Familiarity, recollection, and receiver-operating characteristic (ROC) curves in recognition memory

James F. Juola¹ · Alexandra Caballero-Sanz¹ · Adrián R. Muñoz-García¹ · Juan Botella¹ · Manuel Suero¹

Published online: 4 April 2019
© The Psychonomic Society, Inc. 2019

Abstract

The Atkinson-Shiffrin theory describes and explains some of the processes involved in storing and retrieving information in human memory. Here we examine predictions of related models for search and decision processes in recognizing information in long-term memory. In some models, recognition is presumably based on a test item's familiarity judgment, and subsequent decisions follow from the sensitivity and decision parameters of signal detection theory. Other models dispense with the continuous notion of familiarity and base recognition on discrete internal states such as relative certainty that an item has or has not been previously studied, with an intermediate state of uncertainty that produces guesses. Still others are hybrid models with two criteria located along a familiarity continuum defining areas for rapid decisions based on high or low familiarities. For intermediate familiarity values, the decision can be delayed pending the results of search for, and occasional recollection of, relevant episodic information. Here we present the results from a study of human recognition memory for lists of words using both response time and error data to construct receiver-operating characteristic (ROC) curves derived from three standard methods based on the same data set. Models are evaluated against, and parameters estimated from, group as well as individual subjects' behavior. We report substantially different ROC curves when they are based on variations in target-word frequency, confidence judgments, and response latencies. The results indicate that individual versus group data must be used with caution in determining the appropriate theoretical interpretation of recognition memory performance.

Keywords Recognition memory · Familiarity vs. recollection · Memory models

Introduction

In the late 1960s the Atkinson-Shiffrin (A-S) theory was developed to summarize and explain a large body of data that had been collected since the time of Ebbinghaus (1885/1964) from experiments in which people studied, learned, and eventually tried to remember lists of verbal materials (Atkinson & Shiffrin, 1968, 1971). Verbal learning was important enough to inspire its own journals, conferences, and the beginnings of cognitive science as we know it today. The A-S theory has become known as the “modal model” in cognitive psychology, as most practicing theorists and experimentalists acknowledge it as part of our collective unconscious, driving our

theories, leading to interpretations of our data, and giving us the jargon with which we communicate with our peers.

The theory was more general than the experiments on which it was based, of course, as it had implications for the roles of more primitive processes, such as sensation, perception, and the fleeting storage of iconic and echoic traces of their respective visual and auditory stimuli. It also anticipated some modern ideas of working memory, in that short-term memory was posited to have storage capacities as well as being subject to a host of control processes that determine its contents and functions. In addition, the effectively infinite and permanent long-term store is continuously consulted, reorganized, and updated as a result of control processes such as storage and retrieval. The distinction between memory structures (sensory, short-term, and long-term stores) and control processes (attention, rehearsal, storage, retrieval, etc.) was crucial in giving the theory the flexibility and adaptability that has determined its continuing usefulness. In sum, the model was specific enough to account for the classic results of study time, list length, retention interval, serial position effects, and

✉ James F. Juola
juolas@ku.edu

¹ Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain

interference effects that had dominated 100 years of memory research. Yet it was general enough to form the basis of contemporary theories of attention, consciousness, language use, and artificial intelligence. Here we wish to trace its influence in one specific area of psychological research, namely recognition memory for verbal materials.

In the early 1970s, we developed an extension of the A-S theory that was specific to search for and retrieval of information in long-term memory (Atkinson & Juola, 1973, 1974; Juola, Fischler, Wood & Atkinson, 1971). The Atkinson-Juola (A-J) model applied some of the analytic methods of signal detection theory (SDT) to the results of recognition memory experiments, as earlier work had indicated the power and flexibility of these methods (e.g., Banks, 1970; Egan, 1958; Kintsch, 1967; Parks, 1966). In these applications of SDT, memory strength is represented along a continuum in which Gaussian distributions associated with old (previously studied information) and new (not previously studied items) are separated but overlapping to some extent (see Fig. 1). When confronted with a word in a memory test, the participant presumably compares the test item's familiarity, or subjective memory strength, to a criterion along this continuum to separate "old" (or "yes") and "new" (or "no") responses. The theory was generally quite successful in separating indices of memory strength by the sensitivity parameter, d' , and response bias by the location of the decision criterion, c . Further, d' has been shown to be influenced by variables, such as retention interval and study time, that should theoretically affect memory strength, and c is affected by variables, such as relative target frequencies in test blocks as well as payoffs for correct responses and costs for errors, that should reasonably affect response biases (e.g., Swets, Dawes, & Monahan, 2002; Wickens, 2002).

Recently, there have been several alternative accounts of recognition memory that have challenged the basic assumptions of the A-J model and the application of SDT in general.

For example, Bröder and Schütz (2009) have claimed that the signal-detection basis of the theory is inappropriate, and that recognition memory results could be fit more adequately by a threshold theory, modified from original proposals (e.g., Egan, 1958; Fechner, 1860/1966; Snodgrass & Corwin, 1988). In their model, an underlying continuum of memory strength, if it exists, is mapped onto a set of discrete states, such as one below a threshold that is achieved only by new items, another above a second threshold that is achieved only by old items, and an intermediate state of uncertainty in which participants are forced to guess (see Fig. 2, describing the two high-threshold model, 2HTM).

The result has been a classic, and ongoing debate in the literature about whether recognition performance should best be described as based on continuous memory strength or on discrete memory states (e.g., Dube, Starns, Rotello, & Ratcliff, 2012; Malejka & Bröder, 2019). It might well be the case that both classes of models fit the aggregate data for groups of subjects about equally well, whereas individual subjects might be better characterized as using different strategies or control processes depending on their abilities, strategies, or task demands (Malmberg, 2008; McAdoo, Key, & Gronlund, 2019). This debate has often been couched in terms of differential predictions for the shape of the receiver-operating characteristic (ROC) function that describes the relation between the proportions of hits (responding "yes" to a list item) and false alarms (responding "yes" to a non-list distractor) across several conditions in an experiment. Threshold theories generally predict that the ROC should be linear, rather than the smooth curve predicted by SDT, at least when response bias is manipulated experimentally. Bröder and Schütz (2009) reported the results from several experiments while claiming that the linear ROC functions predicted by a threshold theory give a better fit to the data in many cases than the curved ROC functions predicted by SDT. Such claims have not gone unchallenged in the literature (e.g., Dube & Rotello, 2012), and we

Fig. 1 Signal detection theory (SDT). When applied to recognition memory, the abscissa relates to a hypothetical dimension of memory strength, or familiarity, on which non-studied items (distractors or new items) are represented to the left with a Gaussian distribution having a mean of μ_1 , and studied items (targets or old items) to the right have a mean of μ_2 ; μ_1 is arbitrarily set to 0.0., and the standard deviation, σ_1 is set to 1.0. Then d' , the difference between the means of the two distributions = μ_2 . In the equal variance model, $\sigma_1 = \sigma_2$, but other models

typically allow the target distribution to have a larger standard deviation. Along the familiarity dimension, the subjects are free to place some criterion, c , that separates "Old" (to the right of the criterion) from "New" judgments (to the left), and its location determines the proportions of hits [$P(\text{"Old"}|\text{targets})$] and false alarms [$P(\text{"Old"}|\text{distractors})$]. These proportions are used to plot the receiver-operating characteristic function (ROC curve) showing the relationship between the proportions of hits and false alarms as the criterion moves up or down

Fig. 2 A two-high threshold model (2HTM) for recognition memory. Old items are detected as being targets with probability D_o , and new items are detected as being distractors with probability D_n . If detection fails, subjects are left in an uncertain, non-detect state between the two thresholds

in which they guess “old” with probability g and “new” with probability $(1 - g)$. Detect states automatically lead to the corresponding response

conducted our own analysis of the data from Experiment 3 of Bröder and Schütz, in which the relevant manipulation of relative proportions of targets and distractors was made within subjects. In their model comparisons, they used a maximum-likelihood method to test the relative goodness-of-fit and found that 23 of 40 subjects’ data were fit better by the 2HTM predictions than by SDT. We examined their individual subjects’ data in plots of the proportions of hits against false alarms, and again when the data had been transformed to z -scores. We then found the best-fitting functions by linear regression and compared whether such linear fits were better for the raw proportions (as predicted by threshold theory) or for the z -transformed data (as predicted by SDT). In earlier comparisons of maximum-likelihood versus regression methods, Gronlund and Elam (1994) and Van Zandt (2000) reported very similar results in estimating parameters for recognition memory ROC curves when using both procedures. However, in our case, and in slight contrast to the modeling efforts of Bröder and Schütz, we found that when using linear regression, 24 out of 40 subjects’ data were better fit by SDT, so we view the results as being inconclusive. We will include the results of both linear regression and maximum likelihood procedures in the modeling efforts reported later in the current paper.

In the A-J model developed from the A-S theory, we supplemented the initial familiarity judgment with a secondary memory search process that could be used to determine the best response, especially in those cases in which familiarity was ambiguous in the region of substantial overlap in the underlying target and distractor strength distributions. That is, in a way similar to the 2HTM, we replaced the single criterion c in the original SDT model with two criteria, c_0 and c_1 , to define a central region of relative uncertainty. In

order to test some models of memory search processes, we insured that our subjects could rely on their memories for previously studied information by having them memorize lists of target words and demonstrate perfect serial recall before testing began. When test words were then presented, the individual responses were highly accurate, and response times (RTs) were the main data used for subsequent theoretical analysis (Atkinson & Juola, 1973, 1974; Juola et al., 1971).

The results indicated that, as in the Sternberg task (Sternberg, 1966, 2016), RTs increased linearly with the number of items in the memorized list, but the slopes of the best-fitting linear functions were much smaller than those reported by Sternberg. We argued that recognition of words in a well-learned list would seldom require some kind of memory-scanning process, but more often fast decisions could be made based on the familiarity, or memory strength, of the individual words alone. That is, an initial familiarity judgment could be a reliable indicator of the word’s likely list membership, and words with high or low familiarities were very likely to be list members or distractors, respectively. However, familiarity judgments proved to be fallible, and we assumed that the small list-length effect that we observed resulted from a back-up search process that examined the memorized list for the presence or absence of the test word when the familiarity was of an intermediate value.

To quantify the model, we estimated the relative locations of two decision criteria along the hypothetical familiarity continuum from the false-alarm rates for distractors and the miss rates for targets. These values were quite low and independent of list length, so the estimated values did not differ across list-length conditions (see also Annis, Lenes, Westfall, Criss, & Malmberg, 2015, for a review and discussion of list-length effect in recognition memory). RTs, however, were definitely

not independent of list length, and the RTs were fit by a model that summed encoding and RTs plus fast positive and negative RTs for decisions based on familiarity, and additional search times that depended on list length for decisions based on intermediate familiarity values (see Fig. 3).

The model fit the error data perfectly, within the tolerance levels of experimental noise, and also provided a good fit to the RT data, including predictions of increased memory search times and larger list-length effects for distractors when their familiarities were increased by repeating distractor words during the test phase. Similarly, repetitions of target words during testing reduced both the overall RTs and the effects of list length, as predicted by the theory (Atkinson & Juola, 1973, 1974).

At the time that the A-J model was proposed, we anticipated some of the theoretical developments in memory research arguing for different types of systems on which recognition memory could be based. For example, the distinction between semantic and episodic memories proposed by Tulving (1972) dovetailed with our ideas that the familiarity of a test word is based on whatever information is initially retrieved when a word is encoded, and subsequent decisions are based on retrieval of episodic information stored, in our case, as a specific list. Further developments in theories of recognition memory have emphasized that similar distinctions can be made between familiarity or “feeling of knowing” judgments and those based on recollection of specific episodic details that can sometimes confirm or deny the presence of previous experiences with the test item (e.g., Parks & Yonelinas, 2007; Wixted, 2007; Yonelinas, 1994).

Such hybrid models of recognition performance have been analyzed in terms of the shape of the ROC function relating the observed proportions of hits and false alarms across different conditions and subjects. In the classic SDT model in which signal and noise distributions (corresponding to the

memory strength distributions of target and distractor items, respectively), are equivariant, the ROC function describing choice behavior forms a smooth, symmetric curve lying above the diagonal from the point (0,0) to (1,1). The distance of the curve from the diagonal is determined by d' , the difference between the means of the strength distributions for studied (target) and non-studied (distractor) items, as scores move toward the (1,0) corner when discrimination is easier. Observed data should correspond to points along the curve that are determined by the choice of c , the bias parameter, or the relative willingness to say “yes” or “old” to a test item. In contrast, without additional assumptions, the 2-criterion high-threshold model (2HTM) predicts that the ROC curve should be linear, with data points extending from $(D_o, 0)$ to $(1, D_n - 1)$ depending on the value of g .

Other models are hybrids of SDT and threshold theories, such as the A-J model. Although the A-J model was not used at the time to generate ROC curves, others like it have assumed that there is a familiarity component and a recollection component, with the familiarity part following continuous, Gaussian distributions, and the recollection part being a more discrete, all-or-none process. Thus, the hybrid model reduces to the SDT model when the two criteria converge, and it superficially resembles the 2HTM when the two criteria are separated. The hybrid model differs from the 2HTM in two important ways, however: (1) the model states are not discrete, but rather the underlying familiarity dimension is continuous, allowing for gradations in response speed and confidence, and (2) the state between the two criteria, while resulting in greater levels of uncertainty, produces not only biased guessing strategies, but decisions based on a search of episodic memory that can sometimes produce highly accurate and confident positive responses. If recollection fails, and by definition it must fail for distractor items, a biased decision process determines the final response.

Fig. 3 Theoretical familiarity distributions for distractor words (left) and target words (right) in a recognition memory experiment. The criteria (C_0 and C_1) are determined by the observed miss (false negatives) and false alarm (false positives) rates, respectively, and they also determine the respective proportions of fast, familiarity-based negative and positive

responses. The areas under the familiarity density functions between the criteria are used to estimate the proportions of slower responses that depend on list length as determined by recollective search processes through representations of the studied words in episodic memory (after Atkinson & Juola, 1973, 1974)

Yonelinas and Parks (2007) have argued that such hybrid models generally produce curved ROC functions, but successful recollection increases the relative proportions of high-confidence hits. This result should make the ROC curve asymmetrical, but since the ROC curve is asymmetrical under SDT assumptions when the signal and noise distributions have unequal variances, the distinction might not be critical when considered alone in choosing between the theories.

In summary, enough uncertainty remains about search and decision processes in recognition memory that we decided to make a strong test of the relevant theories by collecting large numbers of observations from a relatively large sample of participants and then employing several different methods to generate ROC curves from the same set of data. In keeping with recently advocated theoretical arguments for analyzing group data as well as observations from individual participants (e.g., Botella, Privado, Suero, Colom, & Juola, 2019; Cohen, Sanborn, & Shiffrin, 2008; Estes & Maddox, 2005; Jang, Wixted, & Huber, 2011; Malejka & Bröder, 2019), we fit models to the data for each subject in order to determine whether some of the differences reported in the literature could be due to heterogeneous abilities or strategic variations in applying control processes across participants. It is also possible that group data could mask underlying behavioral patterns, such as in extreme cases in which individual subjects might generate linear ROC functions in different parts of the probability space, but their averages produce a curved ROC resultant.

For the present study, we used three methods to generate ROC curves: (1) separate blocks of test trials in which target (old item) proportions were varied from low to high versus complementary proportions of distractors (new items), (2) the use of confidence judgments to generate a series of decision criteria, and (3) the analysis of RTs into bins analogous to confidence ratings in which fast responses are assumed to represent confident judgments and slow responses are assumed to correspond more often to guesses. If all three methods converge on a common way of measuring underlying memory search and decision processes, then the recovered ROC curves should be consistent across them. They could then be used as evidence in favor of one or another of the various models proposed for search and decision processes in human recognition memory. As it turns out these predictions were overly optimistic.

Method

Participants

The participants were 58 students and other members of the academic community at the Universidad Autónoma de

Madrid. They varied in age from 18 to 28 years ($\bar{X} = 19.2$; S.D. = 2.1), and 50 of them were females. The data from 11 subjects were excluded from the final analysis due to failure to meet an arbitrary criterion of 66% correct or more in the recognition memory test.

Materials

The words used for study and test lists included 500 common, concrete Spanish nouns selected from word lists provided by Alonso, Fernández, Díez, and Beato (2004), Callejas, Correa, Lupiáñez, and Tudela (2003), and Pérez, Acosta, Megías, and Lupiáñez (2010). The words selected for study were presented on a computer monitor in black on a white background in lowercase 18-point Consolas font. From the participants' viewing distance of approximately 60 cm, a four-letter word subtended about one degree of horizontal visual angle. The experiment was run using E-Prime® 3 (Psychology Software Tools).

Procedure

The subjects were run individually or in small groups in a computer laboratory equipped with several small carrels. After receiving instructions and signing consent forms, they viewed a study list of 250 words presented one at a time. The study lists consisted of two different random samples taken from the original 500-word set assigned to each subject in a counterbalanced order. Each word was presented in the center of the screen for 2.5 s followed by a .5-s blank interval. The participants were asked to pronounce each word aloud as it was presented to be sure that each was attended. The words were presented in sets of 83 or 84 with short rest periods between sets.

After the last study set was presented, the subjects were given an intervention task consisting of making judgments about whether a rotated letter was in its original form or was a mirror image. The intervention task lasted for about 5 min.

The test phase consisted of 500 trials, divided into five blocks of 100 trials each and separated by short rest periods. The tests included one presentation of each of the 250 target words from the study list and one presentation of each of the 250 remaining non-studied, distractor words. The subjects were divided into two groups to receive the test blocks in different orders. Half of the subjects began with a trial block that included 15 targets and 85 distractors, and subsequent blocks contained 35 targets and 65 distractors, 50 of each, 65 targets and 35 distractors, and finally 85 targets and 15 distractors. The order of these blocks was reversed for the other group of subjects. All subjects were informed of the proportions of targets and distractors at the beginning of each trial block.

On each trial, a test word was presented in the same font as in the study phase. A trial began with the presentation of a large, black outline rectangle near the center of the screen for 500 ms. When it disappeared, the test word was shown in the center of where the rectangle had been until the subject responded. Each participant was instructed to make a positive or “yes” recognition response with one finger of the left hand on the “A” key on the computer keyboard. A negative response was to be made with a different finger of the left hand on the “D” key. Subjects were instructed to respond as rapidly as possible while being careful to be correct. When the response was made, instructions appeared on the screen to press a key with a finger of the right hand on the number pad of the keyboard. The subjects were instructed to type “1” if they were certain that their response was correct, “2” if they were relatively sure that they were correct, and “3” if they thought that they were guessing. They were encouraged to use all three numerical responses on some trials throughout the experiment. If the initial response was incorrect, the screen turned red for about 50 ms after the confidence rating was made, and it remained blank for this period following correct responses. Then the rectangle appeared to signal the presentation of the next target word.

Following the final test word, the subjects were debriefed and thanked for their participation. The entire session lasted less than one hour.

Results

The results are presented in several parts, beginning with three sections reporting individual and group analyses of ROC curves based on the proportions of hits and false alarms observed in (1) blocks of trials with different proportions of targets and distractors, (2) categories determined by confidence ratings, and (3) categories determined by RTs.

Effects of target proportions The manipulation of the proportions of targets and distractors in separate trial blocks resulted in the expected changes in response rates. Table 1 shows the average rates of hits and false alarms across subjects for the five trial blocks in which the proportions of targets were varied from .85 to .15. As a result, the number of “yes” responses across blocks (averaging about 75, 61, 47, 35, and 21 across subjects) spanned almost the same range as the actual number of targets shown in each block (85, 65, 50, 35, and 15, respectively). Both hit and false-alarm rates increased significantly with the proportion of targets in each block [$F(4,184) = 40.35$, $p < .001$, and $F(4,184) = 41.21$, $p < .001$, respectively].

We then found d' and c for each subject in each trial block, and subjected these values, also shown in Table 1, to ANOVAs. The average sensitivity values measured by d' did not vary significantly across trial blocks [$F(4,184) <$

Table 1 Summary data of averages across participants for hits and false-alarm rates in five trial blocks with different proportions of targets and distractors (proportions of targets in parentheses). The average signal-detection parameter estimates of d' and c are also indicated, as determined from values of these parameters estimated for each subject. The last column on the right reports the average values across trial blocks

Proportions of targets in each trial block						
	A (.85)	B (.65)	C (.50)	D (.35)	E (.15)	Ave
P(H)	.806	.743	.653	.602	.535	.668
P(FA)	.435	.357	.290	.220	.158	.292
d'	1.118	1.098	1.054	1.142	1.177	1.118
c	-0.359	-0.135	0.102	0.282	0.489	0.076

1.0]. However, the systematic change in the criterion, c , across conditions was significant [$F(4,184) = 52.1$, $p < .001$]. The ROC plots shown in Fig. 4 correspond to the data in the first two rows of Table 1, including the raw, binary proportions in the left panel and their conversions to z -scores in the right panel of the figure. The average d' and c estimates across subjects from Table 1 are shown Fig. 5.

Model fitting to the aggregate data for target proportion blocks We began by first testing the aggregate data across subjects against the predictions from the standard SDT (see Fig. 1) and from the 2HTM (see Fig. 2) using a maximum-likelihood method. In all models it was assumed that the experimental manipulation (target frequency) affected only the response criteria, not the sensitivity. Two 2HTMs were fitted, under the assumption of symmetric states ($D_o = D_n$), and again without that restriction. Two SDT models were also fitted, under the homoscedastic assumption ($\sigma_1 = \sigma_2$), and again allowing for heteroscedasticity. The fits were achieved by using the R package MPTinR (Singmann & Kellen, 2013). Since there are five experimental conditions in which the target probabilities were varied, it is assumed that: (1) in the 2HTM the parameters D_o and D_n are constant across those conditions, while the parameter g can vary depending on the experimental condition; and (2) in SDT, the parameters d' and σ are constant across the conditions, while the criteria c can vary depending on the experimental condition. Consequently, the number of parameters in the symmetrical 2HTM is equal to 6 (because $D_o = D_n$), whereas for the asymmetrical model there are 7. Likewise, for the homoscedastic version of SDT there are 6 parameters (because $\sigma = 1$ for both target and distractor distributions), whereas for the heteroscedastic version there are 7.

The model evaluations were carried out in two steps: (1) their overall fits were quantified by the G^2 statistic (only models with $p > 0.05$ were retained), and (2) the selected model, among the four candidates, was determined using standard statistical criteria. The results are summarized in Table 2.

Fig. 4 Plots of the average proportions of hits against false alarms for the data from the five trial blocks that varied in their relative proportions of targets and distractors reported in Table 1. The left panel shows the ROC function for the raw response proportion data, and the right panel shows

the data for their z-score transforms. The data points in the graphs, from left to right, represent conditions with increasing proportions of targets. The fitted line and curve represent predictions of the symmetric two-high threshold model (2HTM)

The parameters from $g1$ to $g5$ (in the 2HTM), and $c1$ to $c5$ (in SDT), refer to experimental conditions A to E (85% targets to 15% targets), respectively. The D_o and D_n values for the 2HTM-Asymmetric model refer to trials on which old or new items were detected as such with respective average probabilities of 0.373 and 0.377. In the 2HTM-Symmetric model both probabilities equal 0.375. The decreasing value of the g parameter reflects the change in probability of saying “yes” in the uncertain state across the five experimental conditions. As the p -values were higher than .05, the conclusion is that both models provide acceptable fits to the average values found here. The results for the two SDT models do not change very much, as the only difference is that the parameter σ for the target distribution was set to 1.0 in the homoscedastic model whereas it was free to vary in the heteroscedastic model. However, the estimated value (1.05) changed little in that case. As expected, the criterion values ($c1$ to $c5$) shifted from left to the right as the target frequencies decreased. The criterion

values in the SDT heteroscedastic model were obtained assuming $\sigma = 1$ for the noise distribution and were calculated in terms of the noise distribution (see Macmillan & Creelman, 2005; Eq. 3.2, p. 59). As the p -values for both models were also higher than .05, they are retained; so all four models are plausible when they are fitted to the average values. However, taking as a basis both the AIC and BIC indices, the best-fitting model overall is the symmetric 2HTM. The fits to the data shown in Fig. 4 correspond to predictions of the 2HTM.

Model fitting to individual data for trial blocks with varying target proportions As the results of model fitting for the average data in recognition memory experiments sometimes hide important individual differences (e.g., Malmberg & Xu, 2006), we also analyzed the data at the individual level. The simplest way to compare models based on SDT and thresholds of the type proposed by Bröder and Schütz (2009), is to compare the best-fitting linear models to the data of each

Fig. 5 Plots of d' (left panel) and c (right panel) averaged across participants for the five trial blocks as shown in Table 1. Hash marks indicate the 95% confidence intervals. The bars in the graphs, from left to right, represent conditions with decreasing proportions of targets

Table 2 Parameter estimates and goodness-of-fit statistics for the two-high-threshold models (2HTMs) and homo- and heteroscedastic versions of signal detection theory (SDT) for the average data from the target proportion block conditions

<i>2HTM – asymmetric</i>											
D_o	D_n	$g1$	$g2$	$g3$	$g4$	$g5$	G^2	df	p	AIC	BIC
.373	.377	.691	.585	.457	.355	.252	1.48	3	.686	15.48	71.94
<i>2HTM – symmetric</i>											
D_o	D_n	$g1$	$g2$	$g3$	$g4$	$g5$	G^2	df	p	AIC	BIC
.375	=	.690	.583	.455	.353	.251	1.49	4	.228	13.49	61.88
<i>SDT – heteroscedastic</i>											
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$	G^2	df	p	AIC	BIC
1.03	1.05	-.370	-.149	.082	.268	.492	4.39	3	.222	18.39	74.85
<i>SDT – homoscedastic</i>											
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$	G^2	df	p	AIC	BIC
1.01	1.00	-.355	-.144	.081	.263	.486	4.85	4	.303	16.85	65.24

The parameters $g1$ through $g5$ represent the bias for guessing “yes” when in the indeterminate state in the 2HTM across the five target proportion conditions (from high to low target proportions)

σ = the standard deviation of the familiarity distribution of targets or old items in SDT, which is set to 1.0 for the distractors or new items, and the parameters $c1$ through $c5$ indicate the locations of the decision criteria in SDT across the five target proportion blocks (from high to low target proportions)

individual. SDT predicts that the linear fits to the $P(H)$ X $P(FA)$ functions should be better for the z -transformed proportions, and threshold theories (such as the 2HTM) predict that the linear fits should be better for the raw hit and false alarm proportions. Therefore, each subject’s ROC data were fit by linear models for both raw proportions and their z -transforms.

In the data for trial blocks in which target proportions were varied, the data were somewhat unstable, especially in the extreme conditions in which target (or distractor) frequencies were 15 in a trial block. Similar problems have been noted by others who have relied on individual-subject data for model fitting (e.g., Bröder & Schütz, 2009; Malejka & Bröder, 2019; Yonelinas & Parks, 2007).

In the present data collected across monotonic changes in target proportions, six of the 47 subjects actually showed negative slopes in their best-fitting linear functions across the target proportion blocks, although in no cases did the proportions of false alarms exceed the proportions of hits. Also, none of the negative slopes differed significantly from zero, whereas 17 of the 41 positive slopes were significant for the raw scores and 18 were significant for the z -scores. As Bröder and Schütz (2009) and Malejka and Bröder (2019) have reported, these data are unhelpful for discriminating between pure SDT models and two-threshold models with discrete states, as the proportions of variance accounted for by linear functions were inconclusive for distinguishing between the models. In our case, 28 of 47 subjects’ data showed better linear fits over their raw proportions of hits and false alarms, and for the other 19, their z -transforms were better fit by the linear functions. These differences give a slight but non-significant advantage for the 2HTM over SDT ($p = .12$, by a binomial test).

The model-fitting procedure using maximum-likelihoods for individual data was the same as that used for the subject averages, and Table 3 shows the values for the 47 participants. When comparing the individual AIC values, the best-fitting model is the 2HTM-symmetric for 24 individuals (51.1%), SDT-homoscedastic for 13 individuals (27.7%), SDT-heteroscedastic for seven (14.9%), and 2HTM-asymmetric for three (6.4%). When comparing the individual BIC values, the selected model is 2HTM-symmetric for 26 individuals (55.3%), SDT-homoscedastic for 19 individuals (40.4%), and SDT-heteroscedastic for two individuals (4.3%). Although the model selection results are about the same as with the average data for both indices, there is a non-negligible number of subjects for which the selected model is not the 2HTM-symmetric, but the SDT-homoscedastic model. For most individuals the model selected with both indices is the same (38/47; 80.9%). There is insufficient evidence in these data to support further theoretical explorations for the heteroscedastic SDT or for hybrid models, such as the A-J theory, that predict asymmetric ROC curves. Although asymmetric ROC curves consistent with unequal-variance models are clearly the norm in SDT analyses of recognition memory performance (e.g., Parks & Yonelinas, 2007), there are reports supporting the symmetric view, based on z -ROC plots of slopes near 1.0 (e.g., Gronlund & Elam, 1994; Experiment 1), such as the slope of .93 found in the present study.

Effects of grouping responses into confidence rating bins The confidence rating data were obtained by first combining the data from all trial blocks in which target proportions were

Table 3 Parameter estimates and goodness-of-fit statistics for two versions of the two-high-threshold model (2HTM) and two versions of signal detection theory (SDT) for each subject across conditions in which targetproportions were varied. The values listed in **bold** font represent the best-fitting model for each subject

Part.	2HTM-Asymmetric				2HTM-Symmetric				SDT-Heteroscedastic				SDT-Homoscedastic			
	G^2	p	AIC	BIC	G^2	p	AIC	BIC	G^2	p	AIC	BIC	G^2	p	AIC	BIC
1	1.55	.672	15.55	45.05	2.32	.678	14.32	39.60	1.73	.631	15.72	45.23	2.64	.620	14.64	39.92
2	6.12	.106	20.12	49.62	6.19	.185	18.19	43.48	5.22	.157	19.22	48.72	6.17	.187	18.17	43.46
3	4.17	.244	18.17	47.67	6.60	.159	18.60	43.89	3.72	.293	17.72	47.22	6.09	.193	18.09	43.38
4	0.74	.865	14.74	44.24	2.25	.689	14.25	39.54	0.92	.819	14.92	44.43	2.98	.562	14.98	40.26
5	1.52	.679	15.52	45.02	2.51	.643	14.51	39.80	0.61	.894	14.61	44.11	0.93	.920	12.93	38.22
6	1.29	.731	15.29	44.80	1.65	.800	13.65	38.94	2.65	.449	16.65	46.15	12.0	.018	23.96	49.24
7	0.80	.849	14.80	44.30	0.86	.930	12.86	38.15	0.84	.840	14.84	44.34	1.74	.783	13.74	39.03
8	3.55	.314	17.55	47.05	3.71	.447	15.71	40.99	3.38	.337	17.38	46.88	3.51	.476	15.51	40.80
9	3.42	.331	17.42	46.93	3.58	.467	15.58	40.86	5.04	.169	19.04	48.54	5.35	.253	17.35	42.64
10	6.04	.110	20.04	49.54	10.5	.033	22.48	47.76	5.85	.119	19.85	49.35	8.71	.069	20.71	45.99
11	0.96	.811	14.96	44.46	1.32	.857	13.32	38.61	1.25	.740	15.25	44.76	1.50	.826	13.50	38.79
12	7.85	.049	21.85	51.35	8.39	.078	20.39	45.68	10.7	.014	24.65	54.15	27.4	.000	39.37	64.66
13	1.28	.734	15.28	44.78	1.56	.815	13.56	38.85	.099	.803	14.99	44.50	4.92	.295	16.92	42.21
14	0.59	.898	14.59	44.10	1.46	.834	13.46	38.75	0.63	.890	14.63	44.13	1.56	.815	13.56	38.85
15	0.42	.936	14.42	43.92	0.49	.975	12.49	37.78	0.47	.925	14.47	43.97	0.58	.965	12.58	37.87
16	4.84	.184	18.87	48.34	10.5	.033	22.48	47.77	0.99	.804	14.99	44.49	9.54	.049	21.54	46.82
17	7.28	.063	21.28	50.79	12.7	.013	24.72	50.01	5.17	.160	19.17	48.68	12.3	.015	24.31	49.59
18	4.93	.177	18.97	48.43	7.03	.134	19.03	44.32	4.25	.235	18.25	47.76	6.27	.180	18.27	43.56
19	0.94	.816	14.94	44.44	0.95	.917	12.95	38.24	1.34	.719	15.34	44.84	1.51	.824	13.51	38.80
20	2.36	.501	16.36	45.86	5.05	.282	17.05	42.34	2.57	.463	16.57	46.07	4.58	.333	16.58	41.86
21	1.62	.654	15.62	45.12	1.78	.776	13.78	39.07	2.31	.510	16.31	45.81	9.85	.043	21.85	47.13
22	7.76	.051	21.76	51.26	9.32	.053	21.32	46.61	6.38	.095	20.38	49.88	9.28	.054	21.28	46.57
23	4.72	.194	18.72	48.22	6.21	.184	18.21	43.50	4.29	.232	18.29	47.80	6.21	.184	18.21	43.49
24	4.04	.257	18.04	47.54	5.84	.212	17.83	43.12	3.89	.273	17.89	47.40	5.57	.234	17.57	42.86
25	1.06	.787	15.06	44.56	5.55	.235	17.55	42.84	0.33	.954	14.33	43.83	7.28	.122	19.28	44.57
26	4.60	.204	18.60	48.10	9.06	.060	21.06	46.35	4.91	.179	18.91	48.41	7.91	.095	19.91	45.19
27	1.43	.699	15.43	44.93	1.80	.773	13.80	39.09	1.71	.633	15.72	45.22	5.72	.221	17.72	43.00
28	1.61	.658	15.61	45.11	2.05	.726	14.05	39.34	1.71	.633	15.72	45.22	9.31	.054	21.31	46.60
29	2.06	.560	16.06	45.56	2.30	.681	14.30	39.59	1.14	.769	15.14	44.64	1.68	.794	13.68	38.97
30	3.17	.366	17.17	46.67	3.36	.500	15.36	40.64	3.01	.390	17.01	46.51	3.02	.555	15.02	40.30
31	1.04	.792	15.04	44.54	1.04	.904	13.04	38.33	0.44	.931	14.44	43.95	30.1	.000	42.09	67.38
32	1.60	.659	15.60	45.10	2.50	.644	14.50	39.79	1.60	.660	15.60	45.10	6.30	.178	18.30	43.59
33	1.80	.615	15.80	45.30	1.82	.769	13.82	39.11	1.36	.715	15.36	44.86	1.42	.841	13.42	38.70
34	1.44	.697	15.44	44.94	1.89	.756	13.89	39.18	1.31	.727	15.31	44.81	1.34	.855	13.34	38.63
35	2.82	.420	16.82	46.32	3.72	.445	15.72	41.01	2.02	.567	16.02	45.53	10.2	.038	22.15	47.44
36	1.52	.678	15.52	45.02	3.11	.539	15.11	40.40	1.38	.709	15.38	44.89	2.44	.656	14.44	39.72
37	4.31	.230	18.31	47.81	5.84	.212	17.84	43.12	4.19	.242	18.19	47.69	11.4	.022	23.40	48.69
38	1.02	.797	15.02	44.52	1.05	.902	13.05	38.34	1.06	.786	15.06	44.57	17.2	.002	29.15	54.44
39	6.68	.083	20.68	50.19	7.60	.107	19.60	44.89	8.19	.042	22.19	51.69	11.7	.020	23.68	48.97
40	2.32	.508	16.32	45.82	8.75	.068	20.75	46.04	1.86	.601	15.86	45.37	4.11	.391	16.11	41.40
41	0.93	.818	14.93	44.43	2.99	.560	14.99	40.28	1.77	.621	15.77	45.27	4.32	.364	16.32	41.61
42	1.26	.738	15.26	44.76	1.52	.823	13.52	38.81	1.49	.684	15.49	44.99	1.90	.755	13.90	39.18
43	4.33	.228	18.33	47.83	4.37	.358	16.37	41.66	5.58	.134	19.58	49.08	6.14	.189	18.14	43.43
44	1.12	.773	15.12	44.62	8.14	.087	20.14	45.42	0.25	.969	14.25	43.76	8.78	.067	20.78	46.07
45	5.30	.151	19.30	48.80	5.57	.233	17.57	42.86	5.40	.145	19.40	48.90	5.76	.218	17.76	43.04
46	2.60	.459	16.59	46.09	2.65	.619	14.65	39.93	2.57	.463	16.57	46.07	2.58	.630	14.58	39.87
47	1.62	.655	15.62	45.12	2.51	.642	14.51	39.80	1.88	.597	15.88	45.38	4.57	.334	16.57	41.86

Table 4 A. Average proportions of responses on target (T) and distractor (D) trials in each of the confidence rating categories. N = “No” responses and Y = “Yes” responses, with confidence ratings 1 = “certain”, 2 = “relatively sure”, and 3 = “guess”. B. Summary data of averages across participants for hits and false-alarm rates across confidence rating bins (5 = “relatively sure no” + “guess no” + “guess yes” + “relatively sure yes” + “certain yes”; 4 = “guess no” + “guess yes” + “relatively certain yes” + “sure yes”, etc.). The average signal-detection parameter estimates of d' and c are also indicated, as determined from values of these parameters estimated for each subject. The last column on the right reports the average values across all confidence intervals

A. Mean proportions of responses						
	N1	N2	N3	Y3	Y2	Y1
T	.101	.153	.032	.025	.120	.569
D	.394	.313	.048	.026	.094	.125
B. Cumulative response proportions across confidence bins						
	5	4	3	2	1	Ave
P(H)	.899	.746	.714	.689	.569	.723
P(FA)	.606	.293	.245	.219	.125	.298
d'	1.054	1.275	1.316	1.333	1.455	1.287
c	-.852	-.051	.071	.152	.541	-.028

varied. Table 4, part A shows the mean proportions of responses across all subjects and trial blocks for the six confidence categories for target and distractor trials separately. Despite instructions to use all categories about equally often in their responses, it is clear that the most responses were highly confident (59.5%) and guesses were very infrequent (6.6%). However, the instructions to match confidence to their subjective impression of accuracy was fairly successful, as “certain” responses were 81% correct overall, followed by “relatively sure” (64% correct), and “guess” was indeed an accurate description, as the accuracy level for the third response category was barely above chance (56%).

These data were then aggregated into bins beginning with the bin for the highest confidence rating for “yes” responses, then adding the data for the middle and highest confidence rating categories into a second bin, and so on until the 5th bin included all responses that were categorized as “certain yes,” “relatively sure yes,” “guess yes,” “guess no,” and “relatively sure no,” as is typically done when constructing ROC curves from confidence intervals (e.g., Van Zandt, 2000). The mean raw proportions of hits and false alarms across subjects for the confidence rating data are shown in Table 4, part B and also in Fig. 6 along with their conversions to z -scores. The lowest two rows of Table 4 show the average d' and c values that were calculated from the individual subject parameter estimates.

The mean d' for the data averaged across subjects was observed to be slightly higher for the confidence ratings ($d' = 1.29$) than for the same data based on binary, yes-no decisions across the trial blocks with different target proportions ($d' = 1.12$). Unlike the data compared across trial blocks, the values for d' decreased consistently as more bins were added

to the cumulative hit and false-alarm rates. That is, d' was highest for the condition in which only the most-confident “yes” responses were included. Van Zandt (2000) has claimed that “...confidence is not scaled directly from perceived familiarity...” (p. 582), and Bröder and Schütz (2009) have also argued that d' estimates in the literature can be different for binary and rating response measures. There were substantial changes in the c parameter consistent with the use of monotonically-changing decision criteria across a stable set of target and distractor memory strength distributions. Further, the data shown in Table 4 differ from those shown in Table 1 for the separate target proportion blocks. In Table 1 c can be observed to change markedly with changes in the relative prevalence of targets in the test blocks, but d' remains relatively constant. However, across cumulative confidence bins, both c and d' change monotonically and are clearly not independent. Grider and Malmberg (2008) have presented a detailed discussion of this issue and show that non-independence is a consequence of unequal variances in the target (signal) and distractor (noise) distributions in classical signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005). The univariate assumption can be checked against the z ROC data in that the z ROC slope is equal to the ratio of the variances of the distractor and target distributions in SDT.

The ROC plots shown in Fig. 6 correspond to the data in Table 4, as do the d' and c estimates in Fig. 7. In comparing Figs. 4 and 6 we can observe that when the theoretical familiarity distributions for targets and distractors have close to the same variance (z ROC slope = .93 for the target proportion data), d' and c are relatively independent, but the asymmetrical ROC curve in Fig. 6 (z ROC slope = .79) indicates that the independence assumption is likely to be violated. Thus, we expect the positive correlation between c and d' shown in Table 4B.

Model fitting to the aggregate data across confidence levels

We again tested the aggregate data across subjects for each version of the 2HTM and the standard SDT using a maximum-likelihood method, following the same procedures as above. The results are summarized in Table 5.

On the basis of the p -values alone, the only retained model, and the one that provides the best fit to the category-rating data (as shown in Fig. 6), is the heteroscedastic version of SDT.

Model fitting to the individual data for confidence ratings In fitting the data for the confidence rating bins, we began as before by finding the best-fitting linear functions relating the proportions of hits to proportions of false alarms for both the raw data and their z -transformations. In both cases, the linear fits were much better across individual subjects' data for confidence ratings than they were for the target proportion conditions, as the cumulative summing of raw proportions guarantees that both false alarms and hits will increase monotonically

Fig. 6 Plots of the average proportions of hits against false alarms for the data from the cumulative confidence category bins reported in Table 4. The left panel shows the ROC function for the raw response proportion data, and the right panel shows the data for their z -score transforms. The

data points, from left to right, represent bins of increasing size cumulated across confidence bins. The fitted curve and line represent predictions from the heteroscedastic version of signal detection theory (SDT)

across the confidence categories. Nevertheless, 38 of the 47 subjects showed a higher proportion of variance accounted for by linear models in the z -transforms than in the raw data, more consistent with SDT theory than threshold theory.

The ROC plots are clearly non-linear for the raw data for the confidence ratings as shown in Fig. 6. Visual inspection suggests that the confidence rating categories produced asymmetrical curves consistent with the view that the data are best fit by either a heteroscedastic SDT model (corresponding to the fits shown in Fig. 6), with the target or old item distribution having a greater variance, or by a hybrid model, in which responses not based on familiarity alone are followed by a search and possible recollection process which likely to yield a greater proportion of high-confidence “yes” responses to target words.

The maximum-likelihood procedure used was the same as that for the subject averages in the target proportion

conditions, and Table 6 shows the values for the 47 participants. When comparing the individual AIC values, the selected model is SDT-heteroscedastic for 23 individuals (48.9%), SDT-homoscedastic for 15 individuals (31.9%), 2HTM-asymmetric for five (10.6%), and 2HTM-symmetric for four (8.5%). When comparing the individual BIC values, the selected model is SDT-homoscedastic for 29 individuals (61.7%), SDT-heteroscedastic for ten individuals (21.3%), 2HTM-symmetric for five (10.6%), and 2HTM-asymmetric for three (6.4%). In this case, the heteroscedastic SDT yields the best fit for the AIC measure, and the homoscedastic SDT is best using the BIC measure. Both measures indicate that SDT outperforms the 2HTM for the category rating data. For most participants the model selected with both indices was the same (32/47; 68.1%).

Despite the generally better fit obtained from SDT than from threshold models, there are reasons to be concerned

Fig. 7 Plots of d' (left panel) and c (right panel) averaged across participants for the confidence rating data shown in Table 4. Hash marks indicate the 95% confidence intervals. The bars in the graphs, from left to right, represent conditions with decreasing numbers of cumulated bins

Table 5 Parameter estimates and goodness-of-fit statistics for the two two-high-threshold models (2HTMs) and homo- and heteroscedastic versions of signal detection theory (SDT) for the average data arranged into confidence category bins

<i>2HTM – asymmetric</i>												
D_o	D_n	$Y1$	$Y2$	$Y3$	$N3$	$N2$	$N1$	G^2	df	p	AIC	BIC
.537	.254	.153	.149	.041	.067	.390	.200	166	3	<.001	180.2	247.9
<i>2HTM – symmetric</i>												
D_o	D_n	$Y1$	$Y2$	$Y3$	$N3$	$N2$	$N1$	G^2	df	p	AIC	BIC
.419	=	.228	.188	.047	.072	.338	.127	778	4	<.001	789.5	847.5
<i>SDT – heteroscedastic</i>												
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$		G^2	df	p	AIC	BIC
1.40	1.30	.554	.161	.071	-.067	-.876		3.18	3	.365	17.2	84.9
<i>SDT – homoscedastic</i>												
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$		G^2	df	p	AIC	BIC
1.01	1.00	.477	.140	.061	-.059	-.807		156.8	4	<.001	168.8	226.9

The values $Y1$ to $N1$ in the 2HTM models are weights for assigning responses to confidence ratings for items whose familiarities fall between D_o and D_n , corresponding to “Yes, certain” to “No certain”, respectively. Since the weights sum to 1.0, they account for 5 degrees of freedom

σ = the standard deviation of the familiarity distribution of targets or old items in SDT, which is set to 1.0 for the distractors or new items, and the parameters $c1$ through $c5$ indicate the locations of the decision criteria in SDT between the six confidence bins

about the assumption that familiarity directly determines confidence. If this were true, then the confidence rating data should be constant across conditions designed to affect response bias. Van Zandt (2000) demonstrated that the ROC curves obtained in different bias conditions are not identical. Specifically, she found that the slopes of the linear z ROC functions across confidence categories were steeper for conditions in which subjects were biased to make relatively more “old” decisions. In the present study we made a similar analysis of the z ROC plots across confidence ratings and replicated her finding. The z ROC slopes were .832 for the combined 85% and 65% target proportion conditions, .760 for the 50% condition, and .725 for the combined 35% and 15% target proportion conditions. As Van Zandt pointed out, such results are inconsistent with a common familiarity-based continuous recognition model, and she suggested that changes in response bias associated with varying target proportions in a test list should be due to changes in the shapes of the underlying memory strength distributions in continuous recognition memory models, and not due to changes in the criterion, c alone.

Since the data for the confidence ratings were most often consistent with the heteroscedastic SDT, at least for the aggregate data, the results tend to support both unequal-variance SDT models as well as hybrid models. Following the suggestion of Yonelidas (1994; see also Malmberg, 2008) who argued that dual-process, hybrid models can provide a better fit to confidence rating ROCs than single-process models, we attempted to fit the data for confidence bins in each of the target proportion conditions using the A-J model (see Fig. 3). To do this, we first fit the model to the target proportion data for each trial block while temporarily ignoring the confidence rating data. The model has eight parameters; the mean of the target distribution ($d' = 1.0$), the five guess parameters

(g_1 through g_5) for items falling in the intermediate familiarity range when recollection fails, and r = recollection probability for target items falling in the intermediate range. The details of predictions and goodness-of-fit statistics are presented in Table 7. Notice that the model does not fit the data as well as any of the other models described in Table 2, although it provides an acceptable fit to the data overall ($p = .05$) as well as to the data of 36 of the 47 participants by the same criterion.

To determine how the hybrid model handles the data for confidence ratings across the target proportion blocks, we assumed that high-confidence hits could result from two sources, the proportion of the target distribution above the upper criterion, plus a proportion r of the trials in the intermediate, search region that result in successful recollections. Medium-confidence hits presumably result from unsuccessful recollection searches followed by positive responses with probability g . Similarly, high-confidence correct rejections result from the portion of the distractor distribution below the lower criterion, and medium-confidence correct rejections result from unsuccessful recollection searches followed by negative responses with probability $(1 - g)$. Medium-confidence false alarms result from positive responses from the search area, and medium-confidence misses result from unsuccessful searches $(1 - r)$ followed by negative responses $(1 - g)$. Finally, the very small proportions of low-confidence responses (6.6% overall) were assumed to result from items with familiarities near the middle of the search area in which target and distractor familiarities were about the same, producing about 3.3% of guess responses each for hits, correct rejections, misses, and false alarms. The details of the model along with the estimated parameters, data, and model predictions for the confidence rating data are shown in Table 8. In the table we consider the data from the 15% and 85% target proportion

Table 6 Parameter estimates and goodness-of-fit statistics for two versions of the two-high-threshold model (2HTM) and two versions of signal detection theory (SDT) for each subject across confidence ratingcategories. The values listed in **bold** font represent the best-fitting model for each subject

Part.	2HTM-Asymmetric				2HTM-Symmetric				SDT-Heteroscedastic				SDT-Homoscedastic			
	G^2	p	AIC	BIC	G^2	p	AIC	BIC	G^2	P	AIC	BIC	G^2	p	AIC	BIC
1	12.9	.005	26.93	67.70	44.8	.000	56.80	91.74	2.45	.485	16.45	57.21	2.46	.652	14.46	49.40
2	9.49	.023	23.49	64.26	97.8	.000	109.8	144.8	1.53	.676	15.53	56.30	11.3	.023	23.31	58.26
3	2.48	.480	16.48	57.24	2.48	.648	14.48	49.42	0.69	.875	14.69	55.46	3.67	.452	15.67	50.62
4	1.32	.723	15.32	56.09	5.20	.268	17.20	52.14	0.34	.953	14.34	55.10	7.33	.120	19.33	54.27
5	10.9	.012	24.93	65.70	34.1	.000	46.06	81.00	0.29	.962	14.29	55.06	.513	.972	12.51	47.46
6	21.5	.000	35.52	76.29	21.5	.000	33.53	68.47	1.50	.682	15.50	56.27	1.90	.755	13.90	48.84
7	1.12	.772	15.12	55.89	18.7	.001	30.69	65.63	0.44	.933	14.44	55.20	14.5	.006	26.45	61.40
8	0.39	.942	14.39	55.16	1.05	.902	13.05	48.00	0.81	.847	14.81	55.58	1.08	.897	13.08	48.03
9	4.91	.178	18.91	59.68	4.94	.294	16.94	51.88	0.81	.848	14.81	55.58	1.22	.874	13.22	48.17
10	4.92	.178	18.92	59.69	6.17	.187	18.17	53.11	0.21	.976	14.21	54.98	3.06	.548	15.06	50.00
11	0.36	.949	14.36	55.12	7.69	.104	19.69	54.63	0.22	.974	14.22	54.99	.466	.977	12.47	47.41
12	19.8	.000	33.82	74.59	20.4	.000	32.44	67.39	1.07	.783	15.07	55.84	2.72	.605	14.72	49.67
13	0.63	.890	14.63	55.40	41.5	.000	53.49	88.43	0.05	.997	14.05	54.82	21.7	.000	33.68	68.62
14	0.02	.999	14.02	54.79	22.4	.000	34.38	69.32	1.04	.791	15.04	55.81	6.04	.196	18.04	52.99
15	7.98	.046	21.98	62.75	188.	.000	200.6	235.6	0.20	.978	14.20	54.96	11.2	.025	23.19	58.13
16	0.13	.988	14.13	54.90	10.4	.035	22.37	57.31	1.10	.778	15.10	55.87	1.63	.804	13.63	48.57
17	30.7	.000	44.70	85.46	52.1	.000	64.06	99.01	2.88	.410	16.88	57.65	5.73	.221	17.73	52.67
18	29.3	.000	43.16	83.93	61.7	.000	73.67	108.6	1.46	.692	15.46	56.22	13.3	.010	25.30	60.25
19	5.99	.112	19.99	60.76	11.1	.025	23.11	58.06	1.11	.774	15.11	55.88	1.36	.851	13.36	48.30
20	0.74	.865	14.74	55.50	8.44	.077	20.44	55.30	0.47	.926	14.47	55.24	6.59	.159	18.59	53.53
21	0.42	.936	14.42	55.19	45.6	.000	57.60	92.54	0.76	.858	14.76	55.53	2.13	.712	14.13	49.07
22	0.19	.979	14.11	54.96	31.3	.000	43.32	78.26	0.41	.938	14.41	55.18	18.0	.001	29.97	64.92
23	0.32	.956	14.32	55.09	0.32	.988	12.32	47.27	0.26	.968	14.26	55.02	.275	.991	12.23	47.22
24	11.7	.009	25.65	66.42	38.5	.000	50.48	85.42	3.36	.339	17.37	58.14	6.20	.185	18.20	53.14
25	0.74	.864	14.74	55.51	36.0	.000	48.03	82.98	0.25	.975	14.22	54.99	22.1	.000	34.12	69.07
26	0.14	.987	14.14	54.90	18.3	.001	30.32	65.27	0.03	.999	14.03	54.80	7.29	.122	19.29	54.23
27	0.71	.870	14.71	55.48	17.4	.002	29.39	64.34	0.12	.989	14.12	54.89	1.26	.868	13.26	48.21
28	37.7	.000	51.70	92.46	140.	.000	152.4	187.4	4.24	.237	18.24	59.01	16.3	.003	28.25	63.20
29	2.56	.464	16.56	57.33	5.71	.222	17.71	52.65	0.85	.838	14.85	55.62	3.12	.539	15.12	50.06
30	4.98	.173	18.98	59.75	36.5	.000	48.52	83.46	0.75	.861	14.75	55.52	5.79	.215	17.79	52.74
31	0.22	.974	14.22	54.99	1.46	.834	13.46	48.40	0.14	.986	14.14	54.91	2.08	.721	14.08	49.02
32	7.28	.063	21.28	62.05	33.6	.000	45.56	80.50	0.39	.943	14.39	55.16	4.29	.368	16.29	51.24
33	33.8	.000	47.75	88.52	108.	.000	120.5	155.4	0.35	.951	14.35	55.12	2.70	.610	14.70	49.64
34	0.74	.865	14.74	55.50	39.1	.000	51.12	86.05	0.97	.809	14.97	55.74	9.62	.047	21.62	56.57
35	3.75	.290	17.75	58.52	144.	.000	156.0	191.0	0.47	.925	14.47	55.24	31.4	.000	43.38	78.32
36	1.97	.579	15.97	56.74	30.0	.000	42.02	76.97	6.64	.084	20.64	61.41	33.9	.000	45.89	80.84
37	15.4	.001	29.44	70.20	32.8	.000	44.79	79.74	0.21	.976	14.21	54.98	6.17	.187	18.17	53.11
38	9.07	.028	23.07	63.84	9.07	.059	21.07	56.01	3.39	.336	17.39	58.15	3.47	.483	15.47	50.41
39	0.08	.994	14.08	54.81	0.12	.995	12.20	47.14	0.04	.998	14.04	54.81	.213	.995	12.21	47.16
40	0.02	1.00	14.02	54.78	0.19	.996	12.19	47.13	0.02	.999	14.03	54.78	.046	1.00	12.05	46.99
41	1.60	.660	15.60	56.37	8.36	.079	20.36	55.30	0.16	.984	14.16	54.93	.203	.995	12.20	47.15
42	0.41	.938	14.41	55.18	40.8	.000	52.80	87.74	0.04	.998	14.04	54.80	9.22	.056	21.22	56.17
43	1.62	.655	15.62	56.39	2.33	.680	14.30	49.25	0.65	.886	14.65	55.42	1.68	.793	13.69	48.63
44	10.9	.013	24.85	65.62	20.5	.000	32.53	67.48	0.42	.937	14.42	55.18	5.55	.236	17.55	52.49
45	2.58	.462	16.58	57.34	9.26	.055	21.26	56.21	0.60	.896	14.60	55.37	3.85	.427	15.85	50.79
46	0.79	.851	14.79	55.56	8.80	.066	20.80	55.73	0.85	.838	14.85	55.62	4.70	.319	16.70	51.65
47	1.77	.622	15.77	56.54	35.5	.000	47.50	82.44	0.05	.997	14.05	54.82	27.4	.000	39.40	74.35

Table 7 Parameter estimates and goodness-of-fit statistics for the Atkinson-Juola version of the hybrid model for the average data from the target block conditions

<i>A-J hybrid model, homoscedastic version</i>													
d'	$g1$	$g2$	$g3$	$g4$	$g5$	C_0	C_1	r	G^2	df	P	AIC	BIC
1	.869	.633	.375	.179	0	-0.03	1.03	.137	5.98	2	.050	21.98	86.50

The familiarity distributions are assumed to have the same standard deviation (1.0) and to be separated by one standard deviation ($d' = 1.0$), with decision criteria symmetrically imposed over them (since C_0 determines the value of C_1 , their joint estimate uses only one degree of freedom). The parameters $g1$ through $g5$ represent the bias for responding “yes” across the five target proportion conditions (from high to low target proportions) when recollection fails, and r = the probability of a correct recollection for target items of intermediate familiarity (i.e., between C_0 and C_1)

blocks only, since these data sets should show the largest zROC slope discrepancies.

Since the locations of the criteria and the values of g change with each change in the target proportions across trial blocks, the number of parameters to be estimated is very large, even with the value of r fixed as well as the proportion (L) of low-confidence guesses. Nevertheless, our initial aim was to determine whether or not this version of the A-J model could anticipate Van Zandt’s (2000) findings. Therefore, for the purposes of fitting the model, we decided to put all of the effects of changes in target proportions across blocks into changes in the locations of the criteria, C_0 and C_1 , and we chose a fixed value for the positive “guess” parameter g ($= .25$). When we determined the predicted hits and false alarms across confidence ratings for different target proportion conditions (as shown in Table 8), we found that the slopes of the zROC curves were .899 for the 15% target block, and .962 for the combined 85% target conditions. Note that under the general assumptions of SDT with equal standard deviations for the target and distractor distributions (1.0), the expected overall slope of the zROC linear fit should be 1.0. Thus, there is some evidence that the A-J hybrid model can predict slope differences across response-bias conditions without requiring changes in the shapes of the underlying theoretical familiarity distributions, as argued by Van Zandt. In addition, the A-J model predicted the change in d' across cumulative responses in confidence bins as shown in Table 8, generating an expected value of d' of 1.12 for the Y1 target bin, and .98 for the Y1 though N2 combined bins, compared with the respective observed d' values of 1.12 and .77 for the aggregate data (note the d' for the aggregate data, 1.29, is greater than the average d' for each subject’s data, .99, across the confidence bins).

Effects of grouping responses into response time (RT) bins

The RT analysis began by first combining the data from all trial blocks in which target proportions were varied. Then these data were aggregated into bins by first separating all “yes” responses into one bin, and all “no” responses into another. Following this, these two response bins were further divided into thirds to identify the fastest, middle, and slowest responses. Finally, these were organized into six bins ordered

along the continuum beginning with the fastest “yes” responses, the medium “yes” responses, the slowest “yes” responses, the slowest “no” responses, the medium “no” responses, and finally the fastest “no” responses. Dividing the RT continuum into bins provides an index of relative certainty analogous to the confidence ratings, under the assumption that fast responses are generally more confident than slow responses. The mean raw proportions of hits and false alarms across subjects for the RT data are shown in Table 9, and also in Fig. 8 along with their conversions to z -scores.

The mean d' for the data averaged across subjects was observed to be somewhat less for the RT bins ($d' = 1.09$) than for the confidence rating bins ($d' = 1.29$), and it was about the same as that for the data separated into blocks based on target proportions ($d' = 1.12$). However, unlike the data for blocks in which the target proportions were varied, as well as for the data sorted into confidence bins, the values for d' were highest for the bins that included all of the positive responses and decreased in either direction from that point. There were very substantial changes in the c parameter consistent with the use of monotonically-changing decision criteria across a stable set of target and distractor memory strength distributions. The ROC plots shown in Fig. 8 correspond to the data in Table 9, as do the d' and c estimates shown in Fig. 9.

Model fitting for the aggregate data for RT bins We again tested the aggregate data across subjects for the 2HTM and the standard SDT using a maximum-likelihood method, following the same procedures as above. The results are summarized in Table 10. In this case, none of the four models is retained as they all show significant departures from acceptable fits to the data. In addition, we did not attempt to fit any hybrid models, such as the A-J theory, as the fits would be unlikely to be better than the simpler models with fewer parameters, and the RT bins would include complex, perhaps intractable, mixtures of responses from different familiarity, recollection, and guess contributors.

Model fitting to the individual data for RT bins The procedure used was the same as that for the subject averages in the previous conditions. In fitting the data for the RT-binning

Table 8 Derivation of confidence rating predictions from the A-J hybrid model (refer to Fig. 3 and Table 7); D = distractor trials, T = target trials, A_T = area under the target distribution below C_0 (confident misses), B_T = area between C_0 and C_1 for the target distribution (extended search) – L , C_T = area under the target distribution above C_1 (confident hits), A_D = area under the distractor distribution above C_1 (confident false alarms), B_D = area between C_0 and C_1 for the distractor distribution

(extended search) – L , C_D = area under the distractor distribution below C_0 (confident correct rejections), L = sub-areas between C_0 and C_1 that result in low-confidence guesses (= .066 in each case), N = “no” or new response, Y = “yes” or old response, 1 = high confidence, 2 = medium confidence, 3 = guess, g = $P(\text{respond “yes” when recollection fails})$, r = $P(\text{successful recollection of a target word})$

	N1	N2	N3	Y3	Y2	Y1
D	C_D	$B_D(1-g)$	$L/2$	$L/2$	B_Dg	A_D
T	A_T	$B_T(1-r)(1-g)$	$L/2$	$L/2$	$B_T(1-r)g$	$C_T + B_Tr$

Model parameters and data-fitting:

Parameters for the target block with 15% targets: $A_T = .200$, $B_T = (.495 - .066)$, $C_T = .305$, $A_D = .066$, $B_D = (.370 - .066)$, $C_D = .564$, $L = .066$, $r = .137$, and $g = .25$

Parameters for the target block with 85% targets: $A_T = .0495$, $B_T = (.3325 - .066)$, $C_T = .618$, $A_D = .242$, $B_D = (.499 - .066)$, $C_D = .258$, $L = .066$, $r = .137$, and $g = .25$

(Note: in both cases parameter B is not estimated, but is determined since $B = 1 - (A + C + L)$).

Data and fits for the 15% target condition across confidence bins

Data	Raw proportions					
	N1	N2	N3	Y3	Y2	Y1
D	.497	.306	.040	.014	.078	.066
T	.206	.213	.047	.018	.113	.403
	Cumulative proportions					
D	1.00	.504	.198	.158	.144	.066
T	1.00	.794	.581	.534	.516	.403
d'		.81	1.05	1.09	1.10	1.26

Slope of the z ROC regression line: .707

Model predictions	Raw proportions					
	N1	N2	N3	Y3	Y2	Y1
D	.564	.232	.033	.033	.078	.066
T	.200	.278	.033	.033	.093	.364
	Cumulative proportions					
D	1.00	.442	.210	.177	.144	.066
T	1.00	.801	.523	.490	.457	.364
d'		.99	.86	.902	.994	1.16

Slope of the z ROC regression line: .869

Data and fits for the 85% target condition across confidence bins

Data	Raw proportions					
	N1	N2	N3	Y3	Y2	Y1
D	.187	.325	.052	.045	.148	.243
T	.055	.118	.021	.033	.117	.656
	Cumulative proportions					
D	1.00	.813	.488	.436	.391	.243
T	1.00	.945	.827	.806	.773	.656
d'		.71	.97	1.02	1.03	1.10

Slope of the z ROC regression line: .746

Model predictions	Raw proportions					
	N1	N2	N3	Y3	Y2	Y1
D	.258	.325	.033	.033	.108	.243
T	.050	.172	.033	.033	.057	.655

Table 8 (continued)

	Cumulative proportions					
D	1.00	.742	.417	.384	.351	.243
T	1.00	.950	.778	.745	.712	.655
d'		.99	.98	.95	.94	1.10

Slope of the z ROC regression line: .962

procedures, we began as before by finding the best-fitting linear regressions of proportions of hits against proportions of false alarms for both the raw data and their z -transformations. In both cases, the linear fits were much better for individual subjects' data across the RT bins than for the target proportion conditions, as again, the cumulative summing of raw proportions guarantees both false alarms and hits will increase monotonically across the RT bins. For the RT data, all 47 subjects showed better linear fits for their z -scores than for the raw scores. These results led us to reject the unmodified 2HTM model for the RT data. The ROC plots are clearly non-linear for the raw data RT bins, as shown in Fig. 8. Visual inspection suggests that the ROC plot for the raw data based on binning by RT produced a fairly symmetrical ROC curve at least superficially consistent with the homoscedastic SDT model. In support of this conclusion, the z ROC slope was .91, in close agreement with the value of .93 found for the data from the separate target proportion blocks. Therefore, we chose to illustrate the fit to the ROC curves in Fig. 8 with the homoscedastic version of the SDT, as it also gave the best fit overall to the data for individual subjects by the BIC measure, as indicated in Table 12.

The model-fitting procedure using the maximum-likelihood procedure for individual data was the same as that used for the subject averages, and Table 11 shows the values for the 47 participants. When comparing the individual AIC values, the selected model is SDT-heteroscedastic for 29 individuals (61.7%) and SDT-homoscedastic for 18 individuals (31.3%). When comparing the individual BIC values, the selected model is SDT-homoscedastic for 31 individuals (66%) and SDT-heteroscedastic for 16 individuals (34%). In this case, the heteroscedastic SDT yields the best fit for the AIC measure, and the homoscedastic SDT is best using the BIC measure. Both measures indicate that SDT outperforms the 2HTM for fitting the RT data for all subjects. For most participants the model selected with both indices was the same (32/47; 68.1%).

Summary statistics for the data sorted into target proportion trial blocks, confidence rating bins, and RT bins A summary of the model-fitting statistics for all three models is presented in Table 12. Also included are the best-fitting models for the aggregate data (indicated by *).

Discussion

The use of ROC curves to analyze and interpret the data from recognition memory studies has had a long and varied history (e.g., Kellen & Klauer, 2015; Kellen, Klauer, & Bröder, 2013; Malejka & Bröder, 2019; Wixted, 2007; Yonelinas & Parks, 2007). If recognition memory were a unitary process, and several different methods employed to measure it yielded the same general pattern of results, then the choice among models for search and decision processes in recognition memory would converge on a single best interpretation. As it stands, we are left with ambiguous evidence for several alternatives. Recognition could be based upon a single strength continuum with a single response criterion and yield smooth, symmetrical ROC curves above the main diagonal with endpoints at (0.0) and (1.1) (e.g., Dube, Starns, Rotello, & Ratcliff, 2012). Alternatively, recognition could be based on one or two (or more) thresholds delineating discrete states of knowledge about the likelihood that a test word is a target (e.g., Province & Rouder, 2012). Responses generated from such states, without additional assumptions (such as defining response-mapping processes from discrete states that would add additional parameters; e.g., Bröder, Kellen, Schütz, & Rohmeier, 2013; Chen, Starns, & Rotello, 2015), should yield linear ROC functions or functions with separate linear components. Finally, hybrid models based on the presumption that decisions can be based either on the test word's familiarity or on a search of episodic memory for recollectible information predict that the ROC function should be curved, but perhaps highly asymmetric. It is also possible that ROC curves have lost their utility as a diagnostic tool for theory testing (as suggested by Dube, et al., 2012; Malejka & Bröder, 2019; McAdoo, et al., 2019; Malmberg, 2008; and others).

The present study used a single data set, collected from 47 subjects over five 100-trial blocks, in which subjects' memories were tested for 250 previously-studied words among 250 distractors. The relevant variables were the relative proportions of targets in each trial block (varied from 15% to 85%), subjects' RTs when asked to respond as quickly as possible while avoiding errors, and their subsequent confidence judgments about the correctness of their speeded responses. All three methods of measuring recognition memory produced unique aspects in the ROC functions.

Table 9 Probabilities of hits and false alarms across response time (RT) bins

Cumulative RT Bins						
	5	4	3	2	1	Ave
P(H)	.920	.823	.714	.510	.270	.647
P(FA)	.731	.482	.245	.130	.050	.328
d'	.826	1.003	1.316	1.226	1.098	1.094
c	-1.035	-.455	.071	.588	1.167	.067

5 = medium “no” + slow “no” + slow “yes” + medium “yes” + fast “yes”; 4 = slow “no” + slow “yes” + medium “yes” + fast “yes”; etc

The average signal-detection parameter estimates of d' and c are also indicated, as determined from values of these parameters estimated for each subject. The last column on the right reports the average values across RT bins

The manipulation of relative target frequency resulted in predictable results, as the response criterion changed monotonically with the proportion of targets in each block. The subjects used a relatively liberal criterion when targets were frequent and a stricter criterion when targets were rare, closely correlating their overall proportion of “old” responses to the actual target proportions. The resulting ROC curve for the overall data (Fig. 4) was clearly better fit by a linear function than by a curved line, and this conclusion was supported by most but not all of the subjects in different tests applied to individual data. Similar results have been reported by Bröder and Schütz (2009) and others, although curvilinear ROC curves have also been found by researchers using similar

methods (e.g., Dube & Rotello, 2012; Dube, et al., 2012; but see the rejoinder by Bröder & Malejka, 2017). The present results showed changes in the response criterion, c , with changes in target probability in the test lists, but the sensitivity parameter d' was constant across the target frequency manipulation. The inconsistencies observed among individual subjects in our own data, and present also in other studies, render the distinction between discrete-state, threshold theories and continuous signal-detection theories inconclusive.

An examination of the data across confidence judgments yields a different picture. The overall data clearly are better fit by a curved ROC function (Fig. 6), a conclusion supported by the analysis of individual ROC functions. In these data, 38 out of 47 subjects produced ROC functions better fit by SDT than by threshold theory. However, it should be noted that threshold models can predict curved ROC functions with additional assumptions about how the response stage is mapped onto the internal states (e.g., Bröder & Schütz, 2009; Chen, et al., 2015; Kellen, Singmann, Vogt, & Klauer, 2015; Malmberg, 2002). Finally, the analysis across bins determined by RTs yielded data that more closely replicate the traditional SDT findings: smooth, almost symmetrical ROC curves (Fig. 8) that were more consistent than the linear predictions of threshold theory for all 47 subjects. However, these conclusions must be tempered by the relatively poorer fit for all considered models for the RT data, perhaps due to the increased variability of the d' estimates across cumulative RT bins. In this case, d' was neither independent of the decision criterion, c , nor was it monotonically correlated with c , as it was for the data portioned into confidence bins. Rather, d' was greatest at the

Fig. 8 Plots of the average proportions of hits against false alarms for the data from the cumulative response time bins reported in Table 9. The left panel shows the ROC function for the raw response proportion data, and the right panel shows the data for their z-score transforms. The data

points, from left to right, represent bins of increasing size cumulated across response time (RT) bins. The fitted curve and line represent predictions from the homoscedastic version of signal detection theory (SDT)

Fig. 9 Plots of d' (left panel) and c (right panel) averaged across participants for the response time (RT) data shown in Table 9. Hash marks indicate the 95% confidence intervals. The bars in the graphs, from left to right, represent conditions with decreasing numbers of cumulated RT bins

midpoint of the ROC curve, consistent with the “elbow” observed in other derivations of ROC curves from RT data (see Luce, 1986, for a review of such results).

So what can be made of these inconsistencies? A casual inspection of the data indicates that we might be dealing with an artifact. The larger the range of $P(H)$ and $P(FA)$ values, the more curved the data appear. Is it possible that the target proportion manipulation was too weak to induce subjects to change their response criteria sufficiently to demonstrate the extreme values of the ROC function (a point also raised by Dube & Rotello, 2012, and Malmberg & Xu, 2006)? Over any restricted range an arc can be approximated by a straight line. When one looks at the confidence rating data, it can be seen that the ROC curve begins to show its curvilinear tendencies only when the data venture into the upper-right corner of high hit and false-alarm rates. Similarly, the most clearly curved function is that determined by binning the data into RT categories which samples nearly the entire range of hit and false-alarm rates.

This tempting analysis is unfortunately too simple, as there are other differences among the results of the three response measures. The most obvious is that d' is generally higher when measured by confidence ratings than across target proportion manipulations and RT measures (see also Gardner, Macfee & Krinsky, 1975), and it is clearly correlated with c in this case. Further, d' assumed its maximum value for the most certain “yes” category, relative to those that include “probables” and guesses. In contrast, d' is not related in a regular way with c in the RT binning procedure. Rather, d' seems to be at a maximum at the point that divides “yes” and “no” responses, and trails off in either direction as positive or negative responses are made more rapidly. A simple explanation for these results could be that RTs are heavily influenced by familiarity alone, as high or low familiarities lead to fast responses, whereas attempts at recollection for intermediate values produce slower responses (Malmberg, 2008). Still, successful recollection can also lead to highly confident, if slower, responses producing discrepancies between the ROC

Table 10 Parameter estimates and goodness-of-fit statistics for the two two-high-threshold models (2HTMs) and homo- and heteroscedastic versions of signal detection theory (SDT) for the average data arranged into response time bins

2HTM – Asymmetric												
D_o	D_n	FY	MY	SY	SN	MN	FN	G^2	df	p	AIC	BIC
.326	.267	.051	.155	.235	.262	.206	.091	1783	3	<.001	1797	1865
2HTM – Symmetric												
D_o	D_n	FY	MY	SY	SN	MN	FN	G^2	df	p	AIC	BIC
.299	=	.055	.171	.241	.257	.193	.083	1834	4	<.001	1846	1905
SDT – Heteroscedastic												
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$	G^2	df	p	AIC	BIC	
1.13	1.14	1.24	.586	.070	-.473	-1.11	218.2	3	<.001	232.2	299.9	
SDT – Homoscedastic												
d'	σ	$c1$	$c2$	$c3$	$c4$	$c5$	G^2	df	p	AIC	BIC	
1.06	1.00	1.13	.540	.059	-.448	-1.06	334.8	4	<.001	346.8	404.9	

The values FY to FN in the 2HTM models are weights for assigning responses to RT bins for items whose familiarities fall between D_o and D_n ; the RT bins correspond to “Fast Yes” (FY), “Medium Yes” (MY), “Slow No” (SN) etc. Since the weights sum to 1.0, they account for 5 degrees of freedom σ = the standard deviation of the familiarity distribution of target or old items in SDT, which is set to 1.0 for the distractor or new items. The parameters $c1$ through $c5$ indicate the locations of the decision criteria in SDT between the six RT bins

Table 11 Parameter estimates and goodness-of-fit statistics for two versions of the two-high-threshold model (2HTM) and two versions of signal detection theory (SDT) for each subject across response time (RT) bins. The values listed in **bold** font represent the best-fitting model for each subject

Part.	2HTM-Asymmetric				2HTM-Symmetric				SDT-Heteroscedastic				SDT-Homoscedastic			
	G^2	p	AIC	BIC	G^2	p	AIC	BIC	G^2	p	AIC	BIC	G^2	p	AIC	BIC
1	18.92	.000	32.92	73.69	19.24	.001	31.24	66.19	1.186	.756	15.19	55.95	2.710	.607	14.71	49.65
2	85.93	.000	99.93	140.7	86.04	.000	98.04	133.0	12.03	.007	26.03	66.80	15.90	.003	27.90	62.85
3	54.41	.000	68.41	109.2	61.57	.000	73.57	108.5	3.723	.293	17.72	58.49	34.50	.000	46.50	81.44
4	133.5	.000	147.5	188.3	137.7	.000	149.7	184.6	23.24	.000	37.24	78.01	36.99	.000	48.99	83.94
5	34.78	.000	48.78	89.55	35.91	.000	47.91	82.86	3.029	.387	17.03	57.80	3.030	.553	15.03	49.98
6	41.63	.000	55.63	96.40	51.31	.000	63.31	98.26	6.263	.099	20.26	61.03	13.22	.010	25.22	60.16
7	46.93	.000	60.93	101.7	49.07	.000	61.07	96.02	7.197	.066	21.20	61.96	9.955	.041	21.96	56.90
8	24.45	.000	38.45	79.22	24.99	.000	36.99	71.93	3.876	.275	17.88	58.64	3.891	.421	15.89	50.84
9	25.89	.000	39.89	80.66	26.36	.000	38.36	73.30	.4490	.930	14.45	55.22	.4490	.978	12.45	47.39
10	72.35	.000	86.35	127.1	77.47	.000	89.47	124.4	6.808	.078	20.81	61.58	30.97	.000	42.97	77.91
11	21.22	.000	35.22	75.98	21.38	.000	33.38	68.37	2.368	.500	16.37	57.14	2.412	.661	14.41	49.36
12	33.73	.000	47.73	88.50	34.67	.000	46.67	81.61	10.73	.013	24.73	65.50	18.20	.001	30.20	65.14
13	26.88	.000	40.88	81.64	28.46	.000	40.46	75.40	4.968	.174	18.97	59.74	10.27	.036	22.27	57.21
14	44.04	.000	58.04	98.80	46.33	.000	58.33	93.27	2.388	.496	16.39	57.16	7.500	.112	19.50	54.44
15	47.01	.000	61.01	101.8	47.80	.000	59.80	94.75	1.081	.782	15.08	55.85	1.140	.888	13.14	48.08
16	36.22	.000	50.22	90.99	43.91	.000	55.91	90.85	10.62	.014	24.62	65.39	18.71	.001	30.71	65.65
17	29.17	.000	43.17	83.94	29.58	.000	41.58	76.53	3.043	.385	17.04	57.81	3.259	.515	15.26	50.20
18	23.34	.000	37.34	78.11	23.44	.000	35.44	70.38	4.968	.174	18.97	59.74	4.974	.290	16.97	51.92
19	30.34	.000	44.34	85.11	35.56	.000	47.56	82.51	5.993	.112	19.99	60.76	13.15	.011	25.15	60.10
20	180.9	.000	194.9	235.6	187.6	.000	199.6	234.5	11.96	.008	25.96	66.72	39.40	.000	51.40	86.35
21	18.94	.000	32.94	73.71	21.50	.000	33.50	68.44	.4120	.938	14.41	55.18	1.088	.896	13.09	48.03
22	46.88	.000	60.88	101.7	57.80	.000	69.80	104.7	7.064	.070	21.06	61.83	14.75	.005	26.75	61.69
23	44.51	.000	58.51	99.28	44.58	.000	56.58	91.53	5.287	.152	19.29	60.06	5.314	.257	17.31	52.26
24	70.38	.000	84.38	125.1	71.52	.000	83.52	118.5	6.645	.084	20.64	61.41	7.724	.102	19.72	54.67
25	48.68	.000	62.68	103.4	49.06	.000	61.06	96.00	19.52	.000	33.52	74.29	20.66	.000	32.66	67.61
26	64.42	.000	78.42	119.2	66.21	.000	78.21	113.2	12.51	.006	26.51	67.28	12.97	.011	24.50	59.91
27	13.45	.004	27.45	68.22	18.54	.001	30.54	65.49	4.580	.205	18.58	59.35	6.316	.177	18.32	53.26
28	45.68	.000	59.65	100.5	45.76	.000	57.76	92.70	8.722	.033	22.72	63.49	9.269	.055	21.27	56.21
29	48.19	.000	62.11	102.9	55.89	.000	67.89	102.8	12.81	.005	26.81	67.58	24.50	.000	36.05	70.99
30	39.29	.000	53.29	94.06	40.54	.000	52.54	87.49	1.855	.603	15.86	56.62	1.863	.761	13.86	48.81
31	36.72	.000	50.72	91.49	40.83	.000	52.83	87.77	9.365	.025	23.36	64.13	11.87	.018	23.87	58.81
32	12.24	.007	26.24	67.01	13.06	.011	25.06	60.01	1.477	.688	15.48	56.25	1.494	.828	13.49	48.44
33	18.36	.000	32.36	73.13	28.34	.000	40.34	75.29	1.181	.758	15.18	55.95	5.174	.270	17.17	52.12
34	21.35	.000	35.35	76.12	25.40	.000	37.40	72.35	3.466	.325	17.47	58.24	25.66	.000	37.66	72.60
35	22.64	.000	36.64	77.41	22.82	.000	34.82	69.76	3.110	.375	17.11	57.88	5.501	.240	17.50	52.44
36	67.10	.000	81.10	121.9	68.84	.000	80.84	115.8	2.090	.554	16.09	56.86	24.96	.000	36.96	71.90
37	49.13	.000	63.13	103.9	49.16	.000	61.16	96.11	3.590	.309	17.59	58.36	21.09	.000	33.09	68.03
38	18.92	.000	32.92	73.69	36.08	.000	48.08	83.03	3.942	.268	17.94	58.71	19.54	.001	31.54	66.48
39	28.70	.000	42.70	83.47	29.90	.000	41.90	76.84	10.18	.017	24.18	64.95	12.86	.012	24.86	59.81
40	48.92	.000	62.92	103.7	50.06	.000	62.06	97.01	6.689	.083	20.69	61.46	6.966	.138	18.97	53.91
41	32.68	.000	46.68	87.45	38.12	.000	50.12	85.06	2.493	.477	16.49	57.26	6.322	.176	18.32	53.27
42	91.04	.000	105.0	145.8	104.1	.000	116.1	151.1	19.41	.000	33.41	74.18	41.12	.000	53.12	88.07
43	84.44	.000	98.445	139.2	90.34	.000	102.3	137.3	12.24	.007	26.24	67.01	30.43	.000	42.43	77.37
44	59.27	.000	73.272	114.0	68.08	.000	80.08	115.0	6.693	.082	20.69	61.46	22.65	.000	34.65	69.60
45	28.34	.000	42.341	83.11	35.81	.000	47.81	82.76	8.763	.033	22.76	63.53	23.77	.000	35.77	70.71
46	25.42	.000	39.422	80.19	35.96	.000	47.96	82.91	2.113	.549	16.11	56.88	23.12	.000	35.13	70.06
47	17.06	.001	31.055	71.82	22.16	.000	34.16	69.11	1.105	.776	15.10	55.87	14.67	.005	26.67	61.62

Table 12 Summary of model fit statistics for two versions of the two-high-threshold model and signal detection theory for each of the three methods of determining receiver-operating characteristic (ROC) curves from different partitions of the data set

Target proportion blocks				
	2HTM	2HTM	SDT	SDT
	Asymmetric	Symmetric*	Heteroscedastic	Homoscedastic
AIC	3	24	7	13
BIC	0	26	2	19
Confidence ratings				
	2HTM	2HTM	SDT	SDT
	Asymmetric	Symmetric	Heteroscedastic*	Homoscedastic
AIC	5	4	23	15
BIC	3	5	10	29
Response times				
	2HTM	2HTM	SDT	SDT
	Asymmetric	Symmetric	Heteroscedastic*	Homoscedastic
AIC	0	0	29	18
BIC	0	0	16	31

The entries are the numbers of subjects whose data are best fit by the indicated models ($N = 47$). Note: * indicates the best-fitting model for the aggregate data in each case

curves determined from confidence ratings and RTs. The confidence rating data support models that predict asymmetric ROC curves, like the heteroscedastic version of SDT or hybrid SDT and recollection models, like the A-J theory. A similar conclusion has been advocated by Van Zandt (2000) who argued that “...equating confidence with perceived familiarity or likelihood oversimplifies how people estimate confidence” (p. 595). Although some have argued that RTs and confidence ratings measure approximately the same thing, others have shown both theoretically (e.g., Thomas & Myers, 1972) and empirically (e.g., Weidemann & Kahana, 2016) that the overall similarities cannot hide the fact that d' is usually higher after partitioning data into confidence categories than into RT bins. Further, there are idiosyncrasies associated with ROC curves derived from RT binning procedures that can also affect model-fitting (see, e.g., Luce, 1986, pp. 227–8).

Despite some similarities in the ROC functions determined by confidence rating and RT binning procedures, the differences in the d' and c parameters across the two measurement methods indicate that there is not a direct relationship between confidence and RT. In fact, when correlations were calculated between confidence ratings and RTs for the data from individual subjects in the present study, the average correlation was .281 (exceeding the value of .074 which would be significant at the .05 level for 500 pairs of observations), and the r value was significant for 41 of the 47 participants. Still these correlations are low enough that it is obvious that not all rapid responses are confident, and some confident responses result from slow processes. The prediction from hybrid theories that

allow for slower, recollection processes to produce highly-confident responses, especially for old items, could result in the “elbow” in the ROC function for the RT binning procedure representing relatively higher d' values for slower responses than for faster ones.

It should be pointed out that in the context of the A-J model, we presume that familiarity can lead to faster responses, on average, than recollection-based responses (Malmberg, 2008; see especially his application of Shiffrin & Steyvers' (1997) REM model to RT data). This interpretation is also consistent with the “butcher on the bus” anecdote provided by Mandler (1980), in which, when confronted by someone we have seen before, but in a new context (e.g., on a bus) we might first experience a feeling of familiarity, subsequently followed by retrieval of contextual information (a supermarket), and finally recollection of a specific memory (he's the butcher at the supermarket). However, there are other models that presume that the two processes occur in parallel, and that in at least in some cases, recollections might operate before familiarity drives a response (Brainerd, Nakamura, & Lee, 2019; Kellen, Klauer, & Bröder, 2013; Yonelinas, 1997).

In the present case, we believe that high and low familiarities can lead to rapid responses that are usually correct, and slower responses for intermediate familiarities result from attempts at test item recollections that are successful only on a minority of (target) trials. Still, when recollection fails, and unlike the discrete-state threshold models, the A-J model allows for familiarity to guide decisions even in the uncertain region between the two criteria of the model. Only for a very small proportion of trials (about 6.6% in the present case) do subjects seem to guess in a truly random way when both familiarity and recollection fail to discriminate targets from distractors.

We are therefore left with the conclusion that different methods of measuring recognition memory can produce different results, but these differences are not beyond reasonable theoretical interpretation. Fast responses seem to represent a continuous underlying representation of memory strength, and more reasoned responses, such as later indications of confidence, can access supporting evidence from recollections of episodic information stored during the study period. Methods that encourage subjects to produce responses across the range of possible outputs seem to capture a better picture of recognition memory than those based on more constricted domains. Further, analyses of individual participants' data indicate that different strategies or abilities can be obscured by model fitting that examines only aggregate data. It is also likely that participants choose to engage different control processes in order to deal efficiently with variations in task demands and testing conditions (Malmberg, 2008; Malmberg, Raaijmakers, & Shiffrin, 2019). In this regard, it could be argued that: (1) the target proportion manipulation across trials blocks produces a global strategy affecting mainly response bias, (2)

the instructions to respond rapidly emphasizes the role of familiarity in decision-making, and (3) the instructions to add a confidence rating after the speeded response leads to a relatively larger role for recollection in the confidence rating data. If these task variables indeed induce the application of control processes in different stages of the search and decision processes in recognition memory, then it is not surprising that the data tell different stories about behavioral components reflected in the ROC curves generated by the different testing manipulations.

In conclusion, the A-S theory of memory and its extension by the A-J model and other hybrid models involving combinations of SDT and recollective search processes seem flexible enough to remain viable explanations of search and decision processes in human memory (Klauer & Kellen, 2015). The present demonstration that such models can predict the effects of varying target proportions in a test list on the slopes of the zROC curves determined from confidence ratings, and the way in which d' changes across confidence levels, are cases in point. That is not to say that alternative models based on random walks, evidence accumulators, diffusion processes, or multinomial processing trees (e.g., Dube, et al., 2012; Province & Rouder, 2012; Van Zandt, 2000) might ultimately be more successful in accounting for some specific details of recognition memory data, at least at the level of measurement or pragmatic models (Bröder, et al., 2013; Malmberg, 2008). This level of successful data-fitting can exist even if most measurement models lack the elegance and generality of theories of memory storage, retrieval, and decision processes such as the epistemic process models central to the work of Atkinson, Shiffrin, and their colleagues (Atkinson & Shiffrin, 1968; 1971; Atkinson & Juola, 1973, 1974; Shiffrin & Steyvers, 1997; Cox & Shiffrin, 2017).

Author Note JFJ acknowledges support from a research fellowship (2016-T3/SOC-1544) provided by the Community of Madrid, Spain, while the present research was carried out. We are grateful for the useful comments provided by Arndt Bröder, Daniel Heck, Kenneth Malmberg, John Wixted, Andrew Yonelinas, and an anonymous reviewer on an earlier version of this paper. A complete data file for the present research and its description are available on the following link: <https://osf.io/y78mk/>

References

- Alonso, M. A., Fernández, A., Díez, E., & Beato, M. S. (2004). Índices de producción de falso recuerdo y falso reconocimiento para 55 listas de palabras en castellano. *Psicothema*, 16, 357–362.
- Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language*, 85, 27–41.
- Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and Performance IV*. New York: Academic Press.
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. Krantz, R. Atkinson, R. Luce, and P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology*, Vol. 1, San Francisco: W.H. Freeman.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 2). New York: Academic Press.
- Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, 225, 82–90. doi:<https://doi.org/10.1038/scientificamerican0871-82>.
- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, 14, 81–99.
- Botella, J., Privado, J., Suero, M., Colom, R., & Juola, J. F. (2019). Group analyses can hide heterogeneity effects when searching for a general model: Evidence based on a conflict monitoring task. *Acta Psychologica*, 193, 171–179.
- Brainerd, C. J., Nakamura, K., & Lee, W.-F., (2019). Recollection is fast and slow. *Journal of Experimental Psychology: Learning Memory and Cognition*.
- Bröder, A., Kellen, D., Schütz, J., & Rohmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 21, 916–944.
- Bröder, A. & Malejka, S. (2017). On a problematic procedure to manipulate response biases in recognition experiments: The case of “implied” base rates. *Memory*, 25, 736–743.
- 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.
- Callejas, A., Correa, A., Lupiáñez, J., & Tudela, P. (2003). Normas asociativas Intracategoriales para 612 Palabras de Seis Categorías Semánticas en Español. *Psicológica*, 24, 185–214.
- Chen, T., Stams, J. J., & Rotello, C. M. (2015). A Violation of the conditional independence assumption in the two-high-threshold model of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1205–1225.
- Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, 15(4), 692–712.
- Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, 124, 795–860.
- Dube, C. & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 130–151.
- Dube, C., Stams, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language*, 67, 389–406.
- Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie*. Leipzig: Dunker and Humboldt. (Translated by H. A. Ruger & C. E. Bussenius, 1913, and reissued by Dover Publications, 1964.)
- Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCAC-TN-58-51). Bloomington: Indiana University, Hearing and Communication Laboratory.
- Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, 12(3), 403–408.

- Fechner, G. T. (1966). *Elements of Psychophysics* (Vol. 1.). Translated by H. E., Adler, D. H. Howes & E. G. Boring. New York: Holt, Rinehart, & Winston. (Original work published in 1860.)
- Gardner, R. M., Macfee, M., & Krinsky, R. (1975). A comparison of binary and rating techniques in the signal detection analysis of recognition memory. *Acta Psychologica*, 39, 13–19.
- Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
- Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition*, 36, 933–946.
- Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1355–1369.
- Jang, Y., Wixted, J. T., & Huber, D. E. (2011). The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory. *Psychonomic Bulletin & Review*, 18, 751–757.
- Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics*, 10, 8–14.
- Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, 122, 542–557. doi:<https://doi.org/10.1037/a0039251>.
- Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin and Review*, 20, 693–719.
- 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.
- Klauer, K. C., & Kellen, D. (2015). The flexibility of models of recognition memory: The case of confidence ratings. *Journal of Mathematical Psychology*, 67, 8–25.
- Luce, R. D. (1986). *Response Times*. New York: Oxford University Press.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd). New York: Psychological Press.
- Malejka, S., & Bröder, A. (2019). Exploring the shape of signal-detection distributions in individual recognition ROC data. *Journal of Memory and Language*, 104, 83–107.
- Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 380–387.
- Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, 57, 335–384.
- Malmberg, K. J., Raaijmakers, J. G. W., & Shiffrin, R. M. (2019). 50 years of research sparked by Atkinson and Shiffrin (1968). *Memory & Cognition*.
- Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review*, 13, 99–105.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252–271.
- McAdoo, R. M., Key, K. N., & Gronlund, S. D. (2019). Task effects determine whether recognition memory is mediated discretely or continuously. *Memory & Cognition*.
- Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, 114, 188–202.
- Parks, T.E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review*, 73, 44–58.
- Pérez Dueñas, C., Acosta, A., Megías, J. L., & Lupiáñez, J. (2010). Evaluación de las dimensiones de valencia: Activación, frecuencia subjetiva de uso y relevancia para la ansiedad. La depresión y la ira de 238 sustantivos en una muestra universitaria. *Psicológica*, 31, 241–273.
- Province, J. M., & Rouder, J. N. (2012). *Proceedings of the National Academy of Sciences of the United States of America*, 109, 14357–14362.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychological Bulletin & Review*, 4, 145–166.
- Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of Multinomial Processing Tree models with R. *Behavior Research Methods*, 45, 560–575.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117, 34–50.
- Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153, 652–654.
- Sternberg, S. (2016). In defence of high-speed memory scanning. *Quarterly Journal of Experimental Psychology*, 69, 10, 2020–2075. <https://doi.org/10.1080/17470218.2016.1198820>.
- Swets, J.A., Dawes, R.M., & Monahan, J.M. (2002). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1–24.
- Thomas, E. A. C., & Myers, J. L. (1972). Implications of listing data for threshold and nonthreshold models of signal detection. *Journal of Mathematical Psychology*, 9, 253–285.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory*. New York: Academic Press.
- Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 582–600.
- Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, 3: 150670. <https://doi.org/10.1098/rsos.150670>.
- Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford University Press.
- Wixted, J. T. (2007). Dual-process theory and signal detection theory of recognition memory. *Psychological Review*, 114, 152–176.
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1341–1354.
- Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, 25, 747–763.
- Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, 133(5), 800–832.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.