Journal of Experimental Psychology: Learning, Memory, and Cognition

Discrete-State and Continuous Models of Recognition Memory: Testing Core Properties Under Minimal Assumptions

David Kellen and Karl Christoph Klauer

Online First Publication, June 2, 2014. http://dx.doi.org/10.1037/xlm0000016

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000016

RESEARCH REPORT

Discrete-State and Continuous Models of Recognition Memory: Testing Core Properties Under Minimal Assumptions

David Kellen and Karl Christoph Klauer
Albert-Ludwigs-Universität Freiburg

A classic discussion in the recognition-memory literature concerns the question of whether recognition judgments are better described by continuous or discrete processes. These two hypotheses are instantiated by the signal detection theory model (SDT) and the 2-high-threshold model, respectively. Their comparison has almost invariably relied on receiver operating characteristic data. A new model-comparison approach based on ranking judgments is proposed here. This approach has several advantages: It does not rely on particular distributional assumptions for the models, and it does not require costly experimental manipulations. These features permit the comparison of the models by means of simple paired-comparison tests instead of goodness-of-fit results and complex model-selection methods that are predicated on many auxiliary assumptions. Empirical results from 2 experiments are consistent with a continuous memory process such as the one assumed by SDT.

Keywords: recognition memory, mathematical models, signal detection, thresholds, ranking judgments

Supplemental materials: http://dx.doi.org/10.1037/xlm0000016.supp

One of the most prominent debates in the recognition-memory literature (for a review, see Malmberg, 2008) concerns the distinction between continuous and discrete (threshold) processes and the circumstances under which each is expected to occur. Several measurement models assuming continuous and discrete processes (or a mixture of both) have been proposed and discussed in the literature (e.g., Parks & Yonelinas, 2009; Province & Rouder, 2012; Wixted, 2007; Yonelinas & Parks, 2007). Continuous models such as signal detection theory (SDT; Swets, Tanner, & Birdsall, 1961) assume that a latent familiarity variable underlies recognition-memory judgments. Examples of discrete processes are the detection processes in the two-high-threshold model (2HT; e.g., Bröder & Schütz, 2009) or the recollection process in the dual-process SDT model (e.g., Yonelinas & Parks, 2007).

Attempts to discriminate between discrete and continuous processes have almost invariably relied on the shape of receiver operating characteristics (ROC) functions, which plot pairs of hit and false-alarm rates (recognition rates of studied and nonstudied items, respectively) across different response-bias conditions while assuming that memory discriminability is constant (but see Van Zandt, 2000). Different response-bias conditions are implemented by varying the base rate of old and new items in the test phase or alternatively via response-outcome payoff manipulations. Another way of obtaining ROCs is via confidence-rating responses (see Van Zandt, 2000). Because confidence-rating judgments can be collected in a very easy and efficient manner, confidence-rating ROCs constitute the vast majority of ROC data reported in the literature (for a review, see Yonelinas & Parks, 2007).

This reliance on ROC data to compare models and/or determine the nature of particular memory judgments has been shown to be rather problematic and often produce noninformative outcomes (e.g., Bröder & Schütz, 2009; Krantz, 1969; Luce, 1963; Malmberg, 2002; Province & Rouder, 2012; Rouder, Province, Swagman, & Thiele, 2013). This situation encourages the development of alternative experimental approaches that produce more diagnostic empirical evidence.

In the present article, we propose an alternative experimental method for comparing discrete-state and continuous processes. This method, based on *ranking judgments*, is very simple and has several advantages as it does not require (a) any sort of parameter estimation and model fitting, (b) distributional assumptions, (c) exhaustive experimental manipulations, or (d) complex model-selection methods. Although we focus on the comparison between the 2HT and SDT models, which have received a considerable amount of attention in the recent literature (Bröder & Schütz, 2009; Dube & Rotello, 2012; Dube, Starns, Rotello, & Ratcliff, 2012; Kellen, Klauer, & Bröder, 2013; Province & Rouder, 2012), this method can be adjusted to evaluate other models such as the

David Kellen and Karl Christoph Klauer, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg.

The research reported in this article was supported by Grant Kl 614/32-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer. We thank Colleen Parks for providing raw data. Model fitting routines used in this article can be obtained on request.

Correspondence concerning this article should be addressed to David Kellen, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. E–mail: david.kellen@psychologie.uni-freiburg.de

dual-process model (Yonelinas & Parks, 2007). This article is organized as follows: First, the 2HT and the SDT models are described in the context of ROC data as well as multiple-alternative forced-choice judgments. This is followed by a presentation of the new method, along with two experiments implementing it.

Measurement Models of Recognition Memory: SDT and 2HT Models

The SDT model assumes a continuous memory process, often termed *familiarity*, to describe the individuals' decisions on the basis of memory information. Both old and new items evoke some degree of familiarity, with separate familiarity distributions for old and new items. The ability to discriminate between the two kinds of items is determined by the overlap between the two distributions. According to SDT, an item's familiarity is compared with an established response criterion, denoted by parameter τ. If an item's familiarity is larger than the criterion, the response "old" is given; if the familiarity is lower than the criterion, then the response "new" is given instead. The familiarity distributions are usually assumed to be Gaussian, with mean and standard-deviation parameters $\{\mu_o, \sigma_o\}$ and $\{\mu_n, \sigma_n\}$ for old and new items, respectively, with $\mu_o \geq \mu_n$, $\sigma_o > 0$, and $\sigma_n > 0$. Without loss of generality, μ_n and σ_n are fixed to 0 and 1, respectively. A visual depiction of the SDT model is given in Figure 1. Note that the use of the Gaussian distribution is somewhat arbitrary, and other distributional assumptions could have been used instead. According to the (Gaussian) SDT model, the probabilities of an "old" response for an old or a new item, respectively, are given by

$$P(\text{“old”} \mid \text{old}) = \Phi\left(\frac{\mu_o - \tau}{\sigma_o}\right), \tag{1}$$

$$P(\text{“Old”} \mid \text{new}) = \Phi(-\tau). \tag{2}$$

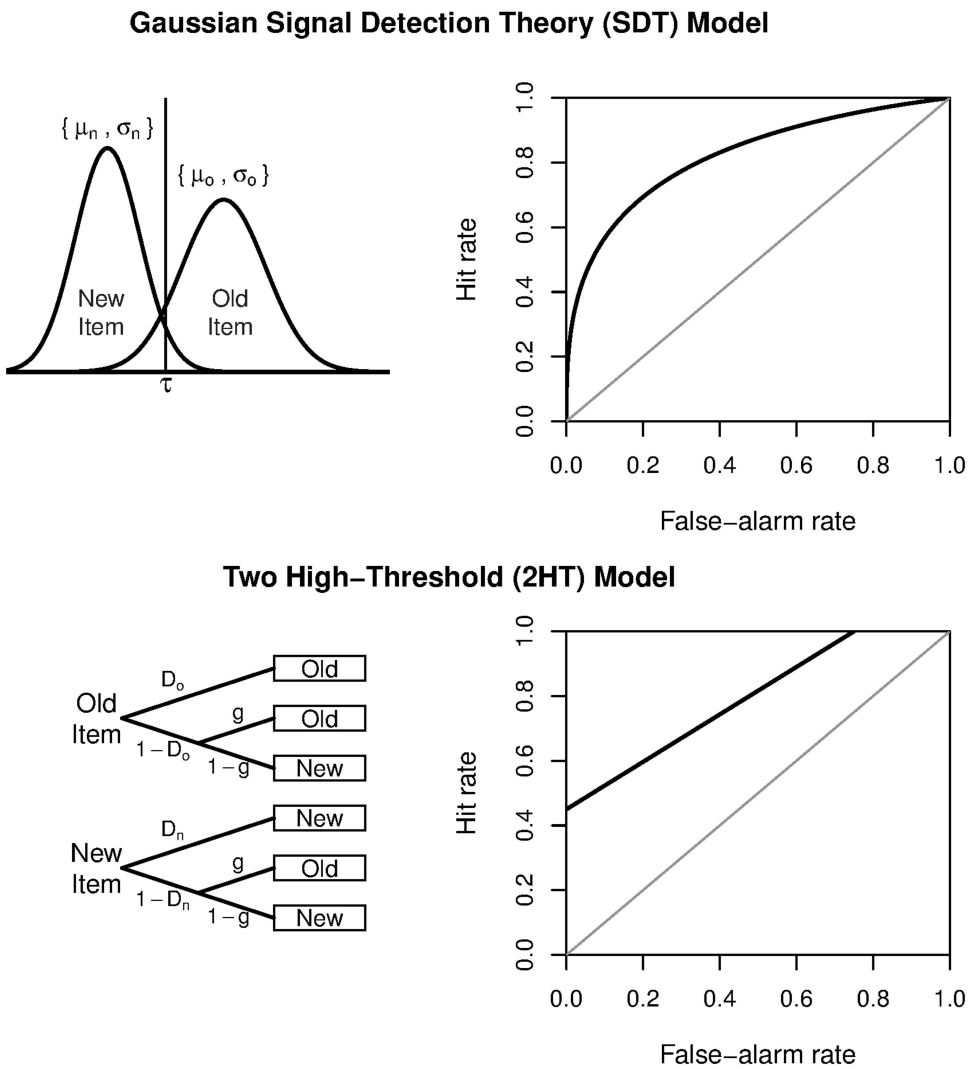

Figure 1. The SDT and 2HT models and their implied binary-response receiver operating characteristics (ROCs). The gray diagonal lines in the graphs on the right depict chance performance.

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The 2HT is a discrete-state model. It assumes that memory judgments are based on "detect" and "guessing" states. According to the 2HT model, the probabilities of an "old" response for an old or a new item, respectively, are given by

$$P(\text{"Old"} \mid \text{old}) = D_o + (1 - D_o)g, \quad (3)$$

$$P(\text{"Old"} \mid \text{new}) = (1 - D_n)g. \quad (4)$$

When presented at test, an old item is detected with probability D_o, invariably leading to an "old" response. If the item's old or new status is not detected, with probability $(1 - D_o)$, then a guessing state is entered: The status of the item is guessed, with response "old" occurring with probability g and response "new" with probability $(1 - g)$. At test, a new item is detected with probability D_n, invariably leading to response "new." Similar to the case of old items, when detection fails with probability $(1 - D_n)$, a guessing process is engaged, with the response "old" occurring with probability g and the response "new" with probability $(1 - g)$. The visual depiction of the 2HT model is also provided in Figure 1.

The 2HT model serves several distinct roles in the literature:

1. It is a model of recognition-memory judgments that makes specific predictions (e.g., Province & Rouder, 2012) that can be contrasted with the predictions by other candidates such as the SDT model.

2. It can represent the recollection processes of dual-process models assuming a mixture of discrete and continuous processes. These dual-process models are often tested in cases where only one process is assumed to underlie performance or one process is expected to be selectively influenced. When it is assumed that (discrete) recollection is the process underlying above-chance performance a model that is equivalent to the 2HT model (or a restricted version of it, where $D_n = 0$) is being postulated (e.g., Parks & Yonelinas, 2009).

3. The 2HT model is also used as tool that provides a rough characterization of recognition-memory data in terms of sensitivity and bias. In the latter case, the model produces accounts that are often in accord with the SDT model (e.g., Bröder, Kellen, Schütz, & Rohrmeier, 2013), and the differences between the two models are downplayed in favor of other attributes such as tractability and generalizability (for a recent discussion, see Batchelder & Alexander, 2013; Dube, Rotello, & Pazzaglia, 2013; Pazzaglia, Dube, & Rotello, 2013).

In the present work, we focus on the first two cases and thus on the 2HT model as a general test bed for discrete-state processes, whether as a stand-alone model or as an instantiation of the recollection process in the dual-process model.

Binary-Response and Confidence-Rating ROC Data

As previously mentioned, these models of recognition-memory judgments have been compared almost invariably by means of ROC data. Although ROCs can be obtained in two different ways—namely, via binary responses or confidence ratings—these two types of ROCs do not have the same diagnostic value for distinguishing between the 2HT and SDT models.

Binary-response ROCs are obtained by manipulating response bias (e.g., the tendency to respond "old") across different test phases, in which different base rates of old and new items (e.g., 90% old items and 10% new items) or outcome payoff matrices (e.g., 10 points per correct "old" response and 5 points per correct "new" response) are implemented. It is assumed that response bias manipulations do not affect memory discriminability, selectively influencing response-criteria/guessing processes (captured by parameters τ and g). The SDT and the 2HT models make different predictions, with the SDT model predicting curvilinear ROCs and the 2HT model linear ROCs (see Figure 1).

Confidence-rating ROCs are simpler to obtain, as they do not require the collection of responses across different response-bias conditions. Instead of plotting the hit and false-alarm rates obtained across conditions, the cumulative proportions obtained across a confidence-rating scale (e.g., from 1 = *sure new* to 8 = *sure old*) for old and new items are used instead. The SDT model accommodates the use of confidence ratings instead of binary responses in a seamless manner, as confidence ratings can be described by simply assuming a set of ordered response criteria along the evidence–familiarity axis. As in the case of binary responses, the SDT model predicts curvilinear confidence-rating ROCs. However, the same does not hold for the 2HT model, which requires the specification of *state-response mapping functions* that determine how the detection and guessing states are mapped onto a confidence-rating scale (e.g., Bröder & Schütz, 2009; Klauer & Kellen, 2010; Malmberg, 2002; Province & Rouder, 2012). Contrary to the case of binary-response ROCs, the 2HT model is able to account for both curvilinear and linear ROCs, with curvilinear ROCs being predicted when detect states are not deterministically mapped onto maximum-confidence responses. This ability compromises any attempt to dismiss discrete-state accounts on the basis of confidence-rating ROC curvilinearity.

Because of the limitations of confidence-rating ROCs, recent work has compared discrete and continuous models—in particular, the SDT and the 2HT models—on the basis of binary-response ROCs (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012; Kellen, Klauer, & Bröder, 2013). However, despite their diagnostic value, individual ROCs cannot be easily obtained via response-bias manipulations as they require multiple study–test phases per individual. Furthermore, to produce diagnostic binary-response ROCs (i.e., ROCs whose shape can be reliably assessed), the response-bias manipulation used by the experimenter needs to produce large differences in response bias. Such an outcome is not easy to accomplish given that individuals are notably reluctant to change their response biases (e.g., Cox & Dobbins, 2011), and it is not entirely clear whether such changes also affect memory discriminability (Van Zandt, 2000). Finally, the noisiness of the data is such that differences in model flexibility still have a nonnegligible weight in model-performance comparisons (see Kellen et al., 2013). Overall, the problems associated with the different types of ROCs indicate the desirability of alternative methods to assess the nature of memory processes.

First- and Second-Choices in Multiple-Alternative Forced-Choice Data

An alternative to ROCs was introduced in the recognition-memory literature by Parks and Yonelinas (2009), who focused on first and second choices in a four-alternative forced-choice (4AFC) task. This approach was originally proposed in the seminal work on SDT by Swets et al. (1961) in which restricted versions of the SDT and 2HT models (with $\sigma_o = 1$ and $D_n = 0$) were compared. In each trial of Parks and Yonelinas' task, one old item and three new items were presented, and individuals were instructed to choose the item they believed to have previously studied. In some of the trials, individuals were additionally requested to provide a second choice among the three remaining alternatives. Across two experiments concerning single-item and pair recognition, Parks and Yonelinas evaluated the accuracy of first choices and the accuracy of second choices conditional on incorrect first choices, particularly their correlation across participants. Parks and Yonelinas argued that according to the SDT model, increases in first choice accuracy should always be accompanied by an increase in (conditional) second-choice accuracy, whereas the dual-process model predicted no correlation in cases where only recollection is expected to drive above-chance performance, as in the case of pair recognition. A simulation study corroborating their predictions concerning the SDT and dual-process model was also reported (see Parks & Yonelinas, 2009, supplemental materials).

The claims of Parks and Yonelinas (2009) were questioned by Kellen and Klauer (2011), who showed that the observed first- and second-choice accuracy could be accounted for by the SDT model as well as by the 2HT model because the data fell inside each models' prediction space. The prediction space of the SDT and 2HT models for first- and second-choice accuracy in a 4AFC task are depicted in Figure 2 along with Parks and Yonelinas's data. As can be seen, the vast majority of data points fall in the regions that can be accounted for by both the SDT and the 2HT models. Kellen and Klauer also estimated parameters (e.g., μ_o and σ_o) on the basis of the observed first- and second-choice accuracies. The individual parameter estimates obtained (as well as their correlations) across data sets were consistent with the ones usually obtained with ROC data, implying that the models account for the data by means of reasonable parameter values. Kellen and Klauer's results show that Parks and Yonelinas's approach is unlikely to provide diagnostic data allowing one to distinguish between continuous and discrete-state models because of the large overlap of 2HT and SDT models in this task. As will be shown in the next section, diagnostic data can be obtained in this framework, if a few appropriate changes are made.

Ranking Judgments

The theoretical work of Iverson and Bamber (1997) articulates the intimate relationship between different forced-choice tasks and ROC data. One of the tasks discussed by Iverson and Bamber is the *ranking task* (e.g., Block & Marschak, 1960; Thurstone, 1931), in which items are ranked according to participants' belief that the items were previously studied (Rank 1 being attributed to the item judged as the most likely to be old). This task is formally equivalent to Parks and Yonelinas' (2009) forced-choice task (i.e., model equations are exactly the same), with first- and (unconditional) second-choice accuracy corresponding to the probabilities of Rank 1 and Rank 2 responses to old items, respectively. One practical advantage of the ranking task is that Rank 2 responses to old items can be potentially observed in any trial, whereas in Parks and Yonelinas's task, second-choice accuracy could only be potentially observed in a subset of trials.

We now characterize the SDT and 2HT models in the context of the ranking task (which subsumes Parks and Yonelinas's, 2009, task). We later show that the data originating from an extension of this task provide a simple nonparametric test to discriminate between these two models. First, consider a *K*-alternative ranking

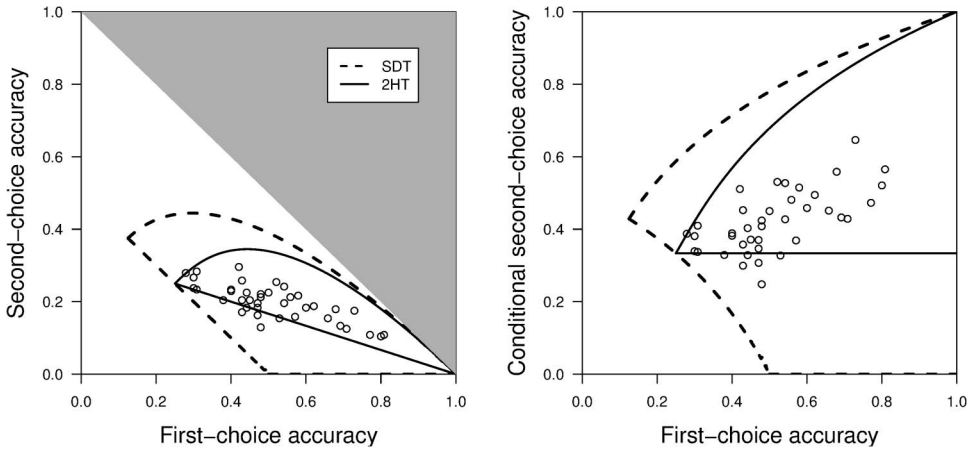

Figure 2. Predictions of the signal detection theory (SDT) and two-high-threshold (2HT) models for Parks and Yonelinas's (2009) four-alternative forced-choice task. In both panels, the lines delimit the regions predicted by the models. The left panel plots first-choice accuracy against second-choice accuracy. The gray area represents the inadmissible region (as first- and second-choice accuracy cannot sum to more than 1). The right panel plots first-choice accuracy against conditional second-choice accuracy. Note that chance performance in conditional second-choice accuracy is 1/3 in this task. The circles in both panels correspond to the individual data from the item memory condition in the study of Parks and Yonelinas.

task, in which participants are presented with sets of K items, one previously studied and $K - 1$ nonstudied, and their task is to rank the items according to the likelihood with which they are believed to have been previously studied. Denote the probability of the old item being assigned rank i among K alternatives by π_i. Furthermore, let $c_2 = \frac{\pi_2}{1-\pi_1}$ denote the *conditional* probability of the old item being assigned Rank 2 given that it was not assigned Rank 1. We show that the continuous SDT model predicts that c_2 increases as a function of the strength of old items in memory, whereas the discrete 2HT model predicts c_2 to be constant as a function of item strength.

For the case of SDT, let f_μ and F_μ denote the density and cumulative distribution functions of the old-item familiarity distribution (parameterized by a parameter μ), with f and F corresponding to the respective functions for the new-item familiarity distribution. The probability π_i of the old item being assigned rank i among K alternatives is given by

$$\pi_i = \binom{K-1}{i-1} \int f_\mu(x) F^{K-i}(x)(1 - F(x))^{i-1} dx. \quad (5)$$

According to the SDT model, π_i simply corresponds to the probability that the old item is the ith most familiar item among the K alternatives.

Let us now consider the predictions of the SDT model concerning c_2 and memory strength. The following theorem describes the conditions under which c_2 increases with increasing item strength μ. These conditions refer to the cumulative distribution function F_μ of the old-item distribution. Specifically, a function H_μ is derived from the distribution function, and the theorem states that a monotonicity property of H_μ, where it is present, entails that c_2 is monotonically increasing as a function of μ. For each familiarity value z, define $H_\mu(z)$ as $\frac{\partial}{\partial \mu} F_\mu(z) \times F_\mu(z)^{-1}$. In the supplemental materials, the following theorem is proved:

Theorem. If $H_\mu(z)$ is monotonically increasing in z for all μ, then c_2 is monotonically increasing in μ.

In the supplemental materials, we show that the monotonicity condition involving H_μ is satisfied for the normal (i.e., Gaussian) familiarity distribution usually postulated by SDT models. The theorem now yields that the SDT model predicts c_2 to increase with increasing μ. But the theorem is, in fact, much more general: In the supplemental materials, we also show that this same prediction is shared by many alternative continuous models defined by familiarity distributions other than the normal distributions, such as the exponential, ex-Gaussian, and gamma distributions, among others. In other words, the monotonicity of c_2 is a shared property of most reasonably regular continuous models that have been considered, and it is not tied to the normality assumptions of the classical SDT model.

We now turn to the 2HT model: Again, let D_o be the probability of a studied item being successfully detected. Now, consider a function $\xi(i)$ denoting the probability (conditional on the absence of old-item detection) of the studied item being attributed rank i among K alternatives on the basis of guessing among nonexcluded alternatives. Note that $\xi(i)$ comprises cases in which new items are detected as new (e.g., with probability D_n) and thus excluded from the set of alternatives. Also, note that we do not need the assumption that the probabilities of detecting new items are independent. The stochastic relation of distractor detection is immaterial for the present purposes and is therefore left unspecified; that is, we just refer to $\xi(i)$. According to the 2HT model, π_i is given by

$$\pi_i = \begin{cases} D_o + (1 - D_o)\xi(i), & \text{if } i = 1 \\ (1 - D_o)\xi(i), & \text{if } 2 \le i \le K \end{cases} \quad (6)$$

The 2HT model states that π_1 corresponds to the probability that the old item is detected as old, plus the probability of the old item being attributed Rank 1 via guessing among nonexcluded alternatives. The probability of the old item being attributed any rank larger than 1 implies the failure of old-item detection and is solely dependent on the guessing among nonexcluded alternatives that takes place in the absence of old-item detection.

Like the SDT model, the 2HT model also makes predictions regarding the relationship between c_2 and memory strength.

Proposition. Let D_o^w and D_o^s represent the probabilities of detecting two different types of studied items (e.g., weak and strong items). Also, assume that there is a common set of distractors that is used for both types of studied items with $\xi(i)$ being a function of guessing and distractor detection. Then the equality $c_2^w = c_2^s$ holds for $0 \le D_o^w, D_o^s \le 1$.

Proof. It is easy to see that $c_2 = \xi(2)/\sum_{i=2}^{K} \xi(i)$, which means that c_2 is independent of the value taken by D_o.

The SDT model and the 2HT model establish distinct hypotheses (Hs) regarding the value of c_2 for weak and strong items. The SDT model (H_{SDT}) expects c_2 to be larger for strong items than for weak items, whereas the 2HT model (H_{2HT}) expects c_2 to be the same for both types of items. By adjusting and extending the experimental approach proposed by Parks and Yonelinas (2009), one is able to test the 2HT and SDT models under rather minimalistic assumptions, as no distributional assumptions have to be made for the SDT model and no assumptions regarding distractor detection have to be made for the 2HT model. The predictions of the two models are reduced to a simple order-restricted hypothesis test, precluding the need of model fitting or the use of complex model-selection methods. In the next section, we report two new experiments that evaluate c_2 values for weak and strong items.

Experiments 1 and 2

The following two experiments tested the above-stated hypotheses concerning c_2 by means of a study-repetition manipulation. Given their similarity, the two experiments are reported together. In these experiments, participants studied a single word list composed of weak (words presented once) and strong items (words presented three times) intermixed. In Experiment 1, the test phase consisted of a four-alternative ranking task. In Experiment 2, a three-alternative raking task was used instead, along with a payoff scheme contingent on the ranking of the old item. There were also differences in the repetition scheme used in the study phase: In Experiment 1, the presentation order was random, but constrained; each third of the study list comprised a single presentation of the strong items randomly intermixed with weak items. In Experiment 2, the presentation of words was completely random. The differences in Experiment 2 were introduced for several reasons: The repetition scheme was relaxed to be equivalent to the one used in Province and Rouder's (2012) experiments (Jeffrey N. Rouder,

personal communication, May 2, 2013). The number of alternatives in the test phase was reduced to make the task less demanding for the participants while increasing the number of overall test trials. The payoff scheme was introduced to further encourage participants to assign ranks to the different alternatives in a way that closely reflects their memory judgments.

Method

Participants. In Experiment 1, 22 individuals (15 university students; M_{age} = 25.23 years, SD = 5.66, range: 19–41 years) served as participants; in Experiment 2, 23 individuals (17 university students; M_{age} = 25.30 years, SD = 4.94, range: 21–40 years). In exchange for their participation, participants in Experiment 1 received €3.5. Participants in Experiment 2 received between €3.5 and €5, depending on their performance. Each participant was tested individually in sessions of approximately 35 min.

Design and procedure. During the study phase, words were presented in a black Arial bold font (letters were all capitalized, 1.7 cm high) against a grey background in the center of a 50.93 × 28.63-cm LCD screen. Participants sat at approximately 60 cm away from the screen. In Experiment 1, 150 words (75 weak and 75 strong) were presented for 600 ms each (100-ms interstimulus interval). In Experiment 2, 270 words (135 weak and 135 strong) were presented for 1,200 ms each (100-ms interstimulus interval). Weak and strong words were presented once and thrice, respectively. For each participant, a randomly generated word list was presented in the study phase. The repetition scheme for strong words differed across experiments, as already described. The test phase started immediately after all items were presented. The presentation of old items in the test phase was randomized anew.

In each trial of the test phase, participants were shown a set of items, one being old and the remaining new. The new items presented along with each old item were randomly selected for each participant. In Experiment 1, the four alternatives were placed in a 2 × 2 arrangement in the center of the screen. In Experiment 2, the three alternatives were placed side by side in the middle of the screen. The position of the old item was randomly determined. Participants were informed of this composition and were requested to rank the items according to their belief that they were previously studied (with Rank 1 assigned to the item the believed to be the one most likely to have been previously studied). As in Kellen et al. (2012), ranks were assigned to the words by clicking on them with the mouse (each word was embedded in a rectangle delimiting the clickable area). The rank assignment was only registered in the screen when the mouse button was released. The first word that was clicked received Rank 1, the second word clicked received Rank 2, and so forth. A number corresponding to the current rank of the item appeared above the word. Participants could also deselect items, which removed their rank and updated the remaining ranks accordingly. Participants could only proceed to the next trial after clicking a button at the bottom of the screen confirming the attributed ranks. Another button at the bottom of the screen deleted all assigned ranks. Figure 3 illustrates test trials in Experiments 1 and 2. To familiarize participants with the task, there was a tryout trial accompanying the test-phase instructions, which participants could repeat as often as they wished before starting with the test trials. In Experiment 2, participants were informed that their final payment would be partly based on their performance. They would receive 1 point if the old item was assigned Rank 1 and lose 2 and 3 points if it was assigned Ranks 2 and 3, respectively. The total number of points determined the final payment that could vary between a fixed minimum of €3.5 and €5. No feedback was provided during the test phase. After finishing the test phase, participants were thanked and debriefed.

Materials. Words were sampled from a selection of words from Lahl, Göritz, Peitrowsky, and Rosenberg (2009), ranging from four to eight letters in length. According to the ratings obtained by Lahl et al., the words were all of medium valence (ranging from 3.50 to 6.50 on an 11-point scale) and low in arousal (ranging from 0.50 to 4.50 on an 11-point scale). Furthermore, all words were of approximately equal word frequency, as indicated by the log frequency ratings obtained for each word via WordGen (ranging from 0.30 to 2.90; Duyck, Desmet, Verbeke, & Brysbaert, 2004).

Results

In Experiment 1, the average values (standard deviations in parentheses) of π_1, π_2, π_3, and π_4 for weak items were, in order, 0.38 (0.08), 0.23 (0.05), 0.21 (0.05), and 0.18 (0.05), whereas for strong items they were 0.55 (0.15), 0.19 (0.07), 0.14 (0.06), and 0.13 (0.05). In Experiment 2, the average values of π_1, π_2, and π_3 for weak items were, in order, 0.62 (0.11), 0.21 (0.06), and 0.17 (0.07), whereas for strong items they were 0.80 (0.11), 0.12 (0.07),

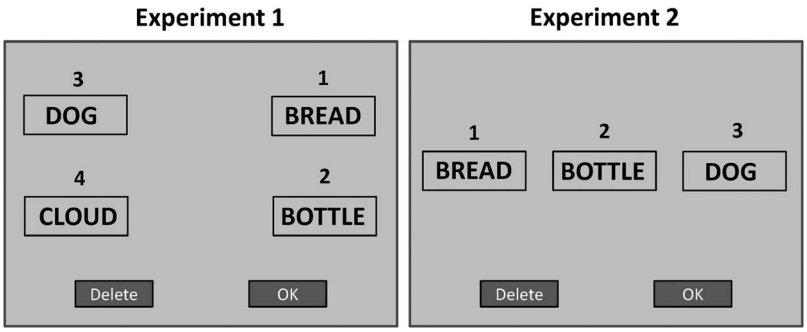

Figure 3. Illustration of ranking trials in Experiments 1 and 2. The numbers on top of the words correspond to the assigned ranks

and 0.08 (0.05).[1] The differences in accuracy (π_1) between weak and strong items was found to be statistically significant in both experiments with a Wilcoxon test (both $ps < .001$). Regarding the critical prediction concerning the individual values of c_2 for weak and strong items, the average c_2^w values—0.37 (0.09) and 0.55 (0.09) in Experiments 1 and 2, respectively—were smaller than c_2^s—0.43 (0.09) and .63 (0.10). This difference, depicted in Figure 4, was found to be statistically significant with a Wilcoxon test in each of the two experiments (smallest $V = 65.5$, largest $p = .02$, one-tailed). This rejection of the null hypothesis suggests that the results are in accord with the predictions made by the SDT model.

The use of a paired-comparison test like the Wilcoxon test on c_2 values is adequate but somewhat nonoptimal because the uncertainty associated to the obtained c_2 values varies within and across participants. For example, if π_1 is large then the number of trials used to estimate π_2 is bound to be rather small, leading to a larger uncertainty regarding c_2. This means that the uncertainty regarding c_2 is expected to be higher for strong than for weak items. Ignoring these differences in uncertainty and taking c_2 values at face value can lead to losses in statistical power and increase the vulnerability to outliers. Another aspect concerns the fact that traditional null-hypothesis testing approaches can overstate the evidence against the null hypothesis (Wagenmakers, 2007). To overcome these issues, we also analyzed the data analyzed with a hierarchical Bayesian model that quantified the relative evidence for the two hypotheses on the basis of individuals' *response frequencies* from both experiments (Rouder & Lu, 2005; Rouder, Morey, & Pratte, in press; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). An effect size parameter δ captured the differences between c_2^w and c_2^s across individuals and experiments such that $H_{SDT} : \delta > 0$ and $H_{2HT} : \delta = 0$. The relative evidence for each hypothesis is quantified by a *Bayes factor*, which corresponds to the ratio $P(\text{data} \mid H_{SDT}) / P(\text{data} \mid H_{2HT})$. This ratio is composed of the marginal likelihood of the data conditional on each model.

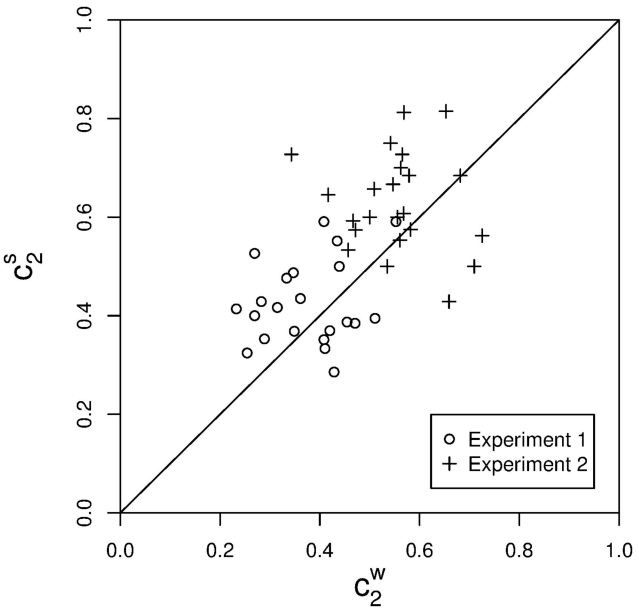

Figure 4. Observed c_2^w and c_2^s in Experiments 1 and 2.

Under a unit information prior on δ (Kass & Wasserman, 1995), the estimated Bayes factor was 34.46, which indicates that the data from both experiments are over 30 times more likely under H_{SDT} than H_{2HT}. A Bayes factor of this magnitude is considered to represent strong evidence in favor of H_{SDT} (Wagenmakers, 2007). The 95% credibility interval of the posterior δ distribution was [0.32, 2.61]. Similar to the null hypothesis testing approach, the hierarchical Bayesian modeling indicates that the c_2^s were reliably higher than the c_2^w, as predicted by the SDT model. A complete description of the hierarchical Bayesian analysis and its implementation can be found in the supplemental materials.

General Discussion

In the recognition-memory literature, it is common to test rather complex models under strong parametric assumptions using ROC data. More recent modeling approaches are beginning to take into account other variables such as response times (e.g., Ratcliff & Starns, 2009). Still, despite the level of sophistication of many of these models and comparison methods, the discussion of which model provides the best characterization is still not settled. The present work follows a rather different approach and abandons the use of ROC data altogether: The core properties of the models were used to derive simple predictions for ranking judgments. These predictions can be tested without distributional assumptions for the models, without implementing costly experimental manipulations and without the need to use more sophisticated forms of hypothesis testing than implemented by a Wilcoxon test, let alone complex model-selection techniques. The results obtained in the two experiments indicate that a continuous process underlies recognition-memory judgments, as estimated c_2 values were larger for strong items than for weak items.

Relating Ranking Results With Evidence for Discrete States

The present results are apparently at odds with recent studies providing evidence for the 2HT model, in particular with Province and Rouder's (2012) work showing that individuals' confidence-rating responses are consistent with the 2HT's prediction of *conditional independence*. This prediction states that behavioral outputs (confidence ratings and response times) are only a function of the discrete memory states generating them but not of the probability of those states being reached. Province and Rouder's results report recognition-memory data in line with conditional independence, both at the level of categorical responses and response times, results that have been subsequently replicated and extended by Kellen, Singmann, Vogt, and Klauer (in press).

[1] It is usually assumed that responses in forced choice or ranking tasks do not manifest any form of spatial bias, an approach that has been recently criticized in the literature (see DeCarlo, 2012). To evaluate the presence of any unaccounted spatial bias, we checked whether π_1 was affected by the spatial position of the old item. Median π_1 estimates were found to be very similar across spatial positions (.44, .48, .45, and .47 for the four positions in Experiment 1 and .72, .71, and .73 for the three positions in Experiment 2), suggesting that spatial bias is absent or negligible in magnitude. Not surprisingly, no significant differences were found across participants in both Experiment 1, $\chi^2(66) = 74.39$, $p = .22$, and Experiment 2, $\chi^2(46) = 48.20$, $p = .38$.

One way of integrating the present results with Province and Rouder's is to follow Rouder et al.'s (2013) theoretical framework and assume that the uncertainty in the mapping of detect states entails that a (small) proportion of detected items can be mapped on any possible response (e.g., a detected old item can be judged "new"; see also Krantz, 1969; Luce, 1963). In the case of the ranking task, this would mean that a detected old item is not invariably assigned Rank 1, making the 2HT model able to account for the increases in c_2.[2] Although we cannot discard this account on the basis of the present data, we find that this relaxation of 2HT's assumptions is implausible in the case of ranking judgments. In the case of ROC data, it is often plausible that participants' attempts to comply with task requirements justifies this assumption relaxation. For example, the response-bias manipulation used by Dube and Rotello (2012) directly instructed participants to give certain responses at extreme rates (e.g., respond "new" 90% of the time) even though the percentage of old and new items in the test phase was always 50% each, suggesting that some items detected as old were nevertheless mapped on "new" responses. In the case of the ranking task, however, it is not clear what sort of task requirements (explicit or tacit) would lead one to assign anything other than Rank 1 to detected old items, especially when there is an explicit payoff scheme encouraging accurate responding (as in Experiment 2).

Another way to integrate both results is to assume that when individuals provide confidence-rating judgments, they partition the evidence/familiarity scale in three regions: two regions with extreme values (very low and very high familiarity), where the status of the items is considered to have been "ascertained" as old or new, and another region (between the other two) where the status of the item is considered uncertain (for a similar proposal, see Malmberg, 2008, pp. 363–364). These regions would then be mapped onto the confidence-rating scale according to a probability distribution (i.e., a state-response mapping function). According to this account, Province and Rouder's (2012) results reflect a *discrete-state mediation* in which a continuous familiarity process is made discrete to engage in the recognition-memory task in an efficient manner (this also consistent with the notion of *task thresholds* introduced by Rouder & Morey, 2009, as well as with the notion of *efficiency* discussed by Malmberg, 2008). In particular, participants only need to maintain two task thresholds in working memory, marking off the regions with extreme values rather than separate response criteria for each level of confidence on the confidence-rating scale. In contrast, discrete-state mediation is not expected to play a role in the ranking judgments, as they only require the comparison of the alternatives' familiarity values, suggesting that ranking judgments provide a more direct evaluation of the memory processes. It is important to note that ranking judgments are not particularly resource demanding, as they do not require the simultaneous comparison of all alternatives and can be obtained from simple paired comparisons (e.g., Block & Marschak, 1960). It should also be noted that the simplification or discretization of continuous (external) information is well known in the judgment and decision-making literature (e.g., Brandstatter, Gigerenzer, & Hertwig, 2006; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011), suggesting that the occurrence of a similar phenomenon in the case of latent mnemonic information is not implausible.

Implications for the Multinomial Processing Tree (MPT) Model Class

The 2HT model is a prominent member of the MPT model class (Batchelder & Riefer, 1999). This situation has led some researchers to associate the successes and failures of the 2HT model to the MPT model class as a whole (see Batchelder & Alexander, 2013; Dube et al., 2013; Pazzaglia et al., 2013). The fact that the 2HT model fails in the present study bears no weight on the suitability of the MPT model class, as this class includes models that go beyond the discrete states assumed by the 2HT model. One such example is the model proposed by Chechile, Sloboda, and Chamberland (2012), which assumes a mixture of states where memory representation of items in memory can be based on different forms of explicit, implicit, and fractional storage. An MPT model like the one proposed by Chechile et al. should have no trouble in accounting for the present results.

Tests on the Dual-Process Model

Although the present work focused on item memory, the ranking approach can be used to test other kinds of memory judgments such as source discrimination or associative recognition. In such cases, the predictions of the 2HT model (or restricted versions of it, with $D_n = D_o$ or $D_n = 0$) correspond to the predictions made by the dual-process model in circumstances where recollection is assumed to be the only process underlying above-chance performance (e.g., in associative recognition and source-discrimination; see Parks & Yonelinas 2009).

Additionally, the study-repetition manipulation can be replaced by experimental manipulations that are expected to selectively influence recollection (e.g., context reinstatement; see Koen, Aly, Wang, & Yonelinas, 2013): Simply note that according to the dual-process model, $\pi_1 = R + (1 - R)\Psi(1)$ and $\pi_2 = (1 - R)\Psi(2)$, with R being the probability of recollection and $\Psi(i)$ the probability of the old item being the ith most familiar item among the k alternatives, with $\Psi(i)$ not being a function of R. It is trivial to see that the proposition stated above for the 2HT model also holds for the dual-process model (simply replace D_o and ξ with R and Ψ, respectively). Note that no parametric assumptions (e.g., assume Gaussian distributions) whatsoever have to be made for $\Psi(i)$. Thus, if an experimental manipulation is expected to only affect recollection, then c_2 should not be affected by that manipulation. The use of ranking judgments sidesteps the problems associated with the estimation of recollection on the basis of confidence-rating ROCs. In particular, it does not lean on the probably auxiliary assumption that recollection is deterministically and invariably mapped onto maximum-confidence responses. Any small violation of this assumption is bound to produce curvilinear ROCs and distort the measurement of recollection and familiarity.[3]

[2] We thank Jeff Rouder for pointing out this possibility.

[3] The equal-variance Gaussian assumption associated to the familiarity component is also likely to produce distortions. This assumption derives from the claim that the familiarity process alone produces symmetrical curvilinear ROCs (Yonelinas & Parks, 2007). However, there are infinitely many kinds of symmetrical curvilinear ROCs produced by distinct parametric forms (Killeen & Taylor, 2004) that can potentially lead to different results if used instead of the Gaussian assumption.

Final Remarks

The excessive reliance on ROC data has led to a neglect of alternative approaches for the study of memory. The present work shows that there is potential in alternative sources of evidence such as ranking judgments. Because alternative approaches such as the one considered here have rarely been explored, it is likely that further work will reveal new additional tests focusing on different models and judgments. Given that simple predictions can be derived under fairly weak assumptions, we believe that these tests on critical properties of models carry greater weight than analyses comparing complex models predicated on strong and somewhat arbitrary distributional assumptions by means of complex statistical techniques and model-selection procedures predicated on many additional auxiliary assumptions (for a similar view, see Birnbaum, 2011). Additionally, ranking judgments can also be useful when considered in combination with other memory judgments (e.g., confidence ratings), as they provide ways of studying processes that so far have been shown to be particularly challenging to measure and test (e.g., response-criteria variability in SDT; see Kellen, Klauer, & Singmann, 2012).

References

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

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

Birnbaum, M. H. (2011, November 15). Testing theories of risky decision making via critical tests. *Frontiers in Psychology, 2,* Article 315. doi:10.3389/fpsyg.2011.00315

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of response. In I. Olkin, S. Ghurye, W. Hoeffding, M. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97–132). Stanford, CA: Stanford University Press.

Brandstatter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113,* 409–432. doi:10.1037/0033-295X.113.2.409

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. doi: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. doi:10.1037/a0015279

Chechile, R. A., Sloboda, L. N., & Chamberland, J. R. (2012). Obtaining separate measures for implicit and explicit memory. *Journal of Mathematical Psychology, 56,* 35–53. doi:10.1016/j.jmp.2012.01.002

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940. doi:10.3758/s13421-011-0090-3

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology, 56,* 196–207. doi:10.1016/j.jmp.2012.02.004

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. doi:10.1037/a0024957

Dube, C., Rotello, C., & Pazzaglia, A. (2013). The statistical accuracy and theoretical status of discrete-state MPT models: Reply to Batchelder and Alexander (2013). *Psychological Bulletin, 139,* 1213–1220. doi:10.1037/a0034453

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

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, German, English, and French. *Behavior Research Methods, Instruments, & Computers, 36,* 488–499. doi:10.3758/BF03195595

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.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90,* 928–934. doi:10.1080/01621459.1995.10476592

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

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

Kellen, D., 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. doi:10.1037/a0027727

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

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology, 48,* 432–434. doi:10.1016/j.jmp.2004.08.005

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. doi:10.3758/PBR.17.4.465

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

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

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

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

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387. doi: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. doi:10.1016/j.cogpsych.2008.02.004

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,* Article 147. doi:10.3389/fpsyg.2011.00147

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences, USA, 106,* 11515–11519. doi:10.1073/pnas.0905505106

Pazzaglia, A., Dube, C., & Rotello, C. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implica-

tions for recognition and beyond. *Psychological Bulletin, 139,* 1173–1203. doi:10.1037/a0033044

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

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

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. doi:10.3758/BF03196750

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review, 116,* 655–660. doi: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*. London, United Kingdom: Cambridge University Press.

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

Swets, J., Tanner, J. W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340. doi:10.1037/0033-295X.68.5.301

Thurstone, L. L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology, 14,* 187–201. doi:10.1037/h0070025

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

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60,* 158–189. doi:10.1016/j.cogpsych.2009.12.001

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. doi: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. doi:10.1037/0033-2909.133.5.800

Received October 2, 2013
Revision received February 12, 2014
Accepted April 4, 2014 ■