Psychological Review
2015, Vol. 122, No. 3, 542–557

0033-295X/15/$12.00 http://dx.doi.org/10.1037/a0039251

THEORETICAL NOTE

Signal Detection and Threshold Modeling of Confidence-Rating ROCs: A Critical Test With Minimal Assumptions

David Kellen
University of Basel

Karl Christoph Klauer
Albert-Ludwigs-Universität Freiburg

An ongoing discussion in the recognition-memory literature concerns the question of whether recognition judgments reflect a direct mapping of graded memory representations (a notion that is instantiated by signal detection theory) or whether they are mediated by a discrete-state representation with the possibility of complete information loss (a notion that is instantiated by threshold models). These 2 accounts are usually evaluated by comparing their (penalized) fits to receiver operating characteristic data, a procedure that is predicated on substantial auxiliary assumptions, which if violated can invalidate results. We show that the 2 accounts can be compared on the basis of critical tests that invoke only minimal assumptions. Using previously published receiver operating characteristic data, we show that confidence-rating judgments are consistent with a discrete-state account.

Keywords: recognition memory, familiarity, discrete states, signal detection, thresholds, ROCs

Supplemental materials: http://dx.doi.org/10.1037/a0039251.supp

The ability to recognize previously experienced information is one of the most basic aspects of our memory faculties. This status places recognition memory at the center stage of many empirical studies and mathematical modeling endeavors (for reviews, see Kahana, 2014; Malmberg, 2008). One fundamental question that arises in modeling recognition memory concerns the representation of the information governing the observed responses. Different proposals have been made, some assuming that recognition judgments result from continuous or graded mnemonic representations, others assuming a discrete-state representation.

The most prominent continuous model of recognition memory is the signal detection theory (SDT) model (Green & Swets, 1966). In the SDT model, mnemonic stimulus information is represented along a so-called familiarity continuum. Before the study phase, items (e.g., words) are assumed to have a baseline familiarity, represented by a continuous distribution. When items are studied their familiarity is assumed to increase, leading to two familiarity distributions, one representing studied (old) items and the other nonstudied (new) items. The familiarity distributions postulated by SDT are not tied to any parametric form but they are for practical reasons traditionally assumed to be Gaussian (Green & Swets, 1966, p. 79). Figure 1 provides a depiction of the Gaussian SDT model, with $\{\mu_o, \sigma_o^2\}$ and $\{\mu_n, \sigma_n^2\}$ being the means and variances of the old and new-item distributions respectively. According to SDT, tested items are evaluated by directly comparing their familiarity value with a set of ordered response criteria τ (see Figure 1) when a confidence-rating scale is used for responding. The observed responses are determined by the criteria exceeded by items' familiarity values. The core assumption of SDT is that stimuli have a *graded information* representation that is *directly mapped* onto the observed responses.

An alternative theoretical account assumes that the representation of mnemonic stimulus information (information that can be graded) is mediated by *discrete mental states* defined by a latent threshold (Rouder & Morey, 2009). Below-threshold information enters a mental state in which the same pattern of observed responses is produced irrespective of the stimulus. This invariance for subthreshold information can be interpreted as a case of *complete information loss*. Observed responses are only a function of stimulus information when that information is above threshold. Complete information loss is a core aspect of several discrete-state models found in the literature (e.g.,

David Kellen, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel; Karl Christoph Klauer, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg.

The research reported in this paper was supported by Grant Kl 614/32-2 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer. We thank Andrew Heathcote, Joshua Koen, and Jeff Rouder for making their data available. We also thank Ken Malmberg and Henrik Singmann for valuable comments.

Correspondence concerning this article should be addressed to David Kellen, Faculty of Psychology, University of Basel, Missionstrasse 60-64, CH-4055 Basel, Switzerland. E-mail: davekellen@gmail.com

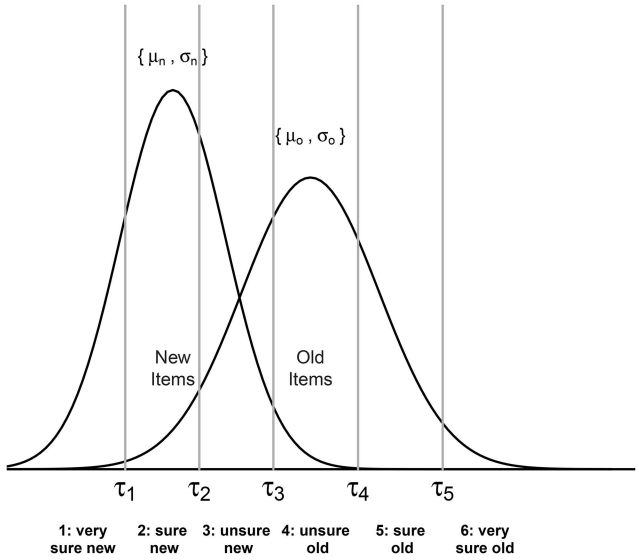

Figure 1. The Gaussian signal detection theory model.

Luce, 1963; Snodgrass & Corwin, 1988; Stevens, Morgan, & Volkmann, 1941; Zhang & Luck, 2009).

The discrete-state model most commonly used in recognition-memory studies is the two-high-threshold model (2HT; Bröder, Kellen, Schütz, & Rohrmeier, 2013; Snodgrass & Corwin, 1988). The 2HT assumes that items can be in one of two states: a detection state in which the item's true status (old or new) is known and an uncertainty state in which no mnemonic information is available.[1]

Figure 2 provides a depiction of the 2HT model: Old and new items are detected with probabilities D_o and D_n, respectively. In the absence of detection, no mnemonic information is available whatsoever and a pure guessing-based response is produced, with responses "old" and "new" occurring with probabilities g and $1 - g$ for both old and new items. Confidence-rating judgments are produced by means of state-response mapping parameters δ_o, δ_n, γ_o, and γ_n (see Figure 2).

Continuous models are far more popular than discrete-state models. A great part of this popularity is due to the high plausibility of continuous representations (e.g., they arise via evidence accumulation processes and in neural responses; see Usher & McClelland, 2001) but also to the successes of continuous models like SDT in describing actual data (for a foundational introduction, see Green & Swets, 1966). However, the plausibility of a continuous representation does not imply that such representations are directly mapped onto the observed responses without any form of mediation or discretization. Moreover, the popularity of continuous models is not accompanied by a clear rejection of a discrete-state account (Rouder & Morey, 2009, provide an example demonstrating the difficulty of rejecting such an account). Luce (1997) discussed this situation in a commentary on unresolved conceptual problems in mathematical modeling, where he referred to the popularity of continuous models like SDT as a "sobering lesson in the sociology of science" (p. 85).

Attempts to distinguish between continuous and discrete-state accounts have almost exclusively relied on receiver operating characteristic (ROC) plots (for a review, see Yonelinas & Parks, 2007; see also Pazzaglia, Dube, & Rotello, 2013, and Batchelder & Alexander, 2013). An ROC is a plot of hit rates ("old" responses to old items) as a function of the false alarm rates ("old" responses to new items). ROCs can be constructed by means of *binary responses* or *confidence ratings*: Binary-response ROCs describe the recognition of old and new items across different response biases (e.g., tendency to respond "old"). Binary-response ROC data allow for the two models to be distinguished given that the SDT model predicts curvilinear ROCs and the 2HT predicts linear ROCs (see Figure 3). Confidence-rating ROCs are based on the responses given on a confidence-rating scale, for example a 6-point scale ranging from 1 (*sure new*) to 6 (*sure old*).

According to the SDT model, confidence rating judgments are produced via a set of ordered response criteria that are placed on the evidence axis, amounting to a straightforward extension of the binary-response case for which there is only one response criterion. In the case of the 2HT model, confidence judgments are generated through state-response mapping functions that establish how the different states are mapped onto the rating scale (Klauer & Kellen, 2010; Malmberg, 2002). Because confidence-rating ROCs are much easier to obtain than ROCs based on binary responses (e.g., only a single study-test block is required in the former), the vast majority of ROCs in the literature has been obtained using confidence ratings (for reviews, see Yonelinas & Parks, 2007; Wixted, 2007). This convenience comes at cost though, as both the SDT and 2HT models are able to account for the curvilinear confidence-rating ROCs that have been almost ubiquitously observed in the literature (e.g., Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Falmagne, 1985; Klauer & Kellen, 2010; Krantz, 1969; Luce, 1963; Malmberg, 2002).

The two models have recently been compared by means of binary-response ROCs (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Starns, Rotello, & Ratcliff, 2012; Kellen, Klauer, & Bröder, 2013). The results obtained with this approach are somewhat unsatisfactory because it is difficult to obtain diagnostic data (i.e., data allowing a reliable assessment of ROC shape; see Kellen et al., 2013). Also, the diagnostic value of binary-response ROCs hinges on the auxiliary assumption that the required response-bias manipulations do not influence memory discriminability, an assumption that has been questioned in the context of both discrete state (Krantz, 1969; Luce, 1963; Rouder, Province, Swagman, & Thiele, 2013) and continuous models (Balakrishnan, 1999; Van Zandt, 2000).

[1] Although most discrete-state models assume the possibility of complete information loss—in particular, the ones within the scope of the present paper—a small minority does not (e.g., Chechile, 2004). The distinguishability of continuous models and models postulating a finite set of partial information states is poor given that the former can always be well approximated by some variant of the latter (e.g., Kaernbach, 1991). Similarly, another minority of discrete-state models (the so-called *low-threshold models*; e.g., Luce, 1963; Malmberg, 2002) assume the possibility of studied items being rejected despite their successful retrieval: the high-threshold discrete-state accounts that are the focus of the present work exclude this possibility. This restriction is not problematic given that the more general low-threshold models have rarely been considered, especially in ways that go beyond a proof of concept (e.g., Swagman, Province, & Rouder, 2015).

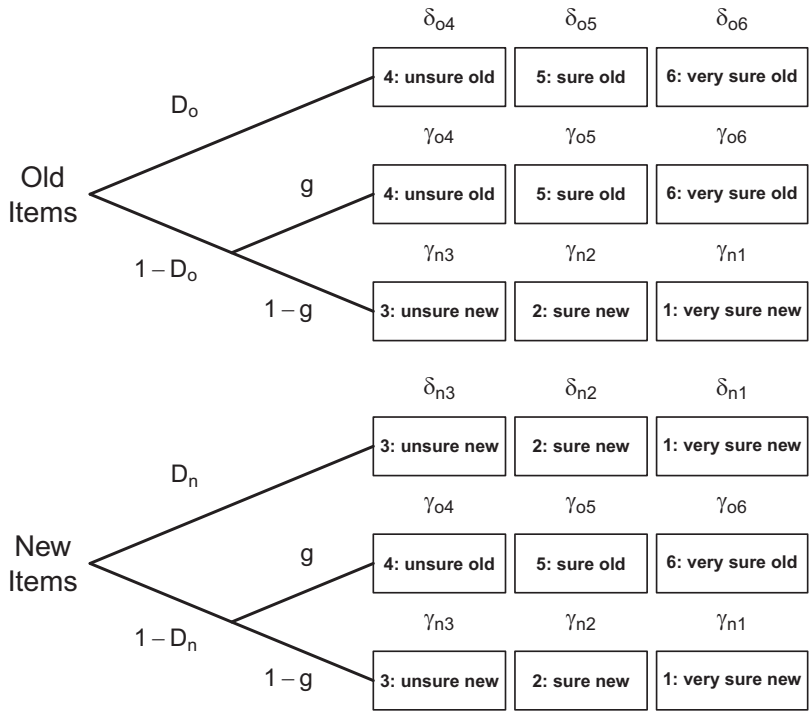

Figure 2. The two-high-threshold model.

An alternative comparison approach has recently been introduced by Rouder and colleagues (Province & Rouder, 2012; Rouder et al., 2013; Swagman et al., 2015). Their approach relies on the use of families of confidence-rating ROCs obtained by means of study-strength manipulations and an overlooked property of discrete-state models, *conditional independence*: The state-response mappings in the 2HT are not a function of the probability of the discrete memory states being reached, which means that study-strength manipulations should *only* affect the detection of studied items (i.e., D_o), but not the mapping of the different states onto a confidence-rating scale (i.e., the δ and γ parameters are unaffected). The response mapping is in this sense *conditionally independent* of detection probability. Province and Rouder (2012) reported results showing that individuals' judgments in a two-alternative forced-choice (2AFC) task are consistent with conditional independence, with the 2HT outperforming the SDT model in terms of model fits. These results were later replicated by Kellen, Singmann, Vogt, and Klauer (in press) and extended by Swagman et al. (2015) to the case of visual word identification. More recently, Chen, Starns, and Rotello (in press) showed that conditional independence is violated for study-strength manipulations that produce virtually perfect performance (but see Krantz, 1969; Rouder et al., 2013).

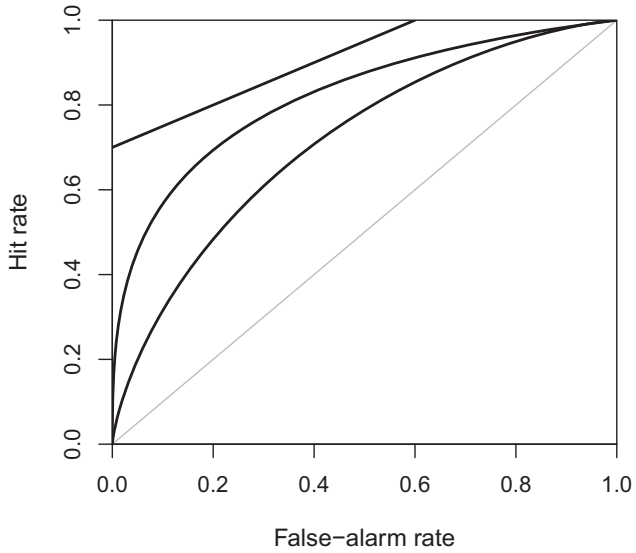

Figure 3. Examples of receiver operating characteristic functions.

These previous approaches have important limitations that can drastically affect results. First, discrete and continuous models can be specified in different ways (i.e., there are no unique implementations): In the case of SDT, different distributional assumptions could be used as an alternative to the traditional Gaussian assumption. In the case of the 2HT, additional memory states could be assumed over and above the detection state entered with probability D_o (see Kellen et al., 2015). In some cases, these memory states are postulated in a principled manner (e.g., source memory; see Klauer & Kellen, 2010) while in others they have been postulated as a way to deal with violations of conditional independence when performance is almost perfect (e.g., Krantz, 1969; see also Rouder et al., 2013). Because of these "degrees of freedom," it is not clear to which extent the failure of a model is due to the inappropriateness of the adopted auxiliary assumptions (e.g., distributional

assumptions, additional memory states, etc.). Second, it is often the case that the flexibility of the chosen model implementations is not taken into account in a sensible way (but see Klauer & Kellen, 2011): Model comparisons usually rely on model-selection indices that take the number of free parameters as a proxy for model flexibility, therefore failing to accurately capture the flexibility emerging from the models' functional form (Pitt, Myung, & Zhang, 2002). Taken together, these limitations often lead to model comparisons in which the relationship between model predictions and the data gets lost in translation. This situation limits researchers to focus their discussions on model-performance statistics that are too detached from the data and the core assumptions of the competing models. For example, finding that a Gaussian SDT model outperforms the 2HT model (or vice versa) provides limited information on the question of whether the responses are based on a model that directly maps graded representations or a model with discrete states and complete information loss given the multitude of models of either kind that were not tested.

A Critical-Test Approach for Comparing Continuous and Discrete-State Accounts

Despite their antagonistic nature, continuous and discrete-state accounts so far have turned out to be extremely difficult to distinguish in a clear-cut manner. Ideally, comparisons would hinge on the two models' core assumptions; namely graded information versus complete information loss. The ongoing discussion between the SDT and 2HT models can be swayed in different directions by several questionable auxiliary assumptions regarding selective influence (e.g., response-bias manipulations do not affect memory sensitivity), model parametrization (e.g., the familiarity distributions are Gaussian), the nature and number of mental states, and how model flexibility should be penalized (e.g., whether the number of parameters is a good proxy for model flexibility). Any of these assumptions can decide over the models' success or failure.

This inability to rely on a model's core assumptions should not be taken lightly given that it has implications for other cognitive-modeling endeavors: If one is unable to distinguish these two very distinct hypotheses, then one should also question the evidence for and against more complex memory representations (e.g., dual-process memory models; Yonelinas & Parks, 2007; Wixted, 2007). One reasonable counterargument or dismissal of this criticism is that models are very often untestable when devoid of parametric assumptions (e.g., Jones & Dzhafarov, 2014). However, we will show that this notion does not hold in the present case and that the models are testable in a very general form.

The goal of the present work is to show that continuous and discrete-state accounts can be distinguished on the basis of their core assumptions independently of auxiliary assumptions regarding, for example, the parametric shape of continuous familiarity distributions in continuous models or the precise nature of detection states in discrete-state models. This implies abandoning the approach to compare models in terms of their fit to a set of observed data given that fitting models to data necessitates the complete specification of the models to be fitted, which, in turn, requires the specification of noncentral auxiliary assumptions.

Instead of focusing on model fits, we try to identify diverging predictions from the class of continuous models and the class of discrete-state models with complete information loss involving only minimal auxiliary assumptions. Critical tests of this kind have a strong tradition in many fields. For example, in the field of judgment and decision making, general classes of models are tested using participants' responses to only a few critical trials, without the need of strong parametric assumptions and without reliance on overall model fits for specific members of the classes of models (e.g., Allais, 1953; Birnbaum, 2008). Ideally, both approaches, (a) testing core predictions derived from properties of competing classes of models and (b) the specification and estimation of a particular well-fitting member of the more successful model class should complement each other, with the first approach taking logical and temporal precedence over the latter.

A first critical test for the SDT and 2HT models was developed by Kellen and Klauer (2014) in the context of a k-*alternative ranking task* (Iverson & Bamber, 1997). In each ranking trial, one old item and $k - 1$ new items are presented. Participants, who were aware of this composition, were requested to rank the items according to their belief that they were previously studied (Rank 1 = item most believed to have been previously studied). An example of a ranking trial with four alternatives is shown in the left panel of Figure 4. The dependent variable of interest, termed c_2, was the probability of the old item being assigned Rank 2, *conditional* on it not being assigned Rank 1. Kellen and Klauer (2014) showed that the SDT and 2HT models make specific predictions regarding c_2 when the study strength of the old items is manipulated (e.g., via repetition or increased study time). Namely, it was shown that the SDT model predicts c_2 to be greater in ranking trials with a strong item than in ranking trials with a weak item. It can be formally shown that this prediction holds across several distributional assumptions (e.g., Gaussian, gamma, Weibull; see Kellen & Klauer, 2014, Supplemental Material), a result that allows us to establish a general SDT hypothesis $\mathcal{H}_{\text{SDT}} : c_2^w \leq c_2^s$. In contrast, the 2HT model predicts that c_2 is invariant across both types of ranking trials. This prediction is also quite general (it does not depend on D_o or D_n, nor is it affected by the existence of additional detection states), leading to the hypothesis $\mathcal{H}_{\text{2HT}} : c_2^w = c_2^s$. By focusing on c_2 predictions, the two models reduce to a comparison between order-constrained nested hypotheses because the null hypothesis $\mathcal{H}_{\text{2HT}}$ is at the boundary of the alternative hypothesis $\mathcal{H}_{\text{SDT}}$ (Hoijtink, 2011; Silvapulle & Sen, 2004). The two hypotheses formalize core notions of the two models: In the case of continuous models with graded information, it is the notion that familiarity distributions with larger means (e.g., strong items, in comparison with weak items) not only imply *fewer errors* but also *fewer extreme errors* (i.e., larger c_2). In the case of discrete-state models assuming complete information loss it is the notion that errors decrease as performance increases but *the error profile remains the same*. Experimental results from two studies showed that c_2 is systematically larger for strong items than for weak items (see the right panel of Figure 4), a pattern that supports $\mathcal{H}_{\text{SDT}}$ and the notion that individuals directly rely on a graded mnemonic information when producing ranking judg-

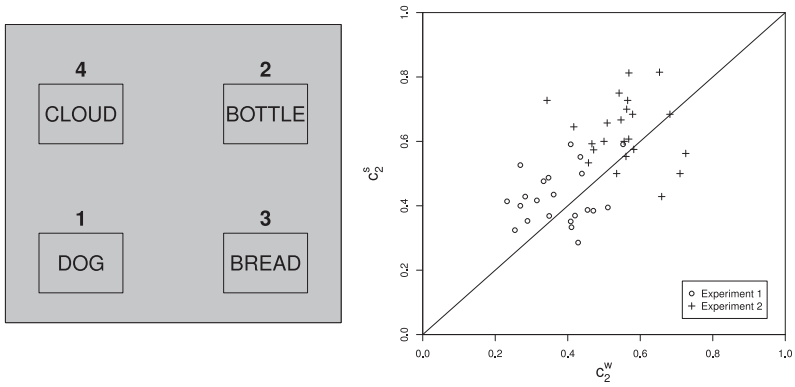

Figure 4. Left panel: Example of a four-alternative ranking task trial. Right panel: c_2 estimates for weak and strong items reported by Kellen and Klauer (2014). Data from Experiments 1 and 2 come from a four- and three-alternative raking task, respectively.

ments.[2] One of the attractive features of these results is that the relationship between the models and the data is transparent (see Figure 4).

A New Critical Test for ROC Data: Study-Strength Effects on Confidence-Rating Judgments

The evidence of a continuous memory representation in ranking judgments does not exclude the possibility that confidence-rating judgments are mediated by discrete states. As previously mentioned, the presence of graded mnemonic information does not imply that such information is directly mapped onto the individuals' judgments. In fact we will argue below that the ranking task encourages respondents to make use of the graded information in memory, whereas this is not the case for the confidence-rating task. Here, discrete states can be seen as nothing more than *task thresholds* (Rouder & Morey, 2009) that simply reflect individuals' attempts to deal with the demands of a typical recognition-memory task in an efficient way by relying on more coarse-grained representations (for a discussion on the role of task efficiency, see Malmberg, 2008). Kellen and Klauer (2014) proposed that a continuous evidence scale can be divided into three regions, two with extreme familiarity values (very low and very high familiarity), where the status of the test item is considered to have been "ascertained" (e.g., old or new), and a third region in between where the status of the test item is classified as "uncertain" (for previous versions of this division concept, see Atkinson & Juola, 1974; Kintsch, 1967; Mandler, Pearlstone, & Koopmans, 1969). The distinction between items whose status has been ascertained and items with an unknown status can be found in research on memory monitoring. For example, Koriat and Goldsmith (1994) have shown that individuals often find the memory evidence available to be unreliable and simply withhold a response. The "uncertain" items may be mapped onto the response set (e.g., the confidence-rating scale) using the same probability distribution irrespective of their true status as old or new items, (i.e., these are chance-level responses with complete information loss).

The new critical test will compare the predictions of continuous and discrete-state accounts for confidence ratings. As in the previous case, the focus will be on how the assumptions of graded information and complete-information loss impact the profiles of incorrect responses. In discussing the new critical test we focus on the old-new task in which incorrect new judgments for old items are the responses of interest (responses to new items are not considered at all). However, note that the test is more general and also applies to incorrect 2AFC judgments along the exact same lines. In fact, the analysis reported below will evaluate old-new and 2AFC data jointly using a hierarchical Bayesian approach (Rouder, Morey, & Pratte, in press).

Consider the SDT depiction in Figure 1: In the case of a 6-point confidence-rating scale, the ordered response criteria $\tau_1 < \tau_2 < \tau_3$ delimit the segments of the familiarity distributions that will be judged as "new" with different degrees of confidence. For any familiarity density f_μ, the *unconditional* probabilities of responses 1 (*very sure new*), 2 (*sure new*), and 3 (*unsure new*) are given, in order, by $\int_{-\infty}^{\tau_1} f_\mu(x)\,dx$, $\int_{\tau_1}^{\tau_2} f_\mu(x)\,dx$, and $\int_{\tau_2}^{\tau_3} f_\mu(x)\,dx$. Now, let $\theta_{[2|3]}$ denote the probability of a 1 (*very sure new*) or a 2 (*sure new*) response, *conditional* on the occurrence of a "new" response (i.e., a rating of 1, 2, or 3). Similarly, let $\theta_{[1|2]}$ denote the probability of response 1 (*very sure new*), *conditional* on the occurrence of a 1 (*very sure new*) or a 2 (*sure new*) response.[3] According to the SDT model,

$$\theta_{[2|3]} = P(1 \cup 2 \,|\, 1 \cup 2 \cup 3) = \frac{\int_{-\infty}^{\tau_2} f_\mu(x)\,dx}{\int_{-\infty}^{\tau_3} f_\mu(x)\,dx} = \frac{F_\mu(\tau_2)}{F_\mu(\tau_3)}, \qquad (1)$$

[2] It should be noted that a discrete-state account could successfully describe the observed increase in c_2 if some of the detected old items were not attributed rank 1 (see Footnote 1). Although possible, such behavior is unlikely here given that the task constraints usually used to justify it (e.g., old-new base rates) do not apply in the case of the ranking task (see Kellen & Klauer, 2014, p. 1802).

[3] A more transparent alternative to subscripts [1|2] and [2|3] would be [1|1 ∪ 2] and [1 ∪ 2|1 ∪ 2 ∪ 3], respectively. Although the latter notation expresses the relevant conditional probabilities in a more explicit way, they are visually more cumbersome.

$$\theta_{[1|2]} = P(1 \mid 1 \cup 2) = \frac{\int_{-\infty}^{\tau_1} f_\mu(x)\,dx}{\int_{-\infty}^{\tau_2} f_\mu(x)\,dx} = \frac{F_\mu(\tau_1)}{F_\mu(\tau_2)}, \quad (2)$$

where F_μ is the cumulative distribution function of the old-item familiarity distribution.

In order to obtain predictions from the SDT model, a function $H_\mu = F_\mu(z)^{-1} \times \frac{\partial}{\partial \mu} F_\mu(z)$ is derived from the distribution function for familiarity values z. This function is somewhat similar to a (reverse) *hazard function* (Chechile, 2003, 2011), $r_\mu(z) = F_\mu(z)^{-1} \times \frac{\partial}{\partial z} F_\mu(z) = \frac{f_\mu(z)}{F_\mu(z)}$, a kind of function that is very often used in model testing (e.g., Balakrishnan & Ratcliff, 1996; Chechile, 2006; Colonius, 1988; Thomas, 1971; Townsend & Ashby, 1983; Townsend & Wenger, 2004). In the present case, H_μ describes how the probability of a familiarity value of at most z changes as μ varies, *conditional* on the occurrence of a familiarity value of at most z.

The theorem here states that a monotonicity property of H_μ, where it is present, entails predictions on the relative values of the conditional confidence-rating probabilities $\theta_{[1|2]}$ and $\theta_{[2|3]}$.

> *Theorem:* If $H_\mu(z)$ is monotonically increasing in z for all μ, then $\frac{F_\mu(a)}{F_\mu(b)}$ is monotonically decreasing in μ for any pair of criterion values a, b with $a < b$.
>
> *Proof:* See Appendix.

This theorem states that conditional confidence-rating probabilities like the ones denoted by $\theta_{[2|3]}$ and $\theta_{[1|2]}$ *decrease* as μ increases, formalizing an intuitive prediction that follows from the direct mapping of graded mnemonic information: Familiarity distributions with larger means not only produce fewer errors but also *fewer extreme errors* (i.e., the confidence associated to these errors decreases as well).[4] In the context of study-strength manipulations, this simply means that these conditional confidence-rating probabilities are greater for weak words than for strong words, $\mathcal{H}_{SDT}: \theta^s_{[1|2]} \leq \theta^w_{[1|2]}$ *and* $\theta^s_{[2|3]} \leq \theta^w_{[2|3]}$. In other words, the probability of rating 1, conditional on the occurrence of rating 1 or 2, is smaller for strong items than weak items. The same prediction applies to the probability of a rating 1 or 2, given the occurrence of a "new" response (rating 1, 2, or 3).

We now turn to the 2HT model: According to this model, incorrect "new" judgments only occur in the absence of item detection, and therefore, these judgments are exclusively produced by the postulated guessing processes. The following proposition establishes the relationship between both $\theta_{[1|2]}$ and $\theta_{[2|3]}$ and the probability of item detection as quantified by D_o:

> *Proposition:* Let $D^w_o, D^s_o \leq 1$ represent the probabilities of detecting two different types of studied items (e.g., weak and strong items). The equalities $\theta^s_{[1|2]} = \theta^w_{[1|2]}$ and $\theta^s_{[2|3]} = \theta^w_{[2|3]}$ hold for $0 \leq D^w_o, D^s_o \leq 1$.
>
> *Proof:* Let $1 - g$ be the probability of guessing "new" in the absence of detection and γ_{n1}, γ_{n2}, and γ_{n3} the response-mapping parameters for the confidence responses (see Figure 2). It is easy to see that $\theta_{[1|2]} = \frac{(1-D_o)(1-g)\gamma_{n1}}{(1-D_o)(1-g)(\gamma_{n1}+\gamma_{n2})} = \frac{\gamma_{n1}}{\gamma_{n1}+\gamma_{n2}}$, which means that $\theta_{[1|2]}$ does not depend on D_o. The same holds for $\theta_{[2|3]}$.

This proposition formalizes the intuitive notion that according to the discrete-state models' assumption of complete information loss the error profile is invariant to changes in the overall proportion of errors. In other words, the conditional-independence property discussed by Province and Rouder (2012) holds. Note that neither D_o nor D_n (nor the potential existence of "higher" detection states) plays a role here. Also, no constraints are imposed on the response-mapping parameters γ_n. Again, the evaluation of the 2HT can be reduced to the evaluation of a null hypothesis $\mathcal{H}_{2HT}: \theta^s_{[1|2]} = \theta^w_{[1|2]}$ *and* $\theta^s_{[2|3]} = \theta^w_{[2|3]}$ that is at the boundary of the alternative hypothesis $\mathcal{H}_{SDT}$ (Hoijtink, 2011; Silvapulle & Sen, 2004). As previously mentioned, this prediction is restricted to incorrect "new" confidence-rating judgments in the case of an old-new task. But in the case of a 2AFC task the critical test can be applied to both incorrect "left" and "right" responses by simply ordering the confidence ratings of both incorrect responses from most extreme to least extreme: 1/6 (*very sure left*/*right*), 2/5 (*moderately sure left*/*right*), and 3/4 (*unsure left*/*right*). Furthermore, note that although the discussion so far focused on 6-point confidence scales, this critical test naturally extends to arbitrarily larger scales. Also, large scales can be collapsed onto smaller scales as the test is not affected by any collapsing of response categories. The only constraint is that there are at least two levels of confidence for incorrect judgments (as in, e.g., a 4-point confidence scale).

Boundary Conditions for $\mathcal{H}_{SDT}$ and $\mathcal{H}_{2HT}$

In order for the new critical test to be valid, there are a number of experimental controls that must be in place. Also, there is a special parametric case under which $\mathcal{H}_{SDT}$ does not necessarily hold. First, the items in the weak and strong-item classes are assumed to be indistinguishable in terms of their external (e.g., word color), internal (e.g., word frequency), or contextual features (e.g., studied/tested in separate experimental blocks). Otherwise, different response criteria or response-mapping functions could be applied to weak and strong items (Stretch & Wixted, 1998), compromising the comparability of the responses to weak and strong items as established by $\mathcal{H}_{SDT}$ and $\mathcal{H}_{2HT}$. This requirement therefore excludes studies on study-strength effects where weak words are distinct from strong words (e.g., weak green and strong red words; see Stretch & Wixted, 1998). By imposing that both item classes are indistinguishable, one is also imposing that the response criteria and response-mapping parameters have to be the

[4] It should be noted that the SDT predictions regarding c_2 in ranking judgments were also based on the same property, namely $H_\mu(z)$ being monotonically increasing in z for all μ (see Kellen & Klauer, 2014).

same when evaluating them (e.g., Dube et al., 2012; Province & Rouder, 2012; Ratcliff, McKoon, & Tindall, 1994).[5]

A second constraint concerns $\mathcal{H}_{SDT}$ when the underlying familiarity distributions are Gaussian: $\mathcal{H}_{SDT}$ holds irrespective of the values taken by the variances of old and new-item familiarity (σ_o^2 and σ_n^2, respectively), demonstrating its compatibility of $\mathcal{H}_{SDT}$ with the unequal-variance Gaussian SDT model that is often assumed in the literature (see Figure 1). However, $\mathcal{H}_{SDT}$ does not necessarily hold if the variances of weak and strong old items are allowed to differ ($\sigma_w^2 \neq \sigma_s^2$). This situation simply reflects the fact that this particular SDT parametrization is able to violate important stochastic ordering relations (Townsend, 1990) that are generally expected to hold in any principled continuous model (Green & Swets, 1966; Rouder, Pratte, & Morey, 2010; Rouder et al., 2013). For instance, this parametrization can predict that items with low familiarity are more likely in the strong-item distribution than in the weak-item distribution and under certain parameter values the model can make bizarre predictions like $\theta^s_{[1|2]} > \theta^w_{[1|2]}$ or that c_2 is below chance or even 0 (see Kellen & Klauer, 2011).[6] Note that none of these predictions emerges from any psychological theory—they are mere byproducts of the adoption of completely unconstrained Gaussian distributions.[7] Fortunately the problems associated to this particular model are minimal given that alternative distributional assumptions compatible with $\mathcal{H}_{SDT}$ are perfectly able to account for the data at large (e.g., they are able to account for increasingly asymmetric ROCs as performance increases; see DeCarlo, 1998; Green & Swets, 1966; Lockhart & Murdock, 1970; Rouder et al., 2010; Wandell & Luce, 1978). In fact, prominent computational models of memory such as REM (Shiffrin & Steyvers, 1997) rely on non-Gaussian evidence distributions.

Evaluating Order-Constrained Hypotheses Using Hierarchical Bayesian Multinomial Processing Tree Models

We now turn to the statistical evaluation of the two hypotheses. Figure 5 provides a characterization of the incorrect "new" confidence-judgment responses using a multinomial processing tree (MPT) model (Riefer & Batchelder, 1988). Any possible probability distribution of the "new" confidence judgments can be perfectly described by parameters $\theta^w_{[1|2]}$, $\theta^w_{[2|3]}$, $\theta^s_{[1|2]}$, and $\theta^s_{[2|3]}$. In order to incorporate both $\mathcal{H}_{SDT}$ and $\mathcal{H}_{2HT}$ into the MPT model, the parameters $\theta^s_{[1|2]}$ and $\theta^s_{[2|3]}$ in the MPT model will be constrained using a simple reparametrization: It is easy to see that order constraints of the form $\theta^s \leq \theta^w$ can be expressed by reparametrizing θ^s as $\theta^\Delta \times \theta^w$, where θ^Δ is an unconstrained parameter ranging between 0 and 1 (Knapp & Batchelder, 2005). The inequality $\theta^s < \theta^w$ holds when $0 \leq \theta^\Delta < 1$ (with $\theta^w > 0$), and the equality $\theta^s = \theta^w$ holds when $\theta^\Delta = 1$. Under this reparametrization the two hypotheses become $\mathcal{H}_{SDT} : 0 \leq \theta^\Delta_{[1|2]}, \theta^\Delta_{[2|3]} < 1$ and $\mathcal{H}_{2HT} : \theta^\Delta_{[1|2]} = \theta^\Delta_{[2|3]} = 1$.

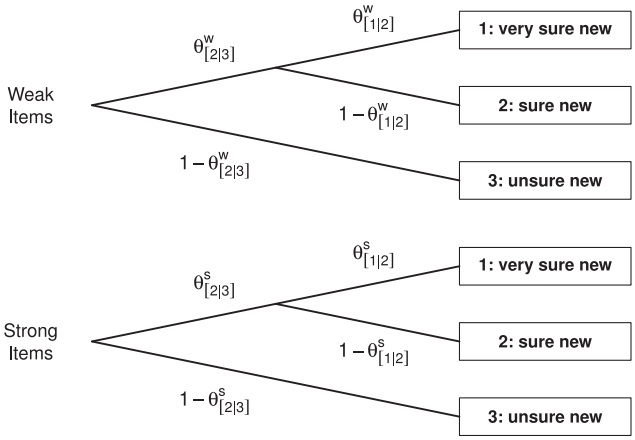

Figure 5. Multinomial processing tree model for "new" judgments.

The MPT model will be implemented using the latent-trait hierarchical-Bayesian approach described by Klauer (2010). Hierarchical extensions of models provide a principled way to estimate individual and group-level parameters, allowing one to make meaningful generalizations taking into account the estimated individual differences and commonalities (for introductions, see Lee & Wagenmakers, 2013; Rouder & Lu, 2005; Rouder et al., in press). In order to see how the hierarchical extension is introduced, let $\theta_{i,j,k}$ be the conditional confidence-rating parameters, with subscripts $j = 1, \ldots, J$ denoting the experimental study, $i = 1, \ldots, I$ the participant in that experimental group, and $k = 1, \ldots, 4$ denote the four different θ parameters in the constrained MPT model, in order, $\theta^w_{[1|2]}$, $\theta^w_{[2|3]}$, $\theta^\Delta_{[1|2]}$, and $\theta^\Delta_{[2|3]}$ (with $\theta^s = \theta^\Delta \times \theta^w$). Using a probit link, the kth parameter of the ith individual in the jth study is given by

$$\theta_{i,j,k} = \Phi(\overline{\mu}_k + \beta_{j,k} + \delta_{i,j,k}), \tag{3}$$

where $\overline{\mu}_k$ is the grand mean (or intercept) of the kth parameter, $\beta_{j,k}$ is the jth group mean displacement from the grand mean of the kth parameter, and $\delta_{i,j,k}$ is the individual-level displacement.[8] A detailed description of the hierarchical MPT model can be found in the Appendix.

Parameters will be estimated within a Bayesian framework, where the available information regarding a model and its parameters are represented by probability distributions (Gelman, Carlin, Stern, & Rubin, 2004). In particular, *posterior* distributions are obtained by using the observed data and Bayes' theorem to update established *prior* distributions. These priors can incorporate knowledge obtained from previous findings, or alternatively be rather noninformative, as in the present case, where they induce uniform

[5] Assuming that participants could still change criteria between weak and strong items in this context would imply the somewhat awkward prediction that participants can also apply different response criteria to old and new items.

[6] Given that these predictions only occur under certain combinations of parameter values, one could still argue that a constrained version of model in which these predictions do not occur would be suitable. We will return to this issue later on in a simulation study.

[7] The Gaussian distributions with unequal variances are usually attributed to variations in the familiarity increments for items presented in the study phase (e.g., Wixted, 2007). However, variable increments do not lead to a Gaussian distribution with an increased variance; such an outcome is only possible when the study of items sometimes paradoxically leads to a *decrease* in their familiarity.

[8] The J group displacements β per kth parameter are replaced via sum-to-zero contrast coding by $J - 1$ independent parameters $\beta^*_{s,k}$ that are linked to β via a J by $J - 1$ design matrix $\mathbf{X}$ that implements the sum-to-zero coding (Rouder, Morey, Speckman, & Province, 2012).

distributions ranging from 0 to 1 on the probability scale. The posterior parameter estimates represent parameter uncertainty in light of the data and the prior. In particular, the posterior $\theta^{\Delta}_{[1|2]}$ and $\theta^{\Delta}_{[2|3]}$ distributions allow one to check whether the data supports $\mathcal{H}_{\text{SDT}}$ by attributing small probabilities to values arbitrarily close to 1 (e.g., values close to 1 not being included in the 95% credibility intervals), or instead supports $\mathcal{H}_{\text{2HT}}$ by being highly concentrated at the upper boundary 1. Potential discrepancies between experimental groups (e.g., differences between the old-new and 2AFC judgments) are captured by the $\beta_{j,k}$ estimates. In the present case, given that several experimental groups will be analyzed together, one is mostly interested in the *posterior grand means* $\overline{\theta}^{\Delta}_{[1|2]} = \Phi(\overline{\mu}^{\Delta}_{[1|2]})$ and $\overline{\theta}^{\Delta}_{[2|3]} = \Phi(\overline{\mu}^{\Delta}_{[2|3]})$.

The relative evidence for each hypothesis brought about by the data will be quantified by means of *Bayes factors* (BFs; Kass & Raftery, 1995) using the *encompassing-prior approach* (Hoijtink, 2011), an approach that been previously employed in tests of similar nature (Davis-Stober & Brown, 2013; Heathcote, Brown, Wagenmakers, & Eidels, 2010; Karabatsos, 2005). Let us associate $\mathcal{H}_{\text{2HT}}$ to very large values of $\overline{\theta}^{\Delta}_{[1|2]}$ and $\overline{\theta}^{\Delta}_{[2|3]}$, for example values larger than $\epsilon = .995$. The natural logarithm of the BF for the two hypotheses corresponds to the following ratio:

$$\log\text{BF} = \log\left(\frac{P\left(\overline{\theta}^{\Delta}_{[1|2]}, \overline{\theta}^{\Delta}_{[2|3]} > \epsilon \,|\, \text{data}\right)}{P(\overline{\theta}^{\Delta}_{[1|2]}, \overline{\theta}^{\Delta}_{[2|3]} > \epsilon)}\right). \tag{4}$$

The numerator corresponds to the probability that paired samples from the *posterior* distributions of $\overline{\theta}^{\Delta}_{[1|2]}$, $\overline{\theta}^{\Delta}_{[2|3]}$ are both larger than ϵ. The denominator corresponds to the probability that paired samples from the *prior* distributions of $\overline{\theta}^{\Delta}_{[1|2]}$, $\overline{\theta}^{\Delta}_{[2|3]}$ are both larger than ϵ. Positive logBF values give support to $\mathcal{H}_{\text{2HT}}$, while negative values provide evidence in favor of $\mathcal{H}_{\text{SDT}}$.

Reanalysis of Previously Published Data

In order to estimate the grand-mean posteriors across datasets we relied on datasets with three confidence levels across two study-strength conditions. A total of nine experimental datasets from different sources were included in this analysis: four old-new datasets (Heathcote, 2003, Experiments 1 and 2; Koen & Yonelinas, 2010; Ratcliff et al., 1994, Exp. 5) and five 2AFC datasets (three from Province & Rouder, 2012, and two from Kellen et al., 2015). These nine experiments provide a total of 325 individual datasets. Contrary to the case of old-new data where only the incorrect "new" responses are relevant, in the case of 2AFC the critical test applies to both "left" and "right" incorrect responses, across which we aggregated. One advantage of this aggregation is that it capitalizes on the symmetry of these judgments in order to improve the precision of the estimates. Furthermore, in order to perfectly match the design of the MPT model, we excluded from the analysis the responses to medium-strength items from Experiments 1 and 2 of Province and Rouder (2012) and collapsed the most extreme confidence-ratings in these studies in order to have only three confidence levels for incorrect responses.[9]

Finally, one concern when implementing this analysis was the potential impact of individual datasets for which there is no clear study-strength effect. These individuals can be seen as nondiagnostic given that in the absence of a study-strength effect both hypotheses become effectively equivalent ($\mathcal{H}_{\text{SDT}}$ reduces to $\mathcal{H}_{\text{2HT}}$). The inclusion of such individuals in the analysis could shift posterior estimates as well as the BF in the direction of $\mathcal{H}_{\text{2HT}}$. In order to minimize such risks, we compared individual hit rates and excluded all individuals for which the hit rates for strong items were not significantly greater than for weak items using a likelihood ratio test ($p < .05$). A total of 115 (35%) individual datasets were excluded this way (for details see Table 1). Also, eight participants had perfect accuracy for strong old-items (their conditional confidence-ratings for "new" responses are not defined), leaving 202 diagnostic individual datasets for testing $\mathcal{H}_{\text{SDT}}$ and $\mathcal{H}_{\text{2HT}}$. This kind of exclusion or preselection of participants is vital in many tests of critical properties (e.g., Luce, 2010) and can often result in the exclusion of a considerable portion of the sample (e.g., Rae, Heathcote, Donkin, Averell, & Brown, 2014; Exp. 1). However, note that the present critical test *does not require* this participant-exclusion procedure; we are simply using it as way to ensure that we minimize the risk of distorted results. As elaborated at the end of this section, including all participants in the analysis has, however, little impact on the results in the present case.

Table 1
Reanalyzed Datasets

Study	Total participants	Diagnostic participants	Task
Heathcote (2003) Exp. 1	63	41 (65%)	Old-new
Heathcote (2003) Exp. 2	67	35 (52%)	Old-new
Kellen et al. (2015) word condition	33	24 (73%)	2AFC
Kellen et al. (2015) picture condition	30	22 (73%)	2AFC
Koen & Yonelinas (2010)	32	7 (22%)	Old-new
Province & Rouder (2012) Exp. 1	36	30 (83%)	2AFC
Province & Rouder (2012) Exp. 2	33	28 (85%)	2AFC
Province & Rouder (2012) Exp. 3	20	12 (60%)	2AFC
Ratcliff et al. (1994) Exp. 5	11	11 (100%)	Old-new

Note. Exp. = experiment; 2AFC = two-alternative forced-choice. The diagnostic individual datasets correspond to those were the hit rate for strong items was significantly *greater* than for weak items ($p < .05$). A likelihood-ratio test was used. The sampling distribution of this test follows a $\bar{\chi}^2$ distribution with a critical value ($p = .05$) of 2.71.

The main results for the hierarchical MPT are presented in Table 2 and Figure 6: The posterior estimates of the grand means $\overline{\theta}^{\Delta}_{[1|2]}$ and $\overline{\theta}^{\Delta}_{[2|3]}$ were found to be highly skewed toward the upper boundary of the parameter space with medians .99 and 1, respectively, a result that is consistent with $\mathcal{H}_{\text{2HT}}$ (the 95% credibility intervals are reported in Table 2). The posterior distribution of $\overline{\theta}^{\Delta}_{[1|2]}$ is not as peaked as the posterior distribution of $\overline{\theta}^{\Delta}_{[2|3]}$, a difference that is expected given that incorrect maximum-confidence responses are rather infrequent. Overall, these results strongly suggest that increases in performance do not lead to changes in the profile of the incorrect confidence ratings. Note that these results are based on noninformative priors that distribute mass uniformly across the unit interval, making the obtained posterior distributions even more impressive. The posterior estimates of the group-level displacements (β^*; see Equation 3 and Footnote 8) for

[9] Data from Province and Rouder (2012, Experiment 1) consists of responses on an 8-point scale from 1 (*very sure left*) to 8 (*very sure right*). We collapsed ratings 1–2 and 7–8 in order to have a 6-point scale. Experiment 2 consists of responses on a 10-point scale and we collapsed ratings 1–3 and 8–10.

Table 2
Posterior Estimates of the Hierarchical Multinomial Processing Tree

$\overline{\theta}^{w}_{[1\vert2]} = .40\ [.35, .43]$	$\overline{\theta}^{w}_{[2\vert3]} = .56\ [.52, .60]$	$\overline{\theta}^{\Delta}_{[1\vert2]} = .99\ [.95, 1]$	$\overline{\theta}^{\Delta}_{[2\vert3]} = 1\ [.99, 1]$
$\beta^{*\Delta}_{1,[1\vert2]} = -0.21\ [-1.46, 1.40]$	$\beta^{*\Delta}_{2,[1\vert2]} = -0.23\ [-1.70, 1.58]$	$\beta^{*\Delta}_{3,[1\vert2]} = .20\ [-0.98, 1.61]$	$\beta^{*\Delta}_{4,[1\vert2]} = -0.05\ [-1.10, 1.09]$
$\beta^{*\Delta}_{5,[1\vert2]} = -0.24\ [-1.32, 1.24]$	$\beta^{*\Delta}_{6,[1\vert2]} = .07\ [-0.79, 1.26]$	$\beta^{*\Delta}_{7,[1\vert2]} = .00\ [-1.28, 1.25]$	$\beta^{*\Delta}_{8,[1\vert2]} = -1.27\ [-2.47, 0.14]$
$\beta^{*\Delta}_{1,[2\vert3]} = -0.65\ [-2.08, 1.09]$	$\beta^{*\Delta}_{2,[2\vert3]} = -0.27\ [-1.37, 1.25]$	$\beta^{*\Delta}_{3,[2\vert3]} = .01\ [-1.59, 1.23]$	$\beta^{*\Delta}_{4,[2\vert3]} = -0.32\ [-1.71, 1.64]$
$\beta^{*\Delta}_{5,[2\vert3]} = -0.57\ [-2.04, 1.38]$	$\beta^{*\Delta}_{6,[2\vert3]} = -0.29\ [-1.01, 1.21]$	$\beta^{*\Delta}_{7,[2\vert3]} = -0.04\ [-1.24, 1.87]$	$\beta^{*\Delta}_{8,[2\vert3]} = -0.38\ [-1.44, 1.46]$

Note. The values inside the square brackets are the 95% credibility intervals. Further details on the model's posterior estimates can be found in the Supplemental Materials.

$\theta^{\Delta}_{[1|2]}$ and $\theta^{\Delta}_{[2|3]}$ corroborate these results: In each case, the value zero is well within the parameter's 95% credibility interval. This shows that $\mathcal{H}_{2HT}$ is consistent with all nine groups, which include both old-new and 2AFC judgments (Kruschke, 2011, Chap. 18). The estimates of the individual-effect variances (the variance of δ; see Equation 3 and the Appendix) for $\theta^{\Delta}_{[1|2]}$ and $\theta^{\Delta}_{[2|3]}$ were 0.32 [0.01, 0.92] and 0.43 [0.02, 1.42], respectively, which indicate the presence of moderate participant heterogeneity. These results suggest that although the data are in general consistent with $\mathcal{H}_{2HT}$, a small portion of the participants might be more in line with $\mathcal{H}_{SDT}$.

The conditional confidence-ratings calculated directly from each individual participant's raw data, $\hat{\theta}^{w}_{[1|2]}$, $\hat{\theta}^{w}_{[2|3]}$, $\hat{\theta}^{s}_{[1|2]}$, and $\hat{\theta}^{s}_{[2|3]}$, fail to show any systematic decrease for strong items (see Figure 6): For the selected participants, $\theta^{w}_{[1|2]}$ is greater than $\theta^{s}_{[1|2]}$ in only 48% of the cases, and $\theta^{w}_{[2|3]}$ is greater than $\theta^{s}_{[2|3]}$ in 50%. For both parameter pairs, the median difference was zero, which exactly corresponds to the invariance predicted by $\mathcal{H}_{2HT}$. For the complete sample of participants with nondiagnostic participants included, the percentages are 47% and 49%, respectively, and the median difference was again zero for both parameter pairs.[10] Note, however, that these plotted differences do not reflect the precision of the empirical values (e.g., $\hat{\theta}^{w}_{[1|2]}$ and $\hat{\theta}^{s}_{[1|2]}$), which vary both within and across participants (see Kellen & Klauer, 2014).

The evaluation of the posterior-parameter distributions was corroborated by a logBF of 9.04 ($\epsilon = .995$), a value that indicates the presence of extremely strong evidence in favor of $\mathcal{H}_{2HT}$ (Jeffreys, 1961). This result reflects the fact that 21% of $\{\overline{\theta}^{\Delta}_{[1|2]}, \overline{\theta}^{\Delta}_{[2|3]}\}$ pairs sampled from the posterior distributions were simultaneously larger than 0.995, in comparison to the 0.0025% expected under the prior distributions.

One valid concern with BFs is their sensitivity to the used priors. The sensitivity of BFs to the choice of priors is an issue that is well documented in the literature (e.g., Liu & Aitkin, 2008), and more moderate BFs would be obtained if priors placing more probability mass above ϵ were used instead of the uniform priors used here. Also, such priors would be more in line with the predictions of many parametric SDT models. But such priors would also lead to $\overline{\theta}^{\Delta}_{[1|2]}$ and $\overline{\theta}^{\Delta}_{[2|3]}$ posteriors that are *even more* skewed toward the upper boundary and therefore even more consistent with $\mathcal{H}_{2HT}$. This situation leads us to argue that the sensitivity to the choice of priors does not affect the overall consistency of the data with $\mathcal{H}_{2HT}$; if anything, it testifies to the strength of the results. The potential impact of the prior will be addressed later on in a robustness analysis in which data generated with a parametric SDT model will be fitted by the hierarchical model.

Moreover, the suitability of the MPT model was evaluated by means of posterior-predictive tests (Gelman & Shalizi, 2012): In such tests, the model's misfit of artificial data generated from the posterior distributions is compared with the misfit of the actual observed data. The proportion of misfits of artificial data that are more extreme than the misfit of the observed data can be interpreted as a Bayesian p value. Two tests were used: A test T_1 quantifying the model's ability to account for the total observed category frequencies, aggregated across individuals, and a test T_2 evaluating the model's ability to account for the variances and covariances in the observed category frequencies (for further details, see Klauer, 2010). The hierarchical MPT model performed well in both tests (smallest Bayesian p value was .45), indicating that it provides a good account of the data.

We also checked whether the present results were affected by excluding participants without a significant study-strength effect: The posterior estimates obtained when including all individual datasets were extremely similar to the ones obtained with the nonexcluded participants, with $\overline{\theta}^{\Delta}_{[1|2]} = .99\ [.97, 1]$, $\overline{\theta}^{\Delta}_{[2|3]} = 1\ [.98, 1]$, $\overline{\theta}^{w}_{[1|2]} = .38\ [.34, .41]$, and $\overline{\theta}^{w}_{[2|3]} = .57\ [.54, .61]$. Not surprisingly, a logBF of 9.84 strongly supporting $\mathcal{H}_{2HT}$ was obtained. This result follows from the fact that 47% of $\{\overline{\theta}^{\Delta}_{[1|2]}, \overline{\theta}^{\Delta}_{[2|3]}\}$ pairs sampled from the posterior distributions was simultaneously larger than 0.995. The similarity between the posterior group-level estimates indicates that besides the magnitude of the strength effect, individuals were overall similar in both groups (see Figure 6). We also checked whether there was any relationship between (a) the individuals' study-strength effect, represented by the difference in hit rates between weak and strong items, and (b) the differences between the conditional confidence-ratings. If the data were consistent with a continuous model then according to the above-established theorem one would expect a negative correlation between the hit-rate differences and conditional confidence-rating differences (e.g., between $\theta^{w}_{[1|2]}$ and $\theta^{s}_{[1|2]}$). No correlation was found (most extreme Spearman $r = -0.03$, smallest $p = .61$), corroborating the visual inspection of Figure 6.

Finally, we conducted a robustness analysis by fitting the hierarchical MPT to data generated from a unequal-variance Gaussian SDT model. This Gaussian SDT model was constrained so as to

[10] For 14 of the diagnostic participants, both ratings 1 and 2 were never selected for strong items, which means that $\hat{\theta}^{s}_{[1|2]}$ is not defined. We removed these cases from Figure 6 and from the calculation of the proportion of times $\hat{\theta}^{w}_{[1|2]}$ is greater than $\hat{\theta}^{s}_{[1|2]}$. Note that these undefined values are not problematic for the Bayesian estimation.

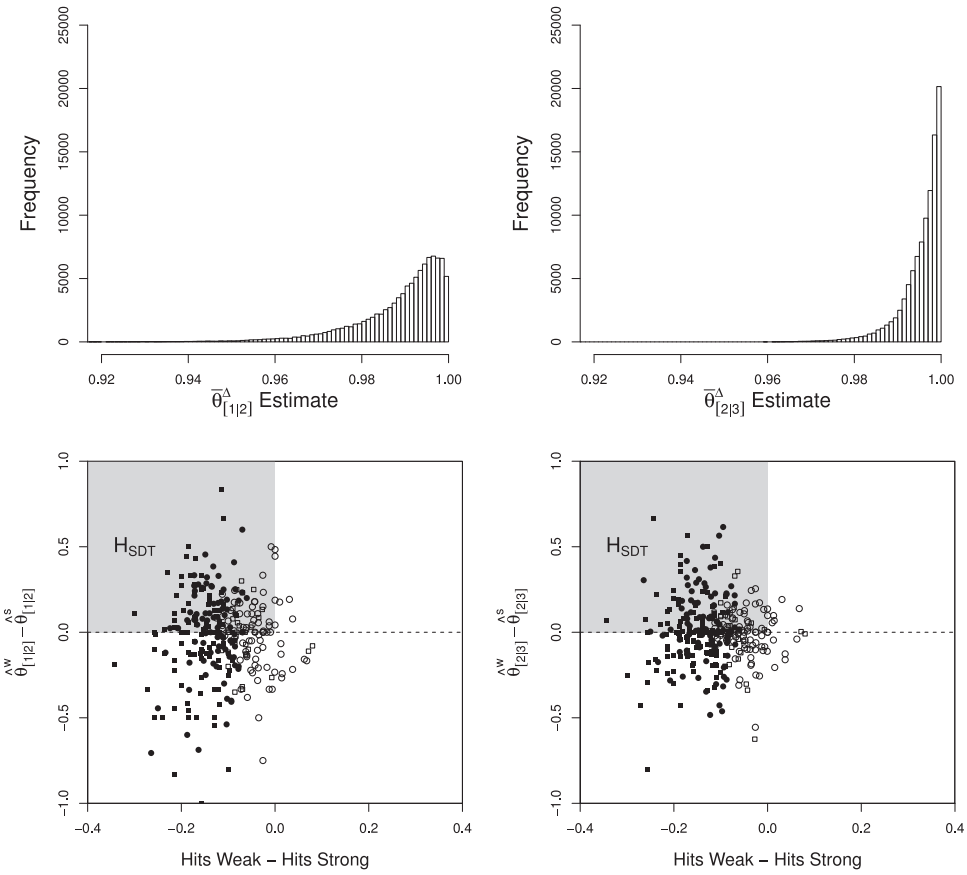

Figure 6. Top panels: Posterior distribution samples of $\overline{\theta}^{\Delta}_{[1|2]}$ and $\hat{\theta}^{s}_{[2|3]}$. Bottom panels: Differences between empirical individual conditional confidence-ratings: $\hat{\theta}^{w}_{[1|2]} - \hat{\theta}^{s}_{[1|2]}$ and $\hat{\theta}^{w}_{[2|3]} - \hat{\theta}^{s}_{[2|3]}$ (calculated directly from the raw data; see Footnote 10). The plotted circles and squares correspond come from old-new and two-alternative forced-choice judgments, respectively. Filled and empty squares/circles correspond to the diagnostic and excluded participants, respectively. Both differences are plotted against the differences between the empirical hit rates for weak and strong items. The shaded area corresponds to the region consistent with $\mathcal{H}_{SDT}$. The dashed horizontal lines correspond to the medians of the differences between the conditional confidence-ratings (the median is the same irrespective of considering the total number of participants or just the diagnostic ones).

avoid the paradoxical predictions that errors increase in confidence as memory strength increases: $\theta^{s}_{[1|2]} > \theta^{w}_{[1|2]}$ and $\theta^{s}_{[2|3]} > \theta^{w}_{[2|3]}$ cannot take place (see Footnote 6). This constraint was imposed by attributing an extremely large misfit to parameter combinations that produced these paradoxical predictions. The data from all participants were fitted with the constrained SDT model and artificial data were subsequently generated from each individual's predicted response probabilities.[11] The fit of the hierarchical MPT to the SDT-generated data found to be diagnostic resulted in a logBF of −23.90, indicating extreme support to $\mathcal{H}_{SDT}$.[12] The posterior estimates of $\overline{\theta}^{\Delta}_{[1|2]}$ and $\overline{\theta}^{\Delta}_{[2|3]}$ were .92 [.83, .98] and .91 [.85, .97], respectively. These results demonstrate the sensitivity of this approach to the data-generating processes (despite the noninformative priors used) and reinforces the results in support of $\mathcal{H}_{2HT}$.

Discussion

The comparison between continuous and discrete-state accounts has a long history that can be traced back to the work of Leibniz (Rouder & Morey, 2009). In the context of modern cognitive-psychological research, work on these comparisons has almost exclusively relied on the overall penalized fit statistics obtained from parametric implementations of the models. In the present paper we explored a different approach and in the context of recognition memory compared the two accounts on the basis of

[11] Although the model was fitted with separate σ^2_w and σ^2_s, the need for separate parameters was hardly justified given that the equality restriction $\sigma^2_w = \sigma^2_s$ was rejected in only 9% of the diagnostic individual datasets, slightly above the 5% nominal rate.

[12] In this case, we were unable to obtain posterior samples of $\overline{\theta}^{\Delta}_{[1|2]}$ and $\overline{\theta}^{\Delta}_{[2|3]}$ that were simultaneously larger than .995. We therefore estimated this probability by fitting the sampled posteriors with a bivariate Gaussian (which provided a suitable approximation) and integrated the estimated distribution across the open rectangle bounded from below by 2.578 on the horizontal and vertical axis (corresponding to the rectangle between .995 and 1 on both axes on a probability scale). For the other logBFs computations, the use of a Gaussian approximation produced negligible deviations from the direct use of sampled proportions above .995.

their core assumptions: *the direct mapping of graded information* versus *complete information loss*. The reanalysis of previously published ROC data provides strong evidence for the presence of discrete states with complete information loss. This reanalysis generalizes the preference for the 2HT reported by Province and Rouder (2012) and Kellen et al. (2014) by corroborating it independently of the chosen model parametrizations and chosen model-selection statistics and extends the results to the classical case of old-new judgments. A major advantage of the critical-test approach pursued here is that the relationship between the model predictions and the data is made transparent and simple. This means that not only do we learn something about the models, but we also acquire direct knowledge on individuals' recognition judgments. In the present case, the support for the 2HT model comes from the fact that individuals' confidence judgments for nonrecognized items do not seem to be related to their memory for the different item classes. In other words, items that one fails to recognize are treated alike, irrespective of the study conditions. This behavioral finding is particularly relevant given that greater accuracy levels for studied items are usually associated with greater confidence, leading many to consider confidence as a proxy for memory strength (Busey, Tunnicliff, Loftus, & Loftus, 2000; Hart, 1967).

It is worth emphasizing that the present evidence in favor of the 2HT model does *not* reject the existence of continuous or graded mnemonic information. The latter notion was strongly supported by Kellen and Klauer's (2014) critical test on ranking judgments. What the present results reject is the core SDT assumption that the observed confidence-rating judgments reflect a *direct mapping* of this continuous mnemonic information via a set of ordered response criteria. This direct-mapping assumption, which has been criticized by others in the past (e.g., Ratcliff & Starns, 2009; Van Zandt, 2000) but is often accepted tacitly (e.g., Mickes, Hwe, Wais, & Wixted, 2011), is independent of the assumption of continuous mnemonic information and should be treated as such. When evaluating the plausibility of the direct-mapping assumption one consideration is how easily it breaks down: For instance, it is extremely unlikely that individuals are able to establish a large set of criteria when using a large confidence-rating scale (e.g., a 20- or 99-point scale; see Mickes, Wixted, & Wais, 2007). Indeed some authors have argued that individuals have problems establishing and maintaining as few as the five response criteria required for using a 6-point scale, with criteria varying randomly across trials in consequence (e.g., Benjamin, Diaz, & Wee, 2009; but see Kellen, Klauer, & Singmann, 2012). If researchers consider individuals' establishment of response criteria so difficult and taxing then it seems fruitful to at least entertain the possibility that individuals act upon the available mnemonic information available in a simplifying (i.e., discretizing) way.[13]

One account that is in line with the present results is that individuals produce confidence-rating judgments through a graded representation mediated by discrete states, perhaps as a way to deal with the demands of a typical recognition-memory task with several levels of confidence (Malmberg, 2008). This mediation is not present in other cases such as ranking judgments (Kellen & Klauer, 2014). Here, the task of ordering the alternatives according to their relative familiarity encourages the use of a graded mnemonic representation. Note that this does not mean that the mnemonic information available is different, what differs is the way mnemonic information is mapped onto the observed responses (see Rouder & Morey, 2009). The literature on judgment and decision-making provides several examples where individuals simplify and/or ignore rich evidence in some contexts but not in others (e.g., Gigerenzer & Goldstein, 1996; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011). The notion that mnemonic information can be acted upon differently depending on the characteristics of the tasks is far from new (e.g., Humphreys, Bain, & Pike, 1989): For instance, both Heathcote, Raymond, and Dunn (2006) and Malmberg and Xu (2007) reported experimental results from different recognition tasks (list/plural discrimination and associative recognition) in which false alarms to similar distractors increased when nonsimilar distractors were included in the test list, a finding that suggests that individuals change the nature of the mnemonic information used in order to satisfy task demands (for an overview of these results, see Malmberg, 2008). A second example can be found in the joint modeling of item and source memory (Hautus, Macmillan, & Rotello, 2008; Klauer & Kellen, 2010; Slotnick & Dodson, 2005): Traditional SDT accounts assume bivariate densities that reflect the graded information on the oldness of items (whether they were studied) and their respective source (in which context did the items occur). On one hand, SDT proponents have argued that the graded nature of mnemonic information can be directly observed in specific cases where individuals are able to correctly identify the source of nonrecognized items (e.g., Starns, Hicks, Brown, & Martin, 2008; see also Ceci, Fitneva, & Williams, 2010). On the other hand, post hoc constraints have been introduced in the implemented SDT models, imposing that source information is completely absent in the confidence-ratings for nonrecognized items (Hautus et al., 2008; Slotnick & Dodson, 2005; but see Starns, Rotello, & Hautus, 2014). If this kind of differences in information use can be induced by slight changes in a specific task, the possibility that individuals act upon the available mnemonic information differently when engaging in distinct tasks does not seem implausible at all.

On a more general level, the present work testifies to the usefulness of critical tests, an approach that so far has been ignored in the case of recognition-memory modeling. The contribution of critical tests to the study of human memory is particularly relevant as it resonates with criticisms that theoretical progress coming from model fits can be limited (Roberts & Pashler, 2000; see also Birnbaum, 1973, 2011). Here, we showed that several problematic issues in the comparison of continuous and discrete-state models can thereby be avoided

[13] One reviewer questioned whether criterion variability in SDT could affect the predictions of $\mathcal{H}_{SDT}$. We inspected the predictions of several plausible forms of criterion variability (Kellen, Klauer, & Singmann, 2012; Klauer & Kellen, 2012; Rosner & Kochanski, 2009) and found the $\mathcal{H}_{SDT}$ predicted pattern of inequalities to remain unchanged. Of course, given that there are infinitely many ways to introduce criterion variability in SDT we cannot simply rule out the possibility that some specific form of criterion variability reduces $\mathcal{H}_{SDT}$ to $\mathcal{H}_{2HT}$ or reverses the predictions of $\mathcal{H}_{SDT}$. But given that this is an unlikely possibility, we believe that the burden of proof should reside with proponents of a criterion-variability account.

and that a test based on each models' core assumptions can be established. For this purpose, we focused on a specific portion of the data that was found to be truly informative and pretty much ignored the rest. Of course, we are not suggesting that attempts to fit the complete data should be abandoned. What is being proposed here is that the most suitable strategy be employed when attempting to answer a specific question, and in the present case the question was which core assumption accounts for confidence-rating judgments better.

References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*. New York, NY: Dover.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society, 21,* 503–546. http://dx.doi.org/10.2307/1907921

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. D. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology, vol. 1, learning, memory and thinking* (pp. 243–293). San Francisco, CA: Freeman.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189–1206. http://dx.doi.org/10.1037/0096-1523.25.5.1189

Balakrishnan, J., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 615–633. http://dx.doi.org/10.1037//0096-1523.22.3.615

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin, 139,* 1204–1212. http://dx.doi.org/10.1037/a0033894

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115. http://dx.doi.org/10.1037/a0014351

Birnbaum, M. H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin, 79,* 239–242. http://dx.doi.org/10.1037/h0033853

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115,* 463–501. http://dx.doi.org/10.1037/0033-295x.115.2.463

Birnbaum, M. H. (2011). Testing theories of risky decision making via critical tests. *Frontiers in Psychology, 2,* 315. http://dx.doi.org/10.3389/fpsyg.2011.00315

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

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

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26–48. http://dx.doi.org/10.3758/bf03210724

Ceci, S. J., Fitneva, S. A., & Williams, W. M. (2010). Representational constraints on the development of memory and metamemory: A developmental–representational theory. *Psychological Review, 117,* 464–495. http://dx.doi.org/10.1037/a0019067

Chechile, R. A. (2003). Mathematical tools for hazard function analysis. *Journal of Mathematical Psychology, 47,* 478–494. http://dx.doi.org/10.1016/S0022-2496(03)00063-4

Chechile, R. A. (2004). New multinomial models for the Chechile-Meyer task. *Journal of Mathematical Psychology, 48,* 364–384. http://dx.doi.org/10.1016/j.jmp.2004.09.002

Chechile, R. A. (2006). Memory hazard functions: A vehicle for theory development and test. *Psychological Review, 113,* 31–56. http://dx.doi.org/10.1037/a0019067

Chechile, R. A. (2011). Properties of reverse hazard functions. *Journal of Mathematical Psychology, 55,* 203–222. http://dx.doi.org/10.1016/j.jmp.2011.03.001

Chen, T., Starns, J. J., & Rotello, C. M. (in press). A violation of the conditional independence assumption in the two-high-threshold model of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Colonius, H. (1988). Modeling the redundant signals effect by specifying the hazard function. *Perception & Psychophysics, 43,* 604–606. http://dx.doi.org/10.3758/BF03207750

Davis-Stober, C. P., & Brown, N. (2013). Evaluating decision maker "type" under *p*-additive utility representations. *Journal of Mathematical Psychology, 57,* 320–328. http://dx.doi.org/10.1016/j.jmp.2013.08.002

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 186–205. http://dx.doi.org/10.1037/1082-989X.3.2.186

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

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

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127,* 83–96. http://dx.doi.org/10.1037/a0013081

Falmagne, J.-C. (1985). *Elements of psychophysical theory*. New York, NY: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.), London, UK: Chapman and Hall.

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

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66,* 8–38. http://dx.doi.org/10.1111/j.2044-8317.2011.02037.x

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior, 6,* 685–691. http://dx.doi.org/10.1016/s0022-5371(67)80072-0

Hautus, M., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15,* 889–905. http://dx.doi.org/10.3758/PBR.15.5.889

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210–1230. http://dx.doi.org/10.1037/0278-7393.29.6.1210

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of*

Memory & Language, 55, 495–514. http://dx.doi.org/10.1016/j.jml.2006.07.001

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

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: CRC Press.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96,* 208–233. http://dx.doi.org/10.1037/0033-295X.96.2.208

Iverson, G. J., & Bamber, D. (1997). The generalized area theorem in signal detection theory. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 301–318). Hillsdale, NJ: Erlbaum.

Jeffreys, H. (1961). *The theory of probability*. New York, NY: Oxford University Press.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121,* 1–32. http://dx.doi.org/10.1037/a0034190

Kaernbach, C. (1991). Poisson signal-detection theory: Link between threshold models and the Gaussian assumption. *Perception & psychophysics, 50,* 498–506. http://dx.doi.org/10.3758/bf03205066

Kahana, M. J. (2014). *Foundations of human memory*. New York, NY: Oxford University Press.

Karabatsos, G. (2005). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. *Journal of Mathematical Psychology, 49,* 51–69. http://dx.doi.org/10.1016/j.jmp.2004.11.001

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.2307/2291091

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology, 55,* 251–266. http://dx.doi.org/10.1016/j.jmp.2010.11.004

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1795–1804. http://dx.doi.org/10.1037/xlm0000016

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

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review, 119,* 457–479. http://dx.doi.org/10.1037/a0027727

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

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review, 74,* 496–504. http://dx.doi.org/10.1037/h0025127

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

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source memory: A discrete-state approach. *Psychonomic Bulletin & Review, 17,* 465–478. http://dx.doi.org/10.3758/PBR.17.4.465

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

Klauer, K. C., & Kellen, D. (2012). The Law of Categorical Judgment (Corrected) extended: A note on Rosner and Kochanski (2009). *Psychological Review, 119,* 216–220.

Knapp, B. R., & Batchelder, W. H. (2005). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology, 48,* 215–229. http://dx.doi.org/10.1016/j.jmp.2004.03.002

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

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123,* 297–315. http://dx.doi.org/10.1037/0096-3445.123.3.297

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308–324. http://dx.doi.org/10.1037/h0027238

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. New York, NY: Academic Press.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York, NY: Cambridge University Press.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52,* 362–375. http://dx.doi.org/10.1016/j.jmp.2008.03.002

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109. http://dx.doi.org/10.1037/h0029536

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61–79. http://dx.doi.org/10.1037/h0039723

Luce, R. D. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology, 41,* 79–87. http://dx.doi.org/10.1006/jmps.1997.1150

Luce, R. D. (2010). Behavioral assumptions for a class of utility models: A program of experiments. *Journal of Risk and Uncertainty, 41,* 19–37. http://dx.doi.org/10.1007/s11166-010-9098-5

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387. http://dx.doi.org/10.1037//0278-7393.28.2.380

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

Malmberg, K. J., & Xu, J. (2007). On the flexibility and fallibility of associative memory. *Memory & Cognition, 35,* 545–556. http://dx.doi.org/10.3758/bf03193293

Mandler, G., Pearlstone, Z., & Koopmans, H. S. (1969). Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8,* 410–423. http://dx.doi.org/10.1016/S0022-5371(69)80134-9

Mickes, L., Hwe, V., Wais, P., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257. http://dx.doi.org/10.1037/a0023007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. http://dx.doi.org/10.1037/e527352012-230

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. http://dx.doi.org/10.3389/fpsyg.2011.00147

Pazzaglia, A., Dube, C., & Rotello, C. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139,* 1173–1203. http://dx.doi.org/10.1037/a0033044

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. http://dx.doi.org/10.1037/0033-295x.109.3.472

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

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1226–1243. http://dx.doi.org/10.1037/a0036801

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785. http://dx.doi.org/10.1037//0278-7393.20.4.763

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83. http://dx.doi.org/10.1037/a0014086

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367. http://dx.doi.org/10.1037//0033-295X.107.2.358

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgment (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116,* 116–128. http://dx.doi.org/10.1037/a0014463

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

Rouder, J., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review, 116,* 655–660. http://dx.doi.org/10.1037/a0016413

Rouder, J. N., Morey, R. D., & Pratte, M. S. (in press). Hierarchical Bayesian models. In W. H. Batchelder, H. Colonius, E. Dzhafarov, & J. I. Myung (Eds.), *New handbook of mathematical psychology, vol. 1*: *Measurement and methodology*. New York, NY: Cambridge University Press.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review, 17,* 427–435. http://dx.doi.org/10.3758/PBR.17.3.427

Rouder, J. N., Province, J. M., Swagman, A. R., & Thiele, J. E. (2013). *From ROC curves to psychological theory* (Manuscript submitted for publication).

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. http://dx.doi.org/10.3758/bf03209391

Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality, and shape constraints*. Hoboken, NJ: Wiley.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33,* 151–170. http://dx.doi.org/10.3758/BF03195305

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

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

Starns, J. J., Rotello, C. M., & Hautus, M. J. (2014). Recognition memory zROC slopes for items with correct versus incorrect source decisions discriminate the dual process and unequal variance signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 685–691. http://dx.doi.org/10.1037/a0036846

Stevens, S. S., Morgan, C. T., & Volkmann, J. (1941). Theory of the neural quantum in the discrimination of loudness and pitch. *The American Journal of Psychology,* 315–335. http://dx.doi.org/10.2307/1417678

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

Swagman, A. R., Province, J. M., & Rouder, J. N. (2015). Perceptual word identification is mediated by discrete states. *Psychonomic Bulletin & Review, 22,* 265–273.

Thomas, E. A. C. (1971). Sufficient conditions for monotone hazard rate an application to latency-probability curves. *Journal of Mathematical Psychology, 8,* 303–332. http://dx.doi.org/10.1016/0022-2496(71)90036-8

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567. http://dx.doi.org/10.1037/0033-2909.108.3.551

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York, NY: Cambridge University Press.

Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111,* 1003–1035. http://dx.doi.org/10.1037/0033-295X.111.4.1003

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological review, 108,* 550–592. http://dx.doi.org/10.1037//0033-295x.108.3.550

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. http://dx.doi.org/10.1037//0278-7393.26.3.582

Wandell, B., & Luce, R. D. (1978). Pooling peripheral information: Averages versus extreme values. *Journal of Mathematical Psychology, 17,* 220–235. http://dx.doi.org/10.1016/0022-2496(78)90017-2

Wei, G., & Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association, 85,* 699–704. http://dx.doi.org/10.2307/2290005

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. http://dx.doi.org/10.1037/0033-295X.114.1.152

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800–832. http://dx.doi.org/10.1037/0033-2909.133.5.800

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20,* 423–428. http://dx.doi.org/10.1111/j.1467-9280.2009.02322.x

(*Appendix follows*)

Appendix

Proof of Theorem

Let F_μ be the cumulative distribution function of the old-item distribution. Note that we are only considering distribution functions that are differentiable in μ and for which $F_\mu(z) > 0$ for all admissible μ and z.

Theorem: If $H_\mu(z)$ is monotonically increasing in z for all μ, then $\frac{F_\mu(a)}{F_\mu(b)}$ is monotonically decreasing in μ for any pair of criterion values a, b with a < b.

Proof: The conditional probability $CP(\mu) = \frac{F_\mu(a)}{F_\mu(b)}$ is monotonically decreasing in μ, if $\frac{\partial}{\partial\mu}CP(\mu) < 0$ for all μ. Taking derivatives, this is equivalent to $\frac{\partial}{\partial\mu}F_\mu(a)F_\mu(b) - F_\mu(a)\frac{\partial}{\partial\mu}F_\mu(b) < 0$. This in turn is equivalent to $\frac{\partial}{\partial\mu}F_\mu(a) \times F_\mu(a)^{-1} < \frac{\partial}{\partial\mu}F_\mu(b) \times F_\mu(b)^{-1}$. This completes the proof.

The function $H_\mu(z)$ is related to the signal distribution's hazard function and can be shown to be increasing for many common distributions. For example, for shift distributions with $f_\mu(y) = f(y - \mu)$, the following corollary relates H_μ to the reverse hazard (Chechile, 2011), given by $r_\mu(y) = \frac{f_\mu(y)}{F_\mu(y)}$:

Corollary: For a family of shift distributions, $H_\mu(y)$ is monotonically increasing in y for all μ, if the reverse hazard is monotonically decreasing in y for all μ under the assumptions of the above theorem.

This follows from the fact that for a shift distribution $H_\mu(y) = -r_\mu(y)$ as is easy to see. Using the results reported by Chechile (2011), it immediately follows that $H_\mu(z)$ is monotonically increasing for signal distributions based on normal distributions (as in the equal- and unequal-variance SDT models), but also for the case of exponential distributions, $f_\mu(y) = k\exp(-k(y - \mu))$, for shift distributions based on ex-Gaussian distributions, Gumbel distributions, and many other distributions.

The distributions satisfying the conditions of the theorem are not limited to distributions with signal strength conceptualized as a shift parameter. Consider, for example, the gamma distribution with shape $\alpha > 0$ and scale $\mu > 0$. Its distribution function is

$$F_\mu(z) = \frac{\gamma\left(\alpha, \frac{z}{\mu}\right)}{\Gamma(\alpha)},$$

where γ is the lower incomplete gamma function. Its derivative with respect to μ is

$$\frac{\partial}{\partial\mu}F_\mu(z) = \frac{1}{\Gamma(\alpha)}\left(-\frac{z}{\mu^2}\right)\left(\frac{z}{\mu}\right)^{\alpha-1}e^{-\frac{z}{\mu}}.$$

Note that this function is bounded in an interval around each μ for all z. Using a series expansion of the lower incomplete gamma function (Abramowitz & Stegun, 1964, p. 262), on the other hand, leads to

$$F_\mu(z) = \left(\frac{z}{\mu}\right)^\alpha e^{-\frac{z}{\mu}}\sum_{i=0}^{\infty}\frac{(z/\mu)^i}{\Gamma(\alpha + i + 1)}.$$

Hence,

$$H_\mu(z) = -\frac{1}{\mu\Gamma(\alpha)}\frac{1}{\sum_{i=0}^{\infty}\frac{(z/\mu)^i}{\Gamma(\alpha + i + 1)}},$$

which is monotonically increasing in $z > 0$.

Hierarchical Bayesian MPT

This subsection describes the hierarchical-Bayesian MPT model used in the analyses. The C++ code implementation can be found in the Supplemental Material.

Let $\theta_{i,j,k}$, with $j = 1, \ldots, J$ denoting the experimental study group, $i = 1, \ldots, I_j$ the participants in group j, and $k = 1, \ldots, K = 4$ denote the four different θ parameters in the MPT model shown in Figure 5, in order, $\theta^w_{[1|2]}$, $\theta^w_{[2|3]}$, $\theta^\Delta_{[1|2]}$, and $\theta^\Delta_{[2|3]}$ (with $\theta^s = \theta^\Delta \times \theta^w$). Using a probit link, the kth parameter of the ith individual in the jth study is given by

$$\theta_{i,j,k} = \Phi(\overline{\mu}_k + \beta_{j,k} + \eta_k\delta_{i,j,k}), \tag{5}$$

where $\overline{\mu}_k$ is the grand mean (or intercept) of the kth parameter, $\beta_{j,k}$ is the jth group mean displacement from the grand mean of the kth parameter, $\delta_{i,j,k}$ is the individual-level displacement. Parameters $\delta_{i,j,k}$ have a zero-centered multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ as prior distribution. The covariance matrix Σ captures any potential correlations between parameters across individuals (Klauer, 2010). Parameter η_k is a redundant multiplicative scale-parameter that only serves to accelerate the rate of convergence in the Markov chain Monte Carlo (MCMC) sampling method used for parameter estimation (see Gelman, Carlin, Stern, & Rubin, 2004, Chap. 15).

(*Appendix continues*)

The J group displacements β per kth parameter are replaced via sum-to-zero contrast coding by $J - 1$ independent parameters $\beta^*_{s,k}$ that are linked to β via a J by $J - 1$ design matrix $\mathbf{X}$ that implements the sum-to-zero contrast coding. $\mathbf{X}$ is the matrix comprised of the $J - 1$ eigenvectors of unit length of $\mathbf{Z} = \mathbf{I_J} - \mathrm{U/J}$, with $= \mathbf{I_J}$ being an identity matrix of size J and $\mathbf{U}$ a square matrix of size J with entries 1.0 (see Rouder, Morey, Speckman, & Province, 2012, p. 363). If $\mathbf{X}$ has rows $\mathbf{x_1}, \ldots, \mathbf{x_J}$, and β^* is the vector of the new parameters, then $\beta_{j,k} = \mathbf{x}_j\beta^*_k$.

The hierarchical MPT parameters have the following hyperpriors:

$$\overline{\mu}_k \sim \mathcal{N}(0, 1),$$

$$\beta^*_{s,k} \sim \mathcal{N}(0, 1),$$

$$\eta_k \sim \mathcal{N}(1, 1),$$

$$\Sigma \sim \mathcal{W}^{-1}(Id(\#K), \#K + 1),$$

where $\mathcal{N}$ is the Gaussian distribution and $\mathcal{W}^{-1}(Id(\#K), \#K + 1)$ is the so-called inverse Wishart distribution with $\#K + 1$ degrees of freedom, with $Id(\#K)$ being the identity matrix with $\#K$ rows and columns, with $\#K = 4$ being the cardinality of set K (see Gelman & Hill, 2007, Chap. 13).

The hierarchical model was implemented in C++ using the NAG library, based on scripts developed by Klauer (2010). In order to obtain posterior-parameter estimates form the hierarchical MPT, four independent streams of MCMC samples were collected using a Gibbs sampler. Rough initial estimates of the parameters were obtained by means of the Monte Carlo EM algorithm (Wei & Tanner, 1990). Chain convergence was assessed via the $\hat{R}$ statistic, which compares within-chain variance to between-chain variance (Gelman et al., 2004, Chap. 11). Sampling with the Gibbs sampler continued until all MCMC streams converged (all $\hat{R} \leq 1.05$), and then went on for 25,000 consecutive samples per stream, for a total of 100,000 draws from the posterior parameter distributions retained for analysis.

Received September 17, 2014
Revision received March 2, 2015
Accepted March 2, 2015 ■