Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 3, 587–606

0278-7393/09/$12.00 DOI: 10.1037/a0015279

Recognition ROCs Are Curvilinear—or Are They? On Premature Arguments Against the Two-High-Threshold Model of Recognition

Arndt Bröder
University of Bonn and Max Planck Institute
for Research on Collective Goods

Julia Schütz
University of Bonn

Recent reviews of recognition receiver operating characteristics (ROCs) claim that their curvilinear shape rules out threshold models of recognition. However, the shape of ROCs based on confidence ratings is not diagnostic to refute threshold models, whereas ROCs based on experimental bias manipulations are. Also, fitting predicted frequencies to actual data is a more sensitive method for model comparisons than ROC regressions. In a reanalysis of 59 published data sets, the 2-high-threshold model (2HTM) fit the data better than an unequal variance signal detection model in about half of the cases. Three recognition experiments with experimental bias manipulation were conducted that yielded linear ROCs and a better fit of the 2HTM in all cases. On the basis of actual data and a simulation, the authors argue that both models are at least equally valid as measurement tools and can perhaps be integrated theoretically.

Keywords: recognition, signal detection, threshold models

One of the oldest methods for testing episodic memory is the recognition method, in which stimuli that have been presented before in a learning list (targets) are later mixed with unlearned items (distractors) and participants have to discriminate old items from new ones. It has been acknowledged very early that deriving valid measures of memory performance from this method can be tricky. The probabilities of responding "old" to old items (hit rate [HR]) or to new items (false alarm rate [FAR]) will be influenced by at least two psychological factors, namely the genuine memory performance and strategic guessing tendencies (biases). Whereas the former factor depends on the participant's ability, item complexity, encoding time, retention interval, and so forth, the latter will be influenced by expectations about the proportion of targets in the test, payoff conditions, or motives of the participant. Clearly, the memory researcher wants a measure of memory capacity that is not contaminated by strategic guessing biases, and different measurement models have been proposed.

In this article, we focus on signal detection theory (SDT) and the two-high-threshold model (2HTM) as potential candidates to solve the problem of disentangling the processes. This endeavor may appear futile at first glance because threshold models have recently been claimed dead in influential publications (Wixted, 2007a, 2007b; Yonelinas & Parks, 2007). However, we challenge this claim because it is based on questionable, although abundant, empirical evidence.

First, we describe the contestants SDT and the 2HTM as well as their predictions concerning so-called receiver operating characteristics (ROCs). Second, following earlier arguments by Erdfelder and Buchner (1998) and Malmberg (2002), we criticize the evidence commonly invoked to refute the threshold model. Third, we report a reanalysis of 59 published data sets with experimental bias manipulations and we fit SDT and the 2HTM directly to the data, which show no clear superiority of either model. Fourth, we report three new recognition experiments with different materials and bias manipulations that show a clear advantage of the 2HTM over SDT in data fitting.

SDT, 2HTM, and Their ROC Predictions

Because the discrimination between targets and foils in a recognition test is structurally similar to a psychophysical detection task including blank trials and signal trials, it has been proposed to model them in a similar way (e.g., Banks, 1970; Egan, Schulman, & Greenberg, 1959; Kintsch, 1967). For example, SDT assumes that the judgment of a participant is based on an internal evidence continuum, often referred to as *familiarity*. We prefer the epistemically more neutral term *evidence strength* because it does not imply a theoretical position about whether this evidence is based on a single process (i.e., familiarity) or on several processes (e.g., familiarity and recollection). A good fit of the SDT model may be compatible with either theoretical position (see Wixted, 2007a). Figure 1 (left panel) shows that targets as well as foils can produce feelings of subjective evidence that the item is old. However, the probability distributions overlap, which implies various degrees of uncertainty. The distance between the distribution means d' (measured in standard deviation units of the foil distribution) is a measure of discriminability and hence, the memory performance. The participant has to place a criterion of evidence strength that

Arndt Bröder, Department of Psychology, University of Bonn, Bonn, Germany, and Max Planck Institute for Research on Collective Goods, Bonn, Germany; Julia Schütz, Department of Psychology, University of Bonn.

This research was supported by Deutsche Forschungsgemeinschaft (DFG) Grant No. BR 2130/3-1. We thank Kristina Adolphs, Sarah Kreuz, and Helen Sauer for their help with the data collection and Lutz Cüpper for teaching us tricks in R. Ute Bayen and Jürgen Bredenkamp provided valuable comments. We are grateful to Trisha van Zandt and Tim Curran for providing the raw data of their studies.

Correspondence concerning this article should be addressed to Arndt Bröder, Department of Psychology, University of Bonn, Kaiser-Karl-Ring 9, D-53111 Bonn, Germany. E-mail: broeder@uni-bonn.de

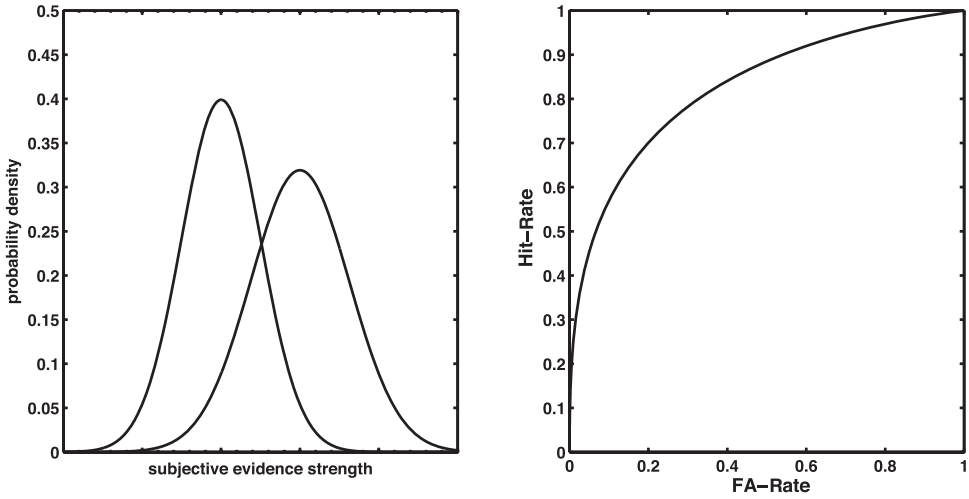

Figure 1. The left panel shows the distributions of subjective evidence for distractors and targets. The targets produce larger subjective evidence on average, and their distribution variance is often estimated to be larger than that of foils. The right panel shows a theoretical ROC that results if the response criterion is moved from left to right in the left panel. FA = false alarm.

suffices for him or her to say "old" to an item. Depending on whether he or she needs a lot of evidence, the criterion will be strict or liberal, which obviously affects HRs and FARs simultaneously. If pairs of HRs and FARs generated by moving the response criterion from left to right are plotted against each other, the theoretical isomemory function in the right panel of Figure 1 results. This function is also referred to as the *receiver operating characteristic* (ROC) or sometimes as the *memory operating characteristic*. If the ROC is plotted in probability space (as in Figure 1, right panel), a curved function will emerge, whereas a plot in z-space (z-transformed probabilities) will be linear when normal distributions are assumed. Asymmetries of the ROC are commonly observed (e.g., Ratcliff, Sheu, & Gronlund, 1992), which can be explained by SDT if one allows for unequal variances of the two normal distributions. Hence, d' can be used as a measure of memory performance that is uncontaminated by guessing biases, given that the SDT model is valid. If one wants to measure the bias as well, several indices exist, for example C, the distance of the criterion from the distractor distribution mean, or logβ, the logarithm of the ratio of probability densities of both distributions at point C. We refer the reader to Macmillan and Creelman (2005) or Wickens (2002) for formulas to obtain the measures and to Snodgrass and Corwin (1988) for arguments as to why these bias indices are preferable to others.

Like SDT, the 2HTM is agnostic about the exact nature of the underlying memory processes producing internal evidence. However, it is assumed that whenever the evidence produced by a target exceeds a threshold (with probability p_o), this results in a state of "target detection," which is always accompanied by an "old" response. The model is depicted in Figure 2 (left panel). A distractor may produce evidence crossing another threshold with probability p_n, which leads to a state of "distractor detection" accompanied by a "new" response. If an item crosses neither threshold, a state of uncertainty results in which strategic guessing processes have to be invoked that can be affected by various biases, expectations, payoff schedules, and so forth. Targets can never cross the "new" threshold, whereas distractors can never cross the "old" threshold. This is why both thresholds are referred to as "high" thresholds.[1] An earlier threshold model that did not include a distractor detection component is known as the one-high-threshold model (Blackwell, 1963) and can be derived from the 2HTM by setting $p_n = 0$. However, it has been dismissed in validation studies (Snodgrass & Corwin, 1988), and it cannot account for the *mirror effect* (e.g., Glanzer, Adams, Iverson, & Kim, 1993), which describes the increase of the HR and the simultaneous decrease of the FAR when memory performance improves. Also, metacognitive processes obviously contribute to the firm rejection of distractors (e.g., Strack & Bless, 1994). Theoretical isomemory ROCs for biases varying from $b = 0$ to $b = 1$ are depicted in the right panel of Figure 2. It is easy to see that this ROC is linear; increments unequal to 1 will result whenever $p_o \neq p_n$. In the 2HTM, p_o and p_n are natural measures of memory performance, uncontaminated by bias, which is measured by the probability b of saying "old" in a state of uncertainty. When probabilities are transformed into standard normal z values, the corresponding ROCs in z-space will be curved upward.

Conceptually, the models differ in one important aspect: According to SDT, every single judgment is influenced by the magnitude of evidence strength relative to the criterion, whereas in the 2HTM, the items not crossing an evidence threshold are exclusively responded to using strategic guessing. Both models predict ROCs of different shapes, so it should be easy to decide between them empirically—or is it?

[1] If viewed from the theoretical point of view involving internal evidence strength, it may be more natural to refer to the 2HTM as a *high-low-threshold model* (e.g., Yonelinas & Parks, 2007), but we stick to the term used in most of the literature.

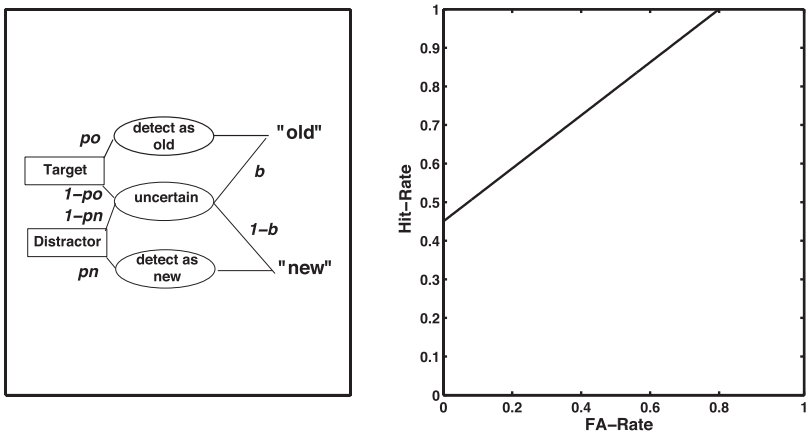

Figure 2. The left panel shows the two-high-threshold model (2HTM) of recognition. Targets cross the detection threshold with probability p_o; foils cross another detection threshold with probability p_n. With the respective complementary probabilities, a state of uncertainty is reached which results in an "old" response with probability b reflecting bias and strategic guessing. The right panel shows a theoretical ROC when bias is varied from 0 to 1. The hit rate at FA = 0 is the estimate for p_o; the FA rate at hit = 1 is the estimate of $(1 - p_n)$. ROCs have a slope of 1.0 when $p_o = p_n$. FA = false alarm.

Reviews of ROC Evidence

In a recent debate about appropriate formal models of recognition memory, Wixted (2007a, 2007b) and Yonelinas and Parks (2007) examined a large number of recognition studies in which ROCs were produced. Although these authors strongly disagreed about the appropriate model of recognition memory, they agreed on one particular conclusion: The generally curvilinear shape of the ROCs refutes threshold models. For example, Wixted (2007a) claimed that "because every recognition memory ROC analyzed between 1958 and 1997 was curvilinear, the high threshold model was abandoned in favor of signal-detection theory" (p. 153). Similarly, Yonelinas and Parks (2007) concluded, "The shape of the ROCs observed in item recognition indicates that threshold models are unable to account for recognition memory performance, and they provide support for signal detection models and hybrid models" (p. 814). The same conclusion has been drawn in the domain of source memory, where "source ROCs are typically curvilinear," too (Qin, Raye, Johnson, & Mitchell, 2001, p. 1110; see also Banks, 2000; DeCarlo, 2003a, 2003b; Glanzer, Hilford, & Kim, 2004; Slotnick & Dodson, 2005; Slotnick, Klein, Dodson, & Shimamura, 2000). This argument is therefore used to dismiss multinomial models of source recognition that are based on the 2HTM. Given this vast amount of empirical evidence, there seems to be no doubt that the case is settled in favor of SDT.[2]

We conjecture that the ROC analyses forming the foundation of this argument are simply invalid. They are for the most part based on confidence ratings, which are nondiagnostic for rejecting threshold models. Hence, we want to emphasize that the conclusion—although based on a large database—may nevertheless be premature and misleading.

Why the Shape of Confidence Rating ROCs Does Not Rule Out Threshold Models

Obtaining ROCs by manipulating bias via payoffs or proportions of item types is laborious, tedious, and expensive: Participants will not understand payoff instructions, at least five groups are needed to produce a curve, and the groups have to be of reasonable size to get stable point estimates. Whereas this is easier to achieve with simple perceptual detection experiments involving a few hundred trials, memory experiments are limited to a reasonable amount of information people can remember, demanding for even more participants. Therefore, confidence rating scales have been used widely because they come as a cheap and easy remedy here. Participants simply judge their confidence that an item was old (vs. new) with k rating categories labeled accordingly. It is assumed that people set their own $k - 1$ decision criteria along the evidence strength dimension. Hence, the experimenter needs only one group and no bias manipulation to get a ROC curve, which is based on cumulated HRs and FARs. The predictions for the form of the curve derived from SDT are identical to those for an experimental bias manipulation. However, contrary to the reasoning used by Wixted (2007a, 2007b), Yonelinas and Parks (2007), Slotnick and Dodson (2005), and others, this is not true for the 2HTM. This model assumes only three discrete latent states: two "detect" states and one "uncertain" state. These states easily map on a binary yes–no response scale, but what follows for a rating scale?

The 2HTM is not originally formulated for ratings, and to derive predictions for the ROC, one has to make assumptions about the *response function* that maps the three memory states onto the response options of a rating scale. The strong prediction of linear ROCs holds *if and only if* one assumes that a detect state always

[2] Multinomial processing tree models are theory-based measurement tools that assume discrete latent cognitive states that can be reached with unknown transition probabilities (see Batchelder & Riefer, 1999, for an extensive overview). The probabilities are measures of the latent processes and can be estimated from experimental data. The 2HTM is a very simple member of this flexible formal model class; conceptual extensions of the 2HTM include discrete state measurement models of source memory (see Bröder & Meiser, 2007, for an overview).

results in the most extreme rating (*certainly old* or *certainly new* for old and new items, respectively). Only in this case, the threshold model implies linear ROCs, irrespective of the response function assumed for the uncertain state. However, it is well known that people often refrain from using extreme judgments on rating scales, that they are influenced by arbitrary numeric anchors (N. Schwarz, Knauper, Hippler, & Neumann, 1991), and that they tend to distribute their judgments across the scale (Herrmann, 1960). Hence, it is psychologically plausible that confidence ratings will not be used in a completely deterministic manner even in detect states. Especially in situations with good discrimination between distractors and targets, participants will probably avoid giving extreme responses exclusively and spread their judgments across the scale.

Malmberg (2002) and Erdfelder and Buchner (1998) have clearly demonstrated that threshold models can generate curvilinear ROCs when participants do not use the rating scale in a strictly deterministic manner in detect states. Even if there is only a portion of individuals with conservative response tendencies (avoid extreme ratings), the averaged ROC will be curved. An illustrative example for a 6-point rating scale is provided in Figure 3. These ROCs were generated assuming a threshold process and a response mapping in which a detect state is accompanied by a .70 probability of an extreme rating and a .30 probability of a less extreme rating. In the state of uncertainty, middle ratings are preferred. Participants with a conservative response style may thus produce curved ROCs, and aggregating across individuals can also lead to curvilinear confidence ROCs even if a threshold process was the basis of recognition.

Malmberg (2008) has emphasized that single item recognition, associative recognition, and source memory must be distinguished and show somewhat different ROC properties. However, the reasoning against confidence ratings also applies to source memory, where the argument against discrete state models based on curvilinear ROCs is equally invalid (see Qin et al., 2001; Slotnick & Dodson, 2005; Slotnick et al., 2000).

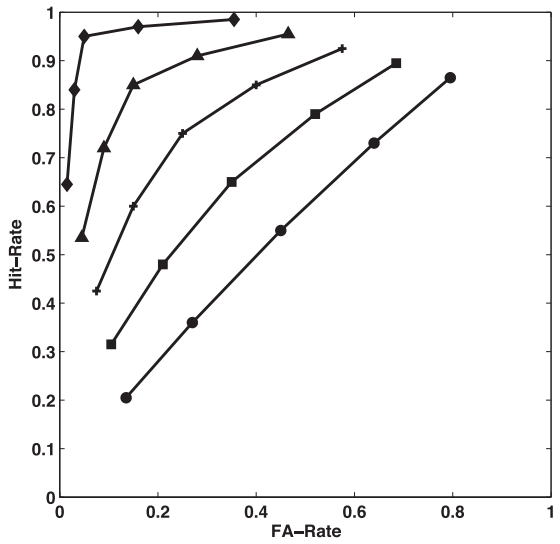

Figure 3. ROCs generated for a 6-point rating scale with the two-high-threshold model (2HTM) for detection probabilities $p_o = p_n = .1, .3, .5, .7,$ and .9 (lowest through highest curves, respectively). It is assumed that participants prefer extreme ratings in "detect" states with probability .7 and less extreme ratings with probability .3. In the state of uncertainty, middle ratings are preferred. FA = false alarm.

Therefore, if one accepts the psychologically plausible and empirically validated notion that response tendencies for confidence scales exist, it follows that the curvilinearity of confidence ROCs cannot be a valid argument against the 2HTM. If one wants ROCs for discriminating between the models, one has to go the laborious way and manipulate response biases experimentally and use a binary answer format. Confidence ratings are cheap and easy, but they are quick and dirty as well.

There are two additional arguments against using rating ROCs for model evaluation. First, hit–false alarm pairs are created using cumulated rating frequencies. Hence, the data points on the ROC are not independent realizations of a random variable, which is the assumption behind the regression procedures typically used to estimate the parameters and assess the model fit (Wickens, 2002). For example, the cumulative procedure forces the HR to increase monotonically with the FAR, thereby truncating potential "downward" random errors. Second, the fit of a linear ROC in *z*-space (which implies a curvilinear ROC in normal probability space) is often used to assess SDT's fit to the data. However, as Van Zandt (2000) observed in numerous simulations, the linearity of *z*-transformed ROCs seems to be a very robust phenomenon, almost independent of the distribution form assumed in SDT (see Van Zandt, 2000, footnote 2). Hence, a failure to find a curvature in the *z*-transformed ROC does not seem to be a strong argument in favor of SDT with normal distributions.[3]

Reexamining Studies With Experimental Bias Manipulations

Because the curvilinearity of ROCs based on confidence ratings does not help to refute threshold models, we searched the literature for recognition studies that involved actual manipulations of response bias via payoffs or the (expected) proportion of old items in the recognition test. Although these methods are described as standard methods to manipulate bias while keeping discrimination constant (Macmillan & Creelman, 2005; McNicol, 1972; Wickens, 2002), we were surprised to find only a limited set of studies. Forty-one data sets in eight publications manipulated bias in two steps (Berch, 1977; Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995; Curran, DeBuse, & Leynes, 2007; Estes & Maddox, 1995; Healy & Jones, 1975; Healy & Kubovy, 1978; Marken & Sandusky, 1974; Rhodes & Jacoby, 2007). Note, however, that an ROC with two data points can be perfectly fit by any curve or straight line. Hence, ROC analyses are impossible. We found 13 data sets in five articles that used a three-step bias manipulation (Allen & Garton, 1969; Curran et al., 2007; Heit, Brockdorff, & Lamberts, 2003; Henriques, Glowacki, & Davidson, 1994; Snodgrass & Corwin, 1988). Although three points cannot be trivially fit by a line, evidence based on one additional error-prone data point is of course not very strong. Finally, we found 13 data sets in three publications with more than three steps of bias

[3] The 2HTM can also be represented as a signal detection model with rectangular distributions of evidence strength.

manipulation, which allows for constructing diagnostic ROC curves (Parks, 1966; Ratcliff et al., 1992; Van Zandt, 2000).

Rather than evaluating the ROCs via the usual regression statistics including a quadratic component to assess curvature, our aim was to fit the models directly to the raw frequencies. With this method, it is also possible to assess the model fits of data sets that include only a two-step manipulation of response bias because we can compare predicted and actual frequencies using the likelihood ratio statistic G^2 as a goodness-of-fit statistic. The two independent HRs and two FARs provide 4 degrees of freedom in the data, which allows for estimating up to four free parameters for each model (see below). Maximum likelihood parameter estimates can be obtained by minimizing G^2 (Wickens, 2002). Although SDT and the 2HTM are nonnested models that cannot be tested against each other directly, one can compare the goodness of fit of both models across a range of studies. The advantages of the method compared to ROC analyses are (a) a greater statistical power to detect model violations and (b) the possibility to include studies with two- and three-step bias manipulations.

Method

We identified 59 data sets in the literature that manipulated response bias in recognition via payoffs or base rates in recognition experiments. The studies of Allen and Garton (1969) and Heit et al. (2003) were excluded because we obtained degenerated parameter estimates (d' or Cs $> 1{,}000$). Maybe this is due to the fact that they investigated people's ability to vary their response criterion trial-by-trial as a function of the length of time given to process the test stimuli (Heit et al., 2003), or as a function of word type (Allen & Garton, 1969). Original raw data were provided by Van Zandt (2000) and for Curran et al.'s (2007) third experiment. For all other experiments, we reconstructed "ideal" raw data from the information given in the experiment descriptions. For example, knowing the HRs and FARs, the number of participants, and the number of items in an experimental condition allowed us to reconstruct the raw frequencies of the responses. Note, however, that these reconstructed data sets are "ideal" in the sense that we obviously were not able to reconstruct the distribution of missing values, if the authors mentioned any. SDT and the 2HTM were then fit to the data with the following specifications: Constant sensitivity parameters (d' and p_o/p_n, respectively) were assumed across bias manipulation conditions, whereas the bias parameters (C_i and b_i, respectively) were allowed to vary.

The data sets with two steps of bias manipulation provide four independent data categories (two HRs and two FARs). Hence, we fit three-parameter versions of SDT and the 2HTM, assuming equal variances in SDT and $p_o = p_n$ in the 2HTM. These restrictions force both models to predict symmetric ROCs. This leaves 1 degree of freedom, and the test statistic G^2 is asymptotically chi-square-distributed with 1 degree of freedom. Data sets with three steps of bias manipulation yield six independent data categories, which allows us to estimate unequal variances in SDT and $p_o \neq p_n$ in the 2HTM, leaving five estimated parameters for each model and, hence, 1 degree of freedom again. Five-step variations of bias yield 10 independent data categories. Estimating seven parameters for each model (d', SD, C_1 to C_5 for SDT and p_o, p_n, b_1 to b_5 for the 2HTM) leaves 3 degrees of freedom.

Parameter Estimation and Model Fitting

We estimated parameters by minimizing G^2 using the routine "optim()" in R (Version 2.4.1; R Development Core Team, 2006). The routine for SDT was cross-checked with the Solver function in Microsoft Excel 2003; the 2HTM routine was compared to results from the software AppleTree (Rothkegel, 1999) for analyzing multinomial models. Results obtained with different methods were identical to the third decimal digit. Multiple starting values were used to avoid local minima.

Results

The model fits and parameter estimates for d', SD, p_o, and p_n are provided in the Appendix. The median G^2 statistic was 2.58 for SDT and 2.87 for the 2HTM. This difference is not significant according to a Wilcoxon test ($Z = 0.11$, $p = .92$).[4] The 2HTM fit the data better in 32 of 59 cases (54%); SDT fit better in the other 27 cases (46%). Figure 4 shows the scatterplot of the G^2 values of both models. According to a conventional significance level of .01, both models did not fare too well: SDT was rejected 15 times, the 2HTM 16 times. Note, however, that some data sets were huge and yielded a high power to detect even tiny model violations. To take this into account we computed compromise power analyses with effect size $w = .10$ (a small effect according to Cohen, 1988) and a β/α ratio $= 1$ for each data set, resulting in 9 rejections of SDT and 11 rejections of the 2HTM. Across the 59 studies, the estimates of d' and p_o correlated $r = .91$.

Discussion

We used a strict test to fit the contestant models to existing data sets examining their ability to reproduce the actual response frequencies. On the basis of a conventional significance criterion, neither model did very well, but taking statistical power into account, the picture looks much better for both models. In any case, there was no apparent advantage for either model, and thus, the available data sets that include a manipulation of response bias at present do not justify a rejection of the 2HTM in favor of SDT.

Hence, the apparent refutation of the 2HTM based on confidence rating data is obviously not supported when ROCs are generated by bias manipulations. One encouraging result from the pragmatist's point of view was the almost perfect convergent validity of SDT's and the 2HTM's sensitivity measures across studies ($r = .91$).

It is quite clear from Figure 4 that the largest misfits of both models are generated by the data sets using five steps of bias manipulation. Typically, these studies collected more data and therefore had a higher statistical power. Apart from that, increasing the number of bias groups boosts the probability that accuracy varies across the groups either due to a lack of experimental control or by chance. Because both models used here assume

[4] The means were 7.49 for SDT and 13.21 for the 2HTM. The higher mean of the 2HTM is driven by three exceptionally bad fits for data sets in which both models were rejected. A t test on log scores of G^2 also did not reveal a difference between the models, $t(56) = 0.90$, $p = .37$.

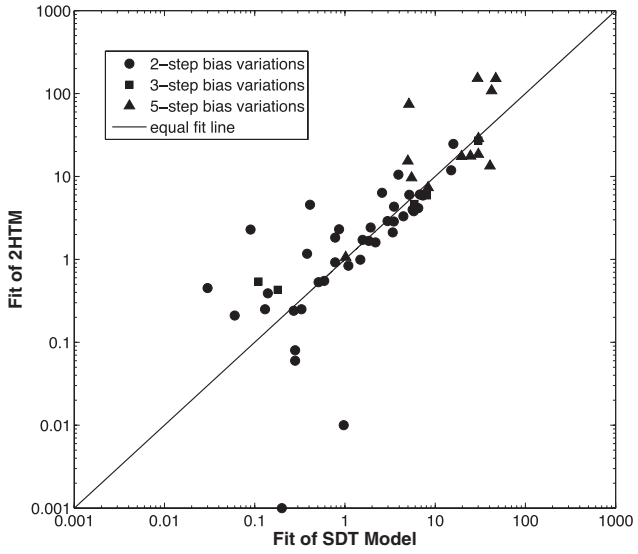

Figure 4. Goodness-of-fit values for signal detection theory (SDT) and the two-high-threshold model (2HTM) in 59 reconstructed data sets. Model versions with equal numbers of free parameters were used in each case. Points below the diagonal indicate a better fit of the 2HTM; points above the diagonal indicate a better fit of SDT. Note the logarithmic axes.

constant discrimination, they both suffer from any violation of this assumption.[5]

The reanalysis of already existing data sets is limited in several ways. First, the data used were only reconstructed. Although it seems improbable that the deviation from the actual data was systematic and distorting, this possibility cannot be ruled out. Second, we mixed studies with very different materials, learning and testing conditions, instructions, and bias manipulations. For example, some studies used a continuous recognition paradigm (Berch, 1977; Estes & Maddox, 1995; Marken & Sandusky, 1974; Parks, 1966) in which participants decided in each trial whether they had encountered the item before. It is unclear whether such a mixed learning and testing procedure entails the same processes as standard recognition (Drake & Hannay, 1992; Richardson, 1994). In addition, Curran et al. (2007, Experiment 3), Healy and Jones (1975), and Van Zandt (2000) asked their participants to give rating responses. It may be possible that binary and rating response techniques lead to different parameter estimates of d' (Gardner, Macfee, & Krinsky 1975; Grasha, 1970).

For these reasons, we conducted three new standard recognition experiments in which we took care to ensure identical learning procedures in each condition.

Experiment 1

The goal of the experiment was to vary response tendencies in five steps and to keep encoding and testing conditions equivalent otherwise. Hence, we expected that only the bias parameter of the respective models would vary across conditions, whereas the sensitivity parameter was assumed to be constant.

Method

Design and procedure. In a standard recognition experiment, the response bias at test was varied between subjects in five steps. Participants learned a list of 60 words that were presented in succession for 2.5 s each on the computer screen. After a 3-min filler task (mental rotation), a test with 60 words was administered that consisted of 10%, 30%, 50%, 70%, or 90% old words and the respective complementary percentages of distractors. Because Rhodes and Jacoby (2007; see also Estes & Maddox, 1995; Healy & Kubovy, 1978) have demonstrated that criterion shifts in recognition memory are more likely when participants are aware of the base rates and when feedback on old–new decisions is provided, an instruction made perfectly clear how many old words were to be expected in the test. During the test, participants received trial-by-trial feedback about the correctness of each response, and points were added to (or subtracted from) a virtual account.[6] Points could be exchanged for candy afterward.

Materials. A pool of 120 German nouns of 4 to 9 letters was selected from a collection of norms on concreteness in Hager and Hasselhorn (1994). These norms comprise ratings from –20 (*very abstract*) to 20 (*very concrete*). We selected 120 concrete nouns (mean ratings > 5), for example, *hotel*, *lettuce*, and *child* (translated from German). For each participant, 60 were drawn randomly for the learning list. From these, 6, 18, 30, 42, or 54 were drawn for the test, depending on the participant's condition; the rest of the test items were randomly drawn from the remaining items not presented in the learning list. The experiment was programmed in E-Prime (Version 1.2; Psychology Software Tools, 2002) and presented on desktop computers and laptops.

Participants. Seventy-five volunteers participated in the experiment. Most of them (73) were students. There were 52 female participants and 23 male participants, and the mean age was 24.84 years ($SD = 6.99$, range = 18–58). Participants were randomly assigned to each bias manipulation condition. Fifteen participants took part in each bias group, and 1 to 4 participants completed the experiment simultaneously in one room at separate computers.

Results

Descriptive results. Figure 5 contains plotted data points from the bias manipulation conditions (left upper panel) together with

[5] Four data sets with particularly bad fits stemmed from Van Zandt (2000, Experiment 2) and Ratcliff et al. (1992, Experiment 2, mixed lists and pure list strong items). In her second experiment, Van Zandt used a payoff manipulation of bias in combination with a confidence rating scale (which we dichotomized for the analysis). In our pilot work, we had bad experiences with payoffs because many participants simply ignored the payoff scheme and failed to show bias effects. Perhaps payoff variations in combination with dichotomized rating data constitute a boundary condition for which neither model is appropriate. The problematic Ratcliff et al. data sets all involved learning lists with repeated presentations of items (five presentations of "strong items"). This was done to increase familiarity. However, in their Experiment 1, strength was varied through presentation time, and the models fit much better (as did the weak items in pure lists). It is conceivable that multiple presentations of one item during learning not only increase its familiarity but also lead to other representational changes that are not captured by either model. These post hoc explanations are speculative, and the recognition experiments reported below attempted to realize simple "standard" recognition conditions.

[6] This method was identified as the most effective method in pilot studies. Varying the amounts of payoffs and punishments for different kinds of errors proved to be too demanding for most participants.

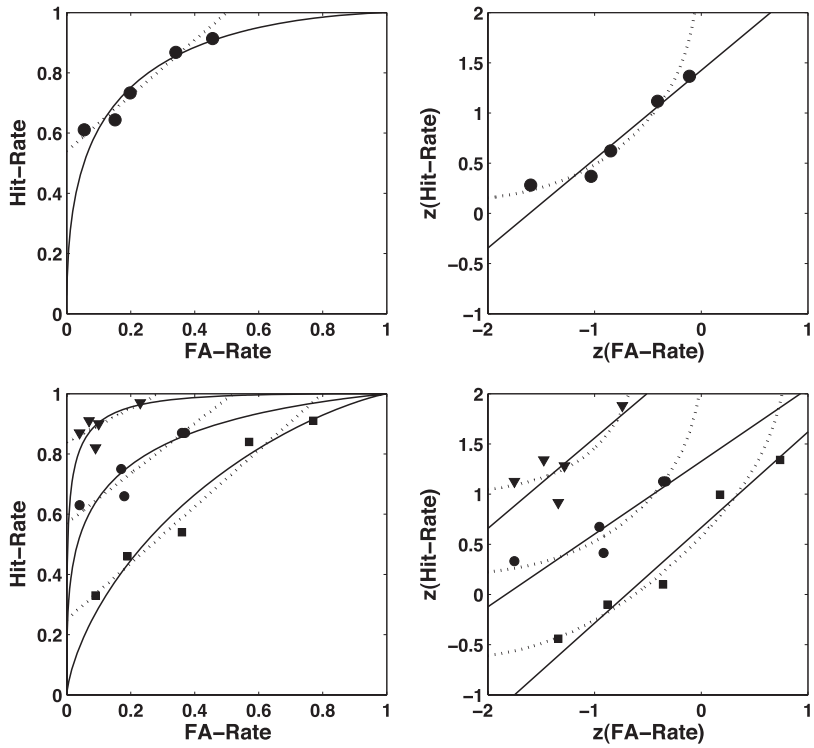

Figure 5. Empirical hit rates and false alarm (FA) rates from Experiment 1 with best-fitting signal detection theory (SDT) curves (solid lines) and best-fitting two-high-threshold model (2HTM) lines (dotted lines). The top panels show data aggregated across all participants. The bottom panels report data for three subgroups classified according to memory accuracy (triangles = best-performing subgroup; circles = medium subgroup; squares = worst-performing subgroup). Left panels are plots in probability space; right panels show the same data plotted in *z*-space.

the best-fitting SDT ROCs (solid lines) and the 2HTM ROCs (dotted lines). The right panels show the same data as the left panels plotted in *z*-space. The bottom row panels contain data from the manipulation conditions for subgroups of the best, medium, and worst performance. Table 1 reports HRs and FARs for all bias manipulation conditions of all experiments. The ROC from the bias manipulation, however, was curved upward in *z*-space as predicted by a threshold model (linear $R^2 = .911$, $\Delta R^2 = .064$). In regular probability space, the quadratic predictor did not add variance to the bias manipulation ROC (linear $R^2 = .960$, $\Delta R^2 = .002$). Hence, the bias ROC did not exhibit the curvilinear shape regularly observed with confidence rating data.

Model fits. More informative than the descriptive ROC analysis are the results of direct model fitting. How well were both models able to reproduce the response frequencies? Seven-parameter versions of both models were fit to the data, including

Table 1
Frequencies of Hits and False Alarms (FA) in Each Condition of the Three Experiments

	Condition									
	1		2		3		4		5	
Experiment	Hit	FA	Hit	FA	Hit	FA	Hit	FA	Hit	FA
Experiment 1 ($N = 75$)										
f	55	45	174	95	330	89	547	92	740	41
%	61	6	64	15	73	20	87	34	91	46
Experiment 2 ($N = 41$)										
f	111	73	275	112	439	91	772	97	896	43
%	69	5	76	10	78	16	80	30	89	38
Experiment 3 ($N = 40$)										
f	145	170	402	211	868	275	1,490	194	1,861	94
%	60	8	67	12	72	23	83	32	86	39

d', SD, and C_1 to C_5 for SDT and p_o, p_n, and b_1 to b_5 for the 2HTM. Note that both versions assume constant sensitivity across groups and allow for asymmetric ROCs. Also, they both include five bias parameters, one for each experimental condition. The G^2 statistic is asymptotically chi-square-distributed with 3 degrees of freedom, and the critical value for model rejection using the conventional significance level .05 is $\chi^2(3, crit) = 7.81$. If all data are combined, the SDT model showed a slightly worse fit, $G^2(3) = 6.22$, $p = .10$, than the 2HTM, $G^2(3) = 2.94$, $p = .40$. However, aggregating across potentially differing levels of performance in a strict sense violates both models. The between-participants variation of response bias did not allow us to provide individual model fits because each participant generates only one HR–FAR pair. In order to reduce the aggregation problem and to at least approximate individual analyses, we stratified the sample in each condition according to performance (HR–FAR difference), selecting the 5 best- and the 5 worst-performing participants in each experimental group. Fitting the models separately for the best, medium, and worst groups yielded a better fit for the 2HTM in all instances (all $G^2_{SDT} > 4.20$, all $G^2_{2HTM} < 4.20$; see Table 2 for details). Note that the bias parameters of both models nicely reflect the experimental manipulation, whereas the sensitivity parameters differ in the expected direction between the stratified groups.[7]

Discussion

In this experiment, data obtained by manipulating the response bias did not produce the curved ROC predicted by SDT. Direct model fits favored the 2HTM. One potential objection to Experiment 1 is that we obtained only 60 data points from each participant. Random variation of memory performance within and between bias manipulation groups may have blurred the ROCs. Also, one cannot rule out that the data accidentally turned out favorably for the 2HTM, and so a replication has to be achieved. Third, we suspected from informal observations in Experiment 1 that some participants might have ignored the instructions intended to induce the bias. This might distort the ROCs and challenge either model. Experiment 2 was intended to address these problems.

Experiment 2

In order to get more data points per participant and to replicate the results from Experiment 1 with different materials, we used pictures in the second experiment. Pictures are easier to learn and recognize than words (Paivio, 1971, 1986). In addition, we added a postexperimental interview in which we asked participants for the percentage of old items in the test. Also, we increased motivation to follow the bias manipulation by introducing direct monetary consequences: Each correct answer was rewarded with €0.03, and a wrong answer was punished by subtracting €0.03 from the account.

Method

Design, materials, and procedure. We used 370 simple line drawings from Snodgrass and Vanderwart (1980) and Szekely et al. (2004) in this experiment. For each participant, 150 of these were randomly selected for the learning list and were presented for 1.5 s each on the computer screen. A 20-min retention interval with various filler tasks followed. Like in Experiment 1, there were five different experimental conditions for inducing different levels of response bias. The recognition test consisted of 150 trials, and depending on the condition, it contained 10%, 25%, 50%, 75%, or 90% old items. This was explained to the participants in an instruction, and they received trial-by-trial-feedback with an update of their account in each trial. In the postexperimental interview, they were asked about the percentage of old items in the test and whether they had used this information for responding under uncertainty.

Participants. Fifty volunteers participated in the experiment. They were recruited via announcements and personal communication. Most of them (30) studied psychology, 2 were trainees, and the others were university students of fields different from psychology. There were 39 female participants and 11 male participants, and the mean age was 23.32 years ($SD = 4.89$, range = 19–41). They were randomly assigned to one of the five conditions.

Results

Descriptive ROC analyses. Nine participants (1 from Condition 2, 3 from Condition 3, 2 from Condition 4, and 3 from Condition 5) were not able to answer the postexperimental question about proportions of targets correctly, so their data were excluded from the analysis. Data on the hits and false alarms for the remaining participants can be found in Table 1. The ROC based on manipulated bias did not show a sign of consistent curvilinearity ($R^2_{linear} = .88$, $\Delta R^2 = .001$). The zROC for the bias manipulations showed a very slight U-shape ($R^2_{linear} = .86$, $\Delta R^2 = .02$). Data points from the bias manipulation conditions, the best-fitting SDT ROCs and the 2HTM ROCs, and the manipulation conditions for subgroups of best, medium, and worst performance are depicted in Figure 6.

Model fits. Table 3 shows the model fits and parameter estimates. As one can see, SDT was rejected on a conventional significance level, $G^2(3) = 9.27$, $p = .026$, whereas the 2HTM was not, $G^2(3) = 5.46$, $p = .14$.[8] If the groups are again split according to memory performance, the 2HTM fits better in all cases. However, both models fit the extreme groups worse than the medium group.

Discussion

Using different materials than in the first experiment, we replicated its general results: No sign of curvilinearity of the ROC in probability space was observed if the bias was manipulated experimentally. Direct model fits again favored the 2HTM over SDT.

However, although base rate variations are recommended for obtaining ROCs in SDT textbooks (Macmillan & Creelman, 2005;

[7] However, regarding the 2HTM this is no surprise because in the restricted model with $p = p_o = p_n$, p is equal to the difference between HR and FAR. This difference was used to stratify the sample.

[8] If the whole sample is used without excluding participants, neither model fit well, $G^2(3) = 28.53$, $p < .001$ and $G^2(3) = 8.45$, $p = .038$, for SDT and the 2HTM, respectively. However, the 2HTM still fit much better. Hence, excluding participants did not give an unfair advantage to the 2HTM.

Table 2
Goodness of Fit and Parameter Estimates for SDT and the 2HTM in Experiment 1

	Signal detection model (SDT)							
Group	$G^2(3)$	d'	SD[a]	C_1	C_2	C_3	C_4	C_5
All	6.22	1.61	1.13	1.54	1.08	0.87	0.37	0.08
Best group	4.35	2.73	1.11	1.74	1.48	1.38	1.30	0.71
Medium group	4.21	1.83	1.38	1.65	0.99	0.93	0.33	0.30
Weakest group	4.28	0.70	1.05	1.34	0.87	0.48	−0.27	−0.69
	Two-high-threshold model (2HTM)							
	$G^2(3)$	p_o	p_n	b_1	b_2	b_3	b_4	b_5
All	2.94	.54	.50	.11	.29	.41	.71	.82
Best group	2.73	.84	.71	.13	.27	.27	.37	.80
Medium group	2.59	.57	.48	.09	.32	.35	.69	.69
Weakest group	4.15	.25	.20	.11	.24	.42	.77	.88

[a] *SD* is the estimated ratio of standard deviations for targets and distractors.

McNicol, 1972; Wickens, 2002), two potential caveats are in order. First, the bias manipulation might also affect sensitivity by influencing retrieval strategies. Second, no individual ROCs can be obtained, which forces one to average across participants with potentially different performance. This might distort the ROC shape. Although we reduced the averaging problem by analyzing homogeneous subgroups, individual data would be preferable. To deal with both potential objections to our conclusions, we intro-

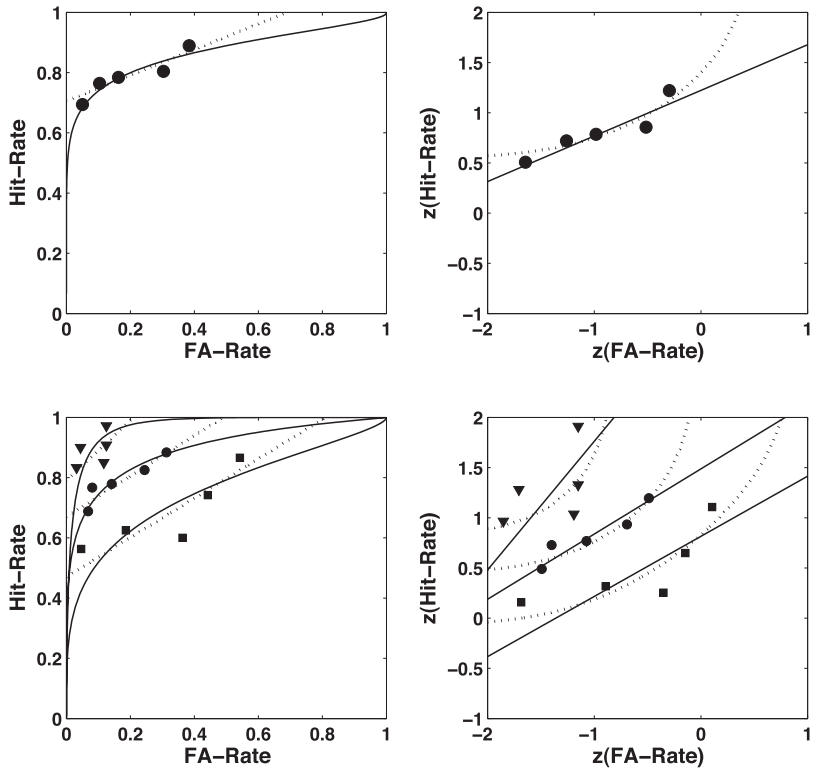

Figure 6. Empirical hit rates and false alarm (FA) rates from Experiment 2 with best-fitting signal detection theory (SDT) curves (solid lines) and best-fitting two-high-threshold model (2HTM) lines (dotted lines). The top panels show data aggregated across all participants. The bottom panels report data for three subgroups classified according to memory accuracy (triangles = best-performing subgroup; circles = medium subgroup; squares = worst-performing subgroup). Left panels are plots in probability space; right panels show the same data plotted in *z*-space.

Table 3
Goodness of Fit and Parameter Estimates for SDT and the 2HTM in Experiment 2

	Signal detection model (SDT)							
Group	$G^2(3)$	d'	SD[a]	C_1	C_2	C_3	C_4	C_5
All	9.27	2.69	2.20	1.63	1.25	0.98	0.62	0.14
Best group	18.73	2.37	0.78	1.77	1.54	1.43	1.47	1.03
Medium group	3.59	2.29	1.54	1.50	1.41	1.12	0.77	0.36
Weakest group	9.35	1.36	1.66	1.62	0.88	0.71	0.21	–0.36
	Two-high-threshold model (2HTM)							
	$G^2(3)$	p_o	p_n	b_1	b_2	b_3	b_4	b_5
All	5.46	.70	.31	.07	.15	.24	.40	.62
Best group	13.30	.79	.79	.15	.24	.48	.40	.81
Medium group	1.31	.67	.51	.14	.15	.26	.48	.69
Weakest group	2.78	.47	.19	.06	.23	.31	.52	.74

[a] *SD* is the estimated ratio of standard deviations for targets and distractors.

duced a within-subjects design in Experiment 3. We conjecture that potential influences on retrieval strategies (if they exist) should be much less pronounced when the same persons respond to stimuli under different base rate conditions. Second, the ability to fit the models to individual participant data will rule out averaging artifacts.

Experiment 3

Method

Design, materials, and procedure. The 370 line drawings from Experiment 2 were used. For each participant, 150 were randomly selected for the learning list. In the learning phase, they were presented for 1.5 s each in random order. After 20 min of diverse filler tasks (e.g., sentence descrambling, mental arithmetic), participants received the instruction that a recognition test would follow that was divided into five parts with 60 trials each. Participants were told that the procedures of these tests were identical but that they were made up from differing percentages of old and new items. Participants were informed that they would receive information about the base rate of old items in advance of each test section. Each test part was then introduced by an instruction that informed participants how many old items they had to expect in the following 60 trials (10%, 25%, 50%, 75%, or 90%). The order of test conditions was randomized for each participant. Participants received €0.02 for each correct answer, and €0.02 was subtracted for each wrong answer. Again, they received trial-by-trial feedback.

Participants. Forty volunteers participated in the experiment; 14 participants were students of psychology, and the others attended different fields of study and different occupations. There were 26 female participants and 14 male participants, and the mean age was 23.7 years (SD = 4.4, range = 20–41).

Results

Descriptive ROC analyses. Data on the hits and false alarms are provided in Table 1. Plotting the ROC (see Figure 7) yielded an almost perfectly linear plot (R^2 = .98, ΔR^2 = .00), whereas the *z*ROC was slightly curvilinear (R^2 = .96, ΔR^2 = .02). However, remember that Van Zandt (2000) emphasized the robustness of linear *z*ROCs, so a direct model fit should be more informative.

Model fits. Model fit statistics and parameter estimates can be found in Table 4. Both models fit the data well, with a slight advantage for the 2HTM (G^2_{SDT} = 7.58, p = .06 vs. G^2_{2HTM} = 2.83, p = .42). When the sample is stratified, the groups with the best and medium performance were fit better by the 2HTM (for the best group, G^2_{SDT} = 4.16, p = .24 vs. G^2_{2HTM} = 1.76, p = .62; for the medium group, G^2_{SDT} = 8.68, p = .03 vs. G^2_{2HTM} = 5.41, p = .14), whereas the group with worst performance was fit somewhat better by SDT (G^2_{SDT} = 1.93, p = .59 vs. G^2_{2HTM} = 2.18, p = .54). In this experimental setting, it is possible to analyze individual (yet noisy) ROCs. Figure 8 shows the G^2 values for both models plotted against each other. In 23 out of 40 cases (58%), the 2HTM fits better. SDT was rejected twice ($p < .05$), whereas the 2HTM was never rejected. There were 6 individuals whose data yielded degenerated SDT parameter estimates ($d' > 10$ and/or $SD < 0.20$ or $SD > 4$). If these data are eliminated, however, the evidence looks even worse for SDT, with only 13 of 34 data sets (38%) fit better by SDT than by the 2HTM.

Discussion

Although bias manipulations might theoretically affect sensitivity as well, this effect should be less severe in a within-subjects design. However, the superiority of the 2HTM was also demonstrated in a within-subjects design aggregated across all participants and in two of three subgroups. The design also allowed for individual model fitting: In these analyses, the 2HTM also faired slightly better than SDT.

General Discussion

We repeated a neglected formal argument of Erdfelder and Buchner (1998) and Malmberg (2002), who claimed that the curvilinearity of confidence-based ROCs does not constitute a valid argument against the 2HTM. Scholars have already warned much earlier to use confidence-based ROCs for evaluating models

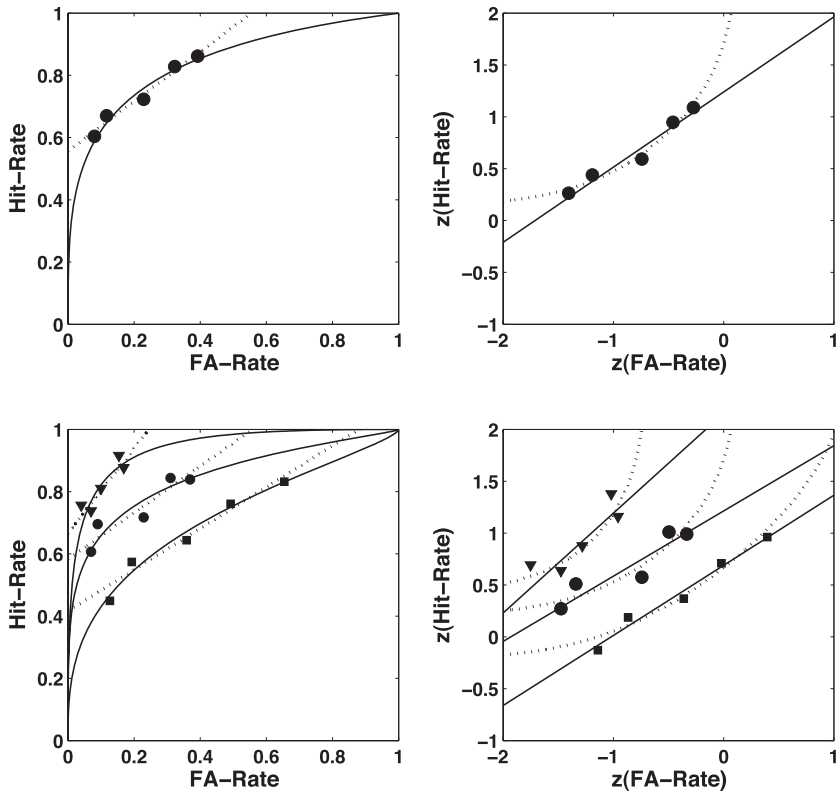

Figure 7. Empirical hit rates and false alarm (FA) rates from Experiment 3 with best-fitting signal detection theory (SDT) curves (solid lines) and best-fitting two-high-threshold model (2HTM) lines (dotted lines). The top panels show data aggregated across all participants. The bottom panels report data for three subgroups classified according to memory accuracy (triangles = best-performing subgroup; circles = medium subgroup; squares = worst-performing subgroup). Left panels are plots in probability space; right panels show the same data plotted in *z*-space.

(Krantz, 1969; Lockart & Murdock, 1970). When response bias is manipulated experimentally—as recommended by SDT textbooks (Macmillan & Creelman, 2005; Wickens, 2002)—a reanalysis of 59 data sets based on direct model fitting showed no superiority of SDT, and the results of three new experiments clearly favored the 2HTM over SDT.

Two potentially important counterarguments against the validity of our findings are possible: First, it may be argued that bias

Table 4
Goodness of Fit and Parameter Estimates for SDT and the 2HTM in Experiment 3

	Signal detection model (SDT)							
	$G^2(3)$	d'	SD[a]	C_1	C_2	C_3	C_4	C_5
All	7.58	1.71	1.38	1.41	1.17	0.80	0.43	0.23
Best group	4.16	2.24	1.04	1.70	1.51	1.31	1.01	0.84
Medium group	8.68	1.93	1.59	1.48	1.30	0.83	0.41	0.35
Weakest group	1.93	1.01	1.48	1.15	0.85	0.40	–0.01	–0.4
	Two-high-threshold model (2HTM)							
	$G^2(3)$	p_o	p_n	b_1	b_2	b_3	b_4	b_5
All	2.83	.56	.45	.14	.22	.40	.60	.69
Best group	1.76	.67	.76	.17	.27	.42	.64	.75
Medium group	5.41	.58	.44	.12	.17	.38	.61	.62
Weakest group	2.18	.42	.12	.14	.22	.40	.58	.72

[a] *SD* is the estimated ratio of standard deviations for targets and distractors.

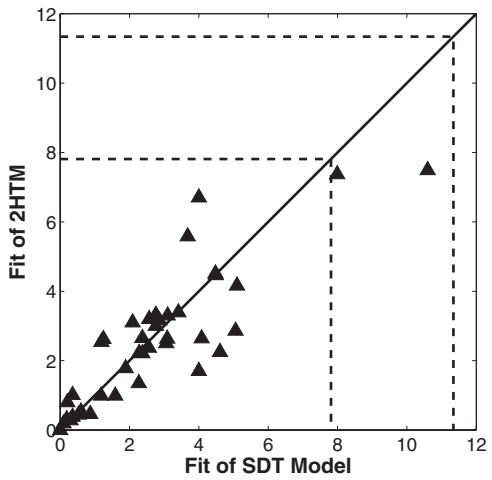

Figure 8. Goodness-of-fit (G^2) values for signal detection theory (SDT) and the two-high-threshold model (2HTM) for the 40 participants in Experiment 3. Points below the diagonal indicate a better fit of the 2HTM; points above the diagonal indicate a better fit of SDT. The dotted lines represent conventional alpha levels of .05 (at $G^2 = 7.81$) and .01 (at $G^2 = 11.34$).

manipulations—as opposed to confidence ratings—may also induce changes in retrieval strategies and hence accuracy (e.g., Yonelinas & Parks, 2007). Although we tried to minimize this possibility by using the within-subjects variation in Experiment 3, the counterargument may still hold even in this situation. Second, the 2HTM and its extension might simply be much more flexible in fitting arbitrary data sets. Despite their identical numbers of parameters, the flexibility of the models and their ability to "overfit" data also depend on their functional form (see Myung & Pitt, 1997). If the 2HTM is more flexible, the conclusions drawn might be artifactual. We deal with both arguments in turn.

Potential Impact of Bias Manipulation on Accuracy

The bias manipulation used might theoretically influence participants' motivation to engage in retrieval attempts. If high accuracy is possible by simple guessing (e.g., in the extreme bias groups), why should one bother about retrieval? Reduced retrieval accuracy and bias might just compensate for each other and flatten the observed ROCs. To assess this possibility, we computed d' and C for SDT as well as p and b for every participant in each condition of all three experiments (note that only the equal variance variant of SDT and the 2HTM assuming $p_o = p_n$ can be used in situations with only one HR–FAR pair). These values were subjected to analyses of variance with the experimental factor bias condition, which was a between-subjects factor in Experiments 1 and 2 and a within-subjects factor in Experiment 3. The bias manipulation had no effect on d' or p in any experiment: all Fs $<$ 1 in Experiments 1 and 2, Greenhouse–Geisser $F(3.42, 133.43) = 2.02$ and $F(1.96, 76.48) = 2.19$ for p and d' in Experiment 3, respectively, both ps $> .10$. However, there were the expected reliable effects on the bias measures C (all Fs > 4, all ps $< .01$) and b (all Fs > 13, all ps $< .001$). These results are paralleled by very large effect sizes for C and b (all η^2s $> .40$) and by rather small effect sizes for d' and p (all η^2s in Experiments 1 and 3 $< .05$, η^2s in Experiment 2 $< .13$). Hence, there is no hint in our data of an accuracy effect of our bias manipulations.

Potential Flexibility of the 2HTM

It has been recognized for a long time that formal models may be more or less flexible in fitting arbitrary random aspects of the data (e.g., sampling error) rather than only the systematic and lawful aspects (Myung & Pitt, 1997, 1998). This phenomenon is called overfitting. One major determinant of model flexibility is the number of free parameters. The often-used Akaike information criterion (Akaike, 1973; Bozdogan, 2000) or Bayes information criterion (G. Schwarz, 1978; Wassermann, 2000) trade off the model fit against the number of model parameters. In our case, this does not help because both models have the same number of parameters. However, model flexibility can also differ between such models (Busemeyer & Wang, 2000) because of the function form defined by each model. Hence, the superiority of the 2HTM in our data sets might be more apparent than real. In order to test for this possibility, we generated 3,000 random data sets reflecting the frequency distributions of old and new items in our experiments. To do so, we generated response frequencies by multiplying five randomly generated HR–FAR pairs with the signal and noise base rates of five conditions. These five conditions differed in their signal and noise base rates (100 signal trials vs. 900 noise trials in Condition 1, 250 vs. 750 in Condition 2, 500 vs. 500 in Condition 3, 750 vs. 250 in Condition 4, 900 vs. 100 in Condition 5). The HR–FAR pairs were generated using the routine "runif()" in R (Version 2.4.1; R Development Core Team, 2006). The minimal restriction in generating the data was HR $>$ FAR in each experimental condition, which is essentially implied by both SDT and the 2HTM. Then we fit our seven-parameter versions of both SDT and the 2HTM to the data. If one model was more flexible in fitting sample noise, this would result in consistently better fit statistics.

Table 5 shows distribution statistics of the goodness-of-fit index G^2 for SDT and the 2HTM in the manipulation conditions, respectively. The results show that SDT obviously is the more flexible

Table 5
Distributions of Fit Statistic G^2 When SDT and the 2HTM Are Fit to Random Data

	Bias manipulation model	
Statistic	SDT	2HTM
% superior fits	67.2	32.8
M	104.90	128.31
SD	90.38	115.24
Minimum	0.15	0.07
Maximum	575.53	780.20
Percentile		
10	16.73	20.57
25	37.10	45.11
50	79.33	93.79
75	146.66	176.80
90	231.40	286.08

Note. There are 3,000 data sets for each column. SDT = signal detection theory; 2HTM = two-high-threshold model.

model with 67.2% better fits and consistently smaller G^2 statistics. So, the 2HTM is even less flexible than SDT and nevertheless fit the data better in most of our data sets. Hence, the superiority of the 2HTM obviously cannot be explained away as a flexibility artifact.

Taken together, the results of our three experiments and our reanalyses do not suggest the superiority of SDT over the 2HTM. There may be other epistemic arguments to favor SDT over threshold models, but obviously, neither model fit nor ROC shape constitute such an argument and hence, Wixted's (2007a, 2007b) and Yonelinas and Parks's (2007) conclusions on this basis are invalid.

Usefulness of Confidence ROC Analysis

Hence, the superiority of SDT still has to be demonstrated because most of the results speaking for SDT are based on rating ROCs. As mentioned before, the curvilinear shape of confidence-based ROCs does not rule out the 2HTM. In addition, Grasha (1970) and Gardner et al. (1975) indicated that the correspondence between rating response format and binary response format in recognition memory has to be demonstrated before analyzing recognition data based solely on rating responses. To our knowledge, there are only two studies in which this attempt has been made, and their results suggest differences between parameter estimates resulting from a rating response format and parameter estimates from a binary response format (Gardner et al., 1975; Van Zandt & Maldonado-Molina, 2004). A comparison of two experiments by Malmberg and Xu (2007) for associative recognition also suggests that yes–no tasks and rating tasks may lead to different results, especially concerning the FARs. Because Egan et al. (1959) demonstrated the equivalence of SDT parameter estimates generated on the basis of rating data and binary response format in perception experiments, this may be a first hint of differences between the applicability of SDT to recognition memory versus perception. However, one should not overestimate the importance of only two results.

Defendants of SDT might argue that there is ample evidence for curved ROCs from binary tasks in perception that favors SDT. This may be true, but we see no a priori reason why the perceptual detection of simple stimuli should necessarily follow the same laws or include similar processes as the recognition of more or less complex stimuli involving the retrieval of an episodic memory. The ingenious idea to apply the successful perceptual model to recognition (e.g., Banks, 1970; Egan et al., 1959; Kintsch, 1967) was probably inspired by the structural similarity of the tasks and the resulting data matrix rather than by a priori reasons to view both processes as similar. This is probably an instance of the so-called "tools to theories" heuristic that Gigerenzer (1991) identified in many discoveries in cognitive psychology. In addition, the threshold concept is not dead in the psychology of perception, and there are also perception phenomena that suggest the existence of thresholds (e.g., Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; King-Smith, 2005).

Theoretical Implications

As pointed out by Lockart and Murdock (1970), when applying SDT epistemically, an additional problem to be solved is the identification of the elements of SDT with constructs constituting a memory theory. As mentioned before, SDT separates decision and discrimination processes in recognition memory by formalizing the input to the decision system as the value of a random variable defined on a memory continuum. There are many heterogeneous attempts to put this memory continuum into concrete terms. One of the earliest attempts is the application of Wickelgren and Norman's (1966) continuous strength theory. Thereby the memory continuum is identified directly with a continuum of memory strength and an item's value on the memory strength continuum can be increased or decreased by processes such as acquisition, forgetting, and generalization. Another attempt at identifying the memory continuum has been the application of global matching models like MINERVA 2 (Hintzman, 1986, 1988), SAM (search of associative memory model; Raaijmakers & Shiffrin, 1980, 1981), and TODAM2 (theory of distributed associative memory; Murdock, 1993). According to global matching models, the memory strength continuum can be conceptualized as a familiarity continuum in which an item's familiarity value results from matching a test item against memory representations. Especially, these particular variants of SDT do not require the signal and noise distributions to have equal variances (e.g., Ratcliff et al., 1992). However, some studies failed to support predictions derived from these variants of SDT too. For example, contrary to the prediction of nearly equal old and new standard deviations (TODAM2) or increasing values of old standard deviations with strength (SAM, MINERVA 2), the standard deviation of new-item familiarity seems to be independent of the strength of the old items (Cleary, 2005; Ratcliff et al., 1992). Although the SDT variants listed above assume a single-process conceptualization of the memory continuum, Wixted (2007a) favored a memory strength variable that arises from the combination of two or more continuous processes, such as familiarity and recollection.

In sum, both because of the lack of convincing empirical data and in the face of the heterogeneous conceptualizations of the memory continuum, the pragmatic and the epistemic superiority of SDT still have to be demonstrated.

A Tentative Explanation

How can this all make sense theoretically? In our view, threshold models are unpopular because (a) discrete states somehow look less plausible than continuous variables and (b) threshold models do not provide a process description that explains how items cross a threshold (if it exists). In this respect, SDT and the abovementioned theories that describe how the evidence continuum is generated are much more explicit and theoretically satisfying. Why, then, does the 2HTM fit the data so well? Rather than presenting a full-fledged theoretical account, we want to propose a simple theoretical idea that might explain the good approximation of recognition data by the 2HTM. This idea entails a two-process view of recognition and is similar to a model originally proposed by Atkinson and Juola (1974). Our model assumes an underlying continuous evidence variable (henceforth called *familiarity* for ease of presentation) and normal distributions of this familiarity, exactly as in SDT. However, in a recognition task, participants set two more or less strict criteria instead of one. When familiarity is very high, they will respond "old" quickly and with high confidence, whereas very low familiarity will lead to a quick "new"

response. However, the region between both criteria is where the distributions overlap, and hence, decisions based on familiarity become uncertain and error prone. We conjecture that people refrain from relying on familiarity in these cases and switch to another strategy that entails attempts of strategic conscious recollection as well as informed guessing. The general idea is very similar to the two-stage process of categorization in the feature-matching model by Smith, Shoben, and Rips (1974), in which a quick decision can be made if the feature overlap between instance and category is either very high or very low, whereas intermediate overlaps demand closer scrutiny of defining features. Because bias manipulations in the binary response format would affect only items that fall in between both criteria, a linear ROC results. Rating responses, however, would lead to more curvilinear ROCs if (a) the criteria attached to rating bins extended outside the two familiarity criteria, or (b) they would push people to rely on familiarity also in uncertain cases, or both. This would explain (a) why rating ROCs are apparently curved whereas bias manipulation curves are not and (b) why rating curves are often less curved than predicted by SDT.

Similar ideas have recently been put forward by Malmberg (2008; see also Ratcliff, 1978), who formalized recognition as a dynamic evidence accumulation process that can reach either of two thresholds that may be mapped onto the "detect" states of the 2HTM. If the process does not reach any threshold after some time, guessing is likely to occur particularly in situations in which familiarity is of low diagnostic value (e.g., frequency judgments, source decisions, associative recognition). This dynamic variant of a two-high-threshold model is attractive because it can explain the good fit of the 2HTM, it provides a description of the recognition process, and furthermore it makes testable assumptions about reaction times.

Furthermore, when considering the retrieval dynamics of response latencies, empirical findings seem to support the existence of two decision criteria in episodic long-term recognition (Malmberg, 2008; Ratcliff, 1978; Ratcliff & Murdock, 1976; Van Zandt & Maldonado-Molina, 2004; Wixted & Stretch, 2004). Some recognition memory models and associated empirical findings even support the assumption that the rejection of distractors is not only based on a simple lack of familiarity but on the accumulation of evidence for a "new" decision that crosses a separate threshold or criterion (Ratcliff, 1978; Ratcliff & Murdock, 1976; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004). In line with this, Mewhort and Johns (2000, 2005) and Johns and Mewhort (2002) demonstrated that correct rejections in short-term recognition memory are based on item features that contradict the study set rather than on item familiarity.

How do the results and our tentative model relate to other recognition paradigms, for instance remember–know judgments? In line with several empirical findings and models of recognition memory (e.g., Donaldson, 1996; Dunn, 2004, 2008; Malmberg, 2008; Malmberg, Holden, & Shiffrin, 2004; Malmberg & Xu, 2006; Shiffrin & Steyvers, 1998; Wixted & Stretch, 2004), we adopt the assumption that "remember" and "know" responses reflect different levels of memory strength or confidence. According to our model, "remember" responses should rather be made whenever an item's memory strength exceeds the upper familiarity criterion, whereas "know" responses are made when an item's memory strength is too weak to exceed the upper familiarity criterion and too strong to undercut the lower familiarity criterion. Consistent with our idea, there are some empirical findings speaking for faster reaction times to "remember" responses and highly confident rating responses and slower reaction times to "know" responses and less confident rating responses (e.g., for remember–know responses, see Dewhurst & Conway, 1994; Dewhurst, Holmes, Brandt, & Dean, 2006; Henson, Rugg, Shallice, Josephs, & Dolan, 1999; Tulving, 1985; Wixted & Stretch, 2004; for confidence ratings, see Ratcliff & Murdock, 1976; Van Zandt & Maldonado-Molina, 2004; Wixted & Stretch, 2004).

Conclusion

Our analyses and findings imply two take-home messages. First, there may be reasons and arguments why researchers prefer signal detection or hybrid models over threshold accounts of recognition memory, but ROCs are obviously no good argument! Contrary to recent claims, the curvilinear shape of rating-based ROCs is uninformative, and ROCs based on bias manipulations do not speak against threshold models. Rather, the 2HTM tends to fit data better than SDT in bias manipulation experiments, and this is no artifact of model flexibility. Hence, critics of discrete state models should invoke better arguments than confidence ROCs or model fits.

Second, even if threshold theories finally for some reason turned out to be epistemically inferior to SDT, they may nevertheless provide good measurement tools under a wide range of conditions. This is an important insight because multinomial models that are based on discrete state assumptions have been successfully used in many domains serving to disentangle various cognitive processes (e.g., Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). These models are flexible and come equipped with a full-fledged statistical machinery for parameter estimation and hypothesis tests (Hu & Batchelder, 1994) as well as convenient analysis software (Rothkegel, 1999; Stahl & Klauer, 2007). Abandoning these useful tools because of their admittedly approximate nature would throw out the baby with the bath water.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Adademiai Kiado.

Allen, L. R., & Garton, R. F. (1969). Detection and criterion change associated with different test contexts in recognition memory. *Perception & Psychophysics, 6*(1), 1–4.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: I. Learning, memory and thinking* (pp. 243–293). Oxford, England: Freeman.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*(2), 81–99.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*(4), 267–273.

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

Berch, D. B. (1977). Effects of stimulus probability and information feedback on response biases in children's recognition memory. *Bulletin of the Psychonomic Society, 10*(4), 328–330.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53,* 129–160.

Bozdogan, H. (2000). Akaike information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44,* 62–91.

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

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124*(2), 137–160.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*(1), 171–189.

Cleary, A. M. (2005). ROCs in recognition with and without identification. *Memory, 13*(5), 472–483.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 2–17.

DeCarlo, L. T. (2003a). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 767–778.

DeCarlo, L. T. (2003b). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology, 47,* 292–303.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204–211.

Dewhurst, S. A., & Conway, M. A. (1994). Pictures, images, and recollective experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1088–1098.

Dewhurst, S. A., Holmes, S., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness and Cognition, 15,* 147–162.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533.

Drake, A. I., & Hannay, H. J. (1992). Continuous recognition memory tests: Are the assumptions of the theory of signal detection met? *Journal of Clinical and Experimental Neuropsychology, 14*(4), 539–544.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111,* 524–542.

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state–trace analysis. *Psychological Review, 115,* 426–446.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31,* 768–773.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127*(1), 83–97.

Estes, W. K., & Maddox, W. T. (1995). Interaction of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1075–1095.

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.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*(3), 546–567.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1176–1195.

Grasha, A. F. (1970). Detection theory and memory processes: Are they compatible? *Perceptual and Motor Skills, 30,* 123–135.

Hager, W., & Hasselhorn, M. (1994). *Handbuch deutschsprachiger Wortnormen* [Handbook of German word norms]. Göttingen, Germany: Hogrefe.

Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3*(3), 233–238.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cut-off location in recognition memory. *Memory & Cognition, 6*(5), 544–553.

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10*(3), 718–723.

Henriques, J. B., Glowacki, J. M., & Davidson, R. J. (1994). Reward fails to alter response bias in depression. *Journal of Abnormal Psychology, 103*(3), 460–466.

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *The Journal of Neuroscience, 19,* 3962–3972.

Herrmann, T. (1960). Über Urteilskonkordanz und Urteilsnuanziertheit [About judgment concordance and judgment accentuation]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 7,* 532–546.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple memory model. *Psychological Review, 93,* 411–428.

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

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59*(1), 21–47.

Johns, E. E., & Mewhort, D. J. K. (2002). What information underlies correct rejections in short-term recognition memory? *Memory & Cognition, 30*(1), 46–59.

King-Smith, P. E. (2005). Threshold nonlinearities and signal detection theory. *Perception, 34*(8), 941–946.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review, 74,* 496–504.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76*(3), 308–324.

Lockart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*(2), 100–109.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 380–387.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated proposal for relating them. *Cognitive Psychology, 57,* 335–384.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on judgments of frequency and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 319–331.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review, 13*(1), 99–105.

Malmberg, K. J., & Xu, J. (2007). On the flexibility and the fallibility of associative memory. *Memory & Cognition, 35*(3), 545–556.

Marken, R. S., & Sandusky, A. J. (1974). Stimulus probability and sequential effect in recognition memory. *Bulletin of the Psychonomic Society, 4*(1), 48–51.

McNicol, D. (1972). *A primer of signal detection theory.* London: Allen & Unwin.

Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: A test of item matching in short-term recognition memory. *Journal of Experimental Psychology: General, 129,* 262–284.

Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory, 13,* 300–307.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*(2), 183–203.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95.

Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 327–355). Hillsdale, NJ: Erlbaum.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental representations.* New York: Oxford University Press.

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73*(1), 44–58.

Psychology Software Tools. (2002). E-Prime Version 1.2 [Computer software]. Pittsburgh, PA: Author.

Qin, J., Raye, C. L., Johnson, M. K., & Mitchell, K. J. (2001). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1110–1115.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 14, pp. 207–262). New York: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190–214.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535.

R Development Core Team. (2006). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320.

Richardson, J. T. E. (1994). Continuous recognition memory tests: Are the assumptions of the theory of signal detection really met? *Journal of Clinical and Experimental Neuropsychology, 16*(3), 482–486.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*(3), 318–339.

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments & Computers, 31*(4), 696–700.

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

Schwarz, N., Knauper, B., Hippler, H. J., & Neumann, E. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570–582.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). London: Oxford University Press.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33,* 151–170.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1499–1517.

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

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

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174–215.

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods, 39*(2), 267–273.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33*(2), 203–217.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, A., Herron, D., et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51,* 247–250.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26,* 1–12.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582–600.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1147–1166.

Wassermann, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44,* 92–107.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York: Oxford University Press.

Wixted, J. T. (2007a). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.

Wixted, J. T. (2007b). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review, 114*(1), 203–209.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11,* 616–641.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800–832.

Appendix

Reanalysis of 59 Data Sets, Model Fits, and Parameter Estimates

Author	Experiment, condition	Type of bias manipulation and no. of manipulation steps	Critical G^2	2HTM fit	2HTM discrimination indices	2HTM bias indices	SDT fit	SDT discrimination index	SDT bias indices
Berch (1977)	Exp. 1, feedback	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 4.199$	$G^2_{emp}(1) = 5.886^*$	$p = .460$	$b_{(25)} = .417$ $b_{(50)} = .609$	$G^2_{emp}(1) = 7.323^*$	$d' = 1.238$	$c_{(25)} = 0.761$ $c_{(50)} = 0.449$
Berch (1977)	Exp. 1, no feedback	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 4.199$	$G^2_{emp}(1) = 6.358^*$	$p = .387$	$b_{(25)} = .522$ $b_{(50)} = .791$	$G^2_{emp}(1) = 2.577$	$d' = 1.109$	$c_{(25)} = 0.511$ $c_{(50)} = 0.001$
Buchner et al. (1995)	Exp. 1, inclusion	Base rate, 2 (50% vs. 75% old items in test)	$G^2_{crit}(1) = 3.588$	$G^2_{emp}(1) = 0.844$	$p = .478$	$b_{(50)} = .454$ $b_{(75)} = .605$	$G^2_{emp}(1) = 1.09$	$d' = 1.293$	$c_{(50)} = 0.723$ $c_{(75)} = 0.478$
Buchner et al. (1995)	Exp. 2, inclusion	Payoff, 2 (liberal vs. conservative)	$G^2_{crit}(1) = 3.588$	$G^2_{emp}(1) = 0.445$	$p = .505$	$b_{(50)} = .528$ $b_{(75)} = .680$	$G^2_{emp}(1) = 0.032$	$d' = 1.383$	$c_{(50)} = 0.652$ $c_{(75)} = 0.385$
Curran et al. (2007)	Exp. 1	Payoff, 2 (liberal vs. conservative)	$G^2_{crit}(1) = 35.078$	$G^2_{emp}(1) = 4.558$	$p = .544$	$b_{(lib)} = .387$ $b_{(con)} = .725$	$G^2_{emp}(1) = 0.414$	$d' = 1.553$	$c_{(lib)} = 0.946$ $c_{(con)} = 0.418$
Curran et al. (2007)	Exp. 2	Payoff, 2 (liberal vs. conservative)	$G^2_{crit}(1) = 19.125$	$G^2_{emp}(1) = 3.805$	$p = .565$	$b_{(lib)} = .335$ $b_{(con)} = .646$	$G^2_{emp}(1) = 5.793$	$d' = 1.604$	$c_{(lib)} = 0$ $c_{(con)} = 1.046$
Curran et al. (2007)	Exp. 3	Payoff, 3 (liberal, neutral, conservative)	$G^2_{crit}(1) = 18.811$	$G^2_{emp}(1) = 27.127^*$	$p_o = .575$ $p_n = .319$	$b_{(lib)} = .645$ $b_{(neu)} = .528$ $b_{(con)} = .38$	$G^2_{emp}(1) = 30.039^*$	$d' = 1.405$ $SD = 1.209$	$c_{(lib)} = 0.148$ $c_{(neu)} = 0.371$ $c_{(con)} = 0.649$
Estes & Maddox (1995)	Exp. 1, digits and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 3.465$	$G^2_{emp}(1) = 0.014$	$p = .298$	$b_{(33)} = .426$ $b_{(67)} = .743$	$G^2_{emp}(1) = 0.969$	$d' = 0.813$	$c_{(33)} = 0.541$ $c_{(67)} = -0.081$
Estes & Maddox (1995)	Exp. 1, digits and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 3.313$	$p = .247$	$b_{(33)} = .545$ $b_{(67)} = .570$	$G^2_{emp}(1) = 4.453$	$d' = 0.635$	$c_{(33)} = -0.041$ $c_{(67)} = 0.226$
Estes & Maddox (1995)	Exp. 1, letters and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 3.465$	$G^2_{emp}(1) = 10.507^*$	$p = .398$	$b_{(33)} = .379$ $b_{(67)} = .810$	$G^2_{emp}(1) = 3.912^*$	$d' = 1.159$	$c_{(33)} = 0.769$ $c_{(67)} = -0.041$
Estes & Maddox (1995)	Exp. 1, letters and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 0.994$	$p = .414$	$b_{(33)} = .572$ $b_{(67)} = .529$	$G^2_{emp}(1) = 1.483$	$d' = 1.093$	$c_{(33)} = 0.426$ $c_{(67)} = 0.496$
Estes & Maddox (1995)	Exp. 1, words and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 2.867$	$p = .744$	$b_{(33)} = .427$ $b_{(67)} = .474$	$G^2_{emp}(1) = 3.482$	$d' = 2.281$	$c_{(33)} = 1.169$ $c_{(67)} = 1.238$
Estes & Maddox (1995)	Exp. 1, words and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 0.391$	$p = .773$	$b_{(33)} = .594$ $b_{(67)} = .772$	$G^2_{emp}(1) = 0.137$	$d' = 2.463$	$c_{(33)} = 1.116$ $c_{(67)} = 0.972$
Estes & Maddox (1995)	Exp. 2, digits and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 24.723^*$	$p = .196$	$b_{(33)} = .416$ $b_{(67)} = .795$	$G^2_{emp}(1) = 15.911^*$	$d' = 0.574$	$c_{(33)} = 0.448$ $c_{(67)} = -0.401$
Estes & Maddox (1995)	Exp. 2, digits and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 6.061$	$p = .121$	$b_{(33)} = .634$ $b_{(67)} = .542$	$G^2_{emp}(1) = 6.756^*$	$d' = 0.315$	$c_{(33)} = -0.146$ $c_{(67)} = 0.066$
Estes & Maddox (1995)	Exp. 2, letters and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 2.420$	$p = .393$	$b_{(33)} = .473$ $b_{(67)} = .631$	$G^2_{emp}(1) = 1.925$	$d' = 1.042$	$c_{(33)} = 0.566$ $c_{(67)} = 0.285$
Estes & Maddox (1995)	Exp. 2, letters and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 11.864^*$	$p = .364$	$b_{(33)} = .539$ $b_{(67)} = .526$	$G^2_{emp}(1) = 15.048^*$	$d' = 0.951$	$c_{(33)} = 0.404$ $c_{(67)} = 0.431$
Estes & Maddox (1995)	Exp. 2, words and feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 2.110$	$p = .799$	$b_{(33)} = .376$ $b_{(67)} = .680$	$G^2_{emp}(1) = 3.375$	$d' = 2.603$	$c_{(33)} = 1.453$ $c_{(67)} = 1.102$
Estes & Maddox (1995)	Exp. 1, words and no feedback	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 6.385$	$G^2_{emp}(1) = 4.156$	$p = .633$	$b_{(33)} = .565$ $b_{(67)} = .649$	$G^2_{emp}(1) = 6.497^*$	$d' = 1.833$	$c_{(33)} = 0.817$ $c_{(67)} = 0.717$
Healy & Jones (1975)	Exp. 2, standard instruction, liberal criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.001$	$p = .250$	$b_{(25)} = .547$ $b_{(50)} = .613$	$G^2_{emp}(1) = 0.020$	$d' = 0.648$	$c_{(25)} = 0.229$ $c_{(50)} = 0.098$
Healy & Jones (1975)	Exp. 2, standard instruction, medium criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.237$	$p = .292$	$b_{(25)} = .432$ $b_{(50)} = .457$	$G^2_{emp}(1) = 0.265$	$d' = 0.752$	$c_{(25)} = 0.503$ $c_{(50)} = 0.457$
Healy & Jones (1975)	Exp. 2, standard instruction, strict criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.551$	$p = .278$	$b_{(25)} = .314$ $b_{(50)} = .325$	$G^2_{emp}(1) = 0.586$	$d' = 0.756$	$c_{(25)} = 0.751$ $c_{(50)} = 0.721$
Healy & Jones (1975)	Exp. 2, strictness instruction, liberal criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 1.717$	$p = .202$	$b_{(25)} = .700$ $b_{(50)} = .735$	$G^2_{emp}(1) = 1.571$	$d' = 0.573$	$c_{(25)} = 0.141$ $c_{(50)} = 0.229$

(*table continues*)

Table (*continued*)

Author	Experiment, condition	Type of bias manipulation and no. of manipulation steps	Critical G^2	2HTM fit	2HTM discrimination indices	2HTM bias indices	SDT fit	SDT discrimination index	SDT bias indices
Healy & Jones (1975)	Exp. 2, strictness instruction, medium criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.248$	$p = .269$	$b_{(25)} = .345$ $b_{(50)} = .350$	$G^2_{emp}(1) = 0.333$	$d' = 0.718$	$c_{(25)} = 0.666$ $c_{(50)} = 0.657$
Healy & Jones (1975)	Exp. 2, Strictness instruction, strict criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.083$	$p = .204$	$b_{(25)} = .114$ $b_{(50)} = .124$	$G^2_{emp}(1) = 0.281$	$d' = 0.784$	$c_{(25)} = 1.329$ $c_{(50)} = 1.299$
Healy & Jones (1975)	Exp. 2, HR/FAR instruction, liberal criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.250$	$p = .219$	$b_{(25)} = .618$ $b_{(50)} = .672$	$G^2_{emp}(1) = 0.130$	$d' = 0.583$	$c_{(25)} = -0.068$ $c_{(50)} = 0.045$
Healy & Jones (1975)	Exp. 2, HR/FAR instruction, medium criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 0.059$	$p = .284$	$b_{(25)} = .351$ $b_{(50)} = .430$	$G^2_{emp}(1) = 0.282$	$d' = 0.750$	$c_{(25)} = 0.666$ $c_{(50)} = 0.507$
Healy & Jones (1975)	Exp. 2, HR/FAR instruction, strict criterion	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 7.030$	$G^2_{emp}(1) = 1.828$	$p = .229$	$b_{(25)} = .151$ $b_{(50)} = .190$	$G^2_{emp}(1) = 0.783$	$d' = 0.760$	$c_{(25)} = 1.193$ $c_{(50)} = 1.056$
Healy & Kubovy (1978)	Exp. 1, standard	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 10.427$	$G^2_{emp}(1) = 2.288$	$p = .279$	$b_{(25)} = .216$ $b_{(50)} = .374$	$G^2_{emp}(1) = 0.094$	$d' = 0.784$	$c_{(25)} = 1.00$ $c_{(50)} = 0.636$
Healy & Kubovy (1978)	Exp. 1, tally	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 10.427$	$G^2_{emp}(1) = 2.309$	$p = .269$	$b_{(25)} = .280$ $b_{(50)} = .382$	$G^2_{emp}(1) = 0.855$	$d' = 0.729$	$c_{(25)} = 0.822$ $c_{(50)} = 0.594$
Healy & Kubovy (1978)	Exp. 1, no feedback	Base rate, 2 (25% vs. 50% old items in test)	$G^2_{crit}(1) = 10.427$	$G^2_{emp}(1) = 1.169$	$p = .275$	$b_{(25)} = .325$ $b_{(50)} = .458$	$G^2_{emp}(1) = 0.379$	$d' = 0.728$	$c_{(25)} = 0.747$ $c_{(50)} = 0.458$
Healy & Kubovy (1978)	Exp. 1, payoff	Payoff, 2 (liberal, conservative)	$G^2_{crit}(1) = 10.427$	$G^2_{emp}(1) = 1.602$	$p = .240$	$b_{(25)} = .487$ $b_{(50)} = .553$	$G^2_{emp}(1) = 2.182$	$d' = 0.612$	$c_{(25)} = 0.332$ $c_{(50)} = 0.200$
Henriques et al. (1994)	Exp. 1 controls	Payoff, 3 (punishment, no consequence, reward)	$G^2_{crit}(1) = 7.835$	$G^2_{emp}(1) = 4.658$	$p_o = .784$ $p_n = .001$	$b_{(punish)} = .399$ $b_{(no\ c)} = .407$ $b_{(reward)} = .284$	$G^2_{emp}(1) = 5.887$	$d' = 4.567$ $SD = 3.585$	$c_{(punish)} = 0.241$ $c_{(no\ c)} = 0.572$ $c_{(reward)} = 0.244$
Henriques et al. (1994)	Exp. 1, dysphorics	Payoff, 3 (punishment, no consequence, reward)	$G^2_{crit}(1) = 8.912$	$G^2_{emp}(1) = 5.915$	$p_o = .770$ $p_n = .001$	$b_{(punish)} = .399$ $b_{(no\ c)} = .259$ $b_{(reward)} = .354$	$G^2_{emp}(1) = 8.083$	$d' = 3.741$ $SD = 3.222$	$c_{(punish)} = 0.265$ $c_{(no\ c)} = 0.648$ $c_{(reward)} = 0.372$
Marken & Sandusky (1974)	Exp. 1	Base rate, 2 (20% vs. 50% old items in test)	$G^2_{crit}(1) = 12.414$	$G^2_{emp}(1) = 0.208$	$p = .391$	$b_{(20)} = .484$ $b_{(50)} = .636$	$G^2_{emp}(1) = 0.060$	$d' = 1.038$	$c_{(20)} = 0.546$ $c_{(50)} = 0.287$
Parks (1966)	Exp. 1, free format	Base rate, 5 (20%, 33%, 50%, 67%, 80% old items in test)	$G^2_{crit}(3) = 4.686$	$G^2_{emp}(3) = 1.057$	$p_o = .486$ $p_n = .516$	$b_{(20)} = .228$ $b_{(33)} = .355$ $b_{(50)} = .533$ $b_{(67)} = .657$ $b_{(80)} = .765$	$G^2_{emp}(3) = 1.020$	$d' = 1.459$ $SD = 1.067$	$c_{(20)} = 0.945$ $c_{(33)} = 0.967$ $c_{(50)} = 0.681$ $c_{(67)} = 0.466$ $c_{(80)} = 0.233$
Parks (1966)	Exp. 1, fixed format	Base rate, 5 (20%, 33%, 50%, 67%, 80% old items in test)	$G^2_{crit}(3) = 4.686$	$G^2_{emp}(3) = 7.337^*$	$p_o = .000$ $p_n = .759$	$b_{(20)} = .629$ $b_{(33)} = .729$ $b_{(50)} = .731$ $b_{(67)} = .832$ $b_{(80)} = .840$	$G^2_{emp}(3) = 8.346^*$	$d' = 1.063$ $SD = 0.188$	$c_{(20)} = 1.010$ $c_{(33)} = 0.945$ $c_{(50)} = 0.943$ $c_{(67)} = 0.887$ $c_{(80)} = 0.876$
Ratcliff et al. (1992)	Exp. 1, mixed list, weak items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 46.281$ (at least)	$G^2_{emp}(3) = 17.632$	$p_o = .336$ $p_n = .304$	$b_{(5)} = .137$ $b_{(17)} = .273$ $b_{(33)} = .416$ $b_{(50)} = .550$ $b_{(83)} = .726$	$G^2_{emp}(3) = 24.568$	$d' = 0.953$ $SD = 1.187$	$c_{(5)} = 1.278$ $c_{(17)} = 0.889$ $c_{(33)} = 0.581$ $c_{(50)} = 0.309$ $c_{(83)} = -0.090$
Ratcliff et al. (1992)	Exp. 1, mixed list, strong items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 46.281$ (at least)	$G^2_{emp}(3) = 18.529$	$p_o = .546$ $p_n = .403$	$b_{(5)} = .158$ $b_{(17)} = .322$ $b_{(33)} = .484$ $b_{(50)} = .640$ $b_{(83)} = .842$	$G^2_{emp}(3) = 30.046$	$d' = 1.457$ $SD = 1.122$	$c_{(5)} = 1.273$ $c_{(17)} = 0.881$ $c_{(33)} = 0.591$ $c_{(50)} = 0.323$ $c_{(83)} = -0.117$

(*table continues*)

Table (*continued*)

Author	Experiment, condition	Type of bias manipulation and no. of manipulation steps	Critical G^2	2HTM fit	2HTM discrimination indices	2HTM bias indices	SDT fit	SDT discrimination index	SDT bias indices
Ratcliff et al. (1992)	Exp. 1, pure list, weak items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 46.281$ (at least)	$G^2_{emp}(3) = 15.370$	$p_o = .380$ $p_n = .286$	$b_{(5)} = .152$ $b_{(17)} = .309$ $b_{(33)} = .461$ $b_{(50)} = .595$ $b_{(83)} = .765$	$G^2_{emp}(3) = 4.975$	$d' = 1.017$ $SD = 1.188$	$c_{(5)} = 1.212$ $c_{(17)} = 0.782$ $c_{(33)} = 0.469$ $c_{(50)} = 0.206$ $c_{(83)} = -0.211$
Ratcliff et al. (1992)	Exp. 1, pure list, strong items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 46.281$ (at least)	$G^2_{emp}(3) = 9.633$	$p_o = .536$ $p_n = .404$	$b_{(5)} = .134$ $b_{(17)} = .309$ $b_{(33)} = .462$ $b_{(50)} = .609$ $b_{(83)} = .805$	$G^2_{emp}(3) = 5.481$	$d' = 1.501$ $SD = 1.208$	$c_{(5)} = 1.373$ $c_{(17)} = 0.907$ $c_{(33)} = 0.631$ $c_{(50)} = 0.377$ $c_{(83)} = -0.062$
Ratcliff et al. (1992)	Exp. 2, mixed list, weak items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 42.680$ (at least)	$G^2_{emp}(3) = 74.474^*$	$p_o = .604$ $p_n = .549$	$b_{(5)} = .071$ $b_{(17)} = .203$ $b_{(33)} = .354$ $b_{(50)} = .470$ $b_{(83)} = .811$	$G^2_{emp}(3) = 5.119$	$d' = 1.963$ $SD = 1.264$	$c_{(5)} = 1.809$ $c_{(17)} = 1.346$ $c_{(33)} = 1.048$ $c_{(50)} = 0.882$ $c_{(83)} = 0.180$
Ratcliff et al. (1992)	Exp. 2, mixed list, strong items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 42.680$ (at least)	$G^2_{emp}(3) = 152.024^*$	$p_o = .861$ $p_n = .634$	$b_{(5)} = .09$ $b_{(17)} = .243$ $b_{(33)} = .424$ $b_{(50)} = .647$ $b_{(83)} = .911$	$G^2_{emp}(3) = 46.938^*$	$d' = 2.822$ $SD = 1.163$	$c_{(5)} = 1.796$ $c_{(17)} = 1.380$ $c_{(33)} = 1.057$ $c_{(50)} = 0.809$ $c_{(83)} = 0.225$
Ratcliff et al. (1992)	Exp. 2, pure list, weak items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 42.680$ (at least)	$G^2_{emp}(3) = 13.414$	$p_o = .636$ $p_n = .515$	$b_{(5)} = .068$ $b_{(17)} = .229$ $b_{(33)} = .419$ $b_{(50)} = .577$ $b_{(83)} = .825$	$G^2_{emp}(3) = 40.547$	$d' = 2.074$ $SD = 1.357$	$c_{(5)} = 1.779$ $c_{(17)} = 1.233$ $c_{(33)} = 0.886$ $c_{(50)} = 0.635$ $c_{(83)} = 0.102$
Ratcliff et al. (1992)	Exp. 2, pure list, strong items	Base rate, 5 (5%, 17%, 33%, 50%, 83% old items in test)	$G^2_{crit}(3) = 42.680$ (at least)	$G^2_{emp}(3) = 108.096^*$	$p_o = .774$ $p_n = .710$	$b_{(5)} = .085$ $b_{(17)} = .292$ $b_{(33)} = .575$ $b_{(50)} = .673$ $b_{(83)} = .912$	$G^2_{emp}(3) = 42.231$	$d' = 2.559$ $SD = 1.092$	$c_{(5)} = 1.907$ $c_{(17)} = 1.409$ $c_{(33)} = 1.048$ $c_{(50)} = 0.913$ $c_{(83)} = 0.395$
Rhodes & Jacoby (2007)	Exp. 1	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 17.947$	$G^2_{emp}(1) = 1.666$	$p = .425$	$b_{(33)} = .479$ $b_{(67)} = .557$	$G^2_{emp}(1) = 1.841$	$d' = 1.124$	$c_{(33)} = 0.598$ $c_{(67)} = 0.466$
Rhodes & Jacoby (2007)	Exp. 2, same response keys, aware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 9.283$	$G^2_{emp}(1) = 4.325$	$p = .375$	$b_{(33)} = .384$ $b_{(67)} = .534$	$G^2_{emp}(1) = 3.502$	$d' = 0.989$	$c_{(33)} = 0.707$ $c_{(67)} = 0.437$
Rhodes & Jacoby (2007)	Exp. 2, same response keys, unaware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 26.593$	$G^2_{emp}(1) = 0.000$	$p = .380$	$b_{(33)} = .403$ $b_{(67)} = .484$	$G^2_{emp}(1) = 0.068$	$d' = 0.999$	$c_{(33)} = 0.672$ $c_{(67)} = 0.529$
Rhodes & Jacoby (2007)	Exp. 2, different response keys, aware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 16.983$	$G^2_{emp}(1) = 3.979$	$p = .320$	$b_{(33)} = .272$ $b_{(67)} = .698$	$G^2_{emp}(1) = 5.624$	$d' = 0.898$	$c_{(33)} = 0.888$ $c_{(67)} = 0.065$
Rhodes & Jacoby (2007)	Exp. 2, different response keys, unaware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 10.729$	$G^2_{emp}(1) = 0.923$	$p = .365$	$b_{(33)} = .496$ $b_{(67)} = .566$	$G^2_{emp}(1) = 0.781$	$d' = 0.953$	$c_{(33)} = 0.483$ $c_{(67)} = 0.357$
Rhodes & Jacoby (2007)	Exp. 3, Feedback Blocks 1 and 2, aware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 16.503$	$G^2_{emp}(1) = 6.009$	$p = .410$	$b_{(33)} = .409$ $b_{(67)} = .526$	$G^2_{emp}(1) = 5.171$	$d' = 1.085$	$c_{(33)} = 0.703$ $c_{(67)} = 0.500$
Rhodes & Jacoby (2007)	Exp. 3, Feedback Blocks 1 and 2, unaware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 19.387$	$G^2_{emp}(1) = 2.899$	$p = .300$	$b_{(33)} = .462$ $b_{(67)} = .548$	$G^2_{emp}(1) = 2.965$	$d' = 0.773$	$c_{(33)} = 0.458$ $c_{(67)} = 0.296$
Rhodes & Jacoby (2007)	Exp. 3, Feedback Blocks 3 and 4, aware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 15.060$	$G^2_{emp}(1) = 0.000$	$p = .290$	$b_{(33)} = .380$ $b_{(67)} = .563$	$G^2_{emp}(1) = 0.062$	$d' = 0.756$	$c_{(33)} = 0.610$ $c_{(67)} = 0.258$
Rhodes & Jacoby (2007)	Exp. 3, Feedback Blocks 3 and 4, unaware	Base rate, 2 (33% vs. 67% old items in test)	$G^2_{crit}(1) = 13.617$	$G^2_{emp}(1) = 0.533$	$p = .380$	$b_{(33)} = .463$ $b_{(67)} = .527$	$G^2_{emp}(1) = 0.514$	$d' = 0.993$	$c_{(33)} = 0.563$ $c_{(67)} = 0.443$

(*table continues*)

Table (*continued*)

Author	Experiment, condition	Type of bias manipulation and no. of manipulation steps	Critical G^2	2HTM fit	2HTM discrimination indices	2HTM bias indices	SDT fit	SDT discrimination index	SDT bias indices
Snodgrass & Corwin (1988)	Exp. 1, high imagery words	Payoff, 3 (liberal, neutral, conservative)	$G^2_{crit}(1) = 5.112$	$G^2_{emp}(1) = 0.489$	$p_o = .596$ $p_n = .692$	$b_{(lib)} = .682$ $b_{(neu)} = .550$ $b_{(con)} = .306$	$G^2_{emp}(1) = 0.087$	$d' = 1.850$ $SD = 0.946$	$c_{(lib)} = 0.783$ $c_{(neu)} = 0.975$ $c_{(con)} = 1.306$
Snodgrass & Corwin (1988)	Exp. 1, low imagery words	Payoff, 3 (liberal, neutral, conservative)	$G^2_{crit}(1) = 5.112$	$G^2_{emp}(1) = 0.300$	$p_o = .216$ $p_n = .400$	$b_{(lib)} = .651$ $b_{(neu)} = .521$ $b_{(con)} = .334$	$G^2_{emp}(1) = 0.076$	$d' = 0.785$ $SD = 0.863$	$c_{(lib)} = 0.268$ $c_{(neu)} = 0.503$ $c_{(con)} = 0.837$
Van Zandt (2000)	Exp. 1, slow trials	Base rate, 5 (20%, 35%, 50%, 65%, 80% old items in test)	$G^2_{crit}(3) = 65.596$	$G^2_{emp}(3) = 17.452$	$p_o = .612$ $p_n = .583$	$b_{(20)} = .168$ $b_{(35)} = .184$ $b_{(50)} = .346$ $b_{(65)} = .394$ $b_{(80)} = .572$	$G^2_{emp}(3) = 19.579$	$d' = 2.086$ $SD = 1.485$	$c_{(20)} = 1.482$ $c_{(35)} = 1.397$ $c_{(50)} = 1.077$ $c_{(65)} = 0.994$ $c_{(80)} = 0.669$
Van Zandt (2000)	Exp. 1, fast trials	Base rate, 5 (20%, 35%, 50%, 65%, 80% old items in test)	$G^2_{crit}(3) = 71.778$	$G^2_{emp}(3) = 28.811$	$p_o = .299$ $p_n = .524$	$b_{(20)} = .315$ $b_{(35)} = .341$ $b_{(50)} = .515$ $b_{(65)} = .516$ $b_{(80)} = .615$	$G^2_{emp}(3) = 30.146$	$d' = 1.076$ $SD = 0.931$	$c_{(20)} = 1.036$ $c_{(35)} = 0.861$ $c_{(50)} = 0.695$ $c_{(65)} = 0.690$ $c_{(80)} = 0.505$
Van Zandt (2000)	Exp. 2	Payoff, 5 matrices	$G^2_{crit}(3) = 64.376$	$G^2_{emp}(3) = 152.658^*$	$p_o = .344$ $p_n = .221$	$b_{(20)} = .091$ $b_{(35)} = .191$ $b_{(50)} = .419$ $b_{(65)} = .690$ $b_{(80)} = .848$	$G^2_{emp}(3) = 29.511$	$d' = 0.999$ $SD = 1.202$	$c_{(20)} = 1.449$ $c_{(35)} = 1.056$ $c_{(50)} = 0.523$ $c_{(65)} = -0.050$ $c_{(80)} = -0.510$

Note. An asterisk denotes significant results ($G^2_{crit} > G^2_{emp}$). 2HTM = two-high-threshold model; SDT = signal detection theory.

Received March 10, 2008
Revision received December 27, 2008
Accepted December 29, 2008 ■