be given. With complementary probability $1 - d$, if the memory strength lies above this second threshold, the object will be in the uncertainty state and, again, the recognition judgment will depend on guessing.

To combine this model with the RH theory, Erdfelder et al. (2011) suggested a new framework - the memory state heuristic (MSH). The MSH is a straightforward extension of the RH, which mainly replaces recognition judgments by memory strengths. That is, it assumes that memory strengths, and not recognition judgments per se, are correlated with the criterion. This simple extension enriches both the predictions that can be drawn and the explanatory scope of the heuristic. Whereas the RH has predictions for recognition pairs only, the MSH has predictions for any pair that involves objects in different memory states. These predictions can be summarized by two simple premises: (1) if objects are in different memory states, there should be a preference for the object in a higher state; (2) the larger the

discrepancy between the memory states of objects in a pair, the higher should be the probability of choosing the object in a higher state. By implication, the probability of choosing the object in a higher memory state should be larger for pairs of one object in the recognition certainty state and the other in the rejection certainty state, than for pairs where the objects are in recognition certainty and uncertainty or uncertainty and rejection certainty. Based on these two principles, Erdfelder et al. managed to both explain previous results that challenged the RH and also draw and test new predictions. To do so, they relied on the fact that multinomial processing tree models like the 2HT model can be interpreted as probabilistic serial processing models (Batchelder & Riefer, 1999; Heck & Erdfelder, in press). By implication, the number of cognitive processing stages in a given branch of the model will influence its total processing time. Specifically, in the case of the 2HT model, whenever an object reaches the memory state of uncertainty and a second cognitive stage is required - guessing - the response time distribution should be stochastically larger than when an object reaches one of the two certainty memory states (Heck & Erdfelder, in press). Following from this interpretation of the 2HT model, a clear prediction can be made: "The larger the recognition judgment latencies, the more likely it is that the judgment originates from guessing and the less likely it is that it originates from memory certainty" (Erdfelder et al., 2011, p. 13).

As mentioned, the MSH offers a simple explanation for previous results that challenged the RH. One example is that, in recognition pairs, recognized objects for which participants claim to have further knowledge are chosen more often than recognized objects that participants claim to merely recognize the name of (e.g., Pohl, 2006). This has been explained by assuming that people are relying on further knowledge, thereby challenging the RH. However, the same result is predicted by the MSH if one makes the reasonable assumption that objects for which participants claim to have further knowledge are more likely to have originated from recognition certainty than objects that are merely recognized. Castela, Kellen, Erdfelder, and Hilbig (2014) tested the three accounts (RH, use of further knowledge, and MSH) and

found support for the MSH.

Along with this, Erdfelder et al. (2011) described other examples of how the MSH can accommodate previously problematic results. But most importantly, they directly tested seven predictions of the MSH, focused on RH accordance rates (proportion of times the recognized object is chosen in recognition pairs) and decision latencies, both as a function of recognition and rejection latencies. The first three predictions, which state that RH accordance rates should increase with decreasing recognition and rejection latencies, and that their effect is additive, were supported in their study. Additionally, they tested whether the decision latency in recognition pairs increases with both the recognition latency of the recognized object and the rejection latency of the unrecognized object, and if their effect is additive. These further three predictions were also supported by their data. Finally, they found support for their seventh prediction, which stated that response bias manipulations (aimed at selectively affecting the guessing probability) in the recognition test should affect recognition judgments but not performance in the comparison task. Since the RH theory assumes that recognition judgments per se influence decisions, it would predict that a bias manipulation will also affect choices. The MSH, in turn, predicts the observed result, since memory-states and not recognition judgments should influence decisions, that is, since biasing the guessing probability does not alter the memory-states distribution, choices should be left unaffected.

The focus of Erdfelder et al. (2011) has been on testing predictions for recognition pairs, but as explained before, the MSH also makes predictions for guessing and knowledge pairs, as long as the objects under comparison are in different memory states. This will be the focus of our first study. As for recognition pairs, the predictions follow from the basic premise of the MSH: If objects are in different memory states, there should be a preference for the one in a higher state. Therefore, in this study we will test two predictions:

1. In knowledge pairs there should be a preference for the object recognized faster

(as this one is more likely in the memory certainty state)

2. In guessing pairs, there should be a preference for the object recognized slower (since this one is more likely in the uncertainty state, which is the highest possible state for unrecognized objects).

However, as outlined above, the MSH also predicts that the preference for the object in a higher state should be strongest the highest the discrepancy between the states. While in recognition pairs the maximal memory state distance can be observed (one object in recognition certainty and the other in rejection certainty), in both knowledge and guessing pairs this can never occur, since objects will either be in the same state or in adjacent states (recognition certainty and uncertainty or rejection certainty and uncertainty, respectively). For this reason, as already noted by Erdfelder et al., we expect weaker effects of recognition latency differences than those found for recognition cases. Additionally, we will also test whether the effect is stronger when the difference in latencies is higher, therefore increasing the probability of the objects being in adjacent states versus in the same state.

We should note at this point that the prediction regarding knowledge cases has already been tested in a different context. Actually, this prediction of the MSH overlaps with what is called the *fluency heuristic* (Hertwig, Herzog, Schooler, & Reimer, 2008; Schooler & Hertwig, 2005), which states that, in knowledge cases, the fastest retrieved option should be chosen. Its premise is that the fluency with which an object is retrieved from memory (approximated by the latency of the recognition judgment) can be used as a single cue and determine inferences. They measured the accordance rate of the fluency heuristic by computing, for each participant, how many times the object retrieved faster is chosen in knowledge pairs (pairs with differences in recognition latency smaller than 100 ms were excluded)[1], and found that it is higher than the individual baseline accordance. Furthermore, they observed

[1]The threshold of 100 ms was shown to be sufficient for discriminating between recognition latencies (Hertwig et al., 2008).

that accordance rates increase with the difference in latencies between objects. While the fluency heuristic can accommodate these results, it is very limited: it only applies to knowledge pairs, and within those, to pairs where the fluency difference is larger than 100ms. The MSH, on the other hand, also predicts these results, and does so while being able to predict much more about the data - predictions for guessing and recognition cases. It is, therefore, a far more parsimonious framework (Erdfelder et al., 2011). Moreover, the MSH predicts that the preference for the faster recognized object should be weak at best, simply because the memory-state discrepancy for knowledge pairs can only be small (i.e., recognition certainty and uncertainty) or even nonexistent (i.e., when both objects are in the same state). The fluency heuristic, in contrast, fails to provide an explanation for the smaller preferences in knowledge pairs compared to recognition pairs (see Pohl, Erdfelder, Michalkiewicz, Castela, & Hilbig, in press).

While the prediction for knowledge pairs seems straightforward and has already been tested in another context, it should be emphasized that the prediction for guessing cases is completely new, and surprising in the sense that it leads to the expectation of a preference for less fluent objects. To the best of our knowledge, no framework other than the MSH makes or can accommodate such prediction.

Besides these predictions for knowledge and guessing cases, in a second study we wanted to test a further prediction of the MSH. Erdfelder et al. (2011) already showed that larger recognition and rejection latencies are associated with smaller RH accordance rates. However, we wanted to test this in a more refined way using a better measure of RH-use. Although the RH accordance rates used by Erdfelder et al. provide an approximation of RH-use, they are a biased measure because counting the number of times choices are in line with recognition does not take into account what led to that choice. An option might have been chosen because it was recognized, or because other information, which points in the same direction, was used. For example, when comparing supposed population sizes of Berlin and Mannheim, a non-european person might chose Berlin because she recognizes it and does not

recognize Mannheim, or because she knows Berlin is the capital of Germany, and therefore likely to be a large city. For this reason, Hilbig, Erdfelder, and Pohl (2010) developed a multinomial processing tree model which estimates RH-use in a more sophisticated way. The r-model (see Figure 2) consists of three trees, which correspond to the three types of pairs. For knowledge and guessing pairs, the trees have only a single parameter that accounts for the accuracy for knowledge and guessing pairs, respectively. For recognition pairs, on the other hand, the model considers the possibility that a recognized option is chosen through use of further knowledge, and provides in this way an unbiased estimate of RH-use (which corresponds to parameter $r$ in the model; see Hilbig et al. (2010, for additional details about the r-model)). By using this model, we can assess in a more precise way how recognition and rejection latencies are associated with reliance on recognition alone. Additionally, we can test whether in the more extreme cases, when the latencies are very short (so that both objects are most likely in recognition and rejection certainty states), people always rely on memory-states only, or whether even then other processes such as integration of further knowledge can take place.

## MSH predictions for guessing and knowledge cases: A reanalysis of published data

We first tested whether choices for guessing and knowledge cases are in accordance with the MSH prediction that there is a preference for the object in a higher state. Specifically, as outlined above, we used recognition and rejection latencies as proxies for underlying memory states. Therefore, we predicted that in knowledge pairs there is a preference for the object with a shorter recognition latency (and therefore a higher probability of being in a recognition certainty state) while in guessing pairs there is a preference for the object with the longest rejection latency (and therefore a higher probability of being in the uncertainty state).

We first reanalyzed the data of 14 published datasets from our lab (see Table

1), in order to look for preliminary evidence for our hypotheses. As shown in Figure 3, we observed that for all 14 datasets the proportion of choosing the object recognized faster in knowledge cases was significantly larger than .5 (smallest $t(21) = 2.78$, all $p < .01$). Regarding guessing cases, in 12 of the 14 datasets the proportion of choosing the object recognized slower was significantly larger than .5 (smallest significant $t(63) = 2.08$, $p = .02$). Clearly, these results are in line with our expectations. However, the studies included in the reanalysis were not conducted with our hypotheses in mind. In order to collect further evidence, we designed a new experiment specifically tailored to our hypotheses. With this new experiment, we primarily aimed at optimizing the proportion of knowledge and guessing cases in order to achieve more powerful tests of the MSH predictions for these cases. Moreover, we were also interested in generalizing the results across different decision domains beyond city-size comparisons.

## Experiment 1

## Material and Procedure

The paradigm we used resembles the city-size paradigm outlined in the *Introduction* but actually involves different types of decisions. This paradigm includes two tasks: (1) a recognition test, where objects are presented and participants must judge whether they have seen them before or not; (2) a comparison task, where participants see pairs of the objects and must infer which scores higher on a given criterion. Since the objects are paired exhaustively, the relative proportion of knowledge, recognition and guessing cases will depend on the proportion of objects recognized. Therefore, in order to optimize the proportion of knowledge and guessing cases, it is important to include in the experiment a condition for which the proportion of recognized objects across participants is larger than .50 (resulting in many knowledge cases) and a different condition in which the proportion of recognized objects is clearly less than .50 (resulting in many guessing cases). A third condition should involve a recognition

rate of about .50, resulting in (almost) equal frequencies of knowledge and guessing cases. Moreover, since we also wanted to generalize our findings across different domains, we made use of different types of objects and inference criteria in the three conditions. Specifically, all participants were presented with objects from three domains: largest world cities (with over 3 million inhabitants; see http://en.wikipedia.org/wiki/List_of_cities_proper_by_population), most successful celebrities (100 most successful celebrities according to the Forbes list of 2015; see www.forbes.com) and longest rivers in the world (over 1900 km long; see https://en.wikipedia.org/wiki/List_of_rivers_by_length). According to pre-tests conducted in our lab, we know that for the domain of world cities normally 50% of the objects are recognized. We included this domain for generalizability, and also because it is one of the most often used domain in the study of the RH and should serve as benchmark. For the domain of celebrities, normally 65% of the objects are recognized. Therefore, this domain is ideal to test the hypothesis regarding knowledge cases. Finally, the rivers domain is ideal for testing the hypothesis regarding guessing cases, since usually 35% of the objects are recognized. The experiment included three blocks, each consisting of the recognition test and the comparison task for each domain. The order of blocks was randomized for all participants. In each block, the recognition test always preceded the comparison task. In the recognition test participants saw all 20 objects (randomly selected from each domain, but the same for all participants) and had to decide whether they have heard of them before or not. Objects were presented one at a time, in random order, and a 500 ms interstimulus fixation-cross followed each response. Response times were recorded along with the recognition judgments. After each recognition test, a comparison task followed. In the comparison task, participants saw 190 pairs, consisting of the exhaustive pairing of the 20 objects, and had to infer which one scored higher on the criterion. Each pair was presented at a time, in random order, and a 500 ms interstimulus fixation-cross followed each response. Response times were recorded along with the responses. For the world cities, the criterion was city-size; for celebrities, the criterion

was how successful they were; and for the rivers, the criterion was their length.

## Participants

We recruited 75 students (50 women) from the University of Mannheim aged between 19 and 46 ($M = 22.00, SD = 5.04$). Participation was compensated monetarily as a function of performance in the comparison task. Every participant received at least two euros, and they could earn up to 7.70. They gained one cent for each correct answer, and lost one cent for each wrong one.

## Results

One participant had to be removed from the analysis for all domains, because he indicated that he did not recognize any object in any domain. Furthermore, one participant was removed from the guessing analysis of the cities domain because he recognized 19 out of the 20 cities, therefore having no guessing pairs. Finally, two additional participants were removed from the knowledge analysis of the rivers domain because they only recognized one river and therefore had no knowledge pairs. For the remaining participants, the proportion of recognized items was on average .68 for celebrities, .58 for the world cities, and .36 for rivers. This was in line with the pre-tests, although a bit higher than what we expected for the world cities domain.

Since our hypotheses refer to the preference for the object recognized fastest in knowledge pairs, and the one judged unrecognized slowest in guessing pairs, we first calculated per participant the proportion of times their choices were in line with those hypotheses (accordance rate). We then performed one-sample t-tests to assess whether the mean accordance rates were larger than .50. As can be seen in Table 2, we found support for both hypotheses in all three domains assessed.

In addition to testing whether there would be an above chance preference for the items more likely to be in a higher state, we also wanted to assess whether this preference would increase with an increasing difference in recognition latencies (i.e.,

latencies of *yes* judgments) or rejection latencies (i.e., latencies of *no* judgments) between objects in a pair (and therefore an increasingly higher probability of being in adjacent states). To do so, we ran a multilevel logistic regression[2] with *Accordance* as a dependent variable. *Accordance* is essentially a binary variable which takes the value one if choices are in line with our hypotheses, and zero when they are not. Specifically, for knowledge pairs, *Accordance* will be one whenever the fastest recognized object is chosen, and zero otherwise. Conversely, for guessing pairs, *Accordance* will be one whenever the slowest unrecognized object is chosen, and zero otherwise. As predictors, we included both the main effects and the interactions of the *RT difference* (difference in recognition or rejection latencies between the objects in a pair) with *Case* (knowledge or guessing) and with *Domain* (celebrities, cities or rivers). Additionally, the model includes a random intercept for each participant and a random slope for each participant regarding the effect of *RT difference*. Our hypothesis would be that *RT difference* has a positive effect on *Accordance* for both cases and in all domains. We find support for our hypothesis. As can be seen in Table 3, *RT difference* has a significant positive effect on *Accordance*. Additionally, there are no differences in *Accordance* between the domains[3]. Moreover, while the effect is present for both knowledge and guessing cases (see Figure 4), we find that it is significantly stronger for knowledge cases. While this was not directly predicted, it does not compromise our findings. This will be addressed in more detail in the *Discussion* section.

[2]The model was estimated using the glmer function of the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2015).

[3]Adding the interaction of *Domain* and *RT difference* does not change the overall pattern of results and the interaction is not significant. Therefore, we opted to present the results of a model without the interaction, so that we can observe the main effect of *RT difference* for all domains and not only for the reference level of the *Domain* variable.

## The influence of removing items with longer recognition/rejection judgment latencies on RH-use: A reanalysis of published data

As mentioned above, in our second study we wanted to test the MSH predictions regarding recognition latencies. This had been already successfully tested by Erdfelder et al. (2011), but by relying on accordance rates. Since we now have access to a better (less biased) method to estimate RH-use – the $r$ parameter of the r-model (Hilbig et al., 2010) –, we can test these predictions in a more precise way. Specifically, we can test if there is an increase in $r$ when we sequentially remove items with longer recognition and rejection latencies and fit the r-model to those subsets of data. The rationale behind this is that by removing those "slow" items we reduce the subset mostly to objects in recognition certainty and rejection certainty states. While doing so, we artificially create the perfect preconditions for relying uniquely on recognition, which should lead to higher $r$ estimates. Additionally, we would like to test if $r = 1$ in the most extreme cases, when only the items with shorter recognition and rejection latencies are involved. To address these questions, we first reanalyzed the data for the 14 published datasets that were also used in our previous study.

For each data set, we first identified for each participant which items where in the first, second, third or fourth quartile of their individual recognition and rejection latency distributions. In a second step, we created (at the aggregate level)[4] four subsets of pairs that consisted only of objects with latencies in each of the quartiles of the latency distributions[5]. Next, we fitted the r-model simultaneously to these four disjoint subsets of data by replicating the r-model trees four times, that is,

[4]While we do the analysis at the aggregate level, it is individual recognition and rejection latency distributions that are considered when assigning the data categories to each subset.

[5]This procedure heavily restricted the amount of available data, since for each subset of data, only pairs where both objects are in the respective quartile of the recognition or rejection latency distributions can be analyzed.

for each subset of pairs. By implication, we ended up with four $r$ estimates. At the level of parameters, our hypothesis can be described as an order restriction such that the $r$ parameters decrease from $r_1$ to $r_4$, with the index 1 corresponding to the first quartile of the distributions (only the fastest recognized and unrecognized objects are included) and 4 the last quartile of the distributions (only the slowest recognized and unrecognized objects are included).

All model-based analyses were performed with MPTinR (Singmann & Kellen, 2013) in R (R Core Team, 2015). We first fitted the model without any restrictions; this baseline model fits the data well for 9 of the 14 datasets (see Table 4). To test our hypothesis, we excluded the 5 datasets that were associated with misfit[6]. In order to evaluate our order restriction we need two tests. First, we test the order restriction, $r_1 \geq r_2 \geq r_3 \geq r_4$, against the baseline model (with no restriction on the four $r$ parameters). Second, we test the model with order restrictions, $r_1 \geq r_2 \geq r_3 \geq r_4$, against a model imposing equality restrictions, $r_1 = r_2 = r_3 = r_4$. If the order restriction corresponds to the most suitable version of the model, the first test should fail to reach statistical significance, while the second test should lead to statistically significant results.

Since our hypothesis involves an order restriction between four parameters, the sampling distribution of the likelihood-ratio test statistic $\Delta G^2$ does not follow a chi-square distribution with the appropriate degrees of freedom. Given the challenge involved in determining the appropriate distribution, we opted for using a double bootstrap method (Van De Schoot, Hoijtink, & Deković, 2010) to compute p-values. For example, when we want to test the order restrictions, $r_1 \geq r_2 \geq r_3 \geq r_4$, against the baseline model, the double bootstrap consists of the following steps: (1) a

[6]In most cases, misfit in the r-model is associated with its inherent restriction in the $b$ parameters, implying that that knowledge validity is the same for knowledge and recognition pairs (Hilbig et al., 2010). Removing this constraint eliminated misfit for 4 out of the 5 datasets, but because the model with two $b$ parameters is saturated, we refrained from including these datasets in the subsequent analysis.

non-parametric bootstrap sample is obtained from a given data set (2) the model imposing the null hypothesis, $r_1 \geq r_2 \geq r_3 \geq r_4$, is fitted to that data set; (3) those parameter estimates are used to obtain a parametric bootstrap sample (4) both models being tested (model imposing the order restriction $r_1 \geq r_2 \geq r_3 \geq r_4$ and the baseline model) are fitted to that sample and the difference in fit is calculated; (5) steps 1 to 4 are repeated many times (we repeated it 1000 times). We then compute the p-value by assessing how many times the difference in fit obtained with the bootstrapped samples is equal or more extreme than the difference in fit obtained with the original data set, and reject the null hypothesis if this proportion is smaller than .05. Additionally, we also compare the models through the model selection measure FIA (Fisher Information Approximation), which takes complexity into account[7].

The results are shown in Table 5 and Figure 5. We find a clear support for the order-restricted model both with the goodness-of-fit test and the FIA comparison. In all except one data set (data set 14) the order restriction did not lead to significant misfit, while the equality restriction did. Accordingly, FIA was for all data sets smaller for the order restricted model than for the baseline or the equality restricted model. Only for Data Set 14, in line with the results from the goodness-of-fit test, the difference in FIA between the baseline and the order restricted model is not sufficient to support the former.

Additionally, to test whether $r$ approaches one for the subset including only the items recognized and rejected fastest, we tested the following restriction: $r_1 = .99$[8]. In all 9 datasets, this restriction led to a significant increase in misfit

[7]When using FIA to compare two models, a difference larger than 1.1 is considered to be substantial evidence in favor of the model with smaller FIA (see Kellen, Klauer, & Bröder, 2013). For comparisons in terms of FIA we additionally made sure that the sample-size of all data sets involved was above the lower-bound recommended by Heck, Moshagen, and Erdfelder (2014).

[8]We tested $r_1 = .99$ instead of $r_1 = 1$ because the latter restriction predicts zero frequencies for some categories of the model and therefore a unique observation in one of those categories would lead to severe misfit. A rejection of such model would therefore be trivial.

(smallest $\Delta G^2(1) = 14.55, p < .001$, smallest $\Delta FIA = 6.11$) suggesting that even under the ideal conditions for use of memory state information alone, people still sometimes rely on other strategies, like use of further knowledge.

While these results lend support to our hypothesis, the reanalyses are not ideal because when creating the subsets of pairs we necessarily limit the data points available for analysis (see Table 5). Therefore, we conducted Experiment 2 that, by being designed specifically to test our hypothesis, allows us to test them with greater power.

## Experiment 2

### Participants

We recruited 52 students (35 women) from the University of Mannheim aged between 18 and 45 ($M = 22.38, SD = 5.49$). Participation was rewarded either with a monetary compensation (2 euros) or with study participation credits. Additionally, for each correct response in the comparison task, participants gained 2.5 cents, and for each incorrect response they lost 2.5 cents.

### Material and Procedure

The experiment consisted of the city-size paradigm, involving two tasks. First, participants had a recognition task, where they saw 60 city names and had to indicate whether they recognize them or not. Naturally, response times were recorded along with the recognition judgments. The 60 cities were a random selection from the largest world cities (with over 3 million inhabitants; see http://en.wikipedia.org/wiki/List_of_cities_proper_by_population). After the recognition task, cities were paired according to their recognition and rejection latencies, with the fastest being paired together, and so on. Specifically, there were four subsamples of pairs, created according to the corresponding four bins of

recognition and rejection latencies. Whenever the number of recognized or rejected objects was not divisible by four, it was randomly defined which bin(s) would have one object more than the other(s). After the pairs were created (the number of pairs varied between participants, being either 420, 421 or 422), participants saw them and had to decide for each pair which city was more populous.

## Results

Before fitting the model, we removed one participant because he recognized only one of the 60 cities, while the remaining participants recognized on average 57% of the objects. With the data from the remaining 51 participants, we determined the frequencies for each category of the model, separately for the four bins of data. Then, we fitted the r-model to the four bins of data. The model performed very well in describing the data ($G^2(4) = 7.44, p = .11, FIA = 65.49$). We repeated the same analysis that we performed with the published data sets, with the goal of testing our order hypothesis on the parameters $r_1$ to $r_4$. As can be seen in Table 5 and Figure 5, we again find support for our hypothesis. Additionally, we again tested whether adding the equality restriction $r_1 = .99$ leads to a significant increase in misfit. We observed a very extreme increase in misfit with this restriction ($\Delta G^2(1) = 1806.84, p < .001, \Delta FIA = 899.78$), which again supports the notion that even under ideal conditions for reliance on memory states alone, other strategies than mere reliance on memory strength take place.

# Discussion

When they introduced the MSH, Erdfelder et al. (2011) contributed to the RH literature by providing an extension of the heuristic which parsimoniously links it with the recognition memory literature. The MSH not only explains a lot of previously problematic results but also provides a set of new predictions. While Erdfelder et al. tested many of these predictions in their original paper, some were

left untested. Our aim was to address this gap by (1) testing the MSH predictions for guessing and knowledge cases and (2) provide further evidence for a crucial prediction regarding recognition cases. We addressed both these issues in two studies by reanalyzing previously published data sets and conducting two new experiments. In this way, we found strong converging evidence in line with the MSH.

In our first study, by relying on recognition and rejection latencies as a proxy for memory states – under the assumption that longer latencies are associated with the uncertainty memory state while shorter latencies are associated with certainty memory states – we found evidence for the MSH prediction that for knowledge and guessing cases people also have a preference for objects that are likely to be in a higher memory state. While for knowledge cases this is not a new prediction – as it can be alternatively explained by the fluency heuristic (Hertwig et al., 2008) – , the prediction regarding guessing cases cannot be accounted by any other framework we are aware of. Furthermore, that latter prediction appears quite counterintuitive, since objects recognized slower should be preferred in guessing cases. Nevertheless, we found evidence for this in all data sets we analyzed.

It is also worth noting that the MSH not only predicts the preference effects, but also predicts they should be smaller than the corresponding effects in recognition cases. This is due to the fact that, in knowledge and guessing pairs, the objects can only be either in the same memory state or in adjacent memory states. Therefore, the preference for the object in a higher state should be less marked than in cases where the distance between states is maximal (pairs of one object in recognition certainty and one object in rejection certainty), a combination that can only occur for recognition pairs. We thus believe the MSH presents itself as the most parsimonious framework for understanding how recognition is used in binary inferences, clearly outperforming other approaches, like the RH and the fluency heuristic, in its explanatory power and predictive reach. While for recognition cases this had already been shown (see Erdfelder et al., 2011), our results extend the support of the MSH to knowledge and guessing cases, thereby closing a gap that was left open.

One result worth noting is the fact that the effect of latencies was stronger for knowledge cases than for guessing cases. While we had not predicted this explicitly, it fits nicely with previous results. Specifically, Castela and Erdfelder (in press) have implemented the MSH in a formal model that accommodates all memory state combinations, and observed that MSH-use is higher for recognition pairs if one object is in recognition certainty and one object in the uncertainty state than for recognition pairs with one object in uncertainty and one object in the rejection certainty state. Since these are the memory state combinations that can underlie adjacent state cases within knowledge and guessing pairs, respectively, our result seems to be exactly in line with was found in Castela and Erdfelder – a stronger tendency to use the MSH in the former cases. Given the converging evidence concerning this effect, future studies should focus on testing possible explanations for it. One such explanation, already suggested by Castela and Erdfelder, is that the distance in memory strength between the recognition certainty and uncertainty memory states might be larger than the corresponding difference between the uncertainty and rejection certainty memory states. This would suggest that a simple ordinal description of the states might be insufficient.

With our second study we aimed at further testing the effect of recognition and rejection latencies in choices for recognition pairs. While this is a conceptual replication of the test carried out by Erdfelder et al. (2011), we relied on a different measure of RH-use, which we believe is far mode adequate. Erdfelder et al. relied on accordance rates to the RH, which, as explained above, are a severely confounded measure since people might choose the recognized option for reasons other than the fact that they are applying the RH, namely because they rely on further knowledge. For this reason, Hilbig et al. (2010) proposed the r-model, and specifically the $r$ parameter of the model, as a better measure of RH-use. The main advantage is that the r-model disentangles choices of the recognized option in recognition pairs that originate from use of the RH from the ones steaming from use of further knowledge. Making use of this superior measurement tool, we investigated the MSH prediction

that RH-use should increase the shorter the recognition and rejection latencies of objects in a pair. We found support for this hypothesis by reanalyzing 9 data sets and, in addition, with a new experiment tailored exactly to this test. Furthermore, we tested whether in the most extreme cases, that is, when the recognition and rejection latencies were shortest and therefore the probabilities that both objects are in recognition and rejection certainty states were highest, MSH-use would be the only strategy used. Our results suggested that this is not the case, therefore indicating that even under perfect conditions for relying on memory strength, people will sometimes resort to other inference strategies and integrate further knowledge.

In sum, with our work we tried to answer some questions left open by Erdfelder et al. (2011), thereby accumulating further support for the MSH. We believe we achieved this goal in two different ways: First, by finding support for its predictions for guessing and knowledge cases and in this way showing how it can parsimoniously explain a much larger chunk of data than the RH or the fluency heuristic; second, by finding converging support for its main prediction while using a more sophisticated measure of MSH use than the one employed by Erdfelder et al. (2011). Finally, our results also show that while the MSH appears to be a more useful framework than the RH, it should not be understood in a deterministic way, since even when the objects are (likely to be) in the two extreme memory states – recognition certainty and rejection certainty – people sometimes resort to strategies other than choosing the option in a higher memory-state.

## References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Castela, M., & Erdfelder, E. (in press). The memory state heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory & Cognition*.

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*(5), 1131–1138.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108–124.

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*(1), 7–22.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100–121.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that makes us smart.* Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90.

Heck, D., & Erdfelder, E. (in press). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*.

Heck, D., 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.

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*(5), 1191–1206.

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

Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition*, *43*(4), 659–671.

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*(4), 693–719.

Michalkiewicz, M., Arden, K., & Erdfelder, E. (2016). *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. (2015). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, 1–15.

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*(5), 983–1002.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology*, *2*(147), 1–14.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review*, *14*(3), 379–391.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(3), 251–271.

Pohl, R. F., Erdfelder, E., Michalkiewicz, M., Castela, M., & Hilbig, B. (in press). The limited use of the fluency heuristic: Converging evidence across different procedures. *Memory & Cognition*.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Van De Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*(3), 443–463.

Table 1: *Source and description of the 14 reanalyzed data sets.*

| Data set | Origin | Materials and criterion | N |
|---|---|---|---|
| | Michalkiewicz & Erdfelder (2015) | | |
| 1 | Exp 1, first session | 100 of 150 largest US cities, size | 19200 |
| 2 | Exp 2, first session | 100 of 150 largest US cities, size | 24900 |
| 3 | Exp 3a | 25 of 100 most successful celebrities, size | 20400 |
| 4 | Exp 3b | 25 of 100 most successful german movies, size | 20400 |
| 5 | Exp 3c | 25 of 60 largest islands, size | 19200 |
| 6 | Exp 3d | 25 of 100 most successful musicians, size | 19200 |
| 7 | Exp 4a | 25 of 100 most successful celebrities, size | 26100 |
| 8 | Exp 4b | 25 of 100 most successful celebrities, pictures, size | 26100 |
| | Michalkiewicz, Arden, & Erdfelder (2016) | | |
| 9 | Exp 1a | 25 of 100 most successful celebrities, success | 13200 |
| | Castela & Erdfelder (in press) | | |
| 10 | Exp 1, first session | 80 of 150 largest US cities, size | 9360 |
| 11 | Exp 2, first session | 80 of 150 largest US cities, size | 7920 |
| | Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder (2015) | | |
| 12 | Exp 1, control group | 20 of 61 largest world cities, size | 4370 |
| 13 | Exp 2, control group | 20 of 61 largest world cities, size | 4180 |
| 14 | Exp 3, control group | 84 of 100 largest world cities, size | 2688 |

Table 2: *Results of one-sample t-test testing if the mean of individual proportion of choices in Accordance with our hypotheses is higher than .50. For knowledge cases, Accordance means choosing the fastest recognized object, while for guessing cases Accordance means choosing the slowest unrecognized object.*

| | Knowledge Cases | | | | Guessing Cases | | | |
|---|---|---|---|---|---|---|---|---|
| | Accordance | $t$ | $df$ | $p$ | Accordance | $t$ | $df$ | $p$ |
| World Cities (size) | .60 | 8.28 | 73 | $< .001^*$ | .55 | 2.85 | 72 | $< .01^*$ |
| Celebrities (success) | .60 | 7.78 | 73 | $< .001^*$ | .55 | 2.13 | 73 | $.02^*$ |
| Rivers (length) | .67 | 9.35 | 71 | $< .001^*$ | .54 | 3.53 | 73 | $.001^*$ |

*Note*: *significant at the .05 $\alpha$ level.

Table 3: *Summary of fixed effects results in multivel logistic regression showing how the difference in latencies between two objects in a pair (RT difference) predicts the Accordance. Accordance is defined as choosing the fastest recognized object in knowledge cases, and the slowest recognized object in guessing cases.*

| Predictor | Coefficient | SE | z value | p |
|---|---|---|---|---|
| Intercept | 0.10 | 0.04 | 2.23 | .03* |
| RT difference | 0.24 | 0.08 | 3.06 | $< .01$* |
| Case (Knowledge vs. Guessing) | 0.14 | 0.04 | 3.28 | $< .01$* |
| Domain Celebrities (vs. Cities) | 0.01 | 0.03 | 0.39 | .70 |
| Domain Rivers (vs. Cities) | .02 | 0.04 | 0.67 | .50 |
| RT difference x Case Knowledge (vs. Guessing) | 0.48 | 0.07 | 6.59 | $< .001$* |

*Note*: For discrete predictors, information in parentheses clarifies the levels of the predictor which are being compared. The RT difference is scaled in seconds. *significant at the .05 $\alpha$ level.

Table 4: *Goodness-of-fit statistics, corresponding degrees of freedom and p-values for all reanalyzed data sets and Experiment 2.*

| Data Set | $G^2$ | df | p-value |
|---|---|---|---|
| 1 | 10.35 | 4 | .03* |
| 2 | 3.87 | 4 | .42 |
| 3 | 10.58 | 4 | .03* |
| 4 | 9.22 | 4 | .06 |
| 5 | 2.51 | 4 | .64 |
| 6 | 0.50 | 4 | .97 |
| 7 | 10.85 | 4 | .03* |
| 8 | 2.74 | 4 | .60 |
| 9 | 4.53 | 4 | .34 |
| 10 | 9.97 | 4 | .04* |
| 11 | 4.62 | 4 | .33 |
| 12 | 12.03 | 4 | .02* |
| 13 | 5.22 | 4 | .27 |
| 14 | 0.79 | 4 | .94 |
| Exp 2 | 7.44 | 4 | .11 |

*Note*: * indicates that the baseline model does not fit the data well, leading to statistically significant misfit.

Table 5: *Maximum likelihood parameter estimates of all r parameters and p-values and differences in FIA for comparisons between the baseline model and the order-restriced model (BO) and between the order-restricted and the equality-restricted model (OE) for all reanalyzed data sets and Experiment 2.*

| Data set | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $p_{BO}$ | $p_{OE}$ | $\Delta\text{FIA}_{BO}$ | $\Delta\text{FIA}_{OE}$ | N |
|---|---|---|---|---|---|---|---|---|---|
| 2 | .71 (.03) | .66 (.03) | .59 (.04) | .46 (.03) | 1 | 0 | 3.18 | -16.44 | 5521 |
| 4 | .86 (.02) | .83 (.03) | .73 (.03) | .70 (.03) | 1 | 0 | 3.18 | -9.18 | 4526 |
| 5 | .82 (.02) | .68 (.03) | .58 (.04) | .51(.03) | 1 | 0 | 3.17 | -25.41 | 4260 |
| 6 | .89 (.02) | .83 (.03) | .71 (.04) | .60 (.03) | 1 | 0 | 3.18 | -28.22 | 4264 |
| 8 | .78 (.02) | .74 (.03) | .62 (.03) | .51 (.03) | 1 | 0 | 3.17 | -27.73 | 5793 |
| 9 | .93 (.02) | .91 (.02) | 86 (.03) | .71 (.04) | 1 | 0 | 3.18 | -16.69 | 2929 |
| 11 | .85 (.04) | .85 (.04) | .69 (.05) | .55 (.06) | .41 | 0 | 3.18 | -11.26 | 1907 |
| 13 | .82 (.05) | .67 (.08) | .63 (.09) | .50 (.07) | 1 | $< .01$ | 3.18 | -3.97 | 902 |
| 14 | .73 (.13) | .64 (.15) | 1 (.55) | .33 (.17) | .05 | .03 | 0.61 | -3.55 | 304 |
| Exp 2 | .70 (.01) | .65 (.02) | .64(.02) | .47 (.02) | 1 | 0 | 3.18 | -53.25 | 21456 |

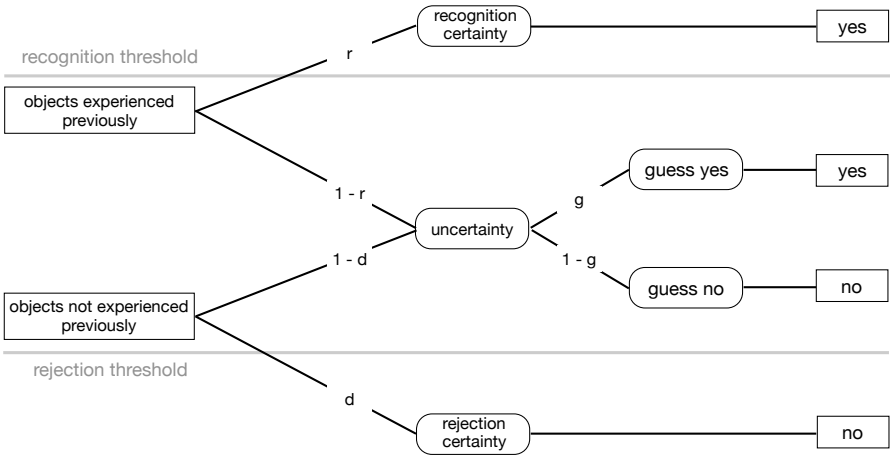



Figure 1: Graphical representation of the two-high-threshold model. Parameter $r$ denotes the probability of old objects exceeding the recognition thresholds. Parameter $d$ denotes the probability of new objects exceeding the rejection threshold. Parameter $g$ denotes the conditional probability of guessing *yes* in the uncertainty state.

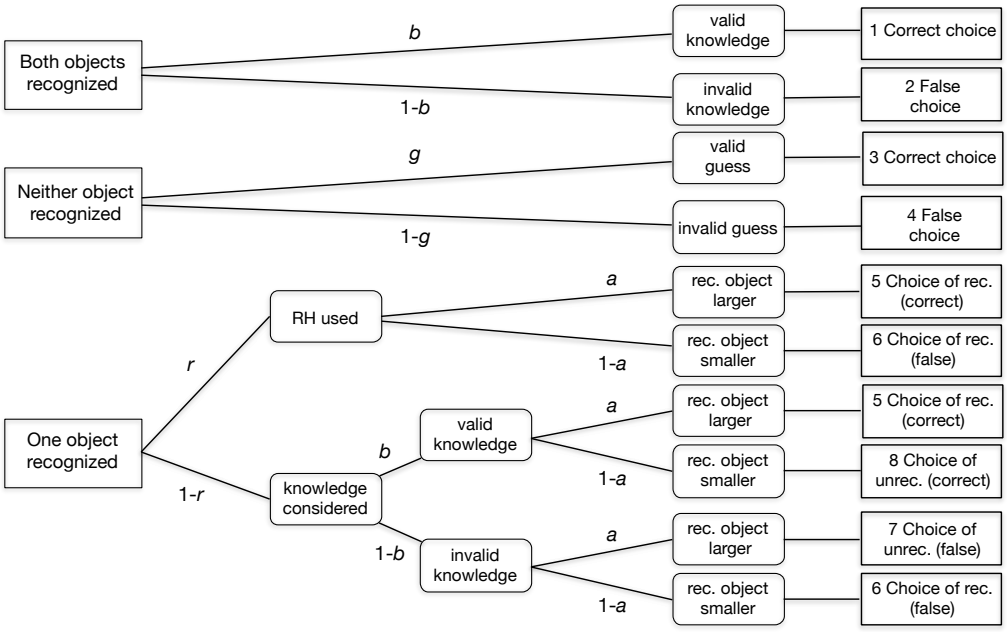



Figure 2: Graphical representation of the r-model: Parameter $r$ denotes the probability of applying the recognition heuristic as originally proposed, that is, by ignoring any knowledge beyond recognition. $a$ = recognition validity (probability of the recognized object representing the correct choice in a recognition case); $b$ = probability of valid knowledge; $g$ = probability of a correct guess; rec. = recognized; unrec. = unrecognized.

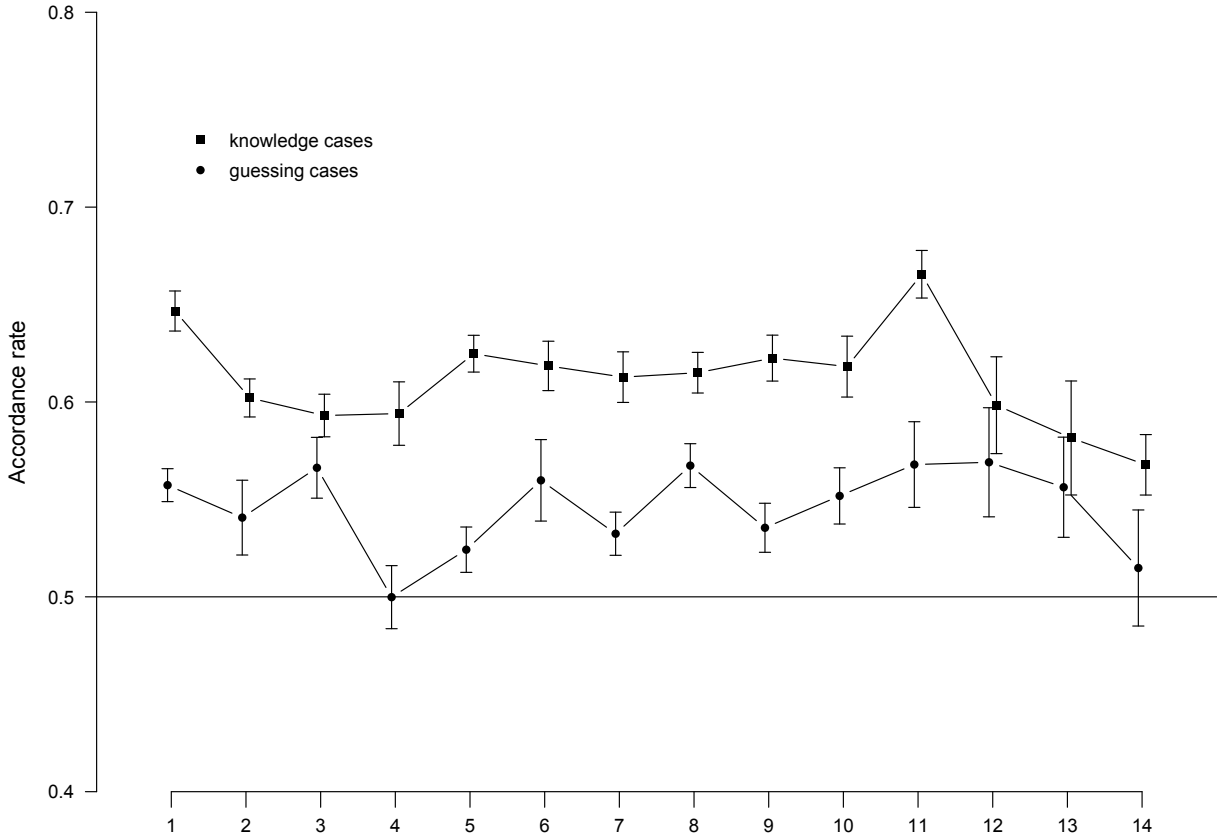



Figure 3: Proportion of choices of the fastest or slowest recognized or unrecognized object for knowledge and guessing cases, respectively, for all 14 reanalyzed datasets. Error bars represent standard error of the mean.

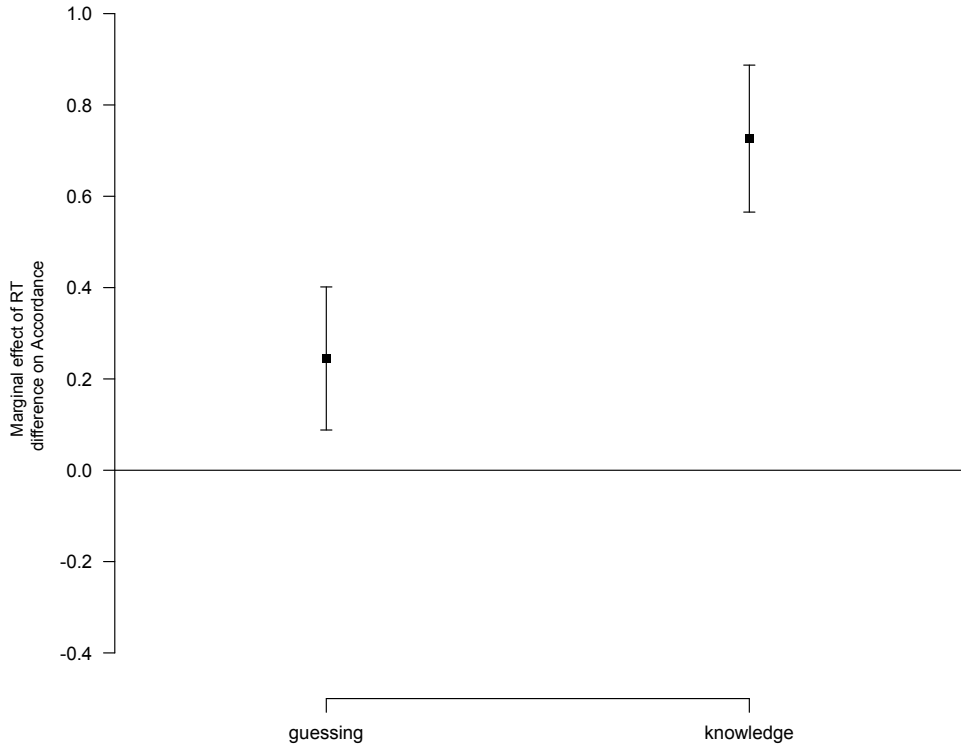



Figure 4: Marginal effect of RT difference on Accordance for guessing and knowledge cases. Error bars represent 95% confidence intervals.

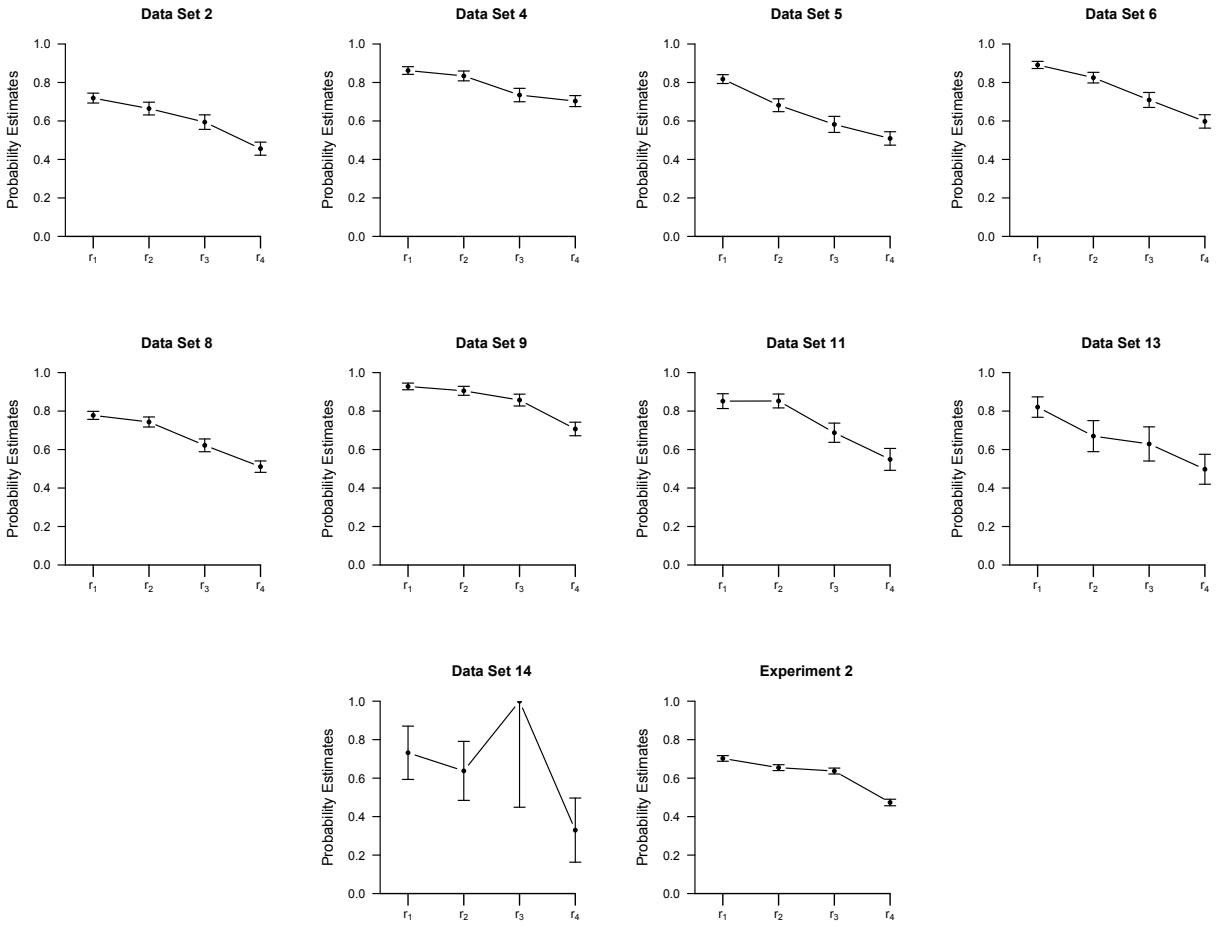



Figure 5: $r$ probability estimates in all four quartiles of recognition and rejection latency distributions for all reanalyzed datasets and for Experiment 2. Error bars represent standard errors.

Psychon Bull Rev (2014) 21:1131–1138
DOI 10.3758/s13423-014-0587-4

BRIEF REPORT

# The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach

**Marta Castela · David Kellen · Edgar Erdfelder · Benjamin E. Hilbig**

Published online: 18 March 2014


**Abstract** The recognition heuristic (RH) theory states that, in comparative judgments (e.g., Which of two cities has more inhabitants?), individuals infer that recognized objects score higher on the criterion (e.g., population) than unrecognized objects. Indeed, it has often been shown that recognized options are judged to outscore unrecognized ones (e.g., recognized cities are judged as larger than unrecognized ones), although different accounts of this general finding have been proposed. According to the RH theory, this pattern occurs because the binary recognition judgment determines the inference and no other information will reverse this. An alternative account posits that recognized objects are chosen because knowledge beyond mere recognition typically points to the recognized object. A third account can be derived from the memory-state heuristic framework. According to this framework, underlying memory states of objects (rather than recognition judgments) determine the extent of RH use: When two objects are compared, the one associated with a "higher" memory state is preferred, and reliance on recognition increases with the "distance" between their memory states. The three accounts make different predictions about the impact of subjective recognition experiences—whether an object is merely recognized or recognized with further knowledge—on RH use. We estimated RH use for different recognition experiences across 16 published data sets, using a multinomial processing tree model. Results supported the memory-state heuristic in showing that RH use increases when recognition is accompanied by further knowledge.

**Keywords** Recognition heuristic · Memory-state heuristic · Recognition memory · Decision making · Multinomial processing tree models

The recognition heuristic (RH) for comparative judgments is among the simplest heuristics proposed by Goldstein and Gigerenzer (2002) within their program of the "adaptive toolbox"—metaphorically standing for decision makers' repertoire of judgment and choice strategies. For pairwise comparisons, the RH can be stated 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" (Goldstein & Gigerenzer, 2002, p. 76). For the RH to be applied, the following preconditions have been proposed: (1) recognition is a valid cue strongly correlated with the criterion; (2) further cues are not openly available; (3) recognition stems from natural encounters in the world (Gigerenzer & Goldstein, 2011).

The typical paradigm for investigating the RH consists of a comparison task in which participants see pairs of objects and must infer, for each pair, which object has a higher value on a *criterion dimension*. The most common example is the *city-size* task in which participants decide which of two cities has the larger population. Additionally, participants engage in a recognition task for each object. That is, they state for each object whether they recognize it or not. On the basis of this information, three types of object pairs can be defined: *recognition pairs* (one object is recognized and the other is not), *knowledge pairs* (both objects are recognized), and *guessing*

**Electronic supplementary material** The online version of this article (doi:10.3758/s13423-014-0587-4) contains supplementary material, which is available to authorized users.

M. Castela (✉) · E. Erdfelder
Department of Psychology, School of Social Sciences, Universität Mannheim, Schloss Ehrenhof Ost, 68161 Mannheim, Germany
e-mail: castela@psychologie.uni-mannheim.de

D. Kellen
Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

B. E. Hilbig
Department of Psychology, School of Social Sciences, Universität Mannheim, 68161 Mannheim, Germany

*pairs* (neither of the objects is recognized). In some experiments, the recognition task additionally asks participants to state whether they merely recognized the name of the object or whether they have further knowledge about it (e.g., Hilbig & Pohl, 2009). However, despite this distinction of *recognition experiences*, participants' judgments are usually simply analyzed as recognized versus unrecognized (some exceptions are Hilbig & Pohl, 2009; Hilbig, Pohl, & Bröder, 2009).

Several studies showed that recognized objects are chosen more often than unrecognized ones in recognition pairs (for reviews, see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011). However, choosing the recognized object does not necessarily involve use of the RH. Whereas the latter implies that recognition alone determined the choice, the former can occur either from consideration of recognition alone or in combination with further knowledge about the recognized object (which will typically be in line with the recognition cue). In this sense, different accounts have been proposed for the observable tendency to choose the recognized object. According to the original RH theory, the recognized object is chosen more often because "if one object is recognized and the other one is not, then the inference is determined; no other information about the recognized object is searched for and, therefore, no other information can reverse the choice determined by recognition" (Goldstein & Gigerenzer, 2002, p. 82). We will refer to this account as the *invariance account*.

An alternative account, which we will designate as the *inhibition account*, presumes that the recognition cue can be overruled by further knowledge. Specifically, the recognized object is chosen more often not for being recognized per se, but because further information about this object leads to the same choice. This account is corroborated by several studies showing that further knowledge affects choices in recognition pairs (e.g., Bröder & Eichler, 2006; Hilbig & Richter, 2011; Newell & Fernandez, 2006). For example, people are more likely to infer that a recognized city is more populous than an unrecognized one if they know that the recognized city has a major league soccer team (Newell & Fernandez, 2006). Naturally, further knowledge can also result in the choice of the unrecognized object when the available information indicates that the recognized object is small. Nevertheless, since nothing is known (and little can be inferred) about unrecognized objects, knowledge will typically support choice of recognized objects.

A third account is given by the *memory-state heuristic* (MSH; Erdfelder, Küpper-Tetzel, & Mattern, 2011). The MSH presumes that individuals tend to choose the object that reaches a "higher" memory state—that is, a higher level of memory strength. Because criterion values are typically strongly correlated with memory strengths (Erdfelder et al., 2011), MSH use will often result in correct inferences. In line with the two-high-threshold model of recognition (e.g., Kellen, Klauer, & Bröder, 2013), the MSH assumes that objects are in one of three memory states: *recognition certainty*, *uncertainty*, or *rejection certainty*. Objects with memory strengths exceeding a *recognition threshold* are in the recognition certainty state and are judged as recognized. If the memory strength falls below this recognition threshold but is still larger than a *rejection threshold*, an object is in the uncertainty state, and the recognition judgment is determined by guessing. Finally, if the memory strength falls below the rejection threshold, an object is in the rejection certainty state and is judged as unrecognized. According to the MSH, reliance on recognition should increase with the "distance" between memory states of the to-be-compared objects. Specifically, if one object is in the recognition certainty state and the other in the rejection certainty state, reliance on recognition should be highest.

## Beyond binary recognition judgments: New predictions

As was previously mentioned, the majority of studies investigating the RH have relied on binary recognition judgments, ignoring the reported subjective recognition experiences. However, when distinguishing between nonrecognition (*U*), mere recognition (*mR*), and recognition with further knowledge ($R^+$) judgments, it can be seen that the different accounts make distinct predictions.

According to the invariance account, RH use should not vary with the composition of the recognition pairs (i.e., pairs judged $R^+$–*U* vs. *mR*–*U*), because only the binary recognition judgment determines choices and the distinction between $R^+$ and *mR* should not matter. In contrast, the inhibition account predicts that RH use will be less frequent for $R^+$–*U* pairs than for *mR*–*U* pairs, since the availability of knowledge should lead to integration of this knowledge and, by implication, decrease reliance on the RH. The MSH account makes the opposite prediction; that is, RH use should be more frequent for $R^+$–*U* than for *mR*–*U* pairs, because it is more likely that the recognized object in the former pair is in the recognition certainty state than that the recognized object in the latter pair is. Note that this prediction assumes only that reported recognition experiences ($R^+$ vs. *mR*) and underlying memory states (recognition certainty vs. uncertainty) are positively correlated. It does not require that all $R^+$ objects be in the recognition certainty state. To derive the MSH prediction, it suffices to assume that $R^+$ objects more likely originate from recognition certainty than *mR* objects do.

The MSH account makes an interesting additional prediction. Specifically, the availability of further knowledge should be used as a cue in $R^+$–*mR* knowledge pairs as well, leading to the $R^+$ object being judged as having a higher criterion value (e.g., being judged as the more populous city). Again, this prediction emerges from the fact that $R^+$ objects are more likely in a recognition certainty state than *mR* objects. The other two accounts make no such prediction, since they predict that choices for knowledge pairs will be based on retrieved knowledge only.

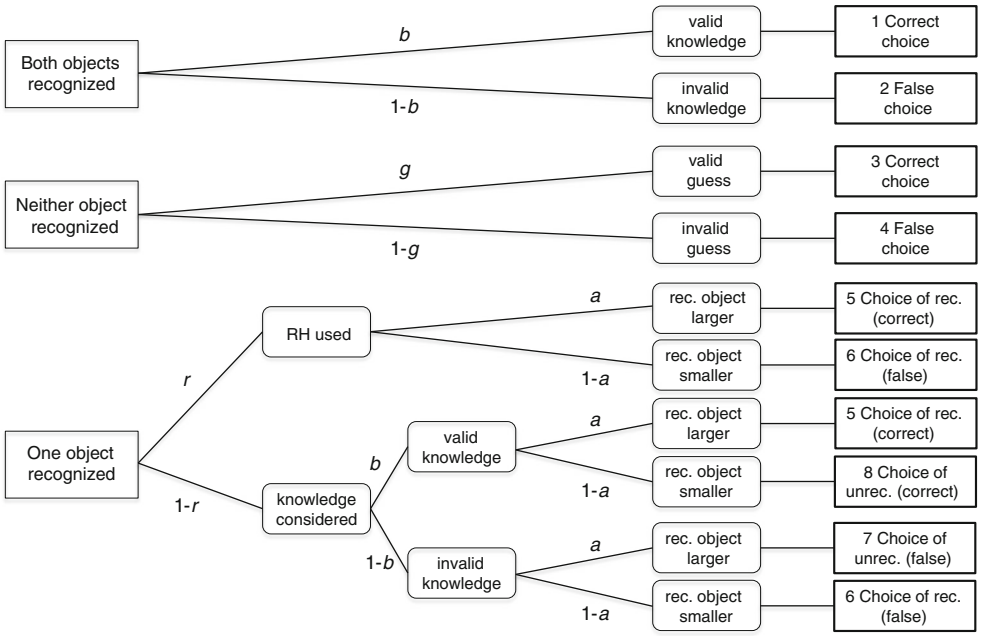



**Fig. 1** Parameter $r$ denotes the probability of applying the recognition heuristic as originally proposed—that is, by ignoring any knowledge beyond recognition. $a$ = recognition validity (probability of the recognized object representing the correct choice when paired with an unrecognized object); $b$ = probability of valid knowledge; $g$ = probability of a correct guess; rec. = R = recognized; unrec. = U = unrecognized

Finally, predictions regarding the ecological validity of the different recognition experiences can also be made. According to the MSH account, objects in the recognition certainty state should have higher criterion values than objects in the uncertainty state (Erdfelder et al., 2011). Thus, the MSH predicts that the probability of the recognized object having the larger criterion value should be greater for $R^+$–$U$ than for $mR$–$U$ pairs. The invariance account predicts no such difference, because $R^+$ and $mR$ objects are treated as equivalent if compared with unrecognized objects.

The evaluation of the above-described predictions requires the ability to disentangle the relative contributions of RH use and reliance on further knowledge. The *r-model* proposed by Hilbig, Erdfelder, and Pohl (2010) provides such a measure of RH use (via parameter $r$), while also taking into account the contribution of further knowledge. However, this model does not distinguish between different types of recognition experiences. In the next section, we first present the r-model and then propose an extension, the *r*-model*, that incorporates different recognition experiences.

## From the r-model to the r*-model

The r-model belongs to the class of multinomial processing tree models (Batchelder & Riefer, 1999; Erdfelder et al., 2009). This class of models assumes that the observed categorical responses are produced by a set of discrete mental states. The probability of each state being entered is determined by the probability of certain cognitive processes taking place or not. The models provide estimates for the probability of each of these processes taking place, producing a characterization of categorical data in terms of latent cognitive processes. Multinomial processing tree models are usually depicted as trees, with each branching presenting the occurrence (or not) of cognitive processes and the terminal nodes representing the observed categorical responses.

The r-model (Hilbig, Erdfelder, & Pohl, 2010) models data from a two-alternative forced choice comparison task and a yes–no recognition task. The recognition judgments are used to categorize the pairs into knowledge, recognition, or guessing cases, defining the three trees of the model (see Fig. 1). They lead to eight outcome categories that are described by four parameters: $r$, the probability of applying the recognition heuristic; $a$, the probability of recognition being a valid cue; $b$, the probability of valid knowledge; and $g$, the probability of a correct guess. While both the knowledge and guessing trees are defined by a single parameter that accounts for accuracy ($b$ and $g$, respectively), the recognition tree is slightly more complex. If the RH is used (with probability $r$), accuracy depends on recognition validity; with probability $a$, the inference will be correct; and with probability $1-a$, it will be false.[1] If further knowledge or any other judgment strategy is used,

[1] The $a$ parameter represents the proportion of recognition cases in which the recognized object has the larger criterion value. This parameter could be placed without loss of generality at the root of the tree or even removed implicitly via the use of two trees (for pairs in which the recognized item has the smaller or larger criterion value, respectively). We find the present parametrization the most convenient one for several (pragmatic) reasons.

the RH is not applied (with probability $1-r$), and accuracy depends on (knowledge) validity. With probability $b$, the answer is correct, and with probability $1-b$, it is false. Again, the choice of either the recognized or the unrecognized object will depend on the recognition validity (but see footnote 1).

To investigate whether use of the RH varies between recognition pairs in which the recognized object is judged as either $R^+$ or $mR$, we extended the r-model to the r*-model (see Fig. 2). The r*-model consists of six trees with 18 outcome categories in total. Because the category probabilities must sum up to one for each tree, only 12 of the 18 probabilities are free to vary. These category probabilities are represented by 10 parameters, resulting in a testable model with $12-10=2$ degrees of freedom. The r*-model comprises three trees for knowledge cases, two trees for recognition cases, and one guessing tree. The three knowledge trees refer to (1) $R^+$–$R^+$ pairs, (2) $R^+$–$mR$ pairs, and (3) $mR$–$mR$ pairs. It could be argued that this is not a knowledge tree, since, according to the participant's judgments, there is no knowledge available. Nevertheless, we refer to the parameter that accounts for accuracy in these pairs as a knowledge parameter, but more for reasons of consistency and simplicity than due to a strong claim about the availability of valid knowledge for these cases. The two recognition trees correspond to simple duplications of the original recognition tree in the r-model (each with its own set of $r$ and $b$ parameters), accounting both for $R^+$–$U$ and $mR$–$U$ pairs. Finally, as in the r-model, the guessing tree includes pairs of two unrecognized objects ($U$–$U$).

As can be seen in Fig. 2, in the $R^+$–$mR$ knowledge tree, we assume that the distinction between merely recognized objects ($mR$) and recognized objects with further knowledge ($R^+$) can be used as a simple cue. In other words, irrespective of the retrieved knowledge, the $R^+$ object would be preferred over the $mR$ object (as measured by parameter $k$). If participants use this strategy (as predicted by the MSH), a correct answer depends on the $R^+$ cue's validity (as measured by parameter $c$)—that is, on the proportion of times the object with the higher criterion value is the one judged as $R^+$. However, if this strategy is not used, participants rely on the knowledge they possess, and a correct answer will depend on the validity of knowledge (as measured by parameter $b_2$). Choice of the $R^+$ or the $mR$ object will again depend on parameter $c$.

## Model-based hypothesis testing

The hypotheses discussed previously can be represented by parameter restrictions in the r*-model:

$$\begin{array}{lll} \text{invariance account} & : r_1 = r_2, & a_1 = a_2, \\ \text{inhibition account} & : \quad r_1 < r_2, & \\ \text{MSH} & : r_1 > r_2, & a_1 > a_2. \end{array}$$

In addition to these restrictions, the MSH predicts that people use the strategy modeled by parameter $k$. Therefore, the MSH predicts that the restriction $k=0$ should produce gross misfits.

The suitability of the different parameter restrictions can be compared by evaluating the relative performance of the models instantiating them. A model selection analysis will allow us to assess which hypotheses are corroborated by the data and which are rejected. Model selection requires a weighting between the ability of each model to account for the observed data and the ability of each model to account for data in general (model complexity or flexibility), since more flexible models provide a better fit to data a priori. The goal is to find the model with the best trade-off between fit and flexibility (see Vandekerckhove, Matzke, & Wagenmakers, in press).

One prominent approach in model selection is based on the minimum description length principle (MDL; Kellen et al., 2013). According to the MDL approach, both models and data are understood as codes that can be compressed. The goal of MDL is to assess models in terms of their ability to compress data. The greater the compression, the better the account of the underlying regularities that are present in the data. One of the indices emerging from the MDL principle is the Fisher information approximation (FIA), which combines a model's goodness of fit with model flexibility penalties:

$$\text{FIA} = -\log f(x, \mathscr{M}) + \frac{p}{2}\log\frac{N}{2\pi} + \log\int \sqrt{\det I(\theta)}\; d\theta. \qquad (1)$$

The first summand of FIA corresponds to the (minus) maximum log-likelihood of observed data $x$ in a particular experiment, quantifying model $\mathscr{M}$'s fit, and the second and third summands correspond to the model penalties. The second summand takes the number of parameters $p$ and sample size $N$ into account. The third summand accounts for the flexibility of the model due to its functional form by integrating over the determinant of the expected Fisher information matrix $I(\theta)$. FIA differences larger than 1.1 already represent substantial evidence in favor of the winning model (Kellen et al., 2013).

## Analysis of data sets

The r*-model requires responses discriminating between objects that were unrecognized, merely recognized, and recognized with further knowledge. Sixteen previously published data sets fulfilled this requirement (Hilbig, Erdfelder, & Pohl, 2010, 2011, 2012; Hilbig & Pohl, 2008, 2009; Hilbig et al., 2009; Hilbig, Scholl, & Pohl, 2010). The choice task used in all data sets was the city-size task. Table 1 provides a description of each data set (additional details can be found in the Supplemental Material). FIA values and parameter estimates were calculated using the MPTinR package (Singmann & Kellen, 2013).

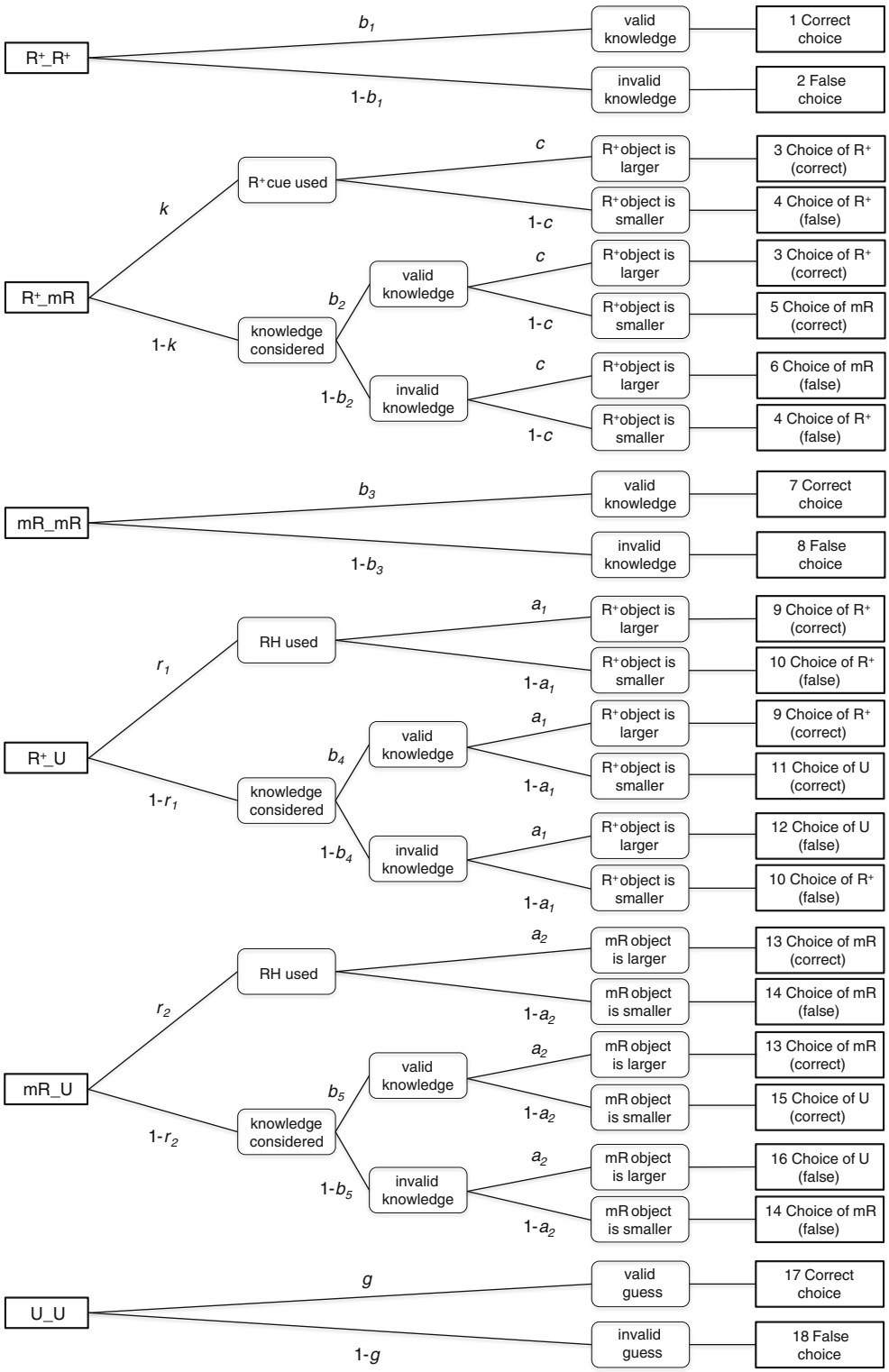



**Fig. 2** Tree representation of the r*-model. $R^+$, object recognized with further knowledge; *mR*, object merely recognized; *U*, object unrecognized; $b_1$, $b_2$, and $b_3$, knowledge validity parameters; *k*, probability of using the further knowledge cue; *c*, validity of choosing the $R^+$ object (probability that it represents the correct choice) in $R^+$–*mR* pairs; $r_1$, probability of applying the recognition heuristic (RH) in pairs for which the recognized object received an $R^+$ judgment; $a_1$, recognition validity (probability of the recognized object representing the correct choice) in pairs for which the recognized object received an $R^+$ judgment; $r_2$, probability of applying the RH in pairs for which the recognized object received an *mR* judgment; $a_2$, recognition validity (probability of the recognized object representing the correct choice) in pairs for which the recognized object received an *mR* judgment; *g*, probability of a valid guess

**Table 1** Data sets

| Data Set | Origin | Materials | *N* |
|---|---|---|---|
| 1 | Hilbig & Pohl, 2009, Experiment 1 | 20 largest Swiss cities | 4,560 |
| 2 | Hilbig & Pohl, 2009, Experiment 2 | 17 random world cities | 9,969* |
| 3 | Hilbig & Pohl, 2009, Experiment 3 | 14 largest Swiss cities | 6,188 |
| 4 | Hilbig & Pohl, 2008, Experiment 5 | 11 random world cities | 5,776* |
| 5 | Hilbig, Erdfelder, & Pohl, 2011 | 14 Polish and 14 Austrian cities | 12,012 |
| 6 | Hilbig, Pohl, & Bröder, 2009 | 14 largest Belgian cities | 7,358* |
| 7 | Hilbig, Erdfelder, & Pohl, 2010 (6a) | 17 random world cities | 2,312 |
| 8 | Hilbig, Erdfelder, & Pohl, 2010 (6b) | 17 random world cities | 2,584 |
| 9 | Hilbig, Erdfelder, & Pohl, 2010 (7a) | 14 largest Italian cities | 1,183 |
| 10 | Hilbig, Scholl, & Pohl, 2010, Experiment 1a | 16 largest Canadian cities | 1,320 |
| 11 | Hilbig, Scholl, & Pohl, 2010, Experiment 1b | 16 largest Canadian cities | 960 |
| 12 | Hilbig, Scholl, & Pohl, 2010, Experiment 2a | 16 largest Canadian cities | 2,400 |
| 13 | Hilbig, Scholl, & Pohl, 2010, Experiment 2b | 16 largest Canadian cities | 2,040 |
| 14 | Hilbig, Erdfelder, & Pohl, 2012, Experiment 1a | 18 random world cities | 3,672 |
| 15 | Hilbig, Erdfelder, & Pohl, 2012, Experiment 1b | 18 random world cities | 3,213 |
| 16 | Hilbig, Erdfelder, & Pohl, 2012, Experiment 1c | 18 random World cities | 3,672 |

*Note.* The sample size corresponds to the aggregate level: total number of trials multiplied by number of participants. For the data sets marked with an *, the total *N* does not match what was reported in the published articles. This is due to missing values in variables required for the analysis.

Following Hilbig, Erdfelder, Pohl (2010), the baseline restrictions $b_1=b_4$ and $b_3=b_5$ were imposed on the model.

The baseline model performed well in describing the data (see Table 2). For 12 of the 16 data sets, it fitted the data according to the standard $G^2$ goodness-of-fit test using $\alpha = .05$ as a criterion of significance. For 4 of the 16 data sets (data sets 5, 13, 15, and 16), there was misfit at this level of significance. However, these misfits did not exceed the critical $G^2$ values obtained in compromise power analysis (i.e., balancing of type I and type II error probabilities) given an effect size of $\omega=0.1$ under $H_1$ (see Erdfelder, 1984; Faul, Erdfelder, Lang, & Buchner, 2007).

**Table 2** Model fit and maximum likelihood parameter estimates

| Data Set | $G^2$ | *p*-value | $b_1$ | $b_2$ | $b_3$ | $k$ | $c$ | $g$ | $r_1$ | $r_2$ | $a_1$ | $a_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.85 | .09 | .75 (.02) | .85 (.02) | .68 (.02) | .37 (.05) | .80 (.02) | .52 (.02) | .77 (.03) | .63 (.03) | .93 (.01) | .79 (.01) |
| 2 | 2.44 | .30 | .70 (.01) | .73 (.02) | .62 (.02) | .46 (.03) | .70 (.01) | .54 (.01) | .73 (.01) | .45 (.03) | .82 (.01) | .74 (.01) |
| 3 | 3.67 | .16 | .74 (.01) | .78 (.02) | .64 (.02) | .42 (.03) | .70 (.01) | .56 (.02) | .84 (.02) | .67 (.02) | .82 (.01) | .73 (.01) |
| 4 | 0.94 | .62 | .65 (.01) | .67 (.03) | .52 (.02) | .60 (.02) | .48 (.02) | .53 (.02) | .70 (.02) | .49 (.03) | .57 (.01) | .62 (.01) |
| 5 | 8.00 | .02 | .66 (.02) | .69 (.02) | .63 (.01) | .50 (.02) | .65 (.01) | .53 (.01) | .82 (.01) | .70 (.02) | .86 (.01) | .81 (.01) |
| 6 | 5.08 | .08 | .69 (.02) | .71 (.04) | .64 (.02) | .61 (.04) | .78 (.02) | .57 (.01) | .82 (.02) | .52 (.02) | .94 (.01) | .78 (.01) |
| 7 | 3.82 | .15 | .64 (.03) | .84 (.04) | .66 (.03) | .50 (.06) | .72 (.03) | .52 (.02) | .74 (.03) | .63 (.04) | .79 (.02) | .70 (.02) |
| 8 | 1.34 | .51 | .63 (.02) | .63 (.04) | .61 (.04) | .50 (.05) | .58 (.03) | .51 (.02) | .84 (.02) | .75 (.04) | .79 (.01) | .77 (.02) |
| 9 | 0.99 | .61 | .71 (.04) | .81 (.05) | .53 (.05) | .41 (.11) | .86 (.03) | .50 (.03) | .75 (.05) | .57 (.06) | .94 (.01) | .69 (.03) |
| 10 | 1.42 | .49 | .52 (.08) | .65 (.09) | .51 (.04) | .64 (.08) | .67 (.04) | .59 (.02) | .98 (.01) | .67 (.04) | .82 (.02) | .74 (.02) |
| 11 | 0.03 | .98 | .58 (.06) | .72 (.10) | .58 (.04) | .67 (.09) | .75 (.04) | .54 (.03) | .95 (.02) | .50 (.06) | .82 (.02) | .70 (.03) |
| 12 | 2.38 | .30 | .62 (.03) | .67 (.04) | .56 (.03) | .40 (.06) | .62 (.03) | .53 (.02) | .77 (.03) | .52 (.04) | .80 (.02) | .68 (.02) |
| 13 | 6.09 | .05 | .63 (.03) | .84 (.05) | .62 (.04) | .60 (.07) | .75 (.03) | .53 (.02) | .85 (.02) | .56 (.05) | .78 (.02) | .68 (.02) |
| 14 | 3.17 | .20 | .66 (.01) | .74 (.03) | .67 (.02) | .30 (.04) | .56 (.02) | .45 (.02) | .56 (.03) | .42 (.04) | .59 (.02) | .59 (.02) |
| 15 | 6.48 | .04 | .68 (.01) | .75 (.03) | .64 (.03) | .39 (.04) | .55 (.02) | .50 (.02) | .58 (.03) | .41 (.05) | .64 (.01) | .57 (.03) |
| 16 | 8.07 | .02 | .64 (.02) | .62 (.03) | .64 (.02) | .13 (.05) | .53 (.02) | .46 (.02) | .69 (.02) | .63 (.03) | .57 (.02) | .56 (.02) |
| Mean | 3.67 | – | .66 | .73 | .61 | .47 | .67 | .52 | .77 | .57 | .78 | .70 |

*Note.* Standard errors in parentheses

**Table 3** Model-Selection Results: FIA indices for different versions of the r*-model applied to 16 data sets

| Data Set | Parameter Restrictions | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | baseline | $r_1=r_2$ | $r_1=r_2$ $a_1=a_2$ | $r_1=r_2$ $a_1=a_2$ $k=0$ | $r_1\leq r_2$ | $r_1\leq r_2$ $k=0$ | $r_1\geq r_2$ $a_1\geq a_2$ | $r_1\geq r_2$ $a_1=a_2$ | $r_1\geq r_2$ $a_1\geq a_2$ $k=0$ |
| 1 | 34.80 | 37.13 | 83.09 | 118.34 | 39.01 | 74.25 | **33.41** | 80.09 | 68.65 |
| 2 | 37.29 | 81.21 | 95.15 | 224.39 | 83.38 | 212.63 | **35.92** | 50.57 | 165.17 |
| 3 | 36.17 | 50.02 | 60.99 | 171.04 | 52.02 | 162.07 | **34.79** | 46.47 | 144.84 |
| 4 | 34.46 | 54.29 | 53.68 | 255.12 | 56.28 | 257.72 | 35.95 | **33.18** | 237.39 |
| 5 | 41.08 | 52.36 | 61.89 | 240.86 | 54.72 | 233.69 | **39.68** | 49.94 | 218.66 |
| 6 | 36.84 | 65.42 | 155.66 | 270.67 | 67.49 | 182.49 | **35.45** | 126.41 | 150.45 |
| 7 | 30.90 | 31.21 | 34.04 | 73.67 | 32.73 | 72.36 | **29.50** | 33.05 | 69.14 |
| 8 | 30.03 | 30.25 | 27.61 | 69.10 | 31.78 | 73.26 | 28.65 | **26.74** | 70.14 |
| 9 | 25.85 | 26.41 | 53.08 | 60.61 | 27.52 | 35.05 | **24.48** | 51.89 | 32.02 |
| 10 | 26.20 | 46.46 | 46.50 | 70.48 | 47.68 | 71.66 | **24.79** | 25.55 | 48.78 |
| 11 | 24.40 | 46.34 | 48.46 | 68.23 | 47.45 | 67.21 | **23.02** | 25.85 | 42.79 |
| 12 | 30.37 | 40.81 | 49.58 | 72.69 | 42.38 | 65.48 | **28.98** | 38.48 | 52.09 |
| 13 | 30.79 | 46.14 | 48.35 | 80.25 | 47.55 | 79.45 | **29.43** | 32.36 | 61.32 |
| 14 | 32.81 | 35.64 | 32.44 | 56.79 | 37.35 | 61.70 | 31.47 | **28.98** | 55.82 |
| 15 | 33.55 | 36.69 | 36.18 | 72.52 | 38.28 | 74.62 | **32.19** | 32.38 | 68.52 |
| 16 | 35.46 | 34.27 | **31.27** | 34.09 | 36.08 | 38.89 | 34.08 | 31.79 | 36.90 |
| Total | 521.00 | 714.65 | 917.97 | 1,938.85 | 741.70 | 1,762.53 | **501.79** | 713.73 | 1,522.68 |

*Note.* FIA indices of the winning model for each data set are set in boldface type. Following Hilbig, Erdfelder, Pohl (2010), all models have the restriction $b_1=b_4$ and $b_3=b_5$. The baseline model had no further restrictions. Extending the set of candidate models by including models without these restrictions does not change the model selection results

The results reported in Table 3 show that for the majority of the data sets (12 out of 16), the FIA metric prefers the model imposing the full set of MSH restrictions, $r_1>r_2$ and $a_1>a_2$, and provides support for $k>0$. These results are corroborated by the parameter estimates obtained with the unrestricted model, which are almost invariably consistent with these parameter restrictions (see Table 2).[2]

Three data sets (4, 7, and 14) were better accounted for by a model imposing the restrictions $r_1>r_2$ and $a_1=a_2$. This departs from the MSH only in terms of the latter's expected ecological validity, since the probability of the recognized object having the larger criterion value was not found to be reliably greater in $R^+$–$U$ pairs than in $mR$–$U$ pairs. Finally, data set 16 was better described by a model imposing the restrictions $r_1=r_2$ and $a_1=a_2$. As can be seen in the Supplemental Material, data set 16 corresponds to a condition in which speeded responses were collected. It is plausible that the retrieval of additional information from memory was impaired by this experimental constraint, leading to the use of fast, familiarity-based recognition judgments (e.g., Pachur & Hertwig, 2006).

[2] The preference for this particular restricted model did not change when including equivalent candidate models that did not include the baseline restrictions $b_1=b_4$ and $b_3=b_5$. Moreover, the FIA-based results were corroborated by order-restricted significance tests on parameter restrictions (see the Supplemental Material).

## General discussion

We tested the predictions of three different accounts about the impact of subjective recognition experiences on RH use. Overall, we found a clear pattern that was predicted by the MSH and is inconsistent with both the invariance and the inhibition accounts. RH use is more frequent when the recognized object is judged as $R^+$ than when judged as $mR$. The MSH predictions about RH use for different recognition experiences rely on the assumption that objects judged as $R^+$ are more likely to have originated from a certainty state than objects judged as $mR$. Despite the plausibility of this assumption, future efforts should be placed on implementing a complete model that associates choice predictions to latent memory states that are themselves estimated from the data (Erdfelder et al., 2011; Pachur et al., 2011). This, however, implies the possibility of distinguishing whether an object (e.g., a city name) was experienced previously or not. One way to achieve this is by inducing recognition experimentally (see Bröder & Eichler, 2006), although it can be argued that this "artificial" recognition is beyond the domain of the RH (Gigerenzer & Goldstein, 2011).

In addition to the main hypotheses, we derived two other predictions from the MSH framework. The first prediction concerns a strategy that was not investigated before—namely, choosing the object judged as "recognized with further

knowledge" ($R^+$) in a heterogeneous $R^+$–$mR$ knowledge pair, irrespective of the retrieved knowledge. The observed use of this strategy suggests that participants are relying on a difference in memory states. The second prediction relates to the recognition validities in the two recognition trees. We observed that recognition validity was (in most data sets) higher in $R^+$–$U$ than in $mR$–$U$ recognition pairs. This shows that the MSH framework reflects the environmental structure better than does the invariance account. Both results reinforce the importance of memory states in adaptive decision making and, thus, the need to go beyond simple binary yes–no recognition judgments.

In sum, we found strong support for the MSH by testing the influence of recognition experiences on RH use. The inhibition account prediction that the availability of knowledge reduces RH use was not supported, and only in one data set (under time pressure conditions) did we find support for the invariance account prediction that RH use should not differ between recognition experiences. We believe that our work shows the importance of focusing on underlying memory processes when investigating memory-based probabilistic inferences and strategies such as the RH.

**Author Note** This research was supported by the Grant Er 224/2-2 from the Deutsche Forschungsgemeinschaft (DFG).



## References

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica, 121,* 275–284.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des beta-Fehlers bei der inferenz statistischen Prüfung log-linearer Modelle [On importance and control of beta errors in statistical tests of log-linear models]. *Zeitschrift für Sozialpsychologie, 15,* 18–32.

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.

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.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making, 6,* 100–121.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90.

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.

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.

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.

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology, 55,* 394–401.

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.

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

Hilbig, B. E., & Richter, T. (2011). Homo heuristicus outnumbered: Comment on Gigerenzer and Brighton (2009). *Topics in Cognitive Science, 3,* 187–196.

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.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, 1–27.

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19,* 333–346.

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.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology, 2,* 147.

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

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (in press). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. Wang, and A. Eidels, editors, *Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press, Oxford.

Running head: A formal model of the Memory State Heuristic

# The Memory State Heuristic: A formal model based on repeated recognition judgments.

Marta Castela and Edgar Erdfelder
University of Mannheim

Author Note

This research was supported by the Grant Er 224/2-2 from the Deutsche Forschungsgemeinschaft (DFG).

Correspondence concerning this article should be addressed to Marta Castela or Edgar Erdfelder at the Department of Psychology, School of Social Sciences, University of Mannheim, L13-15, D-68161 Mannheim, Germany.
Electronic mail may be sent to castela@psychologie.uni-mannheim.de or erdfelder@psychologie.uni-mannheim.de

Word count: 15018

**Abstract**

The recognition heuristic (RH) theory predicts that, in comparative judgment tasks, if one object is recognized and the other is not, the recognized one is chosen. The memory-state heuristic (MSH) extends the RH by assuming that choices are not affected by recognition judgments per se, but by the memory states underlying these judgments (i.e., recognition certainty, uncertainty, or rejection certainty). Specifically, the larger the discrepancy between memory states, the larger the probability of choosing the object in the higher state. The typical RH paradigm does not allow estimation of the underlying memory states because it is unknown whether the objects were previously experienced or not. Therefore, we extended the paradigm by repeating the recognition task twice. In line with high threshold models of recognition, we assumed that inconsistent recognition judgments result from uncertainty whereas consistent judgments most likely result from memory certainty. In Experiment 1, we fitted two nested multinomial models to the data: an MSH model that formalizes the relation between memory states and binary choices explicitly and an approximate model that ignores the (unlikely) possibility of consistent guesses. Both models provided converging results. As predicted, reliance on recognition increased with the discrepancy in the underlying memory states. In Experiment 2, we replicated these results and found support for choice consistency predictions of the MSH. Additionally, recognition and choice latencies were in agreement with the MSH in both experiments. Finally, we validated critical parameters of our MSH model through a cross-validation method and a third experiment.

**Keywords:** recognition heuristic; memory-state heuristic; threshold models; multinomial processing tree models

In everyday life, we continually draw inferences about the world, often with partial knowledge, varying degrees of uncertainty, and limited time. For some of us, like medical doctors or stock market investors, these types of inferences are an integral part of our job, and often must be made under severe time constraints. Therefore, it comes as no surprise that a lot of psychological research has focused on how we arrive at judgments based on the integration of probabilistic cues and how accurate those judgments are. Specifically, in the last decades many researchers were interested in how people manage to make fast and frugal but yet good inferences in typical everyday contexts. One example is the research dedicated to a very simple judgment strategy, the recognition heuristic (RH; Goldstein & Gigerenzer, 1999). For pairwise comparisons, this heuristic can be described 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" (Goldstein & Gigerenzer, 2002, p. 76). Despite the fact that the heuristic assumes that people ignore further cue knowledge they might have, the RH can be very accurate. For example, imagine that a non-expert in German soccer championships is wondering on which team to bet on for the next Bundesliga year, Bayern München or TSG Hoffenheim. Assuming this person recognizes the former but not the latter team, she can (most likely correctly) infer that Bayern München will do better, simply by relying on recognition.

The RH is ecologically rational in the sense that it exploits the structure of the environment by relying on a single cue - recognition - that correlates with the choice criterion (performance of soccer teams in our example). Therefore, it is domain-specific, as it only performs well in environments where recognition correlates with the criterion (Goldstein & Gigerenzer, 1999). Moreover, it applies only to memory-based inferences, and not when information about other cues is readily available. Also, it relies on natural recognition, acquired from experience, and not on experimental recognition, manipulated in the laboratory (Gigerenzer & Goldstein, 2011).

A large body of research has investigated use of the RH and challenged its

boundaries and limitations (for reviews see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler & Goldstein, 2011). Surprisingly, one prominent aspect remains rather unexplored: the nature of the recognition process underlying the application of the RH (see Pachur, 2011). Aside from a few exceptions (e.g., Erdfelder, Küpper-Tetzel, & Mattern, 2011; Pachur & Hertwig, 2006; Pleskac, 2007; Schooler & Hertwig, 2005; Schwikert & Curran, 2014), little attention was devoted to trying to link models of recognition memory with the RH. Within these exceptions, some have notably shown how the consideration of memory processes helps understanding the RH better. For example, Schooler and Hertwig (2005) have shown that a moderate level of forgetting benefits the RH by creating partial ignorance. Pleskac (2007) has shown that as memory sensitivity (ability to distinguish novel from experienced objects) gets worse the accuracy of the RH decreases. Schwikert and Curran (2014) have used event-related potentials to distinguish the separate contributions of familiarity and recollection. While these approaches have made significant contributions, we believe there is still a major gap, namely, directly modeling the recognition process along with the decision process, thereby estimating the influence of recognition memory. We aim at addressing this gap by using a formal model to explore the influence of underlying memory processes on RH-use.

**Extending the recognition heuristic to the memory-state heuristic**

When trying to link the recognition memory literature and the RH, it is important to consider how recognition is understood in both (see Pachur, 2011). In most recognition memory studies, participants first study a list of known words and are later asked to discriminate the studied items from new (but also known) items. Therefore, in this context, all items (new and old) will have some degree of familiarity. In contrast, the typical paradigm in studies of the RH involves two tasks: a comparison task, in which participants see pairs of objects and must infer which one scores higher on a certain criterion (e.g. "Which city is more populous?") and a recognition task, where participants see all objects involved and must say which they

recognize and which they do not (e.g. "Do you recognize this city?"). The judgments from the recognition task allow the definition of three types of object pairs: knowledge pairs (both objects are recognized), recognition pairs (one object is recognized and the other is not), guessing pairs (neither of the objects is recognized). The recognition pairs will be the ones of interest since they allow use of the RH. Note that the recognition task in this case is somewhat different from the one in recognition memory studies. While in a memory recognition test the material consists of known objects for which an episodic judgment must be made, in the RH paradigm some items will have been experienced before (outside the experimental setting) and others not. Due to this difference, Goldstein and Gigerenzer (2002) made the simplifying assumption that the RH acts on the binary output of the recognition process, that is, an all-or-none distinction between the novel and the previously experienced, and that the process itself can be ignored for the purpose of studying the heuristic. Moreover, they asserted that "how often one has been exposed to something is (...) irrelevant for the frugal recognition heuristic" (Goldstein & Gigerenzer, 1999, p. 56). We argue that the ecological rationality of ignoring differences in the degree of familiarity of objects is quite questionable. If recognition is valid because exposure to objects in the natural environment (via newspapers, TV programmes, etc) correlates with the criterion, differences in the memory strength of recognized objects should be relevant.

This is the premise of the memory-state heuristic (MSH). The MSH can be seen as a straightforward extension of the RH (Erdfelder et al., 2011). It incorporates the assumption that memory strength is correlated with the criterion value, and therefore, when comparing two objects, individuals will tend to choose the one associated with a higher level of memory strength. This extension connects models of recognition memory with the RH. Specifically, its predictions follow from the two-high-threshold (2HT) model of recognition memory (Snodgrass & Corwin, 1988), which postulates that objects can be in one of three memory states: recognition certainty, uncertainty, or rejection certainty. The three states are separated by two thresholds, the recognition and the rejection threshold. If the memory strength

associated with an object exceeds the recognition threshold, the object will enter the recognition certainty state. If it lies between the recognition threshold and the rejection threshold, the object will enter an uncertainty state, and the recognition judgment is determined by guessing. Finally, if the memory strength falls below the rejection threshold, the object enters the rejection certainty state in which it will always receive a negative recognition judgment. The combination of these three states leads to different combinations of recognition pairs, depending on the memory states that underlie the recognized and the unrecognized object, respectively. Regardless of the effective $yes - no$ recognition judgment, given a decision criterion that is strongly correlated with memory strength, the two core predictions of the MSH are quite straightforward: First, if objects are in different memory states, there should be a preference for the one in a higher state; and second, the larger the discrepancy between memory states, the larger the probability of choosing the object in the higher state. Thus, the MSH makes predictions beyond recognition pairs, since these two "rules" can be applied also to knowledge pairs or guessing pairs, simply by identifying the memory states that underlie each case.

**Evidence supporting the memory-state heuristic**

To test these predictions, Erdfelder et al. (2011) relied on the fact that multinomial processing tree models like the 2HT model can be interpreted as probabilistic serial processing models of cognition (Batchelder & Riefer, 1999). This means that each branch of the model's tree corresponds to a temporal sequence of processing stages, and therefore the number of cognitive stages in each branch will influence its total processing time (Erdfelder et al., 2011). This interpretation has recently been supported by response time analyses of the 2HT model (Heck & Erdfelder, in press). It follows that in the 2HT model, recognition and rejection latencies originating from uncertainty will be stochastically larger than the ones originating from certainty, since in the former there is an additional processing stage of guessing. This allows direct response time predictions to be drawn. In this way,

Erdfelder et al. (2011) were able to explain previous results that posed a challenge for the RH. For example, it had been found that the RH adherence rate decreases with increasing recognition latencies of the recognized object (e.g., Hertwig, Herzog, Schooler, & Reimer, 2008; Newell & Fernandez, 2006). While posing a serious challenge for the RH, this result is strictly predicted by the MSH, given the assumptions outlined above.

Moreover, it was found that RH accordance rates (the amount of times the recognized object is chosen in recognition pairs) are usually larger when recognition is valid, that is, when recognition leads to a correct response (Hilbig & Pohl, 2008). This was explained by assuming use of further knowledge. However, it can be explained by the MSH without resorting to knowledge-use, since recognition should be more valid for pairs of recognition and rejection certainty than for recognition pairs involving uncertainty.

Another problematic result was observed by Hilbig and Pohl (2009) and Castela, Kellen, Erdfelder, and Hilbig (2014). This refers to the phenomenon that the RH adherence rate is higher for recognition pairs for which participants report having further knowledge about the recognized object ($R^+$) than in cases of mere recognition ($mR$) when only the name of the object is recognized without further knowledge. While the RH predicts the difference should not exist, the result can be explained by assuming that the recognition cue can be overruled by further knowledge (e.g., Hilbig & Pohl, 2008). However, the MSH would predict the same result, but by resting on the assumption that $R^+$ objects are more likely to have originated from the recognition certainty state, while $mR$ objects are more likely to have originated from the uncertainty state. Using an extension of the r-model (Hilbig, Erdfelder, & Pohl, 2010) which allows for an unbiased estimation of RH-use for both types of recognition pairs (involving $R^+$ or $mR$), Castela et al. (2014) found support for the MSH. Specifically, it was shown that the differences between RH adherence for the two types of pairs are due to higher reliance on recognition for pairs involving a $R^+$ object, and not due to use of further knowledge.

A further result that can be similarly explained has been reported by Schwikert and Curran (2014). They observed higher estimates of RH-use for recognition pairs for which the recognized item was said to be recollected than recognition pairs for which the recognized item was said to be only familiar. Assuming that recollected items are more often in the memory certainty state than familiar items, this result is clearly in line with the MSH.

Besides helping explain these previously challenging results, the MSH also led to new qualitative predictions about response latencies that could be tested. These concerned not only recognition and rejection latencies but also choice latencies, as a function of the memory state of the objects in a pair (see Erdfelder et al., 2011 for a detailed description of all predictions and results). Importantly, Erdfelder et al. have shown that RH accordance rates increase with the decreasing recognition and rejection latencies, in an additive manner. Moreover, they have shown that decision latencies in recognition pairs increase with both recognition latency of the recognized object and the rejection latency of the unrecognized object, these effects being, again, additive.

The support found for the MSH's predictions suggests its added value. However, tests of the heuristic were limited since it is only a verbal model so far. Implementing the MSH as a formal model would be ideal for testing it, but this is challenging, since a crucial variable is missing in the data: whether an object was experienced before by a participant or not. This information is missing because, as outlined before, unlike in memory studies, in studies investigating the RH one deals with natural recognition. Therefore, we simply lack knowledge about the actual status of the recognition judgments of the participants. Specifically, we do not know whether a "yes" recognition judgment is a hit or a false alarm, and whether a "no" recognition response is a correct rejection or a miss. It follows that we cannot estimate the probability of an object entering one of the three memory states using a single $yes - no$ recognition task.

**Measuring memory states through repeated recognition judgments**

To overcome this incomplete data problem we wanted to find a way to measure the memory states. Specifically, we were looking for an approach that allows us to integrate all memory state combinations and their corresponding predictions (regarding the preference for a given object) in a single formal model. Erdfelder et al. (2011) relied on response times as a proxy for the memory state and thereby successfully tested the MSH for recognition pairs. However, as already noted, this is a very limited strategy because it relies on an approximation only and does not allow formal testing. It is unclear what the cut-off point should be for a response time to be classified as originating from uncertainty versus certainty, so while response times can be used to test certain predictions, they do not allow the classification of pairs in terms of underlying memory states.

Castela et al. (2014) came closer by including the distinction between recognition with further knowledge ($R^+$) and mere recognition ($mR$) in a formal model, the r*-model. They used these two subjective experiences of recognition as a proxy to the memory state, in the sense that objects classified as $R^+$ are more likely to come from certainty than objects classified as $mR$. While this is a step forward from simply relying on response times and was very useful for critically testing the RH, evidence accumulation models and the MSH (see Castela et al., 2014), the $R^+$ versus $mR$ distinction is a relatively poor estimate of memory states. First, it is a subjective measure as it relies on participants understanding of what $mR$ or $R^+$ represents, and the idiosyncratic criteria they set for $R^+$ responses. Second, they are only informative about the memory state of recognized objects and therefore do not provide a way to model all possible memory state combinations.

Finally, another option would be to use confidence ratings, which seem like a quite appealing option. However, like the $mR$ and $R^+$ judgments, they are subjective measures, depending on the participants understanding of and idiosyncratic way to use the scale. Moreover, there is the question of how to map certain confidence ratings to memory states. While in a standard recognition memory test it has been

shown that the mapping can be done under minimal assumptions (see Bröder, Kellen, Schütz, and Rohrmeier, 2013), it is largely unclear how this could be achieved without the knowledge of which items are old and which are new.

Given all the limitations described above, we were not satisfied with any of the options listed. Therefore, we wanted to come up with a new measure that fits our goals better. This led us to consistency of recognition judgments as a proxy for memory certainty. This measure is associated with a simple extension of the RH paradigm, only involving the addition of two extra recognition tests. Furthermore, it allows for a better (although not perfect) identification of the memory states underlying the recognition judgments. According to the 2HT, if an object is in a certainty state and participants are properly instructed[1], the judgment should be consistent across repeated recognition tests. Therefore, it follows that inconsistent judgments must arise from the uncertainty state. Consistent judgments, on the other hand, will most likely arise from certainty states, although they can also result from consistent guesses. By modeling the probability of both objects in a pair entering certain memory states, we can estimate RH-use for all different memory state combinations.

To estimate MSH-use for different memory state combinations, we rely on the r-model that was originally developed for the standard RH paradigm with a single recognition test (Hilbig et al., 2010). Like threshold models, the r-model belongs to the class of multinomial processing tree models (Batchelder & Riefer, 1999; Erdfelder et al., 2009), that allow for a characterization of categorical data in terms of underlying cognitive processes. The r-model (see Figure 1) consists of three submodels that describe the three possible combinations of recognized and

[1]The assumption only holds if participants are properly instructed not to respond *yes* in subsequent recognition tests because they recognize the item from the previous test. Therefore, instructions have to be very clear, and additional measures to assess possible biases in the participants understanding of the instructions should be considered. In the Methods Section we will describe how we dealt with this problem.

unrecognized objects into pairs (knowledge, guessing, and recognition pairs). The three trees that represent the submodels lead to eight outcome categories, described by four parameters: $r$, the probability of applying the RH; $a$, the probability of recognition being a valid cue (recognition validity); $b$, the probability of valid knowledge (knowledge validity); and $g$, the probability of a correct guess. The knowledge and guessing trees are described by a single parameter each, accounting for the accuracy of the comparative judgment for the object pair ($b$ and $g$, respectively). In the recognition tree, if the RH is applied (with probability $r$), the choice will be correct with probability $a$ if recognition is valid, and it will be incorrect with probability $1 - a$. If the RH is not applied $(1 - r)$, accuracy will depend on the validity of knowledge (or other judgment strategy taking place).

Extending the r-model to our repeated recognition judgments paradigm involves essentially two steps. First, we need to account for the different memory states associated with consistent versus inconsistent recognition judgments. As clarified above, inconsistent judgments imply a memory uncertainty state. In contrast, consistent recognition judgments may arise from either certainty or uncertainty states. For consistent judgments, we model the probabilities that they originate from certainty, and, with complementary probability, that they occurred through consistent guessing. Specifically, $h$ denotes the probability that consistent "yes" judgments originate from recognition certainty whereas $l$ denotes the probability that consistent "no" judgments originate from rejection certainty. Then, depending on the memory state combination, we can use the r-model to estimate MSH-use for this specific combination. Applying the r-model to all possible combinations of memory states is straightforward. When objects are in the same state we use the knowledge tree of the r-model (if both objects are recognized with certainty) or its guessing tree (if objects are in uncertainty or in rejection states). In contrast, when objects are in different states we use the recognition tree of the r-model, here implying MSH-use for different memory-states combinations. We will refer to this extension of the r-model as the latent-states MSH model. This model is

composed of six trees leading to 18 data categories (of which 12 are free to vary)[2] that are described by 14 parameters. Obviously, this version of the model is not identifiable, since it has negative degrees of freedom. Some basic restrictions that render the model identifiable will be described in the Results Section. The six trees of the model correspond to all pair combinations of the three possibilities for repeated recognition judgments: (1) consistently recognized (repeated "yes" judgments); (2) inconsistent judgments (any combination of inconsistent judgments); (3) consistently rejected (repeated "no" judgments).

Figure 2 displays the six trees. To illustrate the logic of the model, let us describe the first one, which corresponds to the tree for pairs of objects that were both consistently recognized. We first model the probability that the consistent recognition judgment for the first object is associated with recognition certainty. With probability $h$ the first object reaches the recognition certainty state, with probability $1 - h$ it is in the uncertainty state. Then, the same is modeled for the second object. At this point, it is possible to determine which tree of the r-model will be appropriate, depending on the memory states of both objects in the pair.

The remaining five trees are built according to the same logic (model equations and a figure with the full model can be found in Appendix A). Aside from the r-model parameters for different memory state combinations, the latent-states MSH model has two parameters $h$ and $l$ that represent the probabilities of an object's consistent judgment being associated with the recognition or rejection threshold, respectively. Regarding the r-model parameters, these capture similar processes as the r-model does. However, now there are different sets of parameters that refer to the combinations of memory states they represent. For clarity, when we refer to specific parameters we will use an index that describes the type of pair, using the subscript R for recognition certainty, U for uncertainty and N for rejection certainty. Most importantly, now there are three $r$ and three $a$ parameters. This allows us to

[2]Because the category probabilities must sum up to one for each tree, only 12 of the 18 probabilities are free to vary.

test differences in MSH-use and memory-state validity between recognition pairs that originate from different memory states. Specifically, we can test the MSH prediction that both the validity and reliance on recognition should be highest when the difference between the memory states of the two objects is highest, that is, when objects are in the recognition certainty and the rejection certainty state, respectively. In our MSH model, this translates into the predictions that $r_{RN} > r_{RU}$ and $r_{RN} > r_{UN}$ for MSH-use. Analogously, we predict $a_{RN} > a_{RU}$ and $a_{RN} > a_{UN}$ for the memory-state specific validities.

Experiment 1 was designed to test these four main hypotheses. In addition, we were also interested in assessing whether an approximate version of the MSH model would be adequate to capture the results and how this affects the accuracy of the parameter estimates. In contrast to the latent-states MSH model introduced above, the approximate model version assumes that consistent judgments always originate from certainty states, thus ignoring the (unlikely) possibility of consistent guesses.

## Experiment 1

### Materials and Procedure

The experiment was conducted using E-Prime 2.0. software (Psychology Software Tools, Pittsburgh, PA). To test the latent-states MSH model, we extended the most common paradigm in the study of the RH: the city-size paradigm. As outlined above, the original paradigm incorporates two tasks, a recognition test and a city-size comparison task. We extended this paradigm by including repeated recognition judgments. Thus, the experiment consisted of three sessions, all spaced by a one-week interval.[3] In the first session, participants performed the city-size task and then the recognition test. In the subsequent two sessions they repeated the recognition test.

[3]We chose a one-week interval to ensure that sufficient time passed and therefore it is unlikely that participants can remember the full list of items from the previous session(s).

In the recognition tests, participants were instructed to indicate whether they recognize the city name or not. The instructions clearly stated that by recognition we meant having at least heard of an object before the experiment started, and not just recognizing its name from a previous session. Besides the clarity in the instructions, we took additional measures to avoid and control for a potential bias to respond *yes* in the second and third recognition tests because of the familiarity of the object associated with the previous session(s). First, we chose a large set of cities to decrease the possibility of participants memorizing the full set. Moreover, we included fillers and lures. Fillers were real city names but changed through the three sessions to add heterogeneity between sessions. Lures were fictitious city names. Since these were necessarily unknown to the participants, their recognition status in the second and third session served as proxy for the bias to respond yes.

In all three recognition tests, participants were presented with 100 city names. Eighty of those corresponded to a random sample drawn from the 150 largest US cities. Additionally, 45 city names were drawn from the remaining set of US cities. These items served as fillers, and 15 were presented at each session. Finally, we included five[4] lures, which were presented at all sessions. Each item was presented at a time, and responses were self-paced. The order of item presentation was randomized. A 500 ms interstimulus fixation-cross followed each response. Reaction times were recorded along with the participants response.

For the city-size task, the 100 items were paired. Fillers and lures were included to prevent making them recognizable as in any way different from the other objects, but we made sure that target objects were always paired with other target objects. Each target object was presented 6 times, resulting in 240 pairs that could be used for the analysis. On each trial participants were presented with a pair (order of pairs was randomized) and had to infer which of the two cities is more populous. A 500 ms interstimulus fixation-cross followed each response. Reaction times were

[4]We only included a small proportion of lures because we wanted to ensure that their presence would not impact the overall proportion of recognized items.

recorded.

### Participants

Forty-four students were recruited from the University of Mannheim. Five participants did not attend all required sessions and were therefore excluded from the analysis. The 39 participants that completed the experiment (29 women) were between 18 and 35 ($M = 22$; $SD = 4.09$) years old. Participation was monetarily compensated at the end of the last session.

### Results

As already mentioned, to minimize the likelihood that there is a *yes* bias in the second and third recognition tests despite the unambiguous instructions, we included fillers and lures in the set of cities. In all three recognition tests, recognition of lures was generally very low and there were no significant differences between sessions ($M_1 = .05$, $SD_1 = .13$, $M_2 = .08$, $SD_2 = .18$, and $M_3 = .06$, $SD_3 = .13$, in Phases 1 to 3, respectively; $F(2, 76) = .44$, $p = .65$). Moreover, the mean proportion of recognized objects was stable between sessions ($M_1 = .50$, $SD_1 = .13$, $M_2 = .50$, $SD_2 = .16$, and $M_3 = .50$, $SD_3 = .14$; $F(2, 76) = .10$, $p = .91$). Taken together, these results indicate there was no considerable *yes* bias.

### Model-based Analysis

Model-based analyses were done with MPTinR (Singmann & Kellen, 2013) in R (R Core Team, 2015), using maximum likelihood parameter estimation and model evaluation based on both the likelihood-ratio goodness-of-fit statistic $G^2$ and the Fisher Information Approximation (FIA) model selection measure that takes model complexity into account (see, e.g., Heck, Moshagen, & Erdfelder, 2014).[5] Some of the hypotheses we wish to test involve inequality restrictions (e.g., $r_{RN} \geq r_{RU}$ or

[5]When comparing two models in terms of FIA, a difference larger than 1.1 represents substantial evidence in favor of the model with smaller FIA (e.g., Kellen, Klauer, & Bröder, 2013). Moreover, for all comparisons in terms of FIA, we ensured the sample-size was above the lower-bound recommended by Heck et al. (2014).

parameters at the boundary of the parameter space (e.g., $h = 1$). For these cases, the sampling distribution of the likelihood-ratio test statistic $\Delta G^2$ under the null hypothesis does not follow a standard $\chi^2$ distribution with the appropriate degrees of freedom, but a mixture of $\chi^2$ distributions (see Iverson, 2006). For simplicity, whenever we test this type of restrictions we will do so by using a double bootstrap procedure (see van de Schoot, Hoijtink, & Dekovic, 2010). For example, if we wanted to test $h = 1$, the procedure goes as follows: 1) a non-parametric bootstrap sample is obtained from our data; 2) we fit the model imposing the null hypothesis $h = 1$ to that sample; 3) a parametric bootstrap sample is obtained from the estimated parameters; 4) both models under test (model with $h = 1$ restriction and model with no restriction on the $h$ parameter) are fitted to that sample and the difference in fit is calculated; 5) steps 1 to 4 are repeated many times (in our case, 1000 times). We then compute the $p$-value by assessing how many times the difference in fit obtained with the bootstrapped samples is equal or more extreme than the one observed with our original data set. Note that, for tests of inequality restrictions, this is a two-step process. First, we test a model imposing the inequality restriction (e.g., $r_{RN} \geq r_{RU}$) against a model imposing no restriction on these parameters. Second, we test the model imposing the inequality restriction (e.g., $r_{RN} \geq r_{RU}$) against a model imposing an equality restriction on those parameters (e.g., $r_{RN} = r_{RU}$). If the inequality restriction (e.g., $r_{RN} > r_{RU}$) is the most suitable parameter restriction, the first test should fail to reach statistical significance while the second test should lead to statistically significant results. For tests involving the double bootstrap method, we will report the $\Delta G^2$ we obtain with our data, and the double bootstrap $p$-value computed by comparing it to the $\Delta G^2$ in our bootstrap samples. For clarity, we will denote misfits and $p$-values obtained through this method with $\Delta\overline{G^2}$ and $\overline{p}$.

We started by determining the frequencies for each data category of the model. As explained before, objects for which the recognition judgments were consistently *yes* or *no* were assumed as most likely originating from recognition and rejection certainty, respectively, whereas objects for which the recognition judgment

varied were assumed to originate from the uncertainty state. The first of the three recognition judgments per object was always used as the binary recognition answer ("yes" or "no") required by the model-based analysis. The mean proportion of consistent recognition judgments (of target items) across sessions was .75 (.47 of which were consistent "no" judgments).

We fitted the latent-states MSH model with three sets of a priori restrictions, (1) $a_{UU} = .5$, (2) $b = b_{RN} = b_{RU}$, and (3) $g = b_{UN} = .5$. The first restriction concerns parameter $a_{UU}$, that is, the memory-state validity in pairs where both objects are in the uncertainty state. Following the logic outlined above, when two unrecognized objects are in the same state we model the choices through a guessing process. However, the combination of two objects in the uncertainty memory state can occur for cases where the data categories distinguish between recognized (there was a "yes" judgment in the first session) and unrecognized (there was a "no" judgment in the first session) objects. In order to model a pathway to those categories while assuming there should be no preference for one or the other, we implemented a recognition-use tree with a $r_{UU}$ parameter fixed to zero. When $r_{UU}$ is fixed to zero, the branch becomes mathematically equivalent to just having a guessing parameter if the $a_{UU}$ parameter is fixed to .5, hence the restriction. The restriction set (2) follows from the original r-model (Hilbig et al., 2010): The probability of a correct judgment based on information other than recognition is invariant whenever at least one of the objects is recognized. As Hilbig et al. (2010) corroborated this assumption empirically, we decided to stick to it in order to render the model as parsimonious as possible. Finally, the last set of restrictions (3) implies that when no object is recognized there is no valid information available and participants should guess, leading to performance approximately at chance level.

All these restriction patterns are reasonable on a priori grounds. In fact, as expected, the baseline model incorporating these restrictions performed very well in describing the data ($G^2(3) = 3.21, p = .36$, FIA = 28.87). Additionally, we tested whether we can further simplify our model by imposing an equality restriction on the

probabilities of consistent recognition judgments originating from certainty states, that is $h = l$. As show in Appendix B, the constraint $h = l$ holds if and only if $\frac{g^3}{(1-g)^3} = \frac{p(111)}{p(000)}$, where $g$ denotes the guessing probability for a "yes" recognition judgment in the memory uncertainty state, $p(111)$ represents the probability of three "yes" judgments and $p(000)$ represents the probability of three "no" judgments in the three repeated recognition tests. In other words, the $h = l$ restriction entails the assumption that people adjust their guessing probabilities such that they mirror the proportions of presumably old and presumably new items in the recognition test. Hence, the higher the proportion of items consistently judged "old" relative to the proportion of items consistently judged "new", the higher the probability of guessing "yes". This behavior corresponds to what is known as probability matching, a strategy that is well documented in many domains (Bayen & Kuhlmann, 2011; Gaissmaier & Schooler, 2008; Koehler & James, 2009; Shanks, Tunney, & McCarthy, 2002; Spaniol & Bayen, 2002). When we add the equality restriction corresponding to this assumption, model misfit increases only slightly and non-significantly ($\Delta G^2(1) = .63, p = .43$). Moreover, the model selection criterion FIA decreases, although not in a substantial amount ($\Delta$ FIA = .40). This indicates that the latent-states MSH model combined with the $h = l$ restriction provides a better balance between model fit and parsimony than a model that allows them to differ. Therefore, we added this restriction to our baseline model.

As can be seen in Figure 3 and Table 1, the pattern of the $r$ and $a$ parameter estimates is consistent with our hypotheses. If two objects are in recognition and rejection certainty states, estimated reliance on recognition is highest ($\hat{r}_{RN} = .88$). When one object is in the uncertainty state, reliance on memory-state differences decreases ($\hat{r}_{RU} = .55, \hat{r}_{UN} = .46$). By implication, when we add the inequality restrictions $r_{RN} \geq r_{RU}$ and $r_{RN} \geq r_{UN}$ to the model, model misfit does not increase ($\Delta\overline{G^2} = 0, \overline{p} = 1$) and FIA decreases ($\Delta FIA = 1.18$). Moreover, when we compare a model including the inequality restrictions with a model including the equality restrictions $r_{RN} = r_{RU} = r_{UN}$, model misfit increases drastically ($\Delta\overline{G^2} = 94.43, \overline{p} = 0$)

and so does FIA ($\Delta$ FIA = 42.64). Importantly, both restrictions lead to a significant increase in misfit ($r_{RN} = r_{RU}$: $\Delta\overline{G^2} = 55.68, \overline{p} = 0$ and $r_{RN} = r_{RU}$: $\Delta\overline{G^2} = 90.58, \overline{p} = 0$), indicating that none of the equality restrictions is compatible with the data. In summary, the results strongly support the MSH hypothesis that reliance on memory states is highest when the discrepancy between memory states of objects under comparison is largest. While we did not have a hypothesis regarding a difference between $r_{RU}$ and $r_{UN}$, it is worth noting that $r_{RU}$ is estimated to be significantly larger than $r_{UN}$ ($\Delta G^2(1) = 7.14, p < .01$, $\Delta$ FIA = .87). This will be addressed in the Discussion Section.

We also see the predicted pattern in the estimated memory-state validities. When we add the inequality restrictions $a_{RN} \geq a_{RU}$ and $a_{RN} \geq a_{UN}$ to the model, model misfit does not increase ($\Delta\overline{G^2} = 0, \overline{p} = 1$) and FIA decreases ($\Delta$ FIA = 1.14). In contrast, when we compare a model including the inequality restrictions with a model imposing the equality restrictions $a_{RN} = a_{RU}$ and $a_{RN} = a_{UN}$, model misfit increases significantly ($\Delta\overline{G^2} = 96.10, \overline{p} = 0$) and FIA also increases ($\Delta$ FIA = 42.06). Again, both restrictions lead to a significant increase in misfit ($a_{RN} = a_{RU}$: $\Delta\overline{G^2} = 36.35, \overline{p} = 0$ and $a_{RN} = a_{UN}$: $\Delta\overline{G^2} = 96.06, \overline{p} = 0$), indicating that none of the restrictions is compatible with the data. Resembling the pattern we observed for the $r$ parameters, we also observe a significant difference between $a_{RU}$ and $a_{UN}$ ($\Delta G^2(1) = 16.87, p < .001$, $\Delta$ FIA = 4.99), a result that will also be addressed in the Discussion Section.

**Approximate MSH model**

In addition to our main hypotheses we wanted to test an approximate version of our latent-states MSH model. The approximate model relies on the simplifying assumption that repeated recognition judgments indicate memory states perfectly and can be directly used to measure them. Specifically, just as inconsistent recognition judgments necessarily indicate memory uncertainty, consistent judgments are assumed to always indicate memory certainty in the approximate model. The

idea behind this model is that it could nicely complement the latent-states version. While the latent-states MSH model has the advantage of directly measuring the probabilities $h$ and $l$ that consistent judgments originate from recognition or rejection certainty states and therefore does not rely on further assumptions, it has the disadvantage that the extra parameters make it more vulnerable to sampling error, resulting in larger standard errors of the parameter estimates. In contrast, the approximate version, by having less parameters, will provide higher stability (i.e., smaller standard errors) of the parameter estimates, and therefore may be more adequate for hypotheses testing, despite the fact that it is based on an assumption that only holds approximately.

Since the approximate model corresponds to a nested version of the latent-states model, testing it simply requires fixing both certainty state probabilities $h$ and $l$ to 1. When we add this restriction $h = l = 1$ to the model, model misfit does not increase significantly ($\Delta\overline{G^2} = .66, p = .19$), although the decrease in FIA is not conclusive by itself ($\Delta$ FIA = 0.09). This indicates that the approximate model can adequately describe the data, and is suitable for testing our hypotheses. With this version of the model, the pattern in the $r$ and $a$ parameters does not change (see Figure 3), and the results perfectly converge with the results from the latent-states model.[6]

**Test of latency predictions**

As outlined above, so far the MSH had only been tested through latency predictions drawn from a serial processing interpretation of the 2HT model (see Erdfelder et al., 2011, p.13). According to this interpretation, when an object's memory strength exceeds one of the high thresholds to either recognition or a rejection certainty state, a fast judgment can be made. However, if the memory strength lies between the two thresholds, a second process (i.e., guessing) is required.

[6]The replication of all results with the approximate version of the model can be found in Appendix C

Therefore, it follows that the recognition latency distributions for responses originating from the uncertainty state should be stochastically larger than recognition latencies for responses originating from certainty states (Erdfelder et al., 2011; Heck & Erdfelder, in press).

Since we have shown that consistency versus inconsistency of repeated recognition judgment patterns is a valid proxy for memory states, we were able to test several latency predictions[7] that can be derived from the MSH more directly than was possible before. Specifically,

**(a)** both recognition and rejection latencies in the first session should be shorter for consistent recognition and rejection patterns (indicating recognition and rejection certainty), respectively, compared to those associated with inconsistent patterns (indicating recognition uncertainty);

**(b)** choice latencies should differ as a function of the distance between memory states of the objects in a pair. More precisely, choice latency should decrease with increasing distance between states.

To test prediction (a), we looked at the response latencies in the first recognition test as a function of whether they correspond to consistent or inconsistent repeated recognition judgments (see Figure 4, left-side, for visualization of the effect with untransformed response times). We then fitted a linear mixed model[8] to predict latency in the first recognition test with recognition status (yes/no) and consistency (consistent/inconsistent) as fixed effects and participant as a random effect. In line with our hypotheses, rather than testing for the main effects and the interaction of

[7]In all these and further analysis involving response times we use log-transformed response times to reduce skewness (the results do not change when we use untransformed response times). The mean of individual response times is used.

[8]The model was estimated using the lmer function of the lme4 package (Bates, Maechler, Bolkner & Walker, 2015) in R (R Core Team, 2015) and $p$-values were obtained by using the lmerTest package (Kuznetsova, Brockhoff & Christensen, 2016) which uses Satterthwaites approximations.

the two fixed factors, we compared the mean latencies as a function of consistency within recognition and rejection cases (simple main effects analyses). As predicted, both recognition and rejection latencies were significantly higher for inconsistent cases than for consistent cases ($\Delta M = 0.24, SE = 0.05, t(114) = 4.51, p < .001$ and $\Delta M = 0.18, SE = 0.05, t(114) = 3.43, p < .001$; for recognition and rejection cases, respectively).

Prediction (b) involves looking at choice latencies for different types of pairs (see Figure 4, right-side, for visualization of the effects with untransformed response times). Therefore, we used the repeated recognition judgments to categorize pairs as a function of the distance between the states. There were three categories: maximal distance between states (recognition certainty and rejection certainty); adjacent memory states (recognition certainty and uncertainty and uncertainty and rejection certainty); and same memory state (both objects in recognition certainty, both in rejection certainty and both in uncertainty). We then calculated the individual mean choice latencies for each participant and type of pair (see Figure 4) and fitted a linear mixed model predicting choice latency with type of pair (maximal, adjacent or same) as a fixed factor and participant as a random factor. As predicted, the results indicate that choices are faster for maximal pairs than for adjacent pairs ($\Delta M = 0.11, SE = 0.02, t(76) = 6.30, p < .001$) and faster for adjacent pairs than for same pairs ($\Delta M = 0.07, SE = 0.02, t(76) = 4.09, p < .001$).

Additionally, another interesting hypothesis is suggested by our model-based results. We found that reliance on recognition is higher for pairs of objects in recognition certainty and uncertainty states than for pairs of objects in uncertainty and rejection certainty states, and the same pattern occurred for the memory-state validity parameters. Therefore, we also tested if, in line with the parameter estimates, choices were faster for the former type of pairs. This was indeed the case ($\Delta M = 0.09, t(38) = 4.14, p < .001$). Thus, choices between objects in adjacent memory states are not equally fast. They are fastest for pairs in recognition certainty and uncertainty states.

**Discussion**

In Experiment 1, we introduced an extension to the city-size paradigm based on repeated recognition judgments. This extended paradigm allowed us to test core predictions of the MSH through a formal model. Our results support the formal MSH model and suggest that patterns of repeated recognition judgments provide excellent indicators for the latent memory states underlying participants' recognition judgments. We tested two nested versions of our MPT model, namely the latent-states MSH model and the approximate MSH model. They are both useful and informative, and complement each other nicely. The latent-states MSH model allows us to test the MSH core predictions without relying on the simplifying assumption of the approximate MSH model that consistent recognition patterns are always associated with certainty memory states. By directly modeling the probabilities that objects with consistent recognition judgments originated from certainty states, the latent-states MSH model takes the possibility of consistent guesses into account and thus provides purer estimates of the processes relevant during paired comparison choices than the approximate MSH model. However, as outlined above, this advantage comes at a cost: As a consequence of the extra parameters $h$ and $l$, the latent-states MSH model is more vulnerable to sampling error, resulting in larger standard errors of the parameter estimates. In contrast, the simplifications implied by the approximate model result in a higher stability (i.e., smaller standard errors) of the parameter estimates, making this parsimonious version of the model more adequate for hypothesis testing, provided that the approximation inherent to this model is at least roughly in line with the data. Therefore, by finding convergent results with both models, we can adequately test our hypotheses while asserting the quality of the repeated recognition judgments as a proxy for the memory states. Indeed, both models fitted our data well, and results based on both models are consistent with the MSH core predictions, suggesting that noncompensatory reliance on recognition in inferential decision making depends on the underlying memory states and not on the recognition judgments per se. The high degree of convergence

between the latent-states and the approximate model is due to the fact that the certainty parameters $h$ and $l$ are estimated to be close to 1 in the former model, suggesting that consistency versus inconsistency across three recognition judgments is an almost perfect empirical indicator of certainty versus uncertainty memory states, respectively.

Besides corroboration of our hypotheses, we found a result that is worth discussing: MSH-use is higher for pairs of an object in recognition certainty and an object in uncertainty than for pairs of an object in uncertainty and an object in rejection certainty. While not explicitly predicted by the MSH, this could be reasonably accommodated through an extension of the theory. In fact, the result is perfectly compatible with the MSH, since it does not contradict any of its core predictions. So far, the MSH (in line with the 2HT model) assumed a simple ordinal relationship between the states, but was silent about the distance between them. However, it is reasonable to question whether this distance is the same between different types of adjacent states. Accordingly, we found that memory-state validity mirrors the pattern in MSH-use. Therefore, at least in this dimension, it seems like the distance between states is different: Rejection certainty appears to be quite close to uncertainty, whereas recognition certainty appears to be clearly distinct from uncertainty. Correspondingly, we also observed that choices are faster for recognition certainty and uncertainty pairs than for uncertainty and rejection certainty pairs. It follows that the difference in the $r_{RU}$ and $r_{UN}$ parameters is in line with several other aspects of our data. This motivates a reconsideration of a plain ordinal assumption about how the three memory states relate, that could be explored in future studies.

Finally, we used consistency versus inconsistency of recognition judgments as an indicator for underlying memory states to test latency predictions that follow from the MSH and found support for them. Specifically, we found that (a) both recognition and rejection latencies are shorter when they originate from a certainty state than when they originate from uncertainty and that (b) choice latencies differ between pairs of different memory states in a way consistent with the MSH.

## Experiment 2

To complement Experiment 1, Experiment 2 was designed around two main goals: (1) replicating the findings of Experiment 1, thereby lending further support to the MSH and our models; (2) testing MSH predictions about choice consistency. To fulfill the first goal, we again relied on the repeated recognition judgments paradigm. Additionally, to address our second goal, we further extended the procedure so that the city-size task was also repeated across sessions. This means that participants attended the lab three times, and in each of the three sessions they performed the recognition test and the city-size task. The advantage of repeating the city-size task is that we have a measure of choice consistency in addition, which allows us to test another prediction of the MSH. Specifically, the MSH predicts that consistency should be higher for pairs of objects in different certainty states. Because reliance on the heuristic will be highest for these cases, the likelihood of consistency should also be higher. For pairs of objects in adjacent memory states, in contrast, the heuristic is not applied as often, and therefore consistency should be lower. Finally, consistency should be lowest for pairs of objects in the same state, for which the heuristic cannot be applied. In addition to the distance between the memory states of objects in a pair, one could also argue that the specific memory-state combinations matters within the adjacent and same state cases, due to state-specific differences in availability of further knowledge. Specifically, for the cases of adjacent memory states, consistency should be higher for recognition certainty and uncertainty pairs than for uncertainty and rejection certainty pairs, since in the former available knowledge about the recognized object could lead to inferences that foster consistent choices across time. Analogously, regarding same state cases, consistency should be highest for pairs of two objects in recognition certainty since, again, retrieval of further knowledge could lead to consistent choices. In contrast, it should be lower for pairs of two objects in uncertainty and two objects in rejection certainty, since, especially in the latter cases, choices rely on guessing in the first place.

### Materials and Procedure

This experiment was conducted using OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). The procedure was very similar to Experiment 1, the main difference being that not only the recognition test but also the city-size comparison task was repeated across sessions. To ensure a full replication, the material was identical to Experiment 1. The pairs for the second and third session were the same as the pairs used in the first session of Experiment 1, with the exception that the 80 pairs that contain fillers and lures included the fillers corresponding to that session.

### Participants

Thirty-nine students were recruited from the University of Mannheim. Four participants did not attend all required sessions and therefore were not included in the analysis. Two additional participants were removed because they recognized none or all of the objects in the second and third session, respectively. The 33 participants (25 women) that were included in the analysis are aged between 17 and 26 ($M = 20.82; SD = 1.89$). Participants were monetarily compensated at the end of the last session.

### Results

We took the same precautions as in Experiment 1 to prevent a *yes* bias in the second and third recognition test. In all three recognition tests, recognition of lures was generally low, but there were significant differences between sessions ($M_1 = .01, SD_1 = .05, M_2 = .10, SD_2 = .10$, and $M_3 = .08, SD_3 = .19$, in Phases 1 to 3, respectively; $F(2, 64) = 3.15, p = .05$). Therefore, we excluded four participants that recognized more than half of the lures in the second or third session. Excluding these participants successfully eliminated the effect of session in proportion of recognized lures ($M_1 = .01, SD_1 = .05, M_2 = .03, SD_2 = .09$, and $M_3 = .04, SD_3 = .10$, in Phases 1 to 3, respectively; $F(2, 56) = 1.19, p = .31$). The mean proportion of recognized objects was stable between sessions ($M_1 = .57, SD_1 = .11, M_2 = .58, SD_2 = .18$, and

$M_3 = .56, SD_3 = .17; F(2, 64) = .43, p = .65$).

**Model-based analysis**

As in Experiment 1, we determined the frequencies of each data category by considering the consistency of the recognition judgments and using the recognition judgment from the first recognition test to determine the recognition status of each object. The mean proportion of consistent recognition judgments across sessions was .78 (of which .43 were no judgments). Again, we first fitted the latent-states MSH model with the following baseline restrictions: $a_{UU} = .5, b = b_{RN} = b_{RU}, g = b_{UN} = .5$. This baseline model performed well in describing the data ($G^2(3) = 2.84, p = .42$, FIA = 27.44). Also, adding an equality restriction in the parameters $h = l$ did not increase misfit significantly and slightly decreased FIA ($\Delta G^2(1) = 0.60, p = .44, \Delta$FIA = .25). Therefore, we again relied on this more parsimonious version of the model.

As clearly shown in Figure 3, we observed the same pattern as in Experiment 1 (see also Table 1). When both objects are in certainty states (recognition and rejection) estimated reliance on recognition is highest ($\hat{r}_{RN} = .86$). When one object is in the uncertainty state reliance on recognition decreases ($\hat{r}_{RU} = .63, \hat{r}_{UN} = .56$). When we add the inequality restrictions $r_{RN} \geq r_{RU}$ and $r_{RN} \geq r_{UN}$ to the model, model misfit does not increase ($\Delta\overline{G^2} = 0, \overline{p} = 1$) and FIA decreases ($\Delta$ FIA = 1.16). When we compare a model including the inequality restrictions with a model including the equality restrictions $r_{RN} = r_{RU}$ and $r_{RN} = r_{UN}$, both model misfit and FIA increase significantly ($\Delta\overline{G^2} = 29.25, \overline{p} = 0, \Delta$FIA = 10.34). Both restrictions lead to a significant increase in misfit ($r_{RN} = r_{RU}$: $\Delta\overline{G^2} = 12.03, \overline{p} = 0$ and $r_{RN} = r_{RU}$: $\Delta\overline{G^2} = 26.49, \overline{p} = 0$), indicating that none of the equality restrictions is compatible with the data. In summary, replicating Experiment 1, the results from Experiment 2 support the MSH hypothesis that reliance on recognition is highest when the distance between memory states of the objects under comparison increases. Again, we observe that $r_{RU}$ is significantly higher than $r_{UN}$ ($\Delta G^2(1) = 4.38, p = .04, \Delta$FIA = 0.36).

Additionally, we see the predicted pattern in estimated memory-state

validities. When we add the inequality restrictions $a_{RN} \geq a_{RU}$ and $a_{RN} \geq a_{UN}$ to the model, model misfit does not increase ($\Delta\overline{G^2} = 0, \overline{p} = 1$) and FIA decreases ($\Delta$ FIA = 1.13). In contrast, when we compare a model including the inequality restrictions with a model based on an equality restrictions $a_{RN} = a_{RU}$ and $a_{RN} = a_{UN}$, model misfit increases significantly ($\Delta\overline{G^2} = 59.55, \overline{p} = 0$) and FIA also increases ($\Delta$ FIA = 23.12). Both restrictions led to an increase in misfit, indicating that none of them is compatible with the data ($a_{RN} = a_{RU}$: $\Delta\overline{G^2} = 7.51, \overline{p} < .01$; $a_{RN} = a_{UN}$: $\Delta\overline{G^2} = 50.65, \overline{p} = 0$). Again, we additionally observe a significant difference between $a_{RU}$ and $a_{UN}$ ($\Delta G^2(1) = 29.63, p < .001, \Delta\text{FIA} = 11.53$) that mirrors the effect evident in the $r$ parameters.

**Approximate MSH model**

In Experiment 2, $h$ and $l$ were estimated as 1 in the baseline latent-states model. Therefore, imposing the restriction $h = l = 1$ leads to no increase in model misfit ($\Delta\overline{G^2} = 0, \overline{p} = 1$). This further validates the adequacy of our proxy and the approximate model as a parsimonious measurement tool.

**Test of latency predictions**

To check full replicability of the results of Experiment 1, we tested the same latency predictions in Experiment 2. Specifically, we predicted that,

**(a)** both recognition and rejection latencies in the first recognition judgment should be shorter for consistent recognition and rejection patterns, respectively, compared to those for inconsistent patterns;

**(b)** choice latencies should differ as a function of the distance between memory states of the objects in a pair. More precisely, choice latency should decrease with increasing distance between states;

**(c)** choice latencies should be faster for recognition certainty and uncertainty pairs than for uncertainty and rejection certainty pairs.

To test prediction (a) we looked at the response latencies of the first recognition test as a function of whether they correspond to consistent or inconsistent repeated recognition judgments (see Figure 4). We then fitted a linear mixed model to predict latency in the first recognition task with recognition status ($yes/no$) and consistency (consistent/inconsistent) as fixed effects and participant as a random effect. Again, we tested the simple main effects of consistency within the two levels of recognition status. As predicted, both recognition and rejection latencies, respectively, were significantly higher for inconsistent cases than for consistent cases ($\Delta M = 0.31, SE = 0.04, t(83.25) = 7.23, p < .001$ and $\Delta M = 0.15, SE = 0.04, t(83.03) = 3.45, p < .01$).

To test prediction (b), we again used the repeated recognition judgments to assign pairs to memory state combinations, and calculated the individual median choice latencies for each participant and type of pair. We fitted a linear mixed model predicting choice latency with type of pair (maximal, adjacent or same) as a fixed factor and participant as random factor. As predicted, the results indicate that choices are faster for maximal pairs than for adjacent pairs ($\Delta M = 0.10, SE = 0.19, t(64) = 5.28, p < .001$) and faster for adjacent pairs than for same pairs ($\Delta M = 0.07, SE = 0.02, t(64) = 3.54, p < .001$, respectively).

Prediction (c), that choice latencies are faster for recognition certainty and uncertainty pairs than for uncertainty and rejection certainty pairs, was also supported by our results ($\Delta M = 0.11, t(28) = 4.08, p < .001$).

**Choice consistency**

In addition to the replication of Experiment 1, the second goal of this Experiment was to analyze the consistency in choices throughout the three sessions. According to the MSH, consistency in choices should relate to the distance between states: it should be highest if both objects in a pair are in different certainty states (maximal distance), less likely if one object is in the uncertainty state (adjacent states), and least likely when both objects are in the same state. Naturally, whenever

the MSH is not used, knowledge may also induce consistency. Therefore, for adjacent states, we expect that recognition certainty and uncertainty pairs will be associated with more consistency than uncertainty and rejection certainty pairs. Regarding same state pairs, we expect highest consistency for pairs when both objects are recognized with certainty. Pairs of objects in uncertainty or in rejection certainty states should be associated with less consistency.

To evaluate our predictions, we first coded consistency as a binary variable: for each pair and participant, choosing the same object in the three sessions versus making a different choice at least once. The mean proportion of consistent choices across participants was .65. We fitted a mixed effects logistic regression predicting consistency with type of pair (memory-state combination) as a fixed effect and participant as a random effect. As summarized in Table 2 (see also Figure 5), the results were in line with our predictions. The maximal distance pairs were associated with the highest consistency and adjacent states pairs were associated with higher consistency than same state pairs. Additionally, within adjacent states, recognition certainty and uncertainty pairs were associated with higher consistency than uncertainty and rejection certainty. Finally, within same state pairs, pairs of two objects in recognition certainty were also associated with higher consistency than pairs of two objects in uncertainty or in rejection certainty, although there is no significant difference between the last two.

**Discussion**

Experiment 2 had two goals: replicating Experiment 1 and assessing the MSH predictions concerning choice consistency. We fully replicated the results of Experiment 1, thereby finding additional support for MSH-use. We found that reliance on recognition is highest when the distance between the memory states is also high, and that this pattern is mirrored by the memory-states validity parameters. Furthermore, we again found that reliance on memory state information is higher for pairs of objects in recognition certainty and uncertainty than pairs of objects in

uncertainty and rejection certainty. The replication of this result adds support to the idea that the distance between these combinations of memory states might not be equivalent. Moreover, we also successfully replicated the validation of latency predictions of the MSH by using our proxy for memory states.

Regarding the second goal, we have tested yet another prediction of the MSH relating choice consistency. We have shown that, as predicted, consistency is higher for cases where the distance between states is maximal, and therefore the MSH is often used; and that consistency decreases with this distance, being smaller for pairs of objects in adjacent states, and lowest when objects are in the same state. Moreover, choices are more consistent when it is more likely that further knowledge is available.

## Validation Studies

In this paper we have introduced a new paradigm and measurement model. Both the paradigm and the MSH model are extensions of the RH paradigm and the r-model. Thus, in many ways, our model involves similar processes as the r-model does. Although the r-model has been validated previously (see Hilbig et al, 2010), we aimed (1) to establish the validity of the new parameters we introduce in the MSH model, $h$ and $l$, and (2) to demonstrate that, just like the $r$ parameter of the r-model, our three $r$ parameters mirror manipulations of MSH validities for different memory state combinations.

### Validation of the $h$ and $l$ parameters

The parameters $h$ and $l$ represent probabilities that consistent recognition judgments originated from a certainty state (recognition certainty or rejection certainty), such that the complementary probabilities ($1 - h$ and $1 - l$) represent the probability that consistency originated from guessing. In other words, these parameters prevent the requirement that consistency of recognition judgments is a perfect indicator of memory states (as the approximate MSH model assumes). While these parameters do not represent psychological processes per se, it is nevertheless

important to demonstrate that they reflect what they are supposed to measure. One clear prediction regarding these parameters is that the larger the number of sessions we include, the higher the probability that consistency of recognition judgments across these sessions is associated with memory certainty. If we increase the number of repetitions of recognition judgments, the probability of consistent judgments will be progressively less likely to be associated with an uncertainty state. Following this logic, we wanted to compare a case where recognition judgments are only repeated once (two sessions cases, 2x) with the case we used in our experiments, where the recognition judgments are repeated twice (three sessions cases, 3x). Our prediction is that $h$ and $l$ estimates are smaller in the 2x case than in the 3x case. Henceforth we will only refer to $h$, since both parameters have an equality restriction and therefore always have the same value. Thus, in a nutshell, we predict $h_{2x} < h_{3x}$.

First, we used our data of Experiment 1 and 2 to estimate $h$ for both cases, by only considering the first two sessions for the 2x case. We found that the estimates of $h$ follow the predicted pattern (Experiment 1, $h_{2x} = .84$ and $h_{3x} = .92$; Experiment 2, $h_{2x} = .85$ and $h_{3x} = 1$). Unfortunately, standard statistical analysis would not be appropriate to test whether these differences are significant, since the data in the 2x and the 3x case are statistically dependent. To overcome this problem, we opted for using a Monte Carlo cross-validation method. We split our original data into two datasets by randomly assigning half of our data points to the 2x case, and the other half to the 3x case. We then fitted the latent-states MSH model to both datasets. This process was repeated 1000 times, so we can asses how often we observe the predicted pattern $h_{2x} < h_{3x}$. Besides the baseline predictions, we included equality restrictions in the $r$ parameters between the two cases. This means that all $r$ parameters were restricted to be equal between the 2x and 3x cases ($r_{RN2x} = r_{RN3x}, r_{RU2x} = r_{RU3x}$ and $r_{UN2x} = r_{UN3x}$). These restrictions are justified to ensure that the model will not adjust the $r$ parameters to accommodate the differences between the cases, but only if they are shown to be reasonable. Therefore, before assessing the pattern in the $h$ parameters we wanted to assess the impact of

these restrictions. We first performed a compromise power-analysis with G*power (Faul, Erdfelder, Lang, Buchner, 2007) to calculate the optimal critical $\Delta G^2$ value (for a $\chi^2(3)$ test detecting small deviations from the null, $w = .1$, and ensuring that type 1 and type 2 error probabilities are equal). The optimum critical value is 28.2 for the data of Experiment 1, and 24.4 for the data of Experiment 2. In all 1000 iterations, $\Delta G^2(3)$ never exceeded the corresponding critical value in Experiment 1, and only 0.1% of the times for Experiment 2. In fact, had we used the standard $\alpha$ level of .05, the increase in misfit due to the constraints $r_{RN2x} = r_{RN3x}, r_{RU2x} = r_{RU3x}$ and $r_{UN2x} = r_{UN3x}$ would be significant in 5% of the samples for Experiment 1 and 7% for Experiment 2, a result that almost perfectly matches the expectation when the constraints hold in both experiments. In light of these results, we decided to assess our hypothesis concerning $h_{2x}$ and $h_{3x}$ under these restrictions. We observe the expected pattern, $h_{2x} < h_{3x}$, in 99% of the cases for Experiment 1 and 96% for Experiment 2. We tested the increase in misfit caused by adding the $h_{2x} \leq h_{3x}$ restriction. The addition of this restriction resulted in no increase in misfit in any of the 1000 iterations, for both experiments. When comparing a model including an inequality restriction, $h_{2x} \leq h_{3x}$, with a model with an equality restriction, $h_{2x} = h_{3x}$, we observe this leads to a significant increase in misfit in 71% of the cases for Experiment 1 and 52% for Experiment 2. Thus, the overall pattern of results is quite consistent and in line with our interpretation of the model parameters $h$ and $l$.

**Experiment 3: Validation of the $r$ parameters**

As mentioned before, the $r$ parameters in our model are borrowed from the r-model, which has been validated. However, given the more complex nature of our model, showing that all our three $r$ parameters reflect manipulations of the degree of reliance on MSH-use should add additionally confidence to our results. Therefore, we conducted an experiment that mimics the second experiment in Hilbig et al. (2010; Data Set 7). In brief, we present one group of participants with a domain where memory strength is a valid cue, and the other with a domain where it is not. In line

with Hilbig et al. (2010; see also Pohl, 2006, Experiment 1) we hypothesized that MSH-use should decrease for the latter case. Within our model, this would translate into the following three hypotheses: (1) $r_{RN1} > r_{RN2}$ ,(2) $r_{RU1} > r_{RU2}$ ,(3) $r_{UN1} > r_{UN2}$, where 1 denotes the condition with a valid domain, and 2 denotes the condition with a non-valid domain.

**Materials and Procedure**

We followed the same procedure as in Experiment 1 but with different materials. The target items consisted of the 30 largest Italian cities. Additionally, we randomly selected 30 more Italian city names by drawing from the $31^{st}$ largest to the $70^{th}$ largest Italian cities. These served as fillers, and 10 were presented at each session. Finally, we used 5 very small (less than 600 inhabitants) Italian *comunes* as lures. In the first session, participants in both groups first had a recognition test with all 45 cities. The cities were paired so that target items are only paired with other target items, while fillers and lures are paired together. Each target city appeared 16 times, creating 240 target pairs. Additionally, each filler and lure was repeated 4 times, resulting in 30 additional pairs. For the comparison task, participants were randomly assigned either to a control condition where they had to judge which of the cities is more populous (domain where recognition is valid) or to an experimental condition where they had to judge which city is higher above sea level (domain where recognition is not valid). Performance on the comparison task was monetarily incentivized.

**Participants**

We initially recruited 54 participants from the University of Mannheim, but unfortunately, 9 participants did not return to the second and/or third session. Of these 45 participants, 6 of them recognized all cities in one or more of the sessions, and therefore had to be removed before the analysis. Of the remaining 39 participants, 23 (12 women; aged between 19 and 41, M = 24, SD = 5.98) had been assigned to the city-size group, while 16 (9 women; aged between 18 and 35, M =

21.31, SD = 4.00) were in the height-above-sea-level group. At the end of the third session, participants were compensated with money or course credit, and had an additional monetary bonus which was a function of their performance in the comparison task.

**Results**

In all three recognition tests, recognition of lures was generally low, and there were no significant differences between sessions ($M_1 = .09, SD_1 = .17, M_2 = .12, SD_2 = .23$, and $M_3 = .08, SD_3 = .14$, in Phases 1 to 3, respectively; $F(2, 88) = 0.79, p = .46$). The mean proportion of recognized objects was also stable across sessions ($M_1 = .67, SD_1 = .18, M_2 = .69, SD_2 = .21$, and $M_3 = .68, SD_3 = .22; F(2, 88) = 0.56, p = .57$). This is in line with the assumption that participants followed our instructions for the repeated recognition tests.

**Model-based Analysis**

We first fit the latent-states MSH model to the data of both conditions. The model performed well with the baseline restrictions, $a_{UU} = .5, b = b_{RN} = b_{RU}, g = b_{UN} = .5$, for both conditions analyzed simultaneously ($G^2(6) = 8.15, p = .23$, FIA = 50.95). Once more, adding the equality restriction $h = l$ in each of the two conditions led to no significant increase in misfit and FIA ($\Delta G^2(2) = 1.35, p = .51, \Delta$FIA $= .32$). Thus, replicating Experiments 1 and 2, the probabilities of originating from recognition and rejection certainty states, respectively, do not differ between consistent *yes* and consistent *no* recognition judgments.

A first test of validity concerns the parameter $h = l$. Since the object domain (i.e., Italian cities) and material does not differ between conditions and participants were randomly assigned to conditions, the $h = l$ parameter must not differ between the city-size (C) and height (H) conditions. To test this straightforward prediction, we added the restriction $h_C = l_C = h_H = l_H$. As predicted, this restriction led to no significant increase in misfit and slightly reduced FIA ($\Delta G^2(1) = 0.57, p = .32, \Delta$FIA

= .21). In light of these results, all additional tests of our hypotheses are based on the baseline model assuming $h_C = l_C = h_H = l_H$. All parameter estimates (and corresponding standard errors) for this baseline model can be found in Table 1.

Second, as a manipulation check, we wanted to make sure that the memory-state validity is higher in the city-size group for all three types of recognition pairs. As can be seen in Table 1, this is the case for all three memory-state validity parameter estimates. Therefore, we compared the baseline model with a model imposing the inequality restrictions, $a_{RN,C} \geq a_{RN,H}$, $a_{RU,C} \geq a_{RU,H}$ and $a_{UN,C} \geq a_{UN,H}$. This leads to no significant increase in misfit or FIA ($\Delta\overline{G^2} = 0, \overline{p} = 1, \Delta\text{FIA} = 2.02$). In contrast, comparing the inequality restricted model with one imposing equality restrictions between conditions in the $a$ parameters, $a_{RN,C} = a_{RN,H}$, $a_{RU,C} = a_{RU,H}$ and $a_{UN,C} = a_{UN,H}$, leads to a significant increase in misfit and FIA ($\Delta\overline{G^2} = 194.29, \overline{p} = 0, \Delta\text{FIA} = 89.92$). Importantly, all three restrictions separately led to a significant increase in misfit ($a_{RN,C} = a_{RN,H}$: $\Delta\overline{G^2} = 32.53, \overline{p} = 0$; $a_{RU,C} = a_{RU,H}$: $\Delta\overline{G^2} = 43.34, \overline{p} = 0$; $a_{UN,C} = a_{UN,H}$: $\Delta\overline{G^2} = 19.08, \overline{p} = 0$), indicating that none of them is compatible with the data.

Third, since our manipulation worked as predicted, we then tested our main hypotheses. As can be seen in Table 1, all three $r$ parameter estimates are larger in the city-size group than in the heigth-above-sea-level group. Therefore, we compared the baseline model with a model imposing the following inequality restrictions, $r_{RN,C} \geq r_{RN,H}$, $r_{RU,C} \geq r_{RU,H}$ and $r_{UN,C} \geq r_{UN,H}$. This led to no increase in misfit and a decrease in FIA ($\Delta\overline{G^2} = 0, \overline{p} = 1, \Delta\text{FIA} = 2.07$). Furthermore, in line with our hypotheses, comparing the inequality restricted model with a model imposing equality restrictions, $r_{RN,C} = r_{RN,H}$, $r_{RU,C} = r_{RU,H}$ and $r_{UN,C} = r_{UN,H}$, resulted in a significant increase in misfit and FIA ($\Delta\overline{G^2} = 242.38, \overline{p} = 0, \Delta\text{FIA} = 116.08$). Importantly, all three restrictions separately led to a significant increase in misfit($r_{RN,C} = r_{RN,H}$: $\Delta\overline{G^2} = 54.48, \overline{p} = 0$; $r_{RU,C} = r_{RU,H}$: $\Delta\overline{G^2} = 87.66, \overline{p} = 0$; $r_{UN,C} = r_{UN,H}$: $\Delta\overline{G^2} = 10.01, \overline{p} = 0$), indicating that none of them is compatible with the data. In sum, the results clearly support our hypothesis, showing that, like RH-use (Hilbig et

al., 2010), MSH-use adaptively adjusts to the memory-state validity.

## General Discussion

In this paper, we aimed at testing a formal model of the MSH with a new paradigm especially suited to acquiring better estimates of the memory states underlying binary recognition judgments. In two experiments, our model-based analyses consistently revealed that, as predicted, reliance on memory states increases with the discrepancy in the underlying states. Moreover, the same pattern was observed in memory-state validity, supporting the ecological validity of the MSH.

Our two nested MSH models constitute the first attempt at formalizing the MSH. With these models, we could independently estimate reliance on recognition for pairs of all different combinations of memory states, allowing a direct test of the core predictions of the MSH. Consistent with previous work (Castela et al., 2014; Erdfelder et al., 2011) we found strong support for the idea that underlying memory states, and not recognition judgments per se, influence reliance on recognition as a single cue in inferential judgments. Importantly, these core predictions of the MSH rely on the assumption that memory strength, and not recognition judgments, correlates with the criterion value in the first place. It follows that memory-state validity should also be higher for recognition pairs of two objects in certainty memory states. We have also successfully tested this prediction, which underlines the fact that following the MSH is an ecologically rational and well-adapted choice strategy.

Furthermore, we successfully validated critical parameters of our model. Specifically, we first focused on the two filter parameters of the latent-states MSH model, $h$ and $l$. These parameters should estimate the probability that consistency in recognition judgments is associated with a recognition or rejection certainty state, respectively. To ensure they correctly serve that purpose, we tested the straightforward prediction that increasing the number of repetitions of the recognition test should lead to higher estimates of $h$ and $l$. This prediction is based on the 2HT: If an object is in a certainty state, the recognition judgment can only be

*yes* for the case of recognition certainty and *no* for the case of rejection certainty; if an object is in the state of uncertainty, the recognition judgment will be based on a guessing process, and can vary. It directly follows that the probability of consistent judgments originating from uncertainty decreases when the number of repetitions increases. We compared the case of one repetition with the cases of two repetitions, and found the predicted pattern: the estimates of $h$ and $l$ are smaller in the former case. In this way, we validated the filter parameters.

In addition, we also wanted to validate the three $r$ parameters. While the $r$ parameter has been validated in the context of the r-model, since we have three different ones to account for all types of recognition cases, it is important to show they all respond to manipulations that should affect MSH-use. Therefore, we compared MSH-use between two conditions with different memory-state validities. Specifically, we predicted that when memory-state validity is very low, MSH-use for all types of recognition pairs should decrease. This same manipulation has been used to validate the $r$ parameters of the r-model (Hilbig et al., 2010), and it should also affect our $r$ parameters. Accordingly, we have shown that all three $r$ parameters are significantly smaller in a condition with low memory-state validity compared to a condition with higher memory-state validity.

Our experiments also revealed a consistent pattern that initially had not been predicted: MSH-use is higher for pairs of objects in recognition certainty and objects in uncertainty states than for pairs of objects in uncertainty and rejection certainty states. As discussed before, we found this pattern in Experiment 1 and 2 and also observed corresponding results for the memory-state validities and choice latencies in both experiments. Additionally, the pattern in MSH-use and memory-state validity is present for both groups in Experiment 3 (MSH-use: $\Delta G^2(1) = 14.33, p < .001, \Delta\text{FIA} = 2.72$; memory-state validity: $\Delta G^2(1) = 24.88, p < .001, \Delta\text{FIA} = 6.53$). Taken together, these results should motivate a reconsideration of a simple ordinal view of the three states. This could be pursued in future studies where specific hypotheses would guide a clear test of its source.

Besides all tests of hypotheses, we also conducted comparative model-based analyses to find the best version of our model. In this process, we found a reasonable set of restrictions. One important restriction is in the $h$ and $l$ parameters. When they are both constrained to be equal to 1, our latent-states model becomes the approximate MSH model, by incorporating the assumption that the memory state is perfectly captured by consistency versus inconsistency in recognition judgments. As becomes clear by inspecting Figure 3, the pattern of results does not change given the model. However, there are differences that, while quite predictable, are still interesting to discuss. While the latent-states MSH model is the superior model in terms of how it captures the processes without further assumptions, the question remains of whether it is the best measurement tool. Despite its inherent misspecification, the parsimonious approximate MSH model does a great job at capturing the pattern of results, and it does not suffer from estimation uncertainty as much as the latent-states model does. This becomes clear by looking at the standard errors of the parameter estimates (see Figure 3). We find that we have the best of two worlds by having both models. Since the results converge, we can surpass the limitations of the approximate model - showing that the pattern of results does not depend on using consistency versus inconsistency of recognition judgments as perfect indicators of memory states. At the same time, we can also surpass the limitations of the latent-states model - ensuring that our results hold even when we test our hypotheses with parameter estimates that are less uncertain.

In addition to the model-based analyses, we used repeated recognition judgments to test latency predictions of the MSH. Assumptions about recognition latency differences between memory states had previously been used to test the MSH (see Erdfelder et al., 2011), but had not been validated so far. With our proxy measure we had an opportunity to validate them using an independent and arguably better indicator of memory states. Based on the data from both experiments, we were able to show that recognition and rejection latencies are generally faster for objects in certainty states (indicated by consistent recognition judgments) than for objects in

the uncertainty state (indicated by inconsistent recognition judgments). Also, we have shown that choices tend to be faster with increasing distance between states.

Finally, since in Experiment 2 we repeated both the recognition test and the inference task, we were able to test the MSH predictions regarding choice consistency, most importantly, that when the distance between memory states of an object pair is maximal, choices are most consistent. Consistency should be lower when objects are in adjacent memory states (and MSH-use is lower), and lowest when they are in the same state, since for this case the heuristic cannot be applied. This is exactly the pattern we observed, lending further support to the MSH and showing its vast predictive potential. Moreover, when we look at the different cases within the adjacent states and same state categories, we observe that choices are more consistent when further knowledge is likely to be available, and less consistent in cases where plain guessing is involved.

We have discussed earlier why we chose the consistency of recognition judgments as our proxy due to its advantages over other options. But naturally, it also has limitations. Perhaps the most evident one is the paradigm. The fact that it requires three separate recognition tasks forces a costly procedure, where participants must be asked to return to the lab for two subsequent sessions. In our case, we have opted for a one-week interval between sessions, since we thought this would provide sufficient time to avoid perfect memory of previous judgments. One evident way to simplify the procedure would be to have all recognition testes in a single session, separated by some distractor tasks. It remains an empirical question whether this would compromise the procedure, and while we think it would lead to a higher number of consistent judgments, it is hard to argue either against or for it without appropriate testing.

Another way to simplify the procedure would be to reduce the number of sessions to two. Given the quality of the proxy with three sessions, we believe it would be sufficient to repeat the judgments once without severe contamination, even if using only the approximate MSH model. But most importantly, we would like to

clarify that we developed this paradigm and new model in order to test the MSH, and we do not wish to suggest that this should be the new standard paradigm for studying the RH. While we strongly recommend that the recognition processes are taken into consideration for drawing new predictions, we do not wish to promote the use of our paradigm as a standard tool. Ideally, we hope our work can inspire the development of even better methods that involve a less costly paradigm and take us one more step forward.

Another aspect worth addressing is how our estimates of MSH-use compare to previous findings of RH-use that ignore differences in underlying memory states. By working with the latent MSH model, we isolated the ideal preconditions for reliance on recognition, that is, the subset of recognition cases where both objects are in different certainty states (i.e., recognition versus rejection certainty). Following the rationale of the MSH, this should lead to comparably higher estimates of reliance on recognition than found before. Indeed, we observed that, for two objects in opposing certainty states, reliance on recognition was .86 and .88 on average (for Experiment 1 and 2, respectively). This is considerably larger than the usual estimates of RH-use previously found with the r-model for similar judgment domains (see Table 1 of Hilbig et al., 2010 for an overview), and even higher than the level of RH-use observed in an experiment where participants were explicitly instructed to use the RH as often as possible ($\hat{r} = .82$, cf. Hilbig et al., 2010, Experiment 6). This suggests that estimating reliance on recognition by only considering binary recognition judgments leads to a gross underestimation of the use of memory strength in inferential decision making when conditions for using it are ideal. When defined in terms of memory states, we see that recognition information is actually used more often, in line with the ecological validity of memory states. Accordingly, Castela et al. (2014) found similarly high estimates of RH-use between .69 and .98 (.77 on average), but only for recognition cases for which there was further knowledge about the recognized object (arguably associated with the recognition certainty state). For recognition cases where there was mere recognition only, RH-use was much lower

(between .45 and .70, .57 on average; see Table 2 of Castela et al., 2014 for an overview of the estimates of RH-use for both types of recognition cases).

The high estimates for parameter $r_{RN}$ bring up the question whether the hypothesis $r_{RN} = 1$ is compatible with our data. This would imply that further knowledge is entirely ignored when one object is certainly recognized and the other is certainly rejected, much in line with a deterministic interpretation of RH use under ideal conditions (cf., Hilbig et al., 2010). We opted to test this hypothesis based on the approximate MSH model because (a) this model is compatible with the data of both experiments and (b) it provides the more powerful test due to smaller standard errors of $\hat{r}$. More precisely, we tested $H_0 : r_{RN} = .99$ rather than $H_0 : r_{RN} = 1$. This slightly weaker null model is more reasonable because the latter restriction would predict zero frequencies for some categories of the model. A single observation in one of these cells would thus result in infinite misfit, rendering a rejection of this model trivial. Notably, for both experiments the hypothesis $r_{RN} \geq .99$ leads to severe increase in both misfit and FIA ($\Delta G^2(1) = 783.09, p < .001, \Delta\text{FIA} = 389.36$, and $\Delta G^2(1) = 388.75, p < .001, \Delta\text{FIA} = 191.47$, for Experiment 1 and 2, respectively)[9]. These results are clear-cut and conceptually replicate corresponding results of Castela et al. (2014). In sum, while our model-based results show that people rely on the MSH quite often when the conditions for successful applications of this strategy are ideal, we still see that this is not the only strategy in play, even under ideal conditions.

Finally, we would like to point out that there are alternatives to the MSH model proposed in the current paper, some of which have already been discussed in the relevant literature. The present model is probably most similar to the r*-model previously proposed by Castela et al. (2014). Compared to the latter model, however, the MSH model has two major advantages: it does not assume perfect empirical indicators for the latent memory states and it accommodates all three memory states (recognition certainty, uncertainty, and rejection certainty) underlying recognized and

[9] A similar, although less extreme, pattern of results is found when testing $r_{RN} = 1$ based on the latent-states MSH model.

unrecognized objects, while the r*-model only approximates the recognition certainty state and the uncertainty state for recognized objects. Moreover, rather than conceptualizing memory strength (or activation, cf. Schooler & Hertwig, 2005) as a discrete variable with three states it is of course also possible to conceive it as a continuous variable. This would be consistent with signal detection theory which, like the 2HT model, is a prominent model in the recognition memory literature (see Kellen & Klauer, in press; Pleskac, 2007) but also with the ACT-R approach (see Schooler & Hertwig, 2005). However, we aimed at formalizing a model of recognition-based inference that is closest to the original idea of the RH (Goldstein & Gigerenzer, 2002), with the single exception that it takes the possibility of memory uncertainty into account. We believe that the MSH model allows for such a generalization of the RH in a more parsimonious way than approaches based on the notion of continuous memory strength. In fact, the MSH-model proposed here contains the original RH-model as a special case that occurs when the probability of memory uncertainty is zero.

In sum, by formalizing the MSH and finding support for its main predictions, the current work takes a new step in bridging the gap between theories of recognition memory and the RH theory. Furthermore, we have shown the potential of our memory states proxy measure (consistent versus inconsistent recognition judgments) to validate previous predictions of the MSH regarding recognition and choice latencies and to derive and test new predictions regarding choice consistency. We believe our vast set of results shows the benefits of extending the RH to the MSH and having a formal model of the latter, allowing us to test several predictions and better understand the processes involved in reliance on recognition for probabilistic inferences.

# References

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

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

Bayen, U. J., & Kuhlmann, B. G. (2011). Influences of source–item contingency and schematic knowledge on source monitoring: Tests of the probability-matching account. *Journal of Memory and Language*, *64*(1), 1–17.

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

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*(5), 1131–1138.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108–124.

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, *6*(1), 7–22.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of

research. *Judgment and Decision Making*, *6*(1), 100–121.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that makes us smart.* Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90.

Heck, D., & Erdfelder, E. (in press). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*.

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.

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*(5), 1191–1206.

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

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology*, *55*(6), 394–401.

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

Iverson, G. J. (2006). An essay on inequalities and order-restricted inference. *Journal of Mathematical Psychology*, *50*(3), 215–219.

Kellen, D., & Klauer, K. C. (in press). Elementary signal detection and threshold theory. In E.-J. Wagenmakers (Ed.), *Stevens handbook of experimental psychology and cognitive neuroscience, fourth edition (vol. V).* New York: John Wiley & Sons, Inc.

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*(4), 693–719.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*(1), 123–127.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). lmertest: Tests in linear mixed effects models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=lmerTest` (R package version 2.0-30)

Matôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experimental bulder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(4), 333–346.

Pachur, T. (2011). The limited value of precise tests of the recognition heuristic. *Judgment and Decision Making*, *6*(5), 413–422.

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*(5), 983–1002.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology*, *2*(147), 1–14.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review*, *14*(3), 379–391.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(3), 251–271.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from

https://www.R-project.org/

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628.

Schwikert, S. R., & Curran, T. (2014). Familiarity and recollection in heuristic decision making. *Journal of Experimental Psychology: General*, *143*(6), 2341–2365.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Spaniol, J., & Bayen, U. J. (2002). When is schematic knowledge used in source monitoring? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 631.

Van De Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*(3), 443–463.

Table 1: Latent-states MSH parameter estimates (and bootstrapped standard errors) of Experiment 1, Experiment 2 and Experiment 3.

| | Experiment 1 | Experiment 2 | Experiment 3: city-size | Experiment 3: height |
|---|---|---|---|---|
| $a_{UU}$ | .50 (-) | .50 (-) | .50 (-) | .50 (-) |
| $a_{RN}$ | .80 (.02) | .76 (.02) | .82 (.02) | .57 (.03) |
| $a_{RU}$ | .69 (.02) | .70 (.03) | .84 (.04) | .64 (.03) |
| $a_{UN}$ | .61 (.01) | .60 (.02) | .70 (03) | .51 (.03) |
| $r_{RN}$ | .88 (.06) | .86 (.05) | 1 (.10) | .26 (.06) |
| $r_{RU}$ | .55 (.05) | .63 (.08) | .74 (.13) | .02 (.05) |
| $r_{UN}$ | .46 (.04) | .56 (.06) | .54 (.07) | .27 (.07) |
| $b$ | .65 (.01) | .64 (.02) | .74 (.02) | .57 (.01) |
| $b_{RN}$ | .65 (.01) | .64 (.02) | .74 (.02) | .57 (.01) |
| $b_{RU}$ | .65 (.01) | .64 (.02) | .74 (.02) | .57 (.01) |
| $b_{UN}$ | .50 (-) | .50 (-) | .50 (-) | .50 (-) |
| $g$ | .50 (-) | .50 (-) | .50 (-) | .50 (-) |
| $h$ | .92 (.06) | 1 (.08) | .76 (.07) | .76 (.07) |
| $l$ | .92 (.06) | 1 (.08) | .76 (.07) | .76 (.07) |

*Note*: Parameters $a_{UU}$, $b_{UN}$, and $g$ are fixed to .50. Additionally, parameters $b$, $b_{RN}$, and $b_{RU}$ are restricted to be equal, as well as parameters $h$ and $l$. In Experiment 3, parameters $h$ and $l$ are additionally restricted to be equal between the city-size and height condition.

Table 2: Summary of mixed effects logistic regression showing how the combination of memory states within a pair predicts choice consistency in Experiment 2.

| Predictor | Coefficient | SE | Wald Z | p |
|---|---|---|---|---|
| Intercept | 0.41 | 0.10 | 4.11 | $< .001$ |
| Maximal vs. all others | 0.27 | 0.01 | 21.29 | $< .001$ |
| Adjacent vs. Same | 0.14 | 0.01 | 10.25 | $< .001$ |
| REC-UNC vs.UNC-REJ | 0.28 | 0.05 | 6.07 | $< .001$ |
| REC-REC vs. UNC-UNC and REJ-REJ | 0.27 | 0.03 | 8.93 | $< .001$ |
| UNC-UNC vs. REJ-REJ | 0.13 | 0.07 | 1.81 | .07 |

*Note*: REC - recognition certainty; UNC - uncertainty; REJ - rejection certainty.

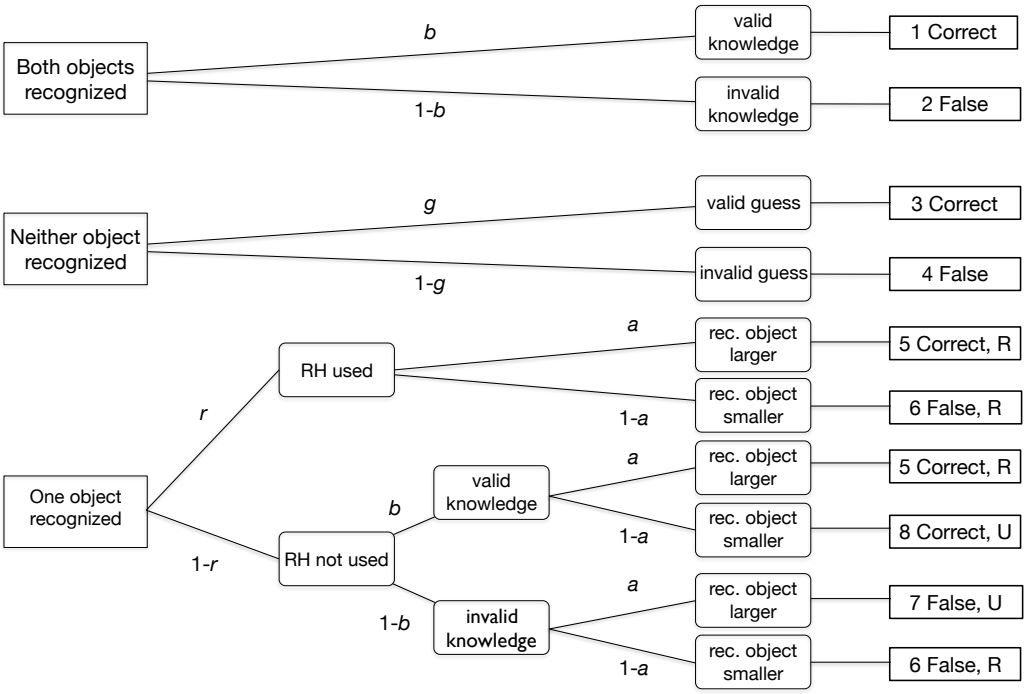



Figure 1: Graphical representation of the r-model: Parameter $r$ denotes the probability of applying the recognition heuristic as originally proposed, that is, by ignoring any knowledge beyond recognition. $a$ = recognition validity (probability of the recognized object representing the correct choice in a recognition case); $b$ = probability of valid knowledge; $g$ = probability of a correct guess; rec. = recognized; unrec. = unrecognized.

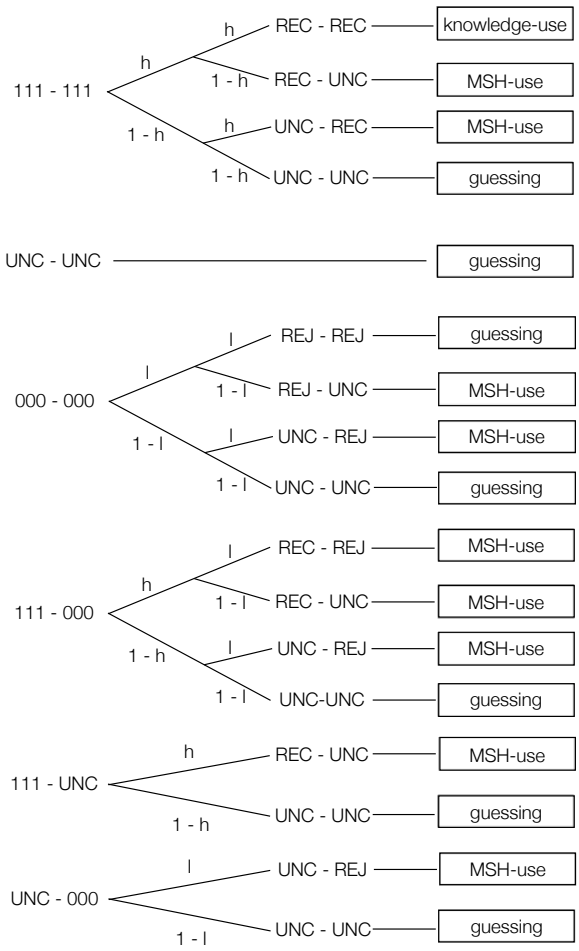



Figure 2: Abstract representation of the latent-states MSH model, denoting how the filter parameters determine the memory-state combination under comparison and, consequently, the appropriate decision process. $h$, probability that consistent recognition judgments originate from recognition certainty; $l$, probability that consistent rejection judgments originate from rejection certainty; 111, consistently recognized objects; 000, consistently rejected objects; REC, recognition certainty; UNC, uncertainty; REJ, rejection certainty. The full model is presented in Appendix A.

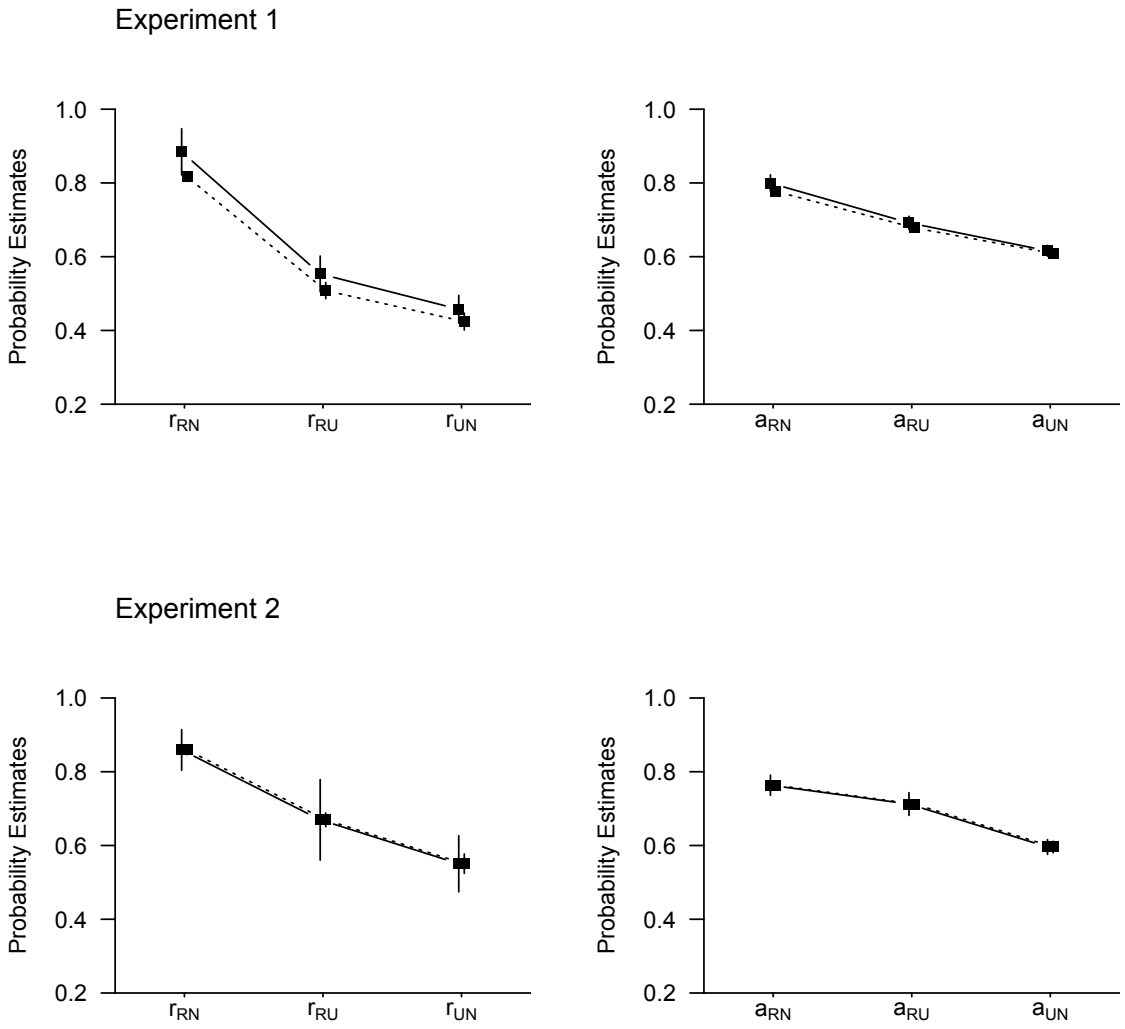



Figure 3: Estimates of the three $r$ and $a$ parameters for Experiment 1 and 2. Solid lines represent the estimates from the latent-states MSH model while dashed lines represent estimates from the approximate MSH model. Error bars represent bootstrapped standard errors of the parameter estimates.

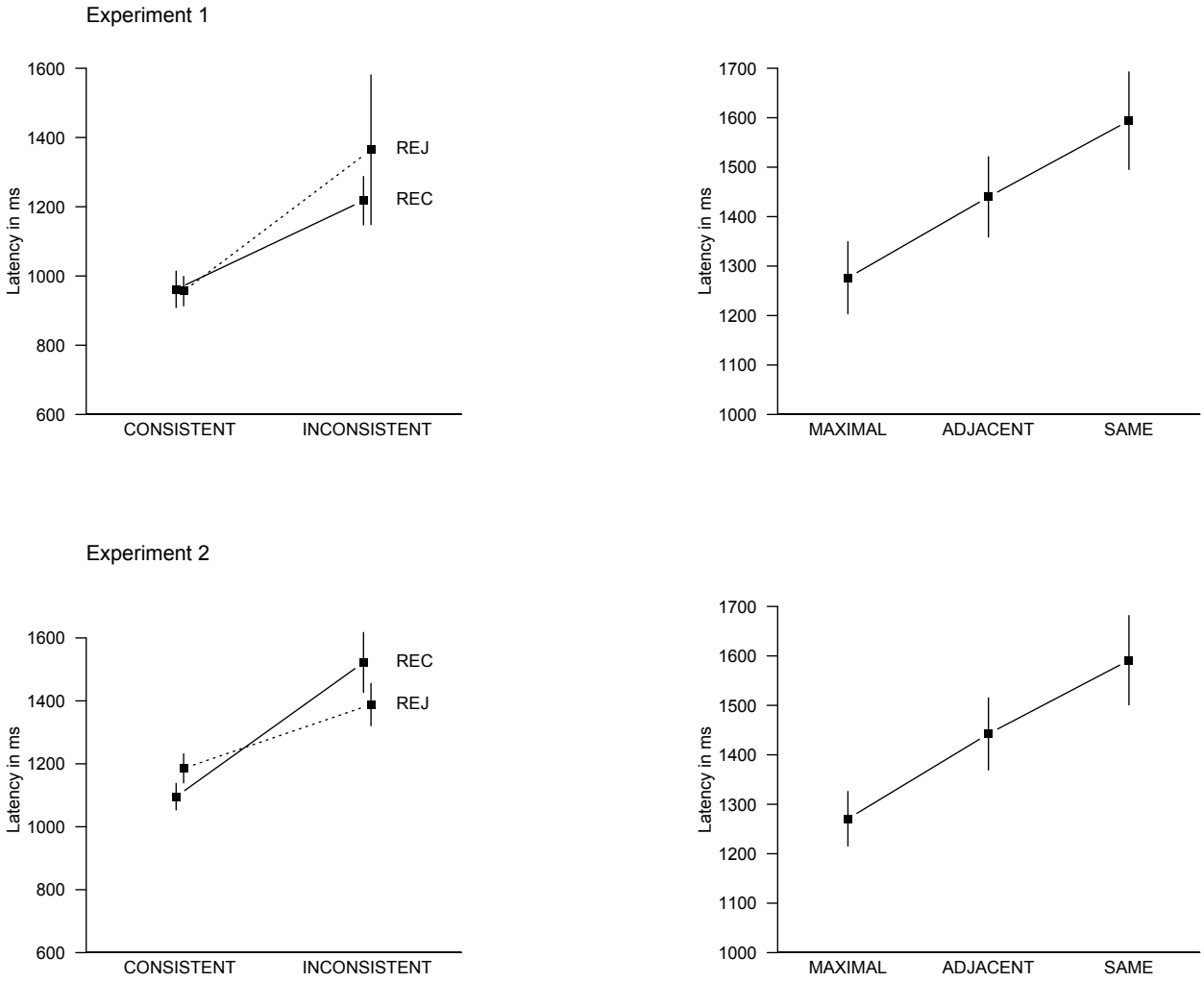



Figure 4: Left-side plots show means of individual median recognition (REC) and rejection (REJ) latencies from the first session, separately for consistent and inconsistent repeated recognition judgments in Experiment 1 and 2. Right-side plots represent means of individual median choice latencies in the first session for pairs where the distance between the states is maximal (recognition certainty and rejection certainty); for pairs where objects are in adjacent memory-states (recognition certainty and uncertainty; rejection certainty and uncertainty); and for pairs in the same state (both objects in recognition certainty, both in rejection certainty and both in uncertainty) for Experiment 1 and 2. Error bars represent standard errors.

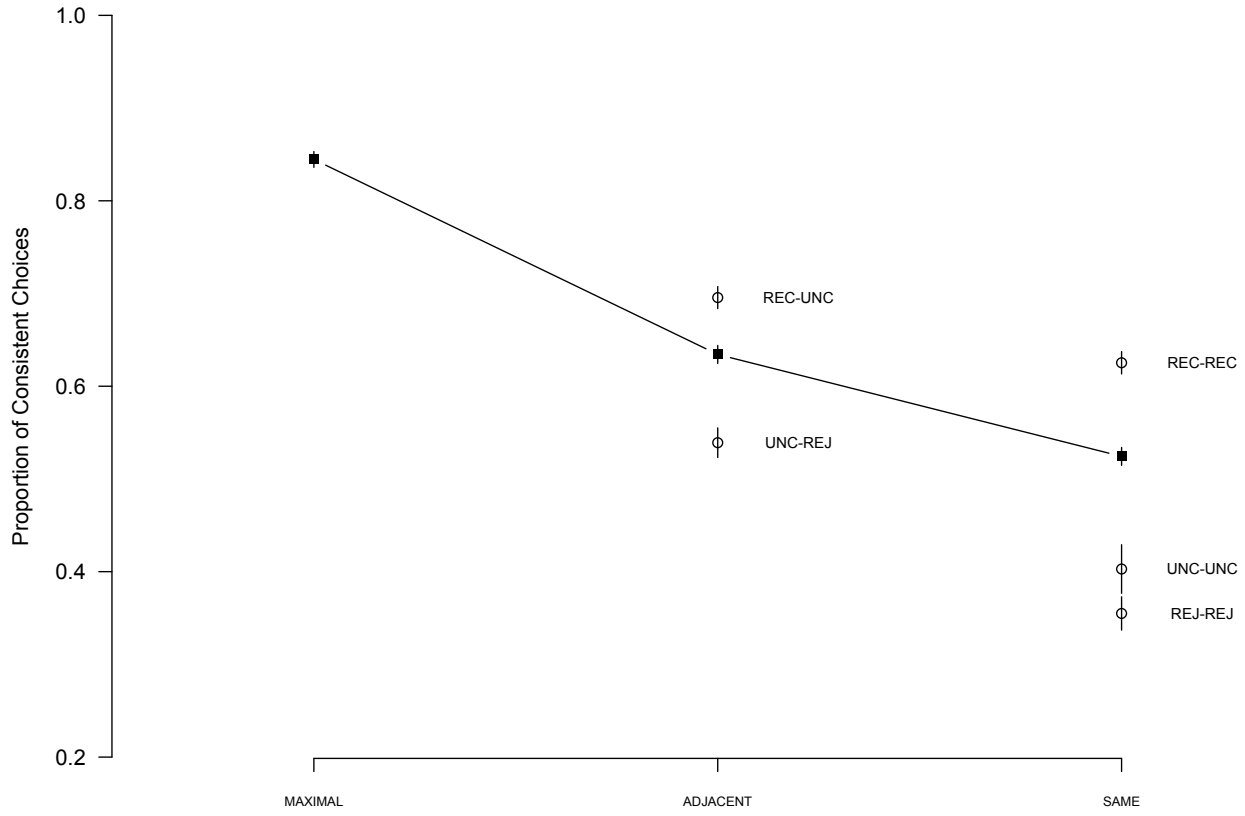


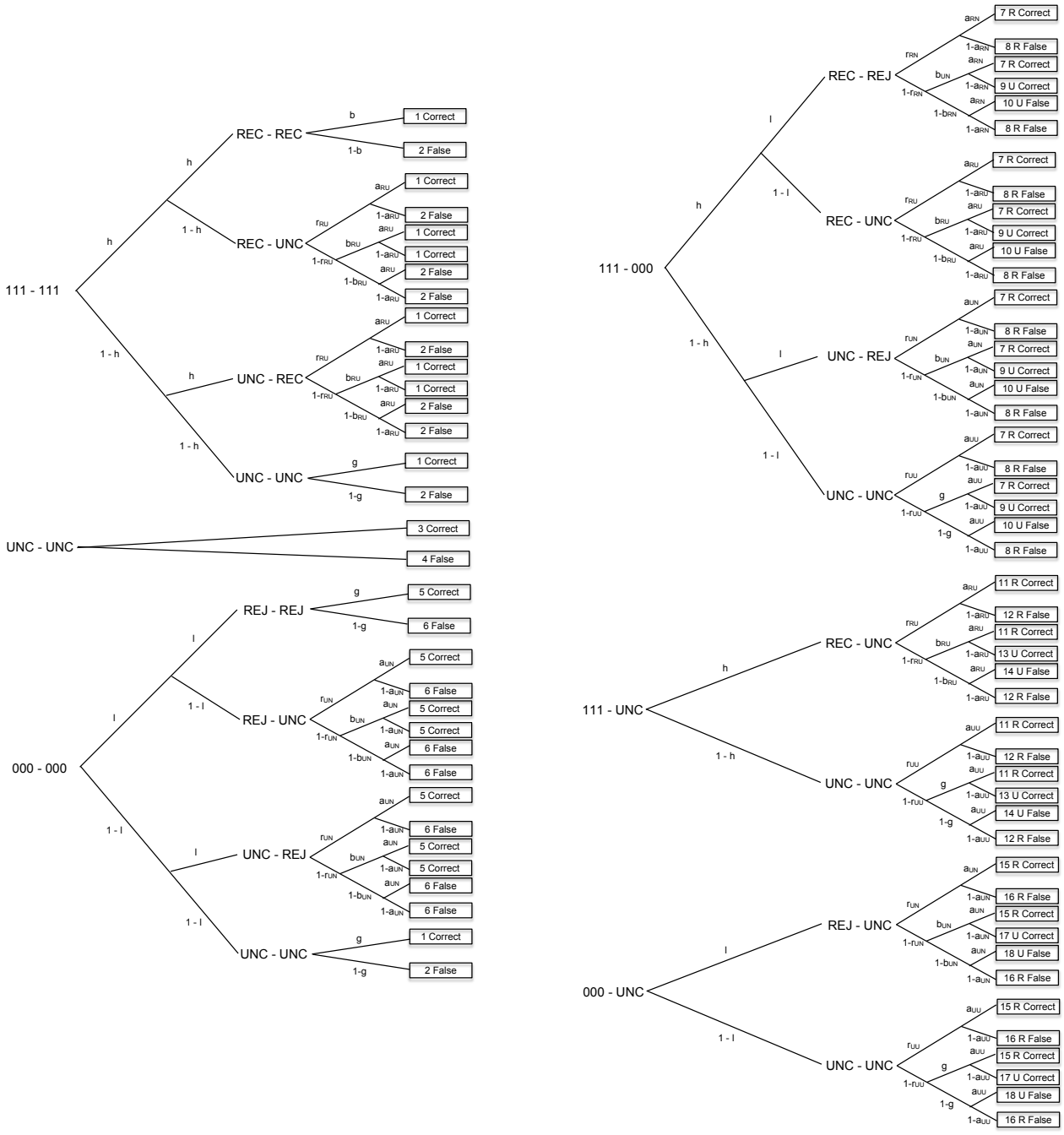

Figure 5: Mean proportion of consistent jugments (across the three sessions) for each type of pair. Error bars represent standard errors.

# APPENDIX A

## Model equations (.eqn format) and full model figure of latent - states MSH model

1 1 h * h * b
1 2 h * h * (1 - b)
1 1 h * (1 - h) * $r_{RU}$ * $a_{RU}$
1 2 h * (1 - h) * $r_{RU}$ * (1 - $a_{RU}$)
1 1 h * (1 - h) * (1 - $r_{RU}$) * $b_{RU}$ * $a_{RU}$
1 1 h * (1 - h) * (1 - $r_{RU}$) * $b_{RU}$ * (1 - $a_{RU}$)
1 2 h * (1 - h) * (1 - $r_{RU}$) * (1 - $b_{RU}$) * $a_{RU}$
1 2 h * (1 - h) * (1 - $r_{RU}$) * (1 - $b_{RU}$) * (1 - $a_{RU}$)
1 1 (1 - h) * h * $r_{RU}$ * $a_{RU}$
1 2 (1 - h) * h * $r_{RU}$ * (1 - $a_{RU}$)
1 1 (1 - h) * h * (1 - $r_{RU}$) * $b_{RU}$ * $a_{RU}$
1 1 (1 - h) * h * (1 - $r_{RU}$) * $b_{RU}$ * (1 - $a_{RU}$)
1 2 (1 - h) * h * (1 - $r_{RU}$) * (1 - $b_{RU}$) * $a_{RU}$
1 2 (1 - h) * h * (1 - $r_{RU}$) * (1 - $b_{RU}$) * (1 - $a_{RU}$)
1 1 (1 - h) * (1 - h) * g
1 2 (1 - h) * (1 - h) * (1 - g)
2 3 g
2 4 (1 - g)
3 5 l * l * g
3 6 l * l * (1 - g)
3 5 l * (1 - l) * $r_{UN}$ * $a_{UN}$
3 6 l * (1 - l) * $r_{UN}$ * (1 - $a_{UN}$)
3 5 l * (1 - l) * (1 - $r_{UN}$) * $b_{UN}$ * $a_{UN}$
3 5 l * (1 - l) * (1 - $r_{UN}$) * $b_{UN}$ * (1 - $a_{UN}$)
3 6 l * (1 - l) * (1 - $r_{UN}$) * (1 - $b_{UN}$) * $a_{UN}$
3 6 l * (1 - l) * (1 - $r_{UN}$) * (1 - $b_{UN}$) * (1 - $a_{UN}$)
3 5 (1 - l) * l * $r_{UN}$ * $a_{UN}$
3 6 (1 - l) * l * $r_{UN}$ * (1 - $a_{UN}$)
3 5 (1 - l) * l * (1 - $r_{UN}$) * $b_{UN}$ * $a_{UN}$
3 5 (1 - l) * l * (1 - $r_{UN}$) * $b_{UN}$ * (1 - $a_{UN}$)
3 6 (1 - l) * l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * $a_{UN}$
3 6 (1 - l) * l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * (1 - $a_{UN}$)
3 5 (1 - l) * (1 - l) * g
3 6 (1 - l) * (1 - l) * (1 - g)
4 7 h * l * $r_{RN}$ * $a_{RN}$
4 8 h * l * $r_{RN}$ * (1 - $a_{RN}$)
4 7 h * l * (1 - $r_{RN}$) * $b_{RN}$ * $a_{RN}$
4 9 h * l * (1 - $r_{RN}$) * $b_{RN}$ * (1 - $a_{RN}$)
4 10 h * l * (1 - $r_{RN}$) * (1 - $b_{RN}$) * $a_{RN}$
4 8 h * l * (1 - $r_{RN}$) * (1 - $b_{RN}$) * (1 - $a_{RN}$)
4 7 h * (1 - l) * $r_{RU}$ * $a_{RU}$
4 8 h * (1 - l) * $r_{RU}$ * (1 - $a_{RU}$)
4 7 h * (1 - l) * (1 - $r_{RU}$) * $b_{RU}$ * $a_{RU}$
4 9 h * (1 - l) * (1 - $r_{RU}$) * $b_{RU}$ * (1 - $a_{RU}$)
4 10 h * (1 - l) * (1 - $r_{RU}$) * (1 - $b_{RU}$) * $a_{RU}$
4 8 h * (1 - l) * (1 - $r_{RU}$) * (1 - $b_{RU}$) * (1 - $a_{RU}$)
4 7 (1 - h) * l * $r_{UN}$ * $a_{UN}$
4 8 (1 - h) * l * $r_{UN}$ * (1 - $a_{UN}$)
4 7 (1 - h) * l * (1 - $r_{UN}$) * $b_{UN}$ * $a_{UN}$
4 9 (1 - h) * l * (1 - $r_{UN}$) * $b_{UN}$ * (1 - $a_{UN}$)
4 10 (1 - h) * l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * $a_{UN}$
4 8 (1 - h) * l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * (1 - $a_{UN}$)
4 7 (1 - h) * (1 - l) * $r_{UU}$ * $a_{UU}$

4 8 (1 - h) * (1 - l) * $r_{UU}$ * (1 - $a_{UU}$)
4 7 (1 - h) * (1 - l) * (1 - $r_{UU}$) * g * $a_{UU}$
4 9 (1 - h) * (1 - l) * (1 - $r_{UU}$) * g * (1 - $a_{UU}$)
4 10 (1 - h) * (1 - l) *(1 - $r_{UU}$) * (1 - g) * $a_{UU}$
4 8 (1 - h) * (1 - l) * (1 - $r_{UU}$) *(1 - g) * (1 - $a_{UU}$)
5 11 h * $r_{RU}$ * $a_{RU}$
5 12 h * $r_{RU}$ * (1 - $a_{RU}$)
5 11 h * (1 - $r_{RU}$) * $b_{RU}$ * $a_{RU}$
5 13 h * (1 - $r_{RU}$) * $b_{RU}$ * (1 - $a_{RU}$)
5 14 h * (1 - $r_{RU}$) * (1 - $b_{RU}$) * $a_{RU}$
5 12 h * (1 - $r_{RU}$) * (1 - $b_{RU}$) * (1 - $a_{RU}$)
5 11 (1 - h) * $r_{UU}$ * $a_{UU}$
5 12 (1 - h) * $r_{UU}$ * (1 - $a_{UU}$)
5 11 (1 - h) * (1 - $r_{UU}$) * g * $a_{UU}$
5 13 (1 - h) * (1 - $r_{UU}$) * g * (1 - $a_{UU}$)
5 14 (1 - h) * (1 - $r_{UU}$) * (1 - g) * $a_{UU}$
5 12 (1 - h) * (1 - $r_{UU}$) * (1 - g) * (1 - $a_{UU}$)
6 15 l * $r_{UN}$ * $a_{UN}$
6 16 l * $r_{UN}$ * (1 - $a_{UN}$)
6 15 l * (1 - $r_{UN}$) * $b_{UN}$ * $a_{UN}$
6 17 l * (1 - $r_{UN}$) * $b_{UN}$ * (1 - $a_{UN}$)
6 18 l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * $a_{UN}$
6 16 l * (1 - $r_{UN}$) * (1 - $b_{UN}$) * (1 - $a_{UN}$)
6 15 (1 - l) * $r_{UU}$ * $a_{UU}$
6 16 (1 - l) * $r_{UU}$ * (1 - $a_{UU}$)
6 15 (1 - l) * (1 - $r_{UU}$) * g * $a_{UU}$
6 17 (1 - l) * (1- $r_{UU}$) * g * (1 - $a_{UU}$)
6 18 (1 - l) * (1 - $r_{UU}$) * (1 - g) * $a_{UU}$
6 16 (1 - l) * (1 - $r_{UU}$) * (1 - g) * (1 - $a_{UU}$)


111 - 111
h
1 - h
REC - REC
b
1 - b
1 Correct
2 False
REC - UNC
$r_{RU}$
1 - $r_{RU}$
$b_{RU}$
1 - $b_{RU}$
$a_{RU}$
1 - $a_{RU}$
UNC - REC
UNC - UNC
g
1 - g
UNC - UNC
3 Correct
4 False
000 - 000
l
1 - l
REJ - REJ
REJ - UNC
UNC - REJ
UNC - UNC
$a_{UN}$
1 - $a_{UN}$
$r_{UN}$
1 - $r_{UN}$
$b_{UN}$
1 - $b_{UN}$
5 Correct
6 False
111 - 000
REC - REJ
UNC - REJ
UNC - UNC
$a_{RN}$
1 - $a_{RN}$
$r_{RN}$
1 - $r_{RN}$
$b_{RN}$
1 - $b_{RN}$
7 R Correct
8 R False
9 U Correct
10 U False
$a_{UU}$
1 - $a_{UU}$
$r_{UU}$
1 - $r_{UU}$
111 - UNC
REC - UNC
UNC - UNC
11 R Correct
12 R False
13 U Correct
14 U False
000 - UNC
REJ - UNC
UNC - UNC
15 R Correct
16 R False
17 U Correct
18 U False

## APPENDIX B

### Necessary and sufficient conditions for $h = l$

Since $h$ represents the probability of a consistent "yes" judgment (111) originating from recognition certainty (REC) and $l$ the probability of a consistent "no" judgment (000) originating from rejection certainty (REJ), it follows that

$$h = Pr(REC|111) \tag{1a}$$

$$l = Pr(REJ|000). \tag{1b}$$

Additonally, this also implies that, for the uncertainty memory state (UNC)

$$1 - h = Pr(UNC|111) \tag{2a}$$

$$1 - l = Pr(UNC|000). \tag{2b}$$

It follows from Bayes' theorem that

$$Pr(UNC|111) = Pr(111|UNC)\frac{Pr(UNC)}{Pr(111)} \tag{3a}$$

$$Pr(UNC|000) = Pr(000|UNC)\frac{Pr(UNC)}{Pr(000)}. \tag{3b}$$

Since we know from the 2HT model that, assuming independent recognition judgments, $Pr(111|UNC) = g^3$ and $Pr(000|UNC) = (1-g)^3$, it follows that

$$Pr(UNC|111) = g^3\frac{Pr(UNC)}{Pr(111)} \tag{4a}$$

$$Pr(UNC|000) = (1-g)^3\frac{Pr(UNC)}{Pr(000)}. \tag{4b}$$

After dividing Equation (4a) by Equation (4b) we see that the equality $h = l \Leftrightarrow P(UNC|111) = P(UNC|000)$ holds if and only if

$$\frac{g^3}{(1-g)^3} = \frac{Pr(111)}{Pr(000)}. \tag{5}$$

Hence, $g$ monotonically increases with $\frac{Pr(111)}{Pr(000)}$. □

# APPENDIX C

## Approximate Model

The baseline model with the restrictions $b = b_{RN} = b_{RU}$ and $g = b_{UN} = .5$ fit quite well (Experiment 1: $G^2(5) = 4.50$, $p = .48$, FIA = 28.37; Experiment 2: $G^2(5) = 3.44$, $p = .63$, FIA = 26.91). When adding the inequality restrictions $r_{RN} \geq r_{RU}$ and $r_{RN} \geq r_{UN}$, model misfit did not increase and FIA decreased (Experiment 1: $\Delta\overline{G^2} = 0$, $\overline{p} = 1$, $\Delta$FIA = 1.20; Experiment 2: $\Delta\overline{G^2} = 0$, $\overline{p} = 1$, $\Delta$FIA = 1.19). When we compared the inequality with an equality restriction $r_{RN} = r_{RU} = r_{RN}$, model misfit increased significantly and so did FIA (Experiment 1: $\Delta\overline{G}^2 = 260.78$, $\overline{p} = 0$ , $\Delta$FIA = 125.53; Experiment 2: $\Delta\overline{G}^2 = 128.75$, $\overline{p} = 0$, $\Delta$FIA = 59.79), indicating that the equality restriction is not compatible with the data. According to the $G^2$ goodness-of-fit statistic, the difference between $r_{RU}$ and $r_{UN}$ was also replicated, but the differences in FIA are too small to allow a preference for any of the two models (Experiment 1: $\Delta G^2(1) = 6.73$, $p < .01$, $\Delta$FIA = .50; Experiment 2: $\Delta G^2(1) = 4.38$, $p = .04$, $\Delta$FIA = $-.53$).

Regarding the $a$ parameter, adding the inequality restrictions $a_{RN} \geq a_{RU}$ and $a_{RN} \geq r_{UN}$ did not increase model misfit and decreased FIA (Experiment 1: $\Delta\overline{G^2} = 0$, $\overline{p} = 1$, $\Delta$FIA = 1.12; Experiment 2: $\Delta\overline{G^2} = 0$, $\overline{p} = 1$, $\Delta$FIA = 1.12). When we compared the inequality with the equality restriction $a_{RN} = a_{RU} = a_{RN}$, we observed a significant increase in model misfit and FIA (Experiment 1: $\Delta\overline{G}^2 = 135.47$, $\overline{p} = 0$ , $\Delta$FIA = 61.53; Experiment 2: $\Delta\overline{G}^2 = 78.71$, $\overline{p} = 0$, $\Delta$FIA = 33.46). The difference between $a_{RU}$ and $a_{UN}$ was replicated with the approximate model, since imposing the equality restriction $a_{RU} = a_{UN}$ led to a significant increase in model misfit and in FIA (Experiment 1: $\Delta G^2(1) = 16.67$, $p < .001$, $\Delta$FIA = 4.77; Experiment 2: $\Delta G^2(1) = 29.63$, $p < .001$, $\Delta$FIA = 11.40).

## APPENDIX D

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exp 1 | 1041 | 549 | 377 | 356 | 611 | 641 | 1901 | 503 | 77 | 120 | 977 | 384 | 177 | 205 | 635 | 391 | 173 | 242 |
| Exp 2 | 1023 | 589 | 183 | 167 | 347 | 363 | 1307 | 390 | 38 | 73 | 932 | 333 | 109 | 134 | 464 | 296 | 92 | 120 |
| Exp 3: size | 1537 | 511 | 78 | 69 | 263 | 240 | 1061 | 266 | 38 | 87 | 608 | 148 | 87 | 89 | 209 | 100 | 58 | 71 |
| Exp 3: heigth | 672 | 506 | 73 | 73 | 215 | 192 | 377 | 257 | 192 | 215 | 239 | 126 | 152 | 182 | 112 | 110 | 73 | 74 |

*Note*: Aggreagated response frequencies for all experiments, separately for each of the 18 response categories.