Psychonomic Bulletin & Review
2008, 15 (5), 906-926
doi: 10.3758/PBR.15.5.906

Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability

Andrew L. Cohen, Caren M. Rotello, and Neil A. Macmillan
University of Massachusetts, Amherst, Massachusetts

Remember–know judgments provide additional information in recognition memory tests, but the nature of this information and the attendant decision process are in dispute. Competing models have proposed that remember judgments reflect a sum of familiarity and recollective information (the one-dimensional model), are based on a difference between these strengths (STREAK), or are purely recollective (the dual-process model). A choice among these accounts is sometimes made by comparing the precision of their fits to data, but this strategy may be muddied by differences in model complexity: Some models that appear to provide good fits may simply be better able to mimic the data produced by other models. To evaluate this possibility, we simulated data with each of the models in each of three popular remember–know paradigms, then fit those data to each of the models. We found that the one-dimensional model is generally less complex than the others, but despite this handicap, it dominates the others as the best-fitting model. For both reasons, the one-dimensional model should be preferred. In addition, we found that some empirical paradigms are ill-suited for distinguishing among models. For example, data collected by soliciting remember/know/new judgments—that is, the trinary task—provide a particularly weak ground for distinguishing models. Additional tables and figures may be downloaded from the Psychonomic Society's Archive of Norms, Stimuli, and Data, at www.psychonomic.org/archive.

Recognition decisions are widely believed to depend on two underlying processes: recollection and familiarity. The *recollection* process retrieves episodes from one's past, bringing along any number of associated details of the prior experience. In contrast, the *familiarity* process reflects a general sense of an item's "oldness" in the absence of any contextual details of prior occurrence. Mandler's (1980) famous "butcher on the bus" example conveys the distinction intuitively: You may have an immediate sense that a man sitting next to you on a bus is someone you have met or seen before, whether or not you can recall that he is the butcher at your favorite grocery store.

An apparently straightforward way to study these processes is to ask subjects whether they used recollection or familiarity when identifying a memory probe as old. This *remember–know* paradigm, first proposed by Tulving (1985), has been used extensively: Meta-analyses conducted in 2004 included about 400 experimental conditions (Dunn, 2004; Rotello, Macmillan, & Reeder, 2004), and the number continues to grow. The vast majority of these experiments have manipulated some empirical factor, such as depth of encoding, and examined the consequences for the fraction of "old" decisions that are followed by "remember" or "know" judgments. Dissociations of all forms have been observed: Remember rates have been influenced without a change in "know" responses, know rates have been affected without a change in "remember" responses, and the two response rates have moved in either opposite directions (one up, the other down) or the same direction (for a review, see Gardiner & Richardson-Klavehn, 2000). Together, these dissociations have led to the common assumption that "remember" and "know" judgments are direct and pure measures of the recollection and familiarity processes that can be manipulated independently.

The Need for Models of Remember–Know

This "process-pure" interpretation of remember–know judgments is undermined by the demonstration that empirical dissociations do not imply multiple underlying processes (Dunn & James, 2003; Dunn & Kirsner, 1988). Several researchers have suggested that remember–know data can be explained by a one-dimensional, signal-detection-based model, illustrated in Figure 1 (Donaldson, 1996; Hirshman & Master, 1997; Inoue & Bellezza, 1998). Old and new items differ in average strength; subjects set a high criterion level of memory strength, above which items are judged to be remembered, and a lower criterion, which determines which items are judged to be old. The know rate is the difference between the old and remember rates. Dunn (2004) showed that this basic model could account for all empirical dissociation patterns that have been observed.

A. L. Cohen, acohen@psych.umass.edu

Copyright 2008 Psychonomic Society, Inc.

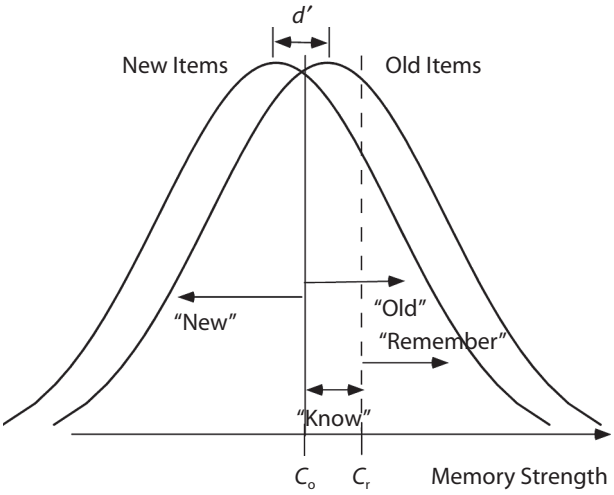

Figure 1. The one-dimensional signal detection model of remember–know judgments. Old and new items differ in average strength. "Remember" responses are given for points above C_r. Points between the two criteria lead to "know" responses. "New" responses are given for points below C_o.

In other current models, two processes or types of information are assumed to contribute to the memory judgments, but the assumption that "remember" and "know" judgments are pure measures of the recollection and familiarity processes is abandoned. As illustrated in Figure 2, old and new items are assumed to generate bivariate distributions, with targets having a greater average value than lures on both axes. The exact nature of the dimensions varies according to the model, with the *x*- and *y*-axes described as familiarity and recollection (Wixted & Stretch, 2004; Yonelinas, 1994), semantic and episodic activation (Reder et al., 2000; Tulving, 1985), item and associative information (Murdock, 2006), or global and specific strengths (Rotello et al., 2004). Macmillan and Rotello (2006) argued that there is little reason to choose any one of these labeling systems over the others. As a convenience, we will use the terms *familiarity* and *recollection*, or *x* and *y*.

All of the models to be considered in this article are derived from this memory space. Figures 3A–3C are two-dimensional representations in which the circles are equal-likelihood contours from the distributions in Figure 2. The bivariate (Gaussian) distributions of strength are uncorrelated, and lures have a smaller standard deviation than targets. To make an old–new or remember–know judgment, decision bounds are applied to the memory space, and three of the models we consider differ only in the nature of this partition. In the *process-pure* model (Figure 3A), "remember" responses are based only on the *y*-axis strength; when that strength is insufficient, "know" responses are based exclusively on *x*-axis strength (Murdock, 2006; Reder et al., 2000). Alternatively, subjects may add the strengths on the two axes (Figure 3B), assigning "remember" judgments to high values of the sum and "know" judgments to somewhat smaller sums (Wixted & Stretch, 2004). This model is a version of the *one-dimensional* model described earlier, with the added assumption that the decision axis is a weighted sum of *x* and *y*. A third possibility is that old–new decisions are based on the sum of the *x* and *y* strengths but that remember–know judgments are based on their difference (Figure 3C; Rotello et al., 2004). According to this "sum–difference theory of remembering and knowing" (*STREAK*), "remember" responses reflect a greater degree of specific strength than of global strength, and "know" responses reflect the opposite. Finally, in the *dual-process* implementation of the process-pure model (Figure 3D; Yonelinas, 2001), the recollective process is assumed to be a high-threshold process. According to this model, if sufficient detail is recollected, the subject makes a "remember" response; otherwise, a new–know decision is made, on the basis of an equal-variance Gaussian distribution of strengths. According to the process-pure (Figure 3A) and dual-process (Figure 3D) models, each memory judgment is based on either recollection or familiarity, never on both. In contrast, the one-dimensional model (Figure 3B) and STREAK (Figure 3C) assume that multiple sources of memory strength contribute to all recognition judgments.

Comparing Competing Models

Signature predictions. These remember–know models have been tested against empirical data in a number of ways. Some approaches have focused on specific aspects of the data, such as the predicted form of the receiver operating characteristic (ROC) curve (Yonelinas, 1994) or the slope of its normal–normal cousin, the *z*ROC (Rotello et al., 2004). Other strategies have used a converging-

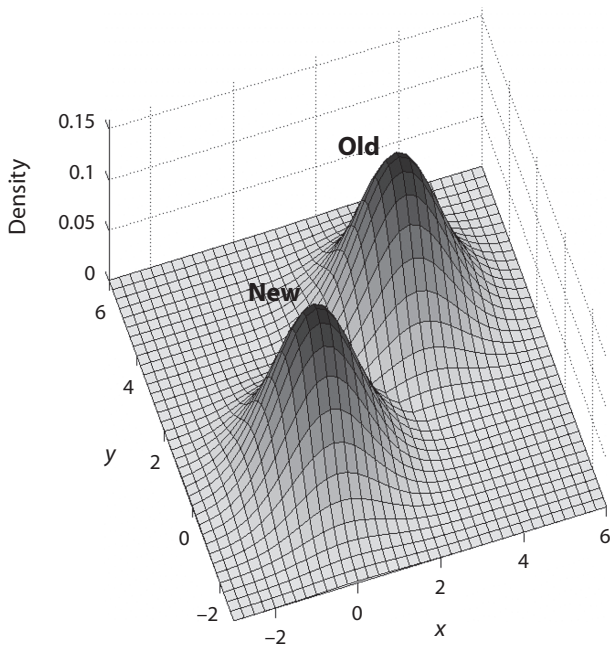

Figure 2. A two-dimensional representation of memory strengths. Old and new items differ in average strength in both *x* and *y*.

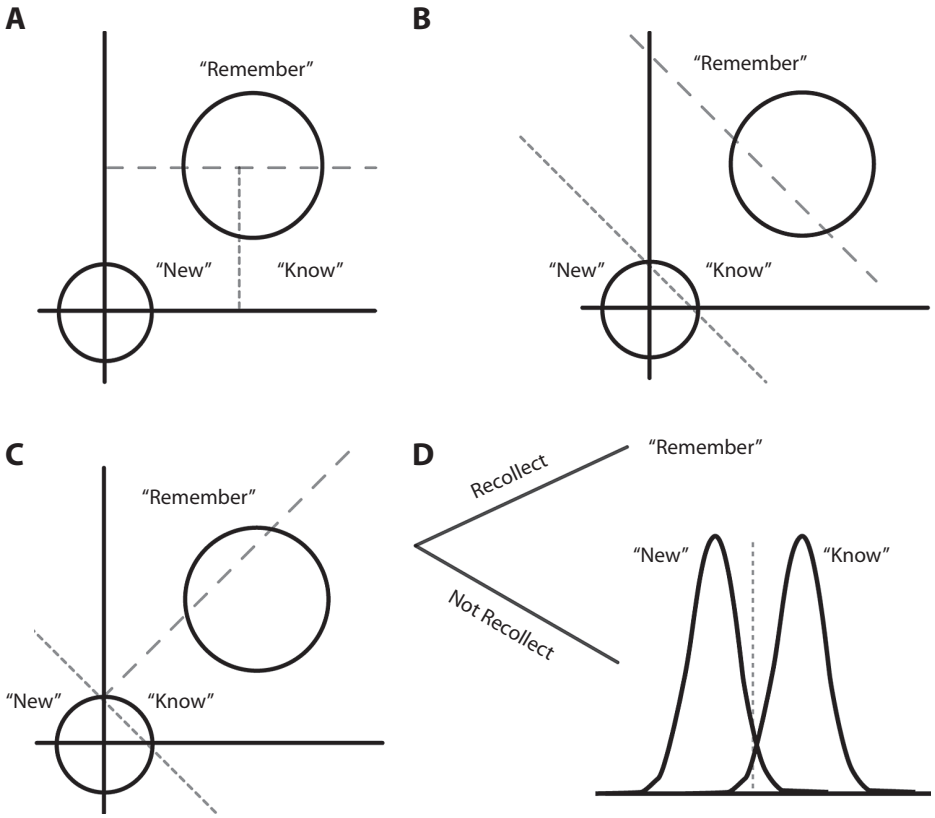

Figure 3. The decision space for the different model classes. The circles are equiprobability contours of the bivariate normal distributions from Figure 2. The lines are decision bounds that separate the memory space into regions that lead to "remember," "know," and "new" responses. (A) The process-pure model. (B) The one-dimensional model. (C) STREAK. (D) The dual-process model, which assumes a high-threshold recollection process and an equal-variance Gaussian familiarity-based process.

operations approach, comparing various parameter estimates across methodologies (Rotello, Macmillan, Reeder, & Wong, 2005; Yonelinas, 2002). In many cases, a model makes a "signature prediction" that seems to distinguish it most clearly from the competition. For example, STREAK uniquely predicts that the proportion of "old" responses that are labeled "remember" is constant when the response bias varies, and this prediction is generally supported when response bias is manipulated across conditions (Rotello, Macmillan, Hicks, & Hautus, 2006). In disputes between proponents of opposing models, it is common for each side to emphasize support for its preferred signature predictions, begging the question of which data patterns are truly the important ones.

Goodness of fit, complexity, and functional form. In fact, we believe that many aspects of the data can provide useful information. Therefore, an arguably superior approach is to fit complete quantitative models to the entirety of the data and compare goodness-of-fit (GOF) measures that quantify the agreement between a model's predictions and the observed data. An immediate problem emerges when applying this strategy to the simplest variants of the remember–know paradigm: There are as many parameters as there are data points (namely, four), and thus all models fit the data perfectly (Macmillan & Rotello, 2006; Murdock, 2006). It is possible to expand the number of data points in a variety of ways—for example, by embedding remember–know judgments in a continuous learning task (Reder et al., 2000). Our solution to the problem has been to solicit confidence ratings and use them to construct multipoint ROCs (Dougal & Rotello, 2007; Kapucu, Rotello, Ready, & Seidl, 2008; Rotello et al., 2006; Rotello et al., 2004). When the number of observations exceeds the number of parameters, the models can be evaluated on the basis of their relative GOF.

Such GOF measures must be adjusted to correct for a model's *complexity*, or the range of different data sets that it can depict. For example, a quadratic model ($y = a_q + b_q x + c_q x^2$) can generate a larger set of curves than a linear model ($y = a_l + b_l x$), so it is the more complex of the two. More precisely, by setting $c_q = 0$, the quadratic model can be made to fit any data set just as well as the linear model, so it is more complex. When the data are noisy—that is, under all realistic conditions—the extra parameter in the quadratic model will usually allow it to account for the data better, even if the data truly are linear.

Model selection measures that take this free-parameter aspect of model complexity into account include Akaike's

information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), both of which penalize models with greater numbers of free parameters. As suggested by the quadratic-versus-linear example, these techniques assume that model complexity increases with the number of parameters. When the AIC or BIC is used as a measure of GOF, a model with more free parameters must fit the data better than one with fewer parameters in order to be considered an equally good description of the data.

Systematic comparisons of remember–know models using the AIC and the BIC have been revealing. A series of ratings experiments have compared the process-pure, STREAK, and one-dimensional models by fitting each of them to the data of individual subjects. These experiments examined the effects of the emotionality of words (Dougal & Rotello, 2007; Kapucu et al., 2008), of experimental paradigm (Rotello & Macmillan, 2006), and of response bias (Rotello et al., 2006). The results are shown in Table 1, which gives the number of subjects best described by each model according to the AIC. The data are quite clear in supporting the one-dimensional model over the others.

A second source of model complexity, beyond the number of free parameters, is functional form (Pitt, Myung, & Zhang, 2002), or the way that the independent variables (x) and parameters (a, b) are combined to predict an outcome (y). Models with the same number of free parameters may nonetheless have greater or lesser complexity if they have different functional forms. For example, a sinusoidal model of the form $y = a_1 \sin b_1 x$ can exactly fit any data points that lie on an arbitrary line, $y = a_2 x + b_2$. The a_1 parameter can be used to scale the sine model to the same range as the linear model, and the b_1 parameter can be set arbitrarily high, so as to interpolate any finite set of points. In contrast, a linear model cannot satisfactorily account for data that lie on a sinusoidal function. The linear model is less complex because it accounts for fewer data sets.

Although the AIC and BIC represent clear advances over simple GOF statistics, they do not directly consider a model's functional form. If the models under consideration have the same number of parameters, the penalties imposed by AIC and BIC are not useful in determining which of the models provides the best account of the data. Nine of the 10 remember–know models to be considered in this article have the same number of parameters.

Beyond goodness of fit. As the sine-versus-linear example illustrates, models may be difficult to distinguish empirically, because one or more of them can mimic the others' data. That is, because of the difference in complexity of two models, a given model may be able to fit empirical data well, even if it does not accurately reflect the cognitive processes that led to the data. An intuition about this point comes from considering research conducted in artificial intelligence, the goal of which is to mimic the behavior of an intelligent being, without concern for accurately describing the processes that led to that behavior. Analogously, models may describe subjects' data without capturing their cognitive processing. (Indeed, Gardiner, Richardson-Klavehn, & Ramponi, 1998, made this claim about remember–know models.)

Dunn (2008) performed an analysis of remember–know models that can be thought of as an evaluation of their complexity. Making few assumptions, Dunn (2008) showed that the one-dimensional model predicts that "remember" and "old" response rates will change in the same direction when the mean of underlying target distribution is changed: For a fixed pair of response criteria, increasing the average strength of the target distribution must increase both "remember" and "old" response rates (see Figure 1). In contrast, the STREAK model does not require any particular relationship between these response rates; increasing the strength of the target distribution will increase its mean value along one or both of the x- and y-dimensions (see Figure 3C), either of which would increase the "old" response rate. The "remember" response rate, on the other hand, could either increase or decrease, depending on whether the x- or the y-dimension is primarily affected. For this reason, STREAK can fit data in which the "old" and "remember" rates change in the same direction (as is predicted by the one-dimensional model), as well as data in which the two response rates move in opposite directions. Therefore, STREAK can fit more old-versus-remember response patterns as target strength changes; it is more complex. Dunn's (2008) analysis of the published data showed that the "remember" and "old" response rates

Table 1
Number of Subjects Whose Data Are Best Described by the Process-Pure, STREAK, and One-Dimensional Models in Several Experiments, Using the Akaike Information Criterion As the Comparative Measure of Fit

Experimental Condition	Process-Pure or Dual-Process	STREAK	One-Dimensional
Emotion[a]	18	0	57
Remember-first[b]	2	7	13
Trinary[b]	2	11	35
30%-old[c]	3	0	19
70%-old[c]	5	0	19
Conservative remember bias[c]	2	2	17
Liberal remember bias[c]	3	3	15
Total	35	23	175

[a]Data from Dougal & Rotello (2007) and Kapucu et al. (2008). [b]Data from Rotello & Macmillan (2006). [c]Data from Rotello et al. (2006).

move in the same direction in all relevant studies, favoring the one-dimensional model over STREAK.

The primary goals of the present research were to evaluate the complexity of current remember–know models under a broader range of conditions and to quantify their ability to mimic each other. We employ a simulation procedure to judge model complexity (Cohen, Sanborn, & Shiffrin, 2008; Myung, Pitt, & Navarro, 2007; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). For each model, we simulated data that could occur if the generating model were the true model, then fit each of the competing models to those simulated data to determine which provided the best fit. Cases in which a model other than the true model was selected would imply high complexity of the chosen model. We ran a very large number of simulations in order to produce distributions of relative fit values; from these distributions, an optimal decision criterion for labeling one or another model as the "best" could be determined. This criterion was then used to select a model on the basis of the data.

Details of the simulations are given below in the Method section. In the remainder of the introduction, we will consider two issues that guided our calculations, one related to analysis and one to experimental design. The analysis question is whether data are better modeled at the level of individual subjects or groups of subjects. The majority of studies in the remember–know literature have been based on group data; only a few studies have fit individual data (Dougal & Rotello, 2007; Kapucu et al., 2008; Rotello & Macmillan, 2006; Rotello et al., 2006). We conducted our calculations in both ways. The question of design arises because remember–know judgments have been collected in several formats. The three paradigms we consider here produce the same data matrix in the absence of ratings, but the addition of ratings allows at least some models to make distinctive predictions.

Group Versus Individual Data

Although group-level analyses have sometimes successfully distinguished among different models' ability to fit data, such analyses do not always agree with the results from individual data. For example, Rotello and Macmillan (2006, Experiment 2) used the AIC and found that most subjects were best described by the one-dimensional model (see Table 1, trinary task). In contrast, their average data were best accounted for by the process-pure model. Other studies have revealed no such discrepancy, supporting the one-dimensional model at both the individual and group levels (Dougal & Rotello, 2007; Kapucu et al., 2008; Rotello et al., 2006).

It is well known that averaging over subjects can distort the underlying form of data functions (see, e.g., Anderson & Tweney, 1997; Ashby, Maddox, & Lee, 1994; Estes, 1956; Estes & Maddox, 2005; Sidman, 1952; Siegler, 1987). Hayes (1953), for example, showed that averaging the data from a group of individuals, each of whom learns a task according to a step function, produces data that show incremental learning. In models of remember–know judgments, distributions of strength for the various classes of memory probes are assumed to be functions of experimental and preexperimental experience with the type of items being used (e.g., words). The form of the decision bounds, in contrast, could well be a strategic choice made by the subject. If so, there is no reason that every person tested on a task should adopt the same bounds. Restricting analysis to group data thus may greatly distort the data and render this aspect of subject control invisible. The situation is not completely clear-cut, however, because in modeling individuals, far fewer data points usually constrain the model fit than when modeling group data. Cohen et al. (2008) have shown that model selection may be superior when based on group data, despite the potential distortion, especially when the data are sparse. We therefore provide analyses of group as well as individual data in this study, allowing for a test of which approach provides the better basis for model selection.

Experimental Paradigms

In the simplest, nonrating, remember–know experiments, the data consist of "remember," "know," and "new" rates for both targets and lures. Because the three response alternatives are exhaustive, there are $(3 - 1) \times 2 = 4$ degrees of freedom. The most direct method for eliciting such data is to require a single choice among the three response alternatives, and this *trinary* paradigm is sometimes used. More often, the responses are broken into two stages, in one of two ways. The most popular design, which we call *old-first*, requires subjects first to choose whether a memory probe is old or new, then, for items judged "old," to characterize their experience as "remembering" or "knowing." The other possibility, which we call *remember-first*, asks first whether an item elicits a "remember" experience, and if not, whether it is best characterized by "know" or "new."

We consider all of these experimental designs here, for three reasons. First, although these paradigms fill the same data matrix, they need not fill it with the same data. Hicks and Marsh (1999) found that the trinary paradigm produces more "remember" responses than the old-first paradigm, and Rotello and Macmillan (2006) found that the remember-first design yields still more "remembers." A second reason for considering all three paradigms is that our application required expanding the standard paradigms to include ratings, so that different remember–know models could be discriminated. This expansion resulted in data matrices that are *not* the same for all designs, as we describe later.

Finally, some models and paradigms seem particularly well suited to one another. For example, the sequence of decisions required in the old-first paradigm appears natural from the perspective of either the one-dimensional model or STREAK. The remember-first task appears especially consonant with the dual-process view, whereas the trinary task seems consistent with any of the models. Rotello and Macmillan (2006) explored this intuition in their comparison of the trinary and remember-first tasks, but they found that the models all performed fairly similarly across tasks for individual data: The differences between the AIC and BIC statistics across models were slight. This result does not, of course, rule out paradigm

differences in model complexity and discriminability, the phenomena under study here.

The remainder of this article is organized as follows. The Method section describes the general simulation procedure used to compare the models. The Results section presents additional simulation details specific to each model–paradigm combination, as well as the results of applying our procedure to relevant models in each of the three remember–know paradigms. A subsequent section assesses the effect of sampling noise on the simulation results. We conclude with recommendations for research design and analysis.

METHOD

Simulation Procedure

The simulation procedure is outlined in Figure 4 for the case in which the one-dimensional and dual-process models are compared on the old-first task (for additional details of this technique, see Wagenmakers et al., 2004, and the Appendix). First, we assumed that the data were generated by one of the models, say the one-dimensional model. Parameters for this model were then selected and allowed to vary across subjects (in a manner described below). In the Simulated Subjects column of Figure 4, θ_i represents the parameters for individual *i*. From the parameter values for a "subject" who behaves in accordance with the one-dimensional model, the probabilities of the various possible responses were calculated. The model yields, for both targets and lures, probabilities for each of seven possible responses: "new," plus "remember" and "know" at three confidence levels apiece. (Details of the data matrices produced by each of the paradigms with ratings are provided in the Results section.) In Figure 4, One-Dimensional(θ_i) represents this set of probabilities for individual *i*.

Second, an experiment was simulated, assuming numbers of subjects and of trials per condition typical of actual experiments (22 or 24 individuals and 60 trials per condition, for the simulations here). This step is illustrated in the Simulated Experiment column of Figure 4. The multinomial distribution was used to convert the model probabilities One-Dimensional(θ_i) into simulated data. If the one-dimensional model with parameters θ_i for individual *i* predicts that the probabilities of making each of the seven responses to a target (T) test probe are $p_{1Ti}, p_{2Ti}, \ldots, p_{7Ti}$, then the probability of obtaining $n_{1Ti}, n_{2Ti}, \ldots, n_{7Ti}$ responses in each response category (where $\Sigma\, n_{jTi} = 60$, the number of target trials) is given by a multinomial with parameters p_{jTi} and $n_{jTi}, j = 1 \ldots 7$. In Figure 4, this vector is denoted M[One-Dimensional(θ_i)]. One response pattern was randomly selected from this multinomial distribution, and the procedure was then repeated for the lure items, resulting in a complete response matrix of simulated data for subject *i* (denoted Data_i in Figure 4). The data from each of the 24 "subjects" were also averaged to produce Data_{Avg}. The result to this point is a completely simulated experiment, assuming that all of the subjects used the one-dimensional model in making their judgments.

To determine whether the generating (one-dimensional) model or the dual-process model better fits the data, GOF values were calculated for each model for both the individual data (Data_i) and the average data (Data_{Avg}). Subtracting these values yielded the differences shown in the Differences of GOF column of Figure 4. For example, GOF(Dual-Process | Data_i) − GOF(One-Dimensional | Data_i) gives the differences of GOF of the dual-process and the one-dimensional models for the data from individual *i*. The fit values from each instantiation of Data_i were combined to form an overall difference of fit values, labeled "Ind" in Figure 4. Because there is only one set of average data, no such combination was necessary; the difference of GOF for the average data is labeled "Avg" in Figure 4. We performed 1,000 experimental simulations of this type.

Next, the entire process was repeated, assuming that subjects behaved in accordance with the dual-process model. In Figure 4, using the dual-process model as the generating model produces probabilities Dual-Process(θ_i) with an associated multinomial M[Dual-Process(θ_i)] for each *i*. The steps in the remainder of the figure are unchanged.

Each of the 2,000 simulations (1,000 for each of the two models) yielded two fit values, one for the average ("Avg") data and one for the individual ("Ind") data. Log likelihood (denoted log *L*) was selected as a GOF measure because it can easily be combined across individuals using simple addition to yield the log of the product of the likelihoods. Log *L* equals 0 when a model accounts for the data perfectly and decreases as the GOF becomes poorer. Values of log *L* can be subtracted to compare fits: If the one-dimensional model accounts for the data better than the dual-process model, the difference log *L*(dual-process) − log *L*(one-dimensional) is negative; if the dual-process model accounts for the data better, the difference is positive.

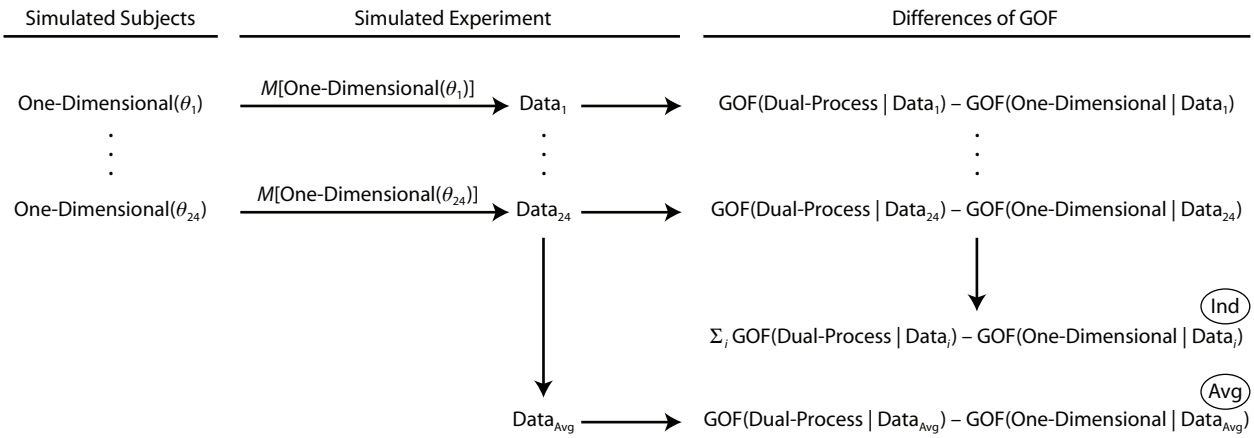

Figure 4. Outline of the simulation procedure. θ_i are the parameters selected for subject *i*. $m(\theta_i)$ are the predicted probabilities of model *m* for each experimental condition using parameters θ_i (in the figure, *m* = one-dimensional). $M[m(\theta_i)]$ is a multinomial distribution that assigns a probability to each possible data set, given the predictions of model *m* with parameters θ_i. Data_i are the data for simulated subject *i* sampled from the multinomial. Data_{Avg} are the data averaged over subjects. GOF(*m* | Data_i) is the goodness of fit of model *m* to the data from subject *i*. GOF(*m* | Data_{Avg}) is the goodness of fit of model *m* to the average data. "Ind" and "Avg" denote the results of a single experimental simulation using individual and average data, respectively.

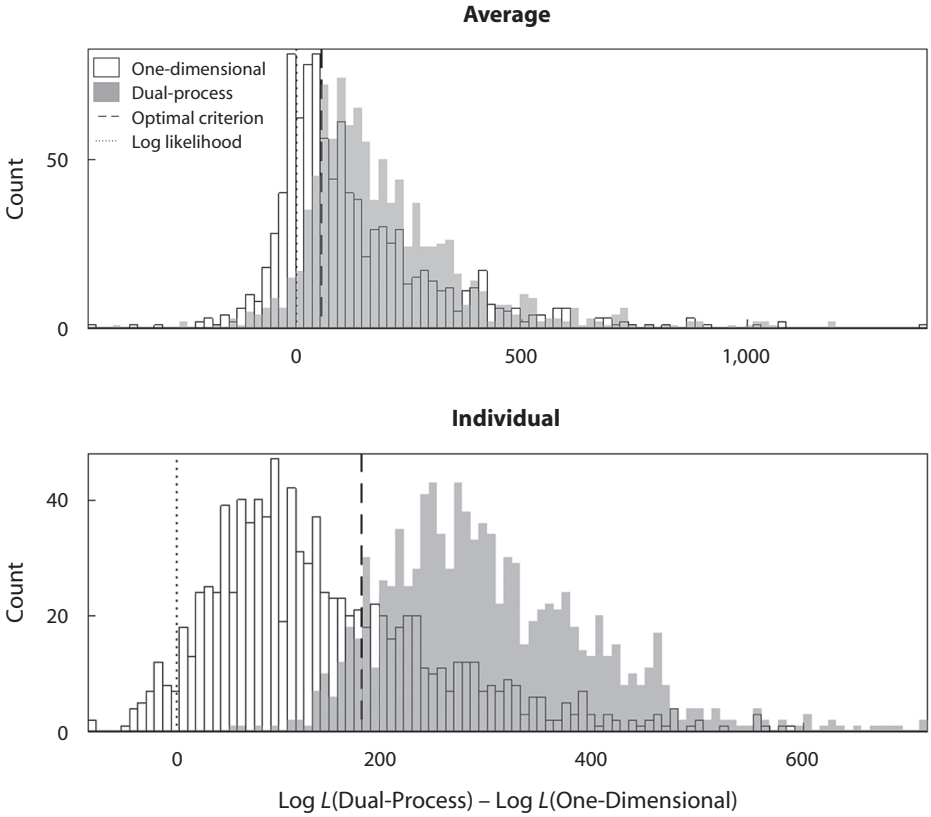

Figure 5. Sample histograms of the difference of goodness of fit (log likelihood, or log *L*) for the extended dual-process and one-dimensional models. The criteria are shown by dotted (log *L*) and dashed (optimal) lines. (Top) Results from fitting average data. (Bottom) Results from fitting individual data. Note that the axes differ in scale across the panels.

Each possible pair of remember–know models was compared within each of the three empirical paradigms (old-first, remember-first, and trinary).

Evaluation Measures

To explain our evaluation measures, we now work through the comparison of the one-dimensional and dual-process models for the old-first paradigm. In Figure 5, the results for the average and individual data are displayed in the upper and lower panels. The *x*-axis of each panel represents the difference in log *L* of the two models (notice that they are scaled differently). The histograms show the difference of log *L* values for the 1,000 simulations in which the one-dimensional (white bars) and the dual-process (gray bars) models generated the data. If the true generating model always fit better than the competitor, all of the white bars would fall to the left of zero (marked with a dotted vertical line) and all of the gray bars would fall to the right of zero. It is clear that this desirable outcome was not obtained. For the average data, about 80% of the white histogram is to the right of zero, meaning that the dual-process model accounts better for the data generated by the one-dimensional model than that model itself can. Thus, in this old-first task, the dual-process model appears to be more complex than the one-dimensional model when describing average data. The situation is much the same when we consider the individual data: Most of the white histogram is to the right of zero, meaning that the dual-process model can still account better for the data from the one-dimensional model.

When the data generated by two models are graphed together, the histograms suggest a signal detection analysis (see, e.g., Green & Swets, 1966; Macmillan & Creelman, 2005). To select a model, the difference in GOF is compared with a decision criterion. If the difference in fit values lies below the criterion, the one-dimensional model is favored; otherwise, the dual-process model is favored. The more the two distributions overlap, the poorer is the discriminability of the models, and the harder it is to select the correct generating model.

Most researchers use zero as a decision criterion, thereby selecting the model with the better fit value. Because we use log *L* as the GOF measure, the zero criterion is also called the log *L* criterion. In the example from Figure 5, every gray bar below zero and every white bar above zero represent a set of fits for which the nongenerating model better accounts for the data than the generating model itself. If the generating model is viewed as the null hypothesis, and both models are equally likely a priori, these areas give an analogue of the Type I error rate: the proportion of experiments for which the nongenerating model is incorrectly accepted as the generating model. In Figure 5, 79% of the one-dimensional model's average data were better fit by the dual-process model (white bars above zero), and 6% of the dual-process model's average data were better fit by the one-dimensional model (gray bars below zero), so the overall error rate for the average data is 42%. A similar calculation shows that the overall error rate for the individual data is 48%: The one-dimensional model was never selected in error (there are no gray bars below zero), and the dual-process model was erroneously chosen on 96% of the simulations (white bars above zero).

Because our goal is to maximize selection of the generating model, zero is not the best choice for a criterion in this example. In the lower panel of Figure 5, for example, all of the gray histogram (data generated by the dual-process model) falls above zero, so zero does not minimize the Type I error rate. A better—indeed, optimal—criterion balances false positives against false negatives; this goal can usually be achieved by placing the criterion near the intersection

point of the two distributions. Thus, for the individual data, a criterion value just below 200 would be more apt. More specifically, the optimal criterion is defined as the point that maximizes the overall probability of choosing the correct model.

Figure 5 shows a dashed line at the optimal criterion location. Examining error rates using the optimal criterion indicates how well two models can be discriminated under ideal conditions. If the two models can fit each other's data perfectly, for example, the two histograms overlap completely, and the overall error rate remains 50% regardless of the location of the criterion. This is very close to the situation in the upper panel of Figure 5. Alternatively, the two histograms could be quite distinct but might also lie completely to the right of zero, as is nearly true in the lower panel of Figure 5. In that case, using zero as a criterion would still result in a 50% overall error rate, but using the optimal criterion would result in many fewer model selection errors. For the individual data in Figure 5, using the optimal criterion increases the fraction of times the one-dimensional model is erroneously selected from 0% to 9%, but at the same time dramatically reduces the percentage of times the dual-process model is mistakenly chosen, from 96% to 33%. These shifts reduce the overall error rate from 48% (zero criterion) to 21% (optimal criterion). The overall error rate for the average data was more modestly affected by the use of the optimal criterion, falling from 42% to 36%.

Simulating Individual Differences

Each simulation began with the selection of parameter values for each subject and model; different simulated subjects were assumed to use different parameter values (e.g., for memory sensitivity). We generated those individual differences in two distinct ways. In the *uninformed* method, each parameter was constrained to lie in a range roughly determined by past experimental results (e.g., d' values between 0.01 and 5). A "central" value for each parameter was randomly and uniformly selected from that range. (Details are available in the supplemental materials, available at www.psychonomic.org/archive.) The parameters for any individual were then selected from a normal distribution with this value as its mean.[1] With this sampling technique, the parameters were selected independently—that is, the value of one parameter did not affect the values of others. However, the most relevant ranges of parameters, including correlations among the parameters, are those found by fitting the models to actual data.

In the *informed* method, the best-fitting parameters for a model when applied to empirical data were used as a population from which the parameters for that model were sampled with replacement and without addition of noise. That is, the simulation parameters were sampled from the best-fitting parameters to empirical data in each paradigm. For the old-first paradigm, we used data from Rotello et al. (2006, Experiment 1, 70% condition); for the remember-first paradigm, we used data from Rotello and Macmillan (2006, Experiment 1); and for the trinary paradigm, we used data from Rotello and Macmillan (2006, Experiment 2). These experiments were selected because the subjects in each reported confidence ratings and the data were otherwise consistent with nonrating results in the literature. The parameters for a subject were sampled as a set; the parameters were not combined across subjects. We report the results of the informed method here; the results of the uninformed method were very similar[2] and are available in the online supplemental materials.

RESULTS

In order to simulate and fit the remember–know models, it was necessary to quantify the ratings versions of each model for each of the three paradigms of interest (old-first, remember-first, and trinary). The manner in which this was accomplished differed slightly for each of the paradigms. In this section, we work through each paradigm in turn, first providing more detail about the models being compared, then describing the results of the simulations of that paradigm.

Table 2
Old-First Paradigm: Matrix of Rating Responses [6 (*Sure Old*) to 1 (*Sure New*)], With Remember (R) or Know (K) Options for Ratings of 6–4

	6		5		4				
Item Class	R	K	R	K	R	K	3	2	1
Target									
Lure									

Paradigm 1: Old-First

The initial binary choice in this commonly used paradigm is between "old" and "new"; confidence ratings qualify this judgment.[3] The complete data matrix records the remember–know decision required after each "old" response (i.e., a rating of 4, 5, or 6), as is shown in Table 2.

Models Considered

We fit four models to this data matrix: two versions of the dual-process model (Yonelinas, 2001), a one-dimensional model, and STREAK (Rotello et al., 2004). We did not fit the process-pure model from Figure 3 because the mapping from confidence rating to criterion location could be implemented in many ways for that model. For example, "sure old" could be implemented as a conservative remember bound (long-dashed lines in Figure 3), a conservative know bound (short-dashed lines), or both.

Dual-process (standard). Following Yonelinas (2001), we assumed that "remember" responses should all be assigned the highest level of confidence and that they should almost never occur for lures. In Figure 6A, "remember" responses to targets arise from the use of a recollection process. "Remember" responses to lures occur because of erroneous keypresses and other nonmemory-based responding; they are not reflected in the figure. In the absence of recollection, high values of familiarity lead to "know" responses with ratings of 6, 5, or 4, and lower values produce "new" responses with ratings of 3, 2, or 1. This *standard dual-process model* allows for a small false remember rate at confidence 6. Equations for this model can be found in the appendix to Rotello et al. (2006).[4]

Dual-process (extended). A more flexible dual-process model allows "remember" responses to be distributed freely across ratings of 6, 5, and 4. Such a modification gives the model more flexibility, at the expense of making it more similar to detection-theoretic models. In Figure 6B, recollection can lead to "remember" responses with any of the top ratings, but because this is a threshold model, the response distribution is random (but not necessarily uniform) within this region. The standard dual-process model is a special case of this *extended dual-process model* in which the "remember–4" and "remember–5" rates are fixed near zero. The predicted values of "know" and "new" proportions are relatively unaffected by this change. Equations for this model can be found in the appendix to Rotello et al. (2006).

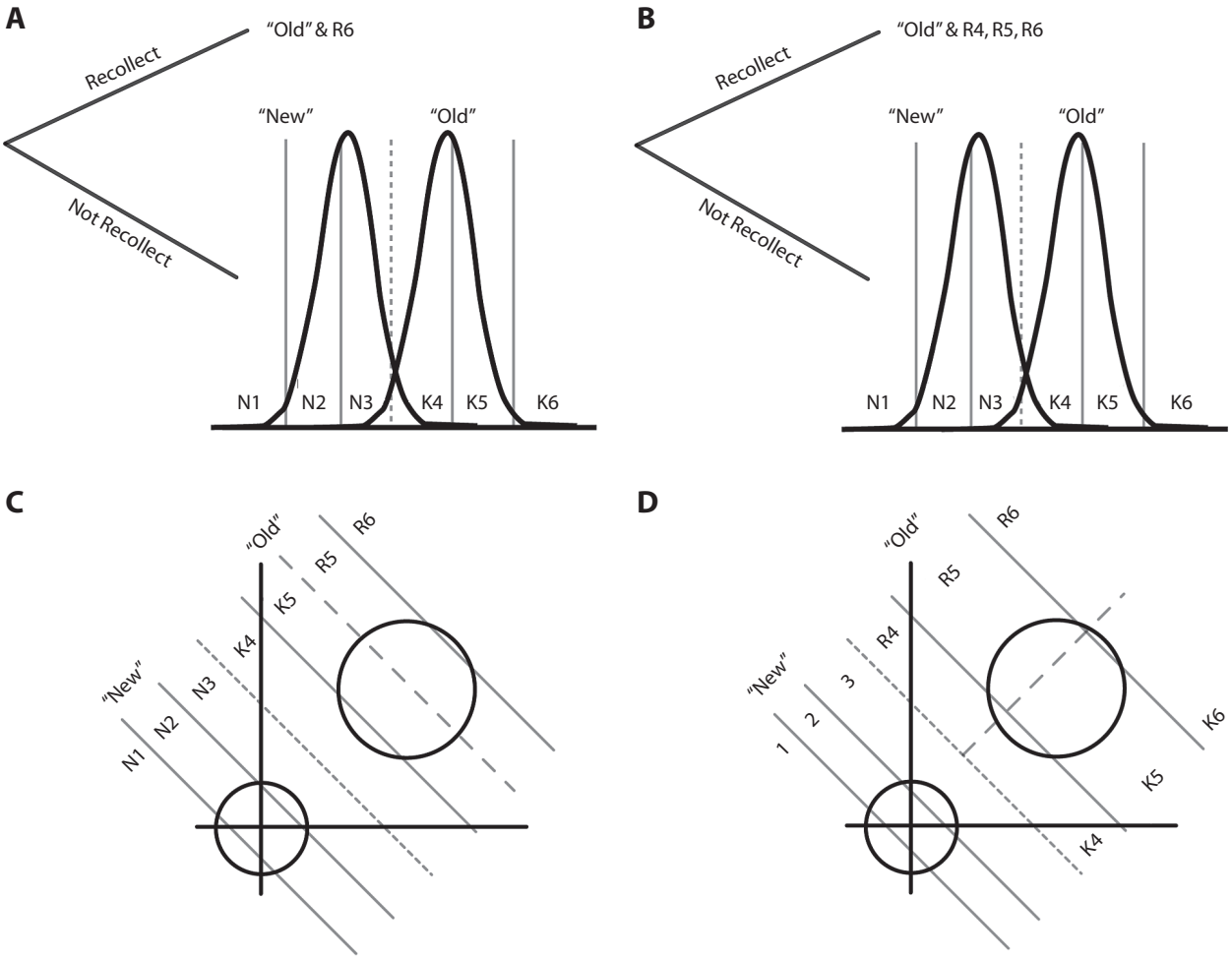

Figure 6. Models of remember–know judgments for the old-first paradigm: (A) standard dual-process, in which recollection results in the highest-confidence responses; (B) extended dual-process, in which recollection is associated with variable levels of confidence; (C) one-dimensional; (D) STREAK. For the models in A and B, recollection leads to a "remember" response and lack of recollection to a signal detection process. In panels C and D, the bivariate normal distributions shown in Figure 2 are represented by equiprobability contours, and responses are determined by linear decision bounds. Dotted lines separate "know" and "new" responses in A and B and "old" and "new" responses in C and D. Dashed decision bounds distinguish "remember" and "know" responses in C and D. Solid gray lines divide levels of confidence in all panels. The response for each region is denoted by R, K, or N ("remember," "know," or "new") and a rating (1–6, with 6 being the highest).

One-dimensional model. This model is shown in the two-dimensional space in Figure 6C, but only the $(x + y)$ decision dimension need be considered. On this dimension, lures have a standard deviation of 1, and targets have a standard deviation of $1/s$. In the simple (nonrating) remember–know task, the remember–know criterion is higher than the old–new criterion. In the rating task, there are multiple criteria for the old–new judgment, and the remember–know criterion can fall within any rating in the "old" region. In Figure 6C, this criterion is shown in the middle of confidence rating 5. The *fixed* version of this model[5] predicts that if the remember–know criterion falls in the region corresponding to rating i, then both "remember"s and "know"s can occur for that rating, only "remember"s can occur for ratings greater than i, and only "know"s can occur for ratings less than i. Equations for this model can be found in the appendix to Rotello et al. (2006).

STREAK. In the STREAK model (Figure 6D), a diagonal decision bound divides the space into "old" and "new" regions, and old–new ratings are defined by bounds parallel to this line. A bound orthogonal to these confidence criteria separates remembering from knowing. Equations for the nonrating version of STREAK can be found in the appendix to Rotello et al. (2004) and are easily generalized.

Simulation Results

The histograms produced by application of the simulation technique to the old-first paradigm are given in Figure 7. Each panel of Figure 7 is of the same form as those in Figure 5; indeed, the data from Figure 5 are repeated in the left-hand panels of Figure 7. Simulations using average data are shown in the upper row, and those from individual data appear in the lower row; each of the

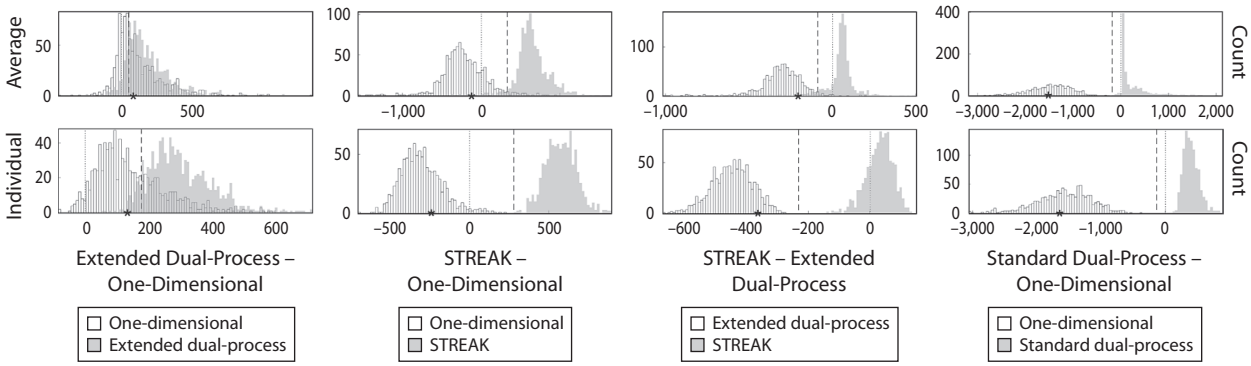

Figure 7. Difference-of-goodness-of-fit histograms comparing the one-dimensional and extended dual-process, one-dimensional and STREAK, extended dual-process and STREAK, and one-dimensional and standard dual-process models, when fit to average data (upper row) and individual data (lower row), using the informed parameter selection method in the old-first paradigm. Criteria are shown by dotted (log *L*) and dashed (optimal) lines. The asterisks on the *x*-axes show the differences in goodness of fit for the empirical data. Extreme outliers (>4 *SD*s from each model's mean) have been excluded from the figure. Note that the axes differ in scale across the panels.

four columns displays a comparison of a particular pair of models. Data points more than four standard deviations from the appropriate model's mean were removed from all histograms in the figures (but not from the analyses).

First, consider the relative complexity of the models. One sensible measure of complexity is obtained by using the log *L* criterion to calculate the proportion of times that the nongenerating model nonetheless produced better GOF. As a rule of thumb, the model with the lower error rate is the more complex of the two. Error rates for the old-first task are shown in Table 3; the four major columns correspond to the four columns of Figure 7. The upper, middle, and lower portions of the table indicate the overall error rates and the error rates for the first and second models of each comparison. On the far left, Models 1 and 2 are the one-dimensional and extended dual-process models. The error rates in this section of Table 3 and Figure 7 are the same as those shown in Figure 5: Using the log *L* criterion, the wrong model was chosen overall 42% of the time with average data and 48% of the time with individual data. Because the error rate for the extended dual-process model is lower than that for the one-dimensional model (compare the lower and middle portions of the table), it is the more complex model.

The other columns in the table reflect other possible model comparisons. Using the relative proportions of incorrect model selections with the log *L* criterion as a measure of model mimicry, the extended dual-process model is far more complex than the one-dimensional model for both average and individual data, as we just demonstrated; it is also more complex than the STREAK model, although the error rates for that comparison are lower. STREAK is more complex than the one-dimensional model, but only marginally so for individual data fits. The one-dimensional and standard dual-process models are approximately equal in complexity (see also Wixted, 2007).[6]

Now consider the discriminability of the models. Under the best possible circumstances, how easy is it to distinguish the generating and nongenerating models? Table 3 indicates the proportions of times that the nongenerating

Table 3
Proportions of Incorrect Model Selections When Using the Log Likelihood (Log *L*) and Optimal (Opt) Criteria in the Old-First Paradigm

	Model Comparison							
	(1) 1-D vs. (2) Extended DP		(1) 1-D vs. (2) STREAK		(1) Extended DP vs. (2) STREAK		(1) 1-D vs. (2) Standard DP	
Data Type	Log *L*	Opt	Log *L*	Opt	Log *L*	Opt	Log *L*	Opt
Overall								
Average	.42	.36	.10	.04	.10	.07	.02	.00
Individual	.48	.21	.01	.00	.12	.00	.00	.00
Generating Model: Model 1								
Average	.79	.57	.20	.04	.01	.03	.00	.00
Individual	.96	.33	.03	.00	.00	.00	.00	.00
Generating Model: Model 2								
Average	.06	.15	.01	.05	.19	.11	.03	.00
Individual	.00	.09	.00	.00	.24	.00	.00	.00

Note—Proportions represent the rates at which simulations selected the nongenerating model as having generated the data. 1-D, one-dimensional model; DP, dual-process model.

model was erroneously selected when using the optimal criterion. This proportion is an indication of the overlap of the two distributions: The more the distributions overlap, the less discriminable they are, and the larger the selection error. For individual data, the one-dimensional and extended dual-process models are relatively difficult to discriminate from one another (the overall error rate is 21%); the same is true for the average data. These high error rates reflect the large overlap of the histograms in the left panels of Figure 7. The other models' histograms do not overlap much or at all (the overall error rates range from 0% to 7%), so they can be discriminated with good to excellent success using the optimal criterion. The locations of these criteria are given in the upper rows of Table 4.

Application of the Simulation Results to Real Data

This model selection procedure can be applied to data from a real experiment, for which the generating model is, of course, not known. To this end, all of the models were fit to real data from the old-first paradigm (Rotello et al., 2006, Experiment 1, 70% condition) using the same method employed when fitting the models to the simulated data, and GOF values were obtained. The differences of GOF for each model comparison are marked as small asterisks in the histograms of Figure 7 (and all analogous histograms) and are shown in the middle rows of Table 4.

Using the log L criterion and average data, the extended dual-process model is selected over the one-dimensional model, because the (extended dual-process − one-dimensional) difference is positive (89.39). The extended dual-process model is also selected over STREAK, because the (STREAK − extended dual-process) difference is negative (−203.10). Similarly, the one-dimensional model is preferred over both STREAK and the standard dual-process model. The individual data support identical conclusions with the log L criterion.

The complexities of these models are far from equal, as discussed previously. For this reason, the models can be discriminated better using the optimal criteria from Table 4. For average data, the extended dual-process model is selected over the one-dimensional model and STREAK. The one-dimensional model is selected over STREAK and the standard dual-process model. For individual data, the one-dimensional model is preferred over STREAK and both versions of the dual-process model, but the extended dual-process model is selected over STREAK. The differences between the average and individual results are discussed in detail below.

These empirical data (Rotello et al., 2006, Experiment 1, 70% condition) may appear biased against the dual-process model, because a liberal response criterion was induced (subjects were falsely told that 70% of test items were studied) and the remember false alarm rate was high (20%). To evaluate this concern, we also fit the empirical data from the conservative condition of the same experiment (Rotello et al., 2006, Experiment 1, 30% condition), for which the remember false alarm rate was much lower (9%). The resulting GOF differences for each model comparison are shown in the lower rows of Table 4. In comparisons using the log L criterion, the average data suggest that the extended dual-process model provides the best fit, and the individual data concur. Rotello et al. (2006, Experiment 1) used the AIC and reached a different conclusion—namely, that most (19 of the 22) subjects' individual data were best fit by the one-dimensional model. When the differences in log L statistics are compared with the optimal criteria in Table 4, the standard dual-process model is preferred for average data, whereas the one-dimensional model is preferred for individual data. Some caution is required with this conclusion, however, because the optimal criteria of Table 4 were derived with informed parameters sampled from the 70% condition, not the 30% condition.

Conclusions: Old-First Paradigm

The extended dual-process model is much more complex than the other models, perhaps because it can account for data in which the "remember" responses to targets occur exclusively at the highest confidence level as well as data in which "remember" responses to targets are distributed over a range of confidence levels. STREAK is more complex than the one-dimensional model, but only for average data. All other comparisons show approximately equal complexity. With the exception of the comparison between the one-dimensional and dual-process models, model discriminability is high.

When using the old-first paradigm, the one-dimensional model is clearly preferred to the others for fitting indi-

Table 4
Optimal Criteria and Observed Goodness-of-Fit Differences in the Old-First Paradigm

		Model Comparison			
Condition	Data Type	(1) 1-D vs. (2) Extended DP	(1) 1-D vs. (2) STREAK	(1) Extended DP vs. (2) STREAK	(1) 1-D vs. (2) Standard DP
Optimal criteria	Average	53.64	327.33	−93.02	−192.90
	Individual	177.51	278.19	−233.96	−145.30
		Data From Rotello et al. (2006, Experiment 1)			
70% condition	Average	89.39†	−113.72*	−203.10*	−1,577.63*
	Individual	130.44*	−240.32*	−370.76*	−1,646.04*
30% condition	Average	39.63*	−68.59*	−108.22*	−161.23†
	Individual	68.97*	−66.02*	−134.99†	−174.90*

Note—1-D, one-dimensional model; DP, dual-process model. *Model 1 selected with optimal criterion. †Model 2 selected with optimal criterion.

vidual data, whereas the extended dual-process model is preferred for average data. These conflicting conclusions may be explained by the presence of sampling noise in the individual subjects' data. As we will show in a later section, noise in the data is particularly problematic when data are averaged and can lead to a reversal of the ordering of model preference, relative to individual data.

Paradigm 2: Remember-First

The initial binary choice in the remember-first paradigm is whether or not an item is "remembered." Ratings are used only for cases in which remembering is denied, as is shown in the data matrix in Table 5.

Models Considered

For this paradigm, we evaluated the process-pure, one-dimensional, and STREAK models.

Process-pure model. The process-pure model for the remember-first paradigm (Figure 8A) is similar to the dual-process model for the old-first task.[7] The essential changes are that "remember" responses do not entail a rating, so that all ratings are levels of "know" or "new," and that the recollection process is continuous rather than high-threshold. This paradigm yields one remember response rate for targets, and one for lures. Given this property of the data, we cannot distinguish recollection that is based on a continuous distribution (with "remember" responses occurring at different strengths) from recollection based on a threshold process that leads to highest-confidence "old" judgments. Thus, the process-pure and dual-process models make the same predictions for the remember-first paradigm. As in the old-first paradigm, a horizontal decision line divides remembering from nonremembering, and five criteria divide the "nonremember" region into "new" to "know" rating categories. Equations for this model can be found in the appendix to Rotello and Macmillan (2006).

One-dimensional model. The one-dimensional model for the remember-first paradigm is similar to that for the old-first paradigm, in that it has a total of six criteria partitioning the decision axis. As shown in Figure 8B, however, all "remember" judgments arise from observations above the highest criterion, which divides remembering from nonremembering, and the other criteria define "new" to "know" rating categories. Equations for this model can be found in the appendix to Rotello and Macmillan (2006).

STREAK. In the STREAK model for the remember-first design (Figure 8C), the partition into "remember," "know," and "new" responses is similar to the STREAK model as applied to the old-first task (Figure 3C). The "remember" region forms a corner of the space that is not subdivided into ratings (a *patio*, in the terminology of Rotello and Macmillan, 2006), and the know criteria are parallel to the old–new bound. If an item is not remembered, a rating from 6 (*sure know*) to 1 (*sure new*) is made; ratings within the "new" region are produced by criteria below the old–new bound. Equations for this model can be found in the appendix to Rotello and Macmillan (2006).

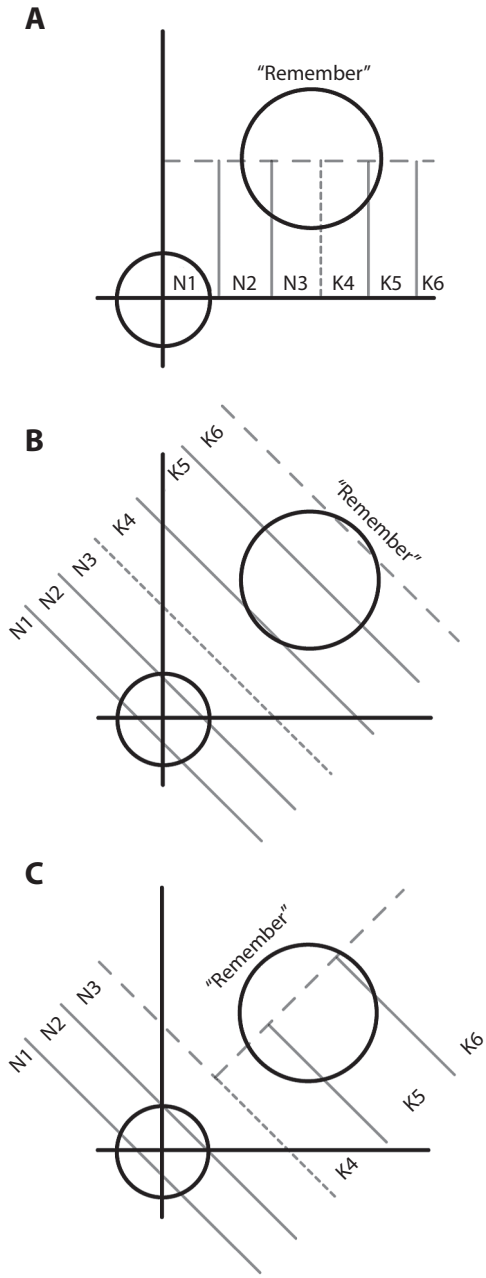

Figure 8. Models of remember–know judgments for the remember-first paradigm: (A) process-pure, (B) one-dimensional, (C) STREAK. The notation is analogous to that in Figure 6.

Table 5
Remember-First Paradigm: Matrix of Remember/Know/New Responses, With Ratings From 6 (*Know Lots of Details*) to 1 (*Sure New*)

	Remember	Know			New		
Stimulus Class	(No Rating)	6	5	4	3	2	1
Target							
Lure							

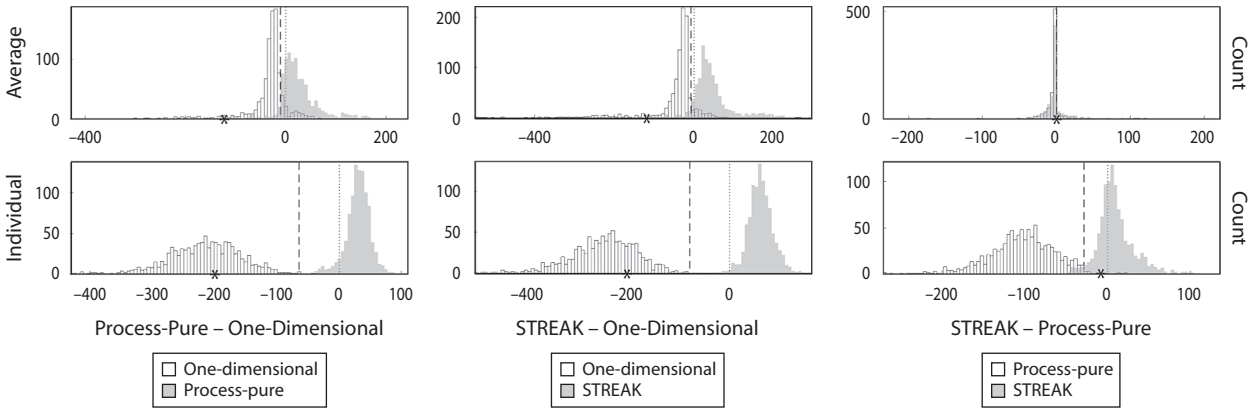

Figure 9. Difference-of-goodness-of-fit histograms comparing the one-dimensional and process-pure, one-dimensional and STREAK, and process-pure and STREAK models, when fit to average data (upper row) and individual data (lower row), using the informed parameter selection method in the remember-first paradigm. Criteria are shown by dotted (log *L*) and dashed (optimal) lines. The asterisks on the *x*-axes show the differences in goodness of fit for the empirical data. Extreme outliers (>4 *SD*s from each model's mean) have been excluded from the figure. Note that the axes differ in scale across the panels.

Simulation Results

The histograms for the remember-first paradigm are given in Figure 9. The overall and separate model error rates for the remember-first paradigm are given in Table 6. Again, the proportions of incorrect model selections using the log *L* criterion can be used as a measure of model complexity. The most striking result is that the process-pure model is much more complex (has a lower error rate) than STREAK in this paradigm. For example, with average data, the process-pure model accounts for 57% of STREAK's data, whereas STREAK accounts for just 28% of the process-pure data. The complexities of the one-dimensional model and STREAK are about equal. The one-dimensional model is slightly more complex than the process-pure model, particularly for average data.

Model discriminability is measured by the overall proportion of incorrect model selections using the optimal criteria (shown in Table 7). Several pairs of models can be discriminated reasonably to extremely well in the remember-first paradigm. When using individual data, the highest overall error rate is a mere 3%, whereas average data present somewhat more of a challenge, yielding higher error rates. The one-dimensional model is fairly easy to distinguish from STREAK with average data, but somewhat more difficult to distinguish from the process-pure model. Discriminability of the process-pure and STREAK models approaches chance.

Application of the Simulation Results to Real Data

We applied this method to experimental data from the remember-first paradigm (Rotello & Macmillan, 2006, Experiment 1); differences in GOF are shown as asterisks in Figure 9 and are given in the lower rows of Table 7. First, consider the log *L* criteria. For average data, these results suggest that the one-dimensional model should be preferred to both the process-pure model and STREAK, and that there should be no preference between STREAK and the process-pure model. A similar conclusion emerges for the individual data: The one-dimensional model is preferred over both the process-pure and STREAK models, and the process-pure model is selected over STREAK by a slim margin.

Using the optimal criteria in Table 7 yields similar conclusions. For average data, the one-dimensional model is preferred to both the process-pure model and STREAK, and there is no preference between STREAK and the process-pure model. A similar story emerges for the individual data: The one-dimensional model is preferred to the process-pure model and STREAK, and STREAK is selected by a slim margin over the process-pure model. Consistent with this conclusion, Rotello and Macmillan (2006) found that the data of 13 (using the AIC) or 17 (using the BIC) of their 22 subjects were best described by the one-dimensional model.

Table 6
Proportions of Incorrect Model Selections When Using the Log Likelihood (Log *L*) and Optimal (Opt) Criteria in the Remember-First Paradigm

	Model Comparison					
	(1) 1-D vs. (2) PP		(1) 1-D vs. (2) STREAK		(1) PP vs. (2) STREAK	
Data Type	Log *L*	Opt	Log *L*	Opt	Log *L*	Opt
Overall						
Average	.15	.13	.09	.08	.43	.44
Individual	.04	.00	.01	.00	.17	.03
Generating Model: Model 1						
Average	.09	.13	.08	.09	.28	.15
Individual	.00	.00	.00	.00	.01	.02
Generating Model: Model 2						
Average	.22	.12	.09	.07	.57	.73
Individual	.08	.00	.02	.00	.33	.04

Note—Proportions represent the rates at which simulations selected the nongenerating model as having generated the data. 1-D, one-dimensional model; PP, process-pure model.

Table 7
Optimal Criteria and Observed Goodness-of-Fit Differences in the Remember-First Paradigm

		Model Comparison		
	Data Type	(1) 1-D vs. (2) PP	(1) 1-D vs. (2) STREAK	(1) PP vs. (2) STREAK
Optimal criteria	Average	−9.83	−7.08	0.12
	Individual	−64.39	−78.92	−30.20
Data from Rotello & Macmillan (2006, Experiment 1)	Average	−123.13*	−123.13*	0.00*
	Individual	−202.79*	−211.47*	−8.68†

Note—1-D, one-dimensional model; PP, process-pure model. *Model 1 selected with optimal criterion. †Model 2 selected with optimal criterion.

Conclusions: Remember-First Paradigm

STREAK and the one-dimensional model are of approximately equal complexity. The process-pure model is less complex than the one-dimensional model but more complex than STREAK, for both individual and average data. When using average data, model discriminability is fair (one-dimensional and STREAK) to poor (process-pure and STREAK). For individual data, model discriminability is excellent. When evaluated against empirical data, the one-dimensional model is clearly preferred.

Paradigm 3: Trinary

In the trinary task, subjects' initial response to each memory probe is a trinary remember/know/new judgment. If the decision is "remember" or "know," a rating response is made on a 3-point scale from *least* to *most confident*.[8] The resulting data matrix has the form shown in Table 8.

Models Considered

Three models of the trinary task are considered: the one-dimensional model, the process-pure model, and STREAK. These models are the three basic models first applied to individual data from the trinary paradigm by Rotello and Macmillan (2006). The dual-process model was not fit because ratings are not typically applied to threshold processes (i.e., remembering in the dual-process model).

Process-pure. The model in its simplest, nonrating, form is displayed in Figure 3A: The stimulus space is divided into three regions that represent the three possible responses. In the version of the model we tested, "remember" responses were followed by a judgment of the degree of remembering. As shown in Figure 10A, this rating depends only on strength from the y-dimension, with higher ratings corresponding to greater strength. Similarly, "know" judgments were supplemented with an evaluation of the strength of that response that depended only on the x-value. Equations for this model can be found in the appendix to Rotello and Macmillan (2006).

One-dimensional model. The one-dimensional model (Figure 10B) is likewise a natural elaboration of the nonrating version (Figure 3B). "Remember" and "know" responses reflect different strengths along the decision axis (which points toward the upper right corner of the stimulus space), with the strongest "remember" responses falling in the extreme upper right and the weakest "know" responses falling just above the old–new bound. Equations for this model are described in the appendix to Rotello and Macmillan (2006).

STREAK. The nonrating version of STREAK is shown in Figure 3C. The rating version (Figure 10C) assumes that the highest-confidence "know" response corresponds to the region farthest from the "remember" regions—that is, it reflects high confidence that the judgment should not be "remember." This model is the same as the model for binary old–new judgments followed by remember–know ratings described by Rotello et al. (2004); Rotello et al. (2006) called it *STREAK parallel*. (They also tested a version of the model—*STREAK parquet*—that assumed that stronger knowing reflected greater certainty that the item was not a lure. We did not include that version in our simulations, because it accounted for fewer subjects' data than did STREAK parallel.)

Simulation Results

The histograms for the trinary paradigm are given in Figure 11; the overall and separate model error rates are in Table 9. In their diagnoses of model complexity (using log L as a criterion), the average and individual data give very different results. With average data, the process-pure model is more complex (has a lower error rate) than both the one-dimensional model and STREAK, and the one-dimensional model is more complex than STREAK. When individual data are used, these results reverse: The one-dimensional model is more complex than the process-pure model, and STREAK is more complex than both. Because averaging can distort the underlying form of the data, such inconsistencies are not necessarily surprising; using average or individual data can shift the relative complexity of models. Both the average and individual results

Table 8
Trinary Paradigm: Matrix of New/Know/Remember Responses, With Ratings From 1 to 3 Within the "Know" and "Remember" Categories

	New	Know			Remember		
Stimulus Class	(No Rating)	1	2	3	1	2	3
Target							
Lure							

Note—For "know" responses, ratings range from 1 (*weak feeling of knowing*) to 3 (*sure was studied, but no details*); for "remember" responses, they range from 1 (*remember few details*) to 3 (*remember lots of details*).

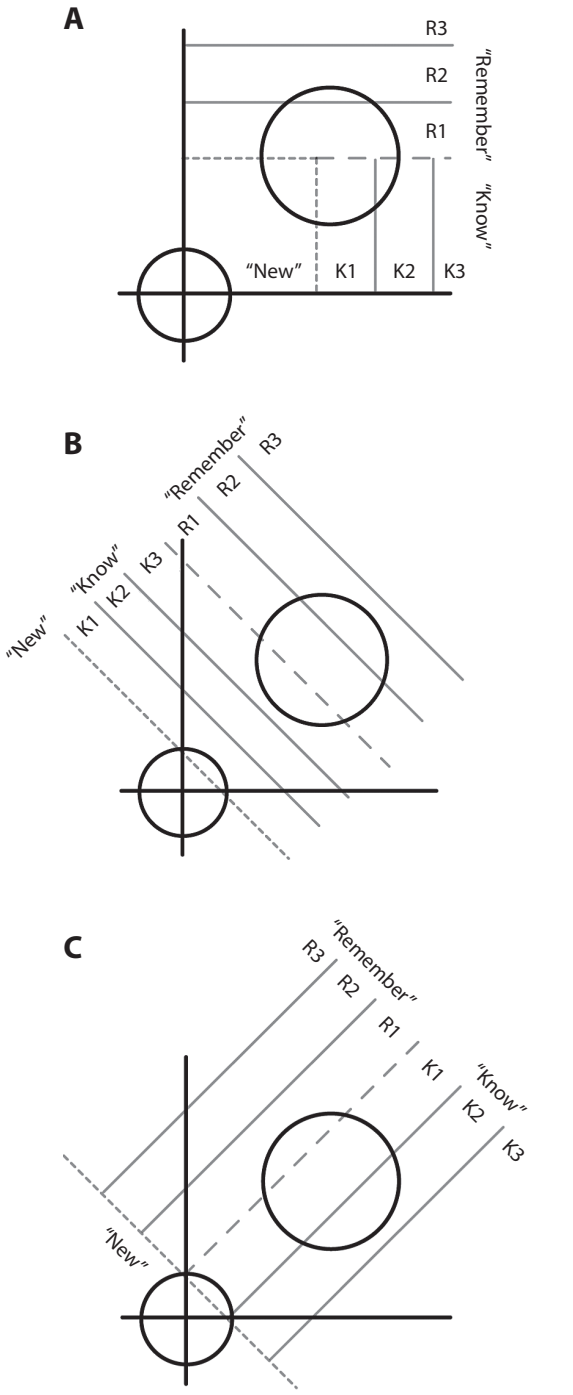

Figure 10. Models of remember–know judgments for the trinary paradigm: (A) process-pure, (B) one-dimensional, (C) STREAK. The notation is analogous to that in Figure 6.

are relevant to teasing apart the remember–know models, since the literature includes many examples of analyses and model fits based on average data, as well as some newer examples of individual-level fits.

Now consider the discriminability of the models, as measured by the overall proportion of incorrect model selections using the optimal criteria, whose locations are given in Table 10. For average data, the overall error rate is between 12% and 25%, whereas for the individual data both the one-dimensional and process-pure models are highly discriminable from STREAK (the overall error rates are 6% and 4%, respectively). In contrast, it is more difficult to distinguish data generated from the one-dimensional and process-pure models (error rate is 25%).

Application of the Simulation Results to Real Data

All three models were fit to real data from the trinary paradigm (Rotello & Macmillan, 2006, Experiment 2); the observed differences in GOF for each model comparison are shown with asterisks in Figure 11 and are given in the lower rows of Table 10.

First, consider model selection using the log L criterion. For average data, the process-pure model is selected over both STREAK and the one-dimensional model (by a slim margin). The one-dimensional model is preferred over STREAK. For individual data, the one-dimensional model is selected over both the process-pure (by a hair) and STREAK models; the process-pure model is selected over STREAK. Consistent with this result, Rotello and Macmillan (2006) concluded, on the basis of the AIC statistic, that the one-dimensional model was preferred in the trinary task when fit to individual data (35 of the 48 subjects).

Models can also be selected by comparing the GOF differences for the experimental data with the optimal criteria in Table 10. For average data, the one-dimensional model has a very slight advantage over the process-pure model; the one-dimensional and process-pure models are preferred over STREAK. The same is true for the individual data.

Conclusions: Trinary Paradigm

The trinary paradigm is not well suited to distinguishing among this set of models. When using average data, the process-pure model is more complex than both the one-dimensional model and STREAK, and the one-dimensional model is more complex than STREAK. When using individual data, these results reverse: The one-dimensional model is more complex than the process-pure model, and STREAK is more complex than both. Regardless of which type of data are considered, the one-dimensional and process-pure models are difficult to discriminate. Using individual data, STREAK is discriminable from the others, but when using average data, STREAK is easily confused with both the one-dimensional and dual-process models. When using empirical data to compare the models, STREAK is never preferred. Given the ability of the one-dimensional and process-pure models to mimic each other, it is perhaps not surprising that there is no clear preference between these two models.

Assessing the Effects of Noisy Data

Our calculations have sometimes led us to clear-cut preferences for one model over the others, but sometimes they led to discrepancies between individual and average data. One possible explanation, evaluated in this section, is that

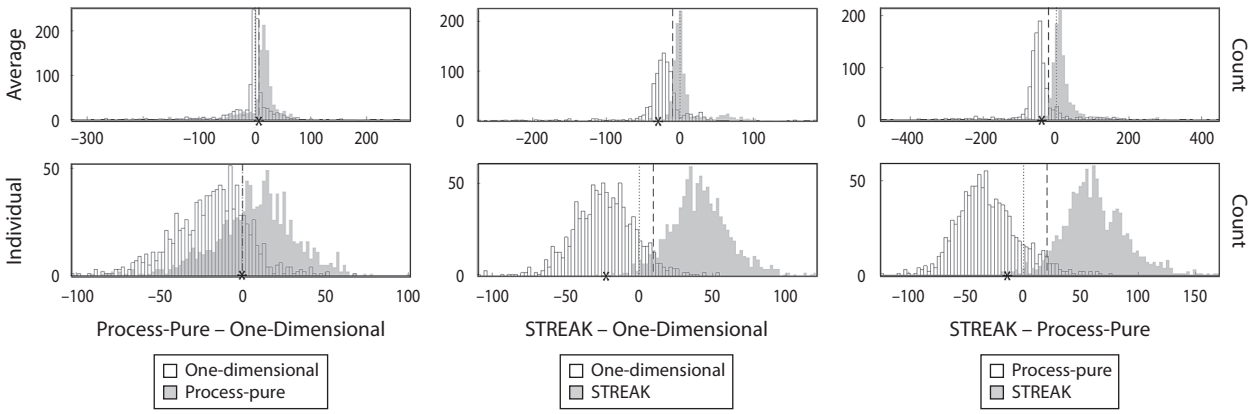

Figure 11. Difference-of-goodness-of-fit histograms comparing the one-dimensional and process-pure, one-dimensional and STREAK, and process-pure and STREAK models, when fit to average data (upper row) and individual data (lower row), using the informed parameter selection method in the trinary paradigm. Criteria are shown by dotted (log *L*) and dashed (optimal) lines. The asterisks on the *x*-axes show the differences in goodness of fit for the empirical data. Extreme outliers (>4 *SD*s from each model's mean) have been excluded from the figure. Note that the axes differ in scale across the panels.

the obtained differences are due to noise in the data, especially when there is high overlap between the fits of the competing models. That is, the data may not strongly suggest that one model is to be preferred over another, so the conclusions appear unstable when analyzed differently. Using a new bootstrapping analysis, the experimental data were resampled and refit with each of the models to determine the stability of the model selection conclusions. We applied the technique to data from each of the three paradigms.

Old-First Paradigm

In the old-first task, a reversal of model dominance was observed when using individual rather than average data: The one-dimensional model was selected when individual data were used, yet the extended dual-process model was preferred when average data were used. How can these seemingly contradictory results be reconciled? To address this question, the individual human subjects' data were treated as a pool from which to sample. Twenty-four individuals were drawn from the pool with replacement. Note that the empirical data were used here, not data generated from a model. As in the simulations described earlier, the competing models were fit to both the individual and average data from these sampled subjects. The differences of GOF were compared with the appropriate optimal criterion to select the better-fitting model. We repeated the process of sampling a set of subjects, fitting the models, and comparing the difference of GOF to the appropriate optimal criterion 100 times. If the data strongly supported one of the models, the proportion of times for which that model was selected out of these 100 samples would be high. At the other extreme, if half of the samples fell on one side of the optimal criterion and half on the other, the data would completely fail to discriminate between the two models.

The proportions of times in which Model 1 was selected from these sampled data are given in the upper section of Table 11. For example, the proportion of these bootstrapped data in which the one-dimensional model was preferred over the extended dual-process model, using average data, was .30. This result is quite stunning. It means that, when bootstrapped *from the same data*, model selection is quite inconsistent: 30% of the time, the one-dimensional model is chosen; 70% of the time, the extended dual-process model is selected. The situation is much clearer when using individual data: For 97% of the resampled data, the one-dimensional model was selected, and the extended dual-process model was selected for only 3% of the data sets. Comparisons of STREAK with the one-dimensional and extended dual-process models give similar results. Regardless of the type of data used, the one-dimensional model is clearly preferred over the standard dual-process model. Although there is a relatively high degree of inconsistency when using average data, the results when using individual data are clear-cut: The one-dimensional model is preferred.

Table 9
Proportions of Incorrect Model Selections When Using the Log Likelihood (Log *L*) and Optimal (Opt) Criteria in the Trinary Paradigm

	Model Comparison					
	(1) 1-D vs. (2) PP		(1) 1-D vs. (2) STREAK		(1) PP vs. (2) STREAK	
Data Type	Log *L*	Opt	Log *L*	Opt	Log *L*	Opt
	Overall					
Average	.35	.25	.34	.17	.20	.12
Individual	.25	.25	.07	.06	.07	.04
	Generating Model: Model 1					
Average	.48	.22	.12	.19	.12	.17
Individual	.18	.18	.12	.05	.13	.04
	Generating Model: Model 2					
Average	.21	.28	.56	.16	.28	.08
Individual	.33	.34	.02	.06	.01	.05

Note—Proportions represent the rates at which simulations selected the nongenerating model as having generated the data. 1-D, one-dimensional model; PP, process-pure model.

Table 10
Optimal Criteria and Observed Goodness-of-Fit Differences in the Trinary Paradigm

		Model Comparison		
	Data Type	(1) 1-D vs. (2) PP	(1) 1-D vs. (2) STREAK	(1) PP vs. (2) STREAK
Optimal criteria	Average	7.70	−10.19	−21.03
	Individual	−0.04	9.25	20.35
Data from Rotello & Macmillan (2006, Experiment 2)	Average	7.51*	−29.08*	−36.59*
	Individual	−0.40*	−23.27*	−14.21*

Note—1-D, one-dimensional model; PP, process-pure model. *Model 1 selected with optimal criterion.

Remember-First Paradigm

When using either average or individual data in the remember-first task, the one-dimensional model is selected over the others. To determine the strength of this conclusion, we applied the data-bootstrapping procedure to the models in the remember-first paradigm (see Table 11, middle section). The fits to average data show less consistency than the fits to the individual data, although the effect here is not so great. When using individual data, the one-dimensional model is clearly preferred.

Trinary Paradigm

There was no clear preference between the one-dimensional and process-pure models in the trinary paradigm. For this reason, we expected that our data-bootstrapping procedure would show little preference for either model, and the results support this expectation (see Table 11, lower section).

DISCUSSION

To understand the implications of the present research, situating it in the history of remember–know research will be helpful. Early experiments (and, to be sure, many later ones) adopted a process-pure approach that identified the "remember" and "know" responses directly with underlying processes. This interpretation has been undercut by a variety of results, such as the strong effects of response bias on both "old" responses (Rotello et al., 2006) and "remember" responses (Rotello et al., 2005). Models that incorporate decision processes have been developed and are often pitted against each other. All existing models can be interpreted as a combination of familiarity and recollective variables, coupled with a response rule.

Such models have been compared in several ways. A particularly popular approach is to identify what we have called a "signature" prediction, an aspect of the data that is uniquely predicted (or not predicted) by one model or another. The *z*ROC, for example, is often examined for curvature, because the dual-process and one-dimensional models predict different shapes. Myung et al. (2007) argued that such signature predictions allow for the most powerful experimental designs, but the results of such comparisons have been inconclusive in distinguishing models of recognition memory (see the recent exchange between Parks & Yonelinas, 2007, and Wixted, 2007). Our results suggest that one reason these analyses have not been incisive is that they have been based primarily on average data, for which the best-fitting model is relatively unstable (see Table 11). Moreover, the success of models cannot depend completely on signature predictions unless the data sets in question are fully characterized by those aspects.

A potentially better way to compare models is by fitting them to complete data sets for individual subjects. Because recent experiments designed to test remember–know models have usually included rating responses, these models must generate a conditional probability for every combination of stimulus type, remember–know decision, and rating in the data matrix. Competing models may not have the same number of parameters, but measures such as the AIC and BIC attempt to level the field.

In previous research, we compared models in exactly this way and, as Table 1 illustrates, generally found sup-

Table 11
Proportions of Samples Favoring Model 1 in Each Paradigm

		Model Comparison			
Paradigm	Data Type	(1) 1-D vs. (2) Extended DP	(1) 1-D vs. (2) STREAK	(1) Extended DP vs. (2) STREAK	(1) 1-D vs. (2) Standard DP
Old-first	Average	.30	.82	.73	.98
	Individual	.97	1.00	.98	1.00
		(1) 1-D vs. (2) PP		(1) 1-D vs. (2) STREAK	(1) PP vs. (2) STREAK
Remember-first	Average	.87		.84	.51
	Individual	1.00		1.00	.00
Trinary	Average	.72		.76	.56
	Individual	.51		1.00	1.00

Note—1-D, one-dimensional model; DP, dual-process model; PP, process-pure model.

port for the one-dimensional model. But the AIC and BIC do not necessarily equate models for complexity, which raises the possibility that the one-dimensional model may have had an advantage in being able to mimic data from subjects operating under an entirely different model. The present simulations were designed to evaluate the relative complexity of current remember–know models.[9]

The outcome of these calculations is easily summarized: Far from having an unfair advantage, the one-dimensional model is often the least complex of the models under evaluation. Thus, its support when evaluated by GOF measures is not an artifact of its complexity. In cases in which other models do better (or almost as well), the differences in fit are small and—according to our simulations—largely the result of empirical "noise." Our calculations are very comforting for proponents of the one-dimensional model.

These results are clear when based on individual data. Averaging data across individuals in order to compare these models, however, is often counterproductive. Aside from the risk of averaging out the critical decision components of the models, our simulations show that averaging can increase the effect of noise and lower model discriminability. The use of group data to compare models, along with the focus on signature predictions mentioned earlier, has contributed to the inconclusiveness of direct comparisons of remember–know models.

Our simulations also support a strong, simple recommendation about experimental paradigms: Do not use the trinary response set. We found this design to be plagued by noise and low discriminability. It is fortunate that the design is not popular (only 54 of the 400 conditions in Dunn's [2004] database used it), and the present results should discourage any enthusiasm for a trinary revival.

Implications for Future Research

Our main recommendations are for those testing models or applying them to remember–know data.

1. Extend the data beyond hits, false alarms, and remember–know response rates, so that the number of data points exceeds the number of parameters in the model. Our preferred method is to use confidence ratings and ROC curves, both because a number of models of remember–know judgments have been extended to consider ratings (Rotello & Macmillan, 2006; Rotello et al., 2006) and because the availability of ROC data allows appropriate performance measures to be determined (Macmillan & Creelman, 2005; Rotello, Masson, & Verde, 2008)

2. Deal with all data simultaneously. Testing only particular parts of the data in an experiment—the signature predictions—has led to successes and failures for all remember–know models, and therefore to no strong conclusions. A complementary approach is to test the predictions of entire classes of models (i.e., one- or two-process) against large numbers of data sets simultaneously. Dunn (2008) has adopted this elegant method with good success; his results are in agreement with ours, in suggesting that one-dimensional models provide a better description of the literature than two-process models.

3. Fit individual data. Our analyses assume that one of the models under consideration is the correct generating model. A minority of subjects is nearly always best described by some other model (see Table 1). Although these individual differences may be due to the mimicry properties of the models and/or to noise in the data, another possibility is that different subjects may act in accord with different models. The difficulty in discriminating the one-dimensional and process-pure models in the trinary paradigm may also be attributed to such individual differences.[10] Thus, not only does the use of average data lower model discriminability, it also obscures such potential differences across subjects. Our analyses of the individual subjects' fits may also obscure individual differences, because we summarized model fits by summing across subjects. A set of fit value differences preferring one model can be offset by one or more such differences strongly preferring the competing model. We find it comforting that a visual inspection of the fits of the most successful models to individual subjects' data in each of the three paradigms reveals that the fit values were tightly clustered with no clear outliers. Additional analyses could be done using techniques designed to uncover groups of subjects utilizing similar processes (see, e.g., Lee & Webb, 2005).

4. Use either the old-first or the remember-first paradigm; avoid the trinary task. The trinary task is much less able to distinguish among models than the other designs.

The remaining recommendations make more explicit use of our simulation approach:

5. In experiments similar to those we simulated (including the experimental paradigm, numbers of trials and subjects, etc.), our results can be used to qualify GOF conclusions. In particular, models can be discriminated with an optimal criterion, rather than the conventional zero criterion.

6. In experiments that are too dissimilar from those presented here, new simulations could be generated following the method reported above, and the new results could be used in the manner illustrated here. Although it is best to perform new simulations for different experimental designs, our conclusions are likely to extend to experiments with different numbers of trials per condition and to experiments with different numbers of subjects. In a set of comparisons like ours (but using different models), Cohen et al. (2008) systematically varied the numbers of trials and subjects in the simulated experiments. Across three different sets of models, their results changed very little when the numbers of trials and individuals were in the range used in most remember–know experiments.

Of course, a variety of other methods take the functional form of a model into account and could also be used to address the issue of model mimicry. These include the Fisher information approximation (Rissanen, 1996) and normalized maximum-likelihood (Barron, Rissanen, & Yu, 1998; Grünwald, Myung, & Pitt, 2005; Rissanen, 2001) implementations of minimum descriptive length, Bayesian model selection (Kass & Raftery, 1995), and Bayesian nonparametric model selection (Karabatsos, 2006). One of our goals, however, was to explore model mimicry using a method that could plausibly be used by a large number of researchers. Because the computational challenges inherent in applying these very powerful tech-

niques put them out of the reach of many researchers, we leave their application to models of remember–know judgments for future research.

The simulation approach reported here has been shown to produce excellent results in other domains (Cohen et al., 2008) and, at the same time, is easy to implement. Beyond their quantitative results, this and other simulation methods (e.g., Navarro, Pitt, & Myung, 2004) have the advantage of producing an informative visual description of the relative generality and distinguishability of the models under consideration, an important advantage over methods such as cross-validation (e.g., Berger & Pericchi, 1996; Browne, 2000) and generalization (Busemeyer & Wang, 2000).

AUTHOR NOTE

This research was supported by a grant from the National Institutes of Health (R01 MH60274) to C.M.R. and N.A.M. We thank John Dunn, Andy Yonelinas, and E.-J. Wagenmakers for their comments. Correspondence may be addressed to A. L. Cohen, University of Massachusetts, Department of Psychology, Box 37710, Amherst, MA 01003-7710 (e-mail: acohen@psych.umass.edu).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, **25**, 724-730.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144-151.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, **44**, 2743-2760.

Berger, J. O., & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108-132.

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

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, **15**, 692-712.

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

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, **14**, 423-429.

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.

Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, **47**, 389-416.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, **95**, 91-101.

Egan, J., 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.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, **12**, 403-408.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229-244). Oxford: Oxford University Press.

Gardiner, J. M., Richardson-Klavehn, A., & Ramponi, C. (1998). Limitations of the signal detection model of the remember–know paradigm: A reply to Hirshman. *Consciousness & Cognition*, **7**, 285-288.

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

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press, Bradford Books.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review*, **60**, 269-275.

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory & Language*, **55**, 495-514.

Hicks, J. L., & Marsh, R. L. (1999). Remember–know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, **6**, 117-122.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember–know paradigm. *Memory & Cognition*, **25**, 345-351.

Inoue, C., & Bellezza, F. S. (1998). The detection model of recognition using know and remember judgments. *Memory & Cognition*, **26**, 299-308.

Kapucu, A., Rotello, C. M., Ready, R. E., & Seidl, K. N. (2008). Response bias in "remembering" emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 703-711.

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, **50**, 123-148.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal of Optimization*, **9**, 112-147.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.

Macho, S. (2004). Modeling associative recognition: A comparison of two-high-threshold, two-high-threshold signal detection, and mixture distribution models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 83-97.

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

Macmillan, N. A., & Rotello, C. M. (2006). Deciding about decision models of remember and know judgments: A reply to Murdock (2006). *Psychological Review*, **113**, 657-664.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, **87**, 252-271.

Murdock, B. (2006). Decision-making models of remember–know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, **113**, 648-656.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, **14**, 1043-1050.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, **49**, 47-84.

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, **114**, 188-201.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computation model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 294-320.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40-47.

Rissanen, J. [J.] (2001). Strong optimality of the normalized ML mod-

els as universal codes and information in data. *IEEE Transactions on Information Theory*, **47**, 1712-1717.

Rotello, C. M., & Macmillan, N. A. (2006). Remember–know models as decision strategies in two experimental paradigms. *Journal of Memory & Language*, **55**, 479-494.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models. *Memory & Cognition*, **34**, 1598-1614.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum–difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review*, **111**, 588-616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, **12**, 865-873.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, **70**, 389-401.

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

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263-269.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250-264.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, **26**, 1-12.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, **48**, 28-50.

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

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. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, **130**, 361-379.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory & Language*, **46**, 441-517.

NOTES

1. For all simulations, the standard deviation of the normal was 5% of the range. A parameter value was resampled if it fell outside the range.

2. The uninformed and informed simulations led to only three substantively different conclusions. According to the uninformed simulations, in the remember-first paradigm, the one-dimensional model was less complex than the process-pure model, and the process-pure model and STREAK were better distinguished using average data. In the trinary paradigm, the relative complexities of the models reversed for the individual data. Because the model preference remained unchanged for the remember-first task and the trinary task yielded poor discriminability overall, we are not overly concerned by these fluctuations.

3. Confidence ratings have been used to expand binary responses in a wide range of domains. Systematic comparisons of rating and nonrating data have shown the results to be consistent (see, e.g., Egan, Schulman, & Greenberg, 1959). For each of the remember–know models fit in this article, we used the most straightforward assumptions possible to add ratings to the binary choices. Our previous work compared different assumptions about ratings in cases of ambiguity within a model (Dougal & Rotello, 2007; Rotello & Macmillan, 2006; Rotello et al., 2006); only the most successful versions are considered here.

4. In this and all other models, cells predicted to have a response proportion of 0 were replaced with predicted values of .01. The other cell proportions were then adjusted so the sum of all response proportions was 1.

5. Wixted and Stretch (2004) suggested that the location of the remember criterion might vary across trials. This *extended* version of the model was tested by Rotello et al. (2006), but because it was for the most part not successful, we consider only the fixed version here.

6. Because the one-dimensional model and STREAK have 8 parameters and the extended dual-process model has 10, the AIC and BIC will produce results different from those with the log L criterion. Applying the AIC to the one-dimensional and extended dual-process comparison produces overall average and individual error rates of 42% and 41%, respectively. When applied to the extended dual-process and STREAK comparison, the overall average and individual error rates are 10% and 3%, respectively. Because the one-dimensional, standard dual-process, and STREAK models (and all further models discussed below) have the same number of parameters in this paradigm, errors for these comparisons cannot be reduced from those based on the log L criterion by using the AIC.

7. Following Yonelinas (2001), we call models *dual-process* if the recollection process has a threshold character. Similar models in which recollection is continuous are referred to as *process-pure*. Process-pure *models* are to be distinguished from the (model-free) process-pure *approach* criticized in the introduction.

8. Knowing was defined as a feeling of familiarity in the absence of any recollected details. If a subject responded "know," he or she was told to "rate your feeling of knowing about reading this word" using a 3-point scale: (1) *weak feeling of knowing*, (2) *moderate feeling of knowing*, or (3) *strong feeling of knowing*.

9. Although other researchers have actively compared remember–know models, we are not aware of any direct comparisons of complete, rating-based remember–know models aside from our own efforts. For recent examples of this approach in studies without a remember–know component, see Heathcote, Raymond, and Dunn (2006) and Macho (2004).

10. Although individual differences almost certainly do play a role in this inconsistency found for the one-dimensional and process-pure comparison of Table 9, it is likely that a more subtle force is also at work: The empirical data for the one-dimensional and process-pure models are in a location of data space in which the two models can well mimic each other. Regardless, the conclusion remains the same: It is difficult to declare that either of these models is the "correct" one.

ARCHIVED MATERIALS

Additional materials associated with this article (details of the uninformed parameter simulations) may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive.

To access this file, search the archive for this article using the journal name (*Psychonomic Bulletin & Review*), the first author's name (Cohen), and the publication year (2008).

File: Cohen3-PB&R-2008.zip

Description: The compressed archive file contains two files:

SupplementaryMaterials.doc, containing additional data for the uninformed simulations, as well as Tables A1–A9 and Figures A1–A3, in Microsoft Word format.

SupplementaryMaterials.pdf, containing the same information in pdf format.

Author's e-mail address: acohen@psych.umass.edu.

(*Continued on next page*)

APPENDIX

Fitting the Data

All simulations were run in MATLAB using the simplex search method (Lagarias, Reeds, Wright, & Wright, 1998) to find the maximum-likelihood parameters. The simulated numbers of trials and subjects matched those in the real experiments (22 subjects for the remember-first paradigm and 24 subjects for the other paradigms, and 60 trials per condition for all three paradigms). To reduce potential problems with local minima, each fit was repeated three times using different starting parameters.

Optimal Criterion

The optimal criterion was found using a simple step-search (using 1,000 steps between the minimum and maximum difference of GOF). Occasionally (in 4 of the 54 cases in the text) the grid search did not include the log L criterion and yielded a higher error rate than the zero criterion. In those cases, we adopted the log L criterion as optimal. The deviations between the step-search and log L criterion error rates were typically very small (mean of 1.75% and a maximum of 3%) and did not affect the results in any significant way. If the two distributions are widely separated, a range of "optimal" criterion locations will all yield the same, minimal error rate. In that case, we defined the optimal criterion to be the mean of that range.

To estimate the variability of the optimal criterion, 900 of the 1,000 points in each histogram were used to estimate the criterion location; the remaining 100 points (0–100, 100–200, etc.) were used to determine the error rate. This estimation procedure was repeated 10 times for each model comparison. The standard deviation of the criterion location was generally well below 0.05.

(Manuscript received July 26, 2007;
revision accepted for publication February 27, 2008.)