Memory & Cognition
2006, 34 (8), 1598-1614

Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models

CAREN M. ROTELLO and NEIL A. MACMILLAN
University of Massachusetts, Amherst, Massachusetts

JASON L. HICKS
Louisiana State University, Baton Rouge, Louisiana

and

MICHAEL J. HAUTUS
University of Auckland, Auckland, New Zealand

In recognition memory experiments, the tendency to identify a test item as "old" or "new" can be increased or decreased by instructions given at test. The effect of such response bias on remember–know judgments is to change "remember" as well as "old" responses. Existing models of the remember–know paradigm (based on dual-process and signal detection theories) interpret this effect as a shift in response criteria, but differ on the nature of the dimension along which the changes take place. We extended the models to account simultaneously for remember–know and confidence rating data and tested them using old–new (Experiment 1) and remember–know (Experiment 2) rating designs. Quantitative fits show that the signal detection models provide the best overall description of the data.

The remember–know paradigm for recognition memory elicits direct reports of the subjective experiences underlying "old" judgments. The appearance of an item on the study list may be specifically "remembered," or participants may merely "know" that it is familiar and attribute their knowledge to a prior presentation. The paradigm, first proposed by Tulving (1985), has been used to provide evidence for Mandler's (1980) hypothesis that both recollection and familiarity contribute to recognition performance. In the most common interpretation of remember–know data, "remember" and "know" responses are presumed to reflect these separate recollection and familiarity processes (e.g., Gardiner & Richardson-Klavehn, 2000). Support for such process purity is often adduced from dissociations, in which an independent variable affects one response but not the other, or affects both responses in opposite directions. Many experiments of this type have been conducted; recent surveys by Dunn (2004) and Rotello, Macmillan, and Reeder (2004) identified about 400 separate experimental conditions in the literature. In themselves, however, dissociations provide unconvincing evidence for the dual-process view: Dunn and Kirsner (1988) have shown that many dissociation patterns are quite consistent with single-process models, and Dunn (2004) successfully fit a unitary strength model to data from several classic remember–know dissociation results.

Dunn's model-based accounts of these data allowed for changes across conditions in response bias as well as sensitivity. A completely process-pure view would seem to assume that no such response bias exists for remember–know judgments, but empirically it is known that bias can be manipulated. Participants in a number of experiments have been told that a larger (say, 70%) or smaller (30%) fraction of the test items had been studied, when the true ratio was 50% (Gardiner, Richardson-Klavehn, & Ramponi, 1997; Hirshman & Henzler, 1998; Strack & Förster, 1995). The consistent finding in these experiments is that participants in the putative *70% Old* condition make a higher proportion of "old" decisions than participants in the *30% Old* condition, and they also make a significantly greater number of remember judgments. Postma (1999) obtained analogous results in an experiment in which participants were encouraged to be either "very certain" or to have "even only a weak notion that they had studied" a test word before calling it "old."

How should such response-bias effects be understood? This question can be answered with reference to quantita-

C.M.R. and N.A.M. were supported by Grant R01 MH60274 from the National Institutes of Health. Thanks to Satomi Imai and Mungchen Wong for collecting the data in Experiments 1 and 2, respectively. We thank an anonymous reviewer for helpful comments on an earlier version of this article, and Andy Yonelinas for both his comments on the manuscript and extensive discussions of the dual-process model. Correspondence should be addressed to C. M. Rotello, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003-7710 (e-mail: caren@psych.umass.edu).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

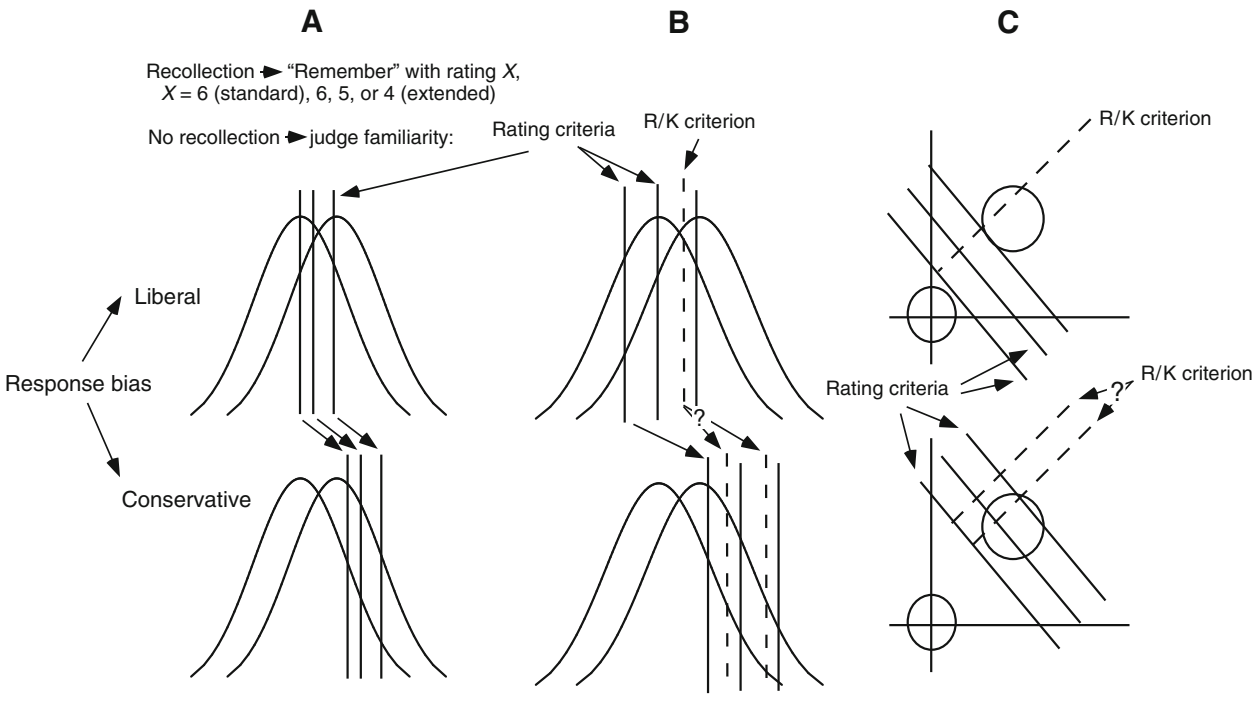

Figure 1. How old–new response bias operates according to three models of remember–know (R/K) judgments. (A) Dual-process model. If recollection occurs, a "remember" response is made; otherwise a judgment is based on familiarity. Response bias affects only familiarity. (B) The one-dimensional signal detection model. Old and New items vary in their average memory strength, and two kinds of criteria are used to determine responses: The remember–know criterion divides "remember" and "know" responses, and old–new criteria determine rating responses. Response bias affects the rating criteria and may affect the remember–know criterion. (C) STREAK. Old and New items differ in both specific (d_y) and global (d_x) strength. Circles represent equal-likelihood contours from bivariate distributions; the upper circle is the Old distribution and the lower is the New distribution. The old–new decision bound distinguishes "old" from "new" responses on the basis of a weighted sum of the two axes, and rating responses are determined by bounds parallel to it. An orthogonal bound distinguishes "remember" from "know" responses on the basis of a weighted difference. Response bias affects the old–new bounds and may affect the remember–know bound.

tive models that include both sensitivity and response-bias components. Several current models describe different roles for response bias and therefore make discrepant predictions about the patterns of data to be expected when response bias is manipulated. In all of them, a response-bias mechanism is characterized by a parameter that changes value when response bias shifts, leaving sensitivity parameters unchanged. Bias manipulations can thus provide an efficient tool for distinguishing competing models. In this article, we take advantage of this power to compare models for the remember–know experiment that fall into two categories: dual-process accounts that incorporate a threshold process, and those based on signal detection theory (SDT).

A useful tool for examining the predictions of models about response-bias manipulations is the receiver operating characteristic (ROC), in which hit rates are plotted against false alarm rates as response bias (usually inferred from rating responses) varies. All the models we consider make predictions about the form of these curves, and the experiments we report use rating designs.

Dual-Process Models

Quantitative models of the dual-process interpretation of remember–know have been offered by Yonelinas and his colleagues (Yonelinas, 2001; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). The assumptions of these dual-process models are illustrated in Figure 1A. Recollection is a high-threshold process that cannot lead to memory-based false alarms. If an item is (correctly or incorrectly) recollected, a "remember" response is given; in a rating design, this judgment is made with high confidence. If recollection does not occur, the response is based on Old and New distributions on a familiarity dimension that is partitioned into two or (in a rating design) more regions by criteria. Response bias operates on the location of these criteria.

The ROCs predicted by dual-process models follow directly from these assumptions. The simplest model (Yonelinas, 1994) assumes that the highest confidence rating response is dedicated to the recollection process; the resulting ROC has a positive hit rate when the false alarm rate equals 0. The solid line in Figure 2A is an example of such a curve. The presence of a few false alarms at this high confidence level (the leftmost point on the ROC) can be accommodated by the assumption of a nonzero false remember rate; the dashed curve in the figure incorporates this possibility. When response bias becomes more liberal, most ROC points move to the right along a

fixed curve, but if the leftmost point depends only on recollection, it remains fixed. Later in this article, we discuss modifications of this essential idea. Malmberg (2002) has discussed other factors that could affect the shape of ROCs generated by a high-threshold model.

The source activation confusion (SAC) model of Reder et al. (2000; Diana, Reder, Arndt, & Park, 2006) has a decision structure that is similar to that of the Yonelinas dual-process model: "Remember" responses occur if sufficient episodic information is available; if not, a "know" or "new" response is made on the basis of semantic information. However, the episodic information is in a continuous rather than a threshold form. The SAC model has not applied directly to ROC data, and we will not fit it here.

Dual-process models are close to the process-pure view, in that a relatively unprocessed aspect of the data ("remember" responding) results uniquely from a specific underlying process (recollection). They (1) allow separate estimation of the accuracy of the two postulated processes, (2) make specific, testable assumptions about the nature of those processes (discrete for recollection, continuous for familiarity), and (3) include a mechanism for response bias (shifting of criteria) in the familiarity but not the recollection process.

The One-Dimensional Signal Detection Model

Signal detection models postulate continuous rather than discrete representations. In the simpler of the two signal detection approaches, the *one-dimensional model* (Donaldson, 1996),[1] remember and know judgments result from higher and lower strengths on a single continuum. In a conventional (nonrating) experiment, two criteria partition the continuum; "remember" responses result from values above the upper (remember–know) criterion, "new" responses result from values below the lower (old–new) criterion, and "know" responses result from the region in between. When a rating scale is used, multiple criteria partition the continuum into regions corresponding to the possible responses, as shown in Figure 1B. The predicted ROC is concave downward, as shown in Figure 2B, and is linear if plotted on *z*-coordinates.

What should happen when old–new response bias is manipulated? Clearly, the rating criteria are expected to shift (Figure 1B), and ROC points should move along the curve (Figure 2B). It is less clear whether the remember–know criterion should change: The model neither requires nor forbids this outcome. The ambiguous status of the effect is indicated in Figure 1B by a question mark. Also unknown is whether the criteria in the one-dimensional model have fixed or variable locations. Wixted and Stretch (2004) have proposed that the remember–know criterion in particular is variable, and we elaborate this assumption of the model later.

The one-dimensional model offers an internal test, in that the ratio of the standard deviations of the New and Old item distributions can be estimated from the slope of *z*ROCs obtained from two distinct aspects of the data. First, a *rating ROC* like that in Figure 2B can be constructed from confidence ratings in a standard way (Macmillan & Creelman, 2005). Second, a *two-point ROC* can be plotted by treating the "remember" and Old false alarm and hit rates as two points in ROC space; the points correspond to the old–new and remember–know criteria, which have the same status as those thought to underlie confidence judgments. The "remember" and Old points should therefore fall directly on the rating ROC, and the *z*-slopes of the two-point and rating ROCs should be equal. In a meta-analysis, Rotello et al. (2004) found that this prediction of the one-dimensional model was violated across experimental conditions.

On the other hand, Wixted and Stretch (2004) summarized a number of studies in which participants provided both confidence ratings and remember–know judgments

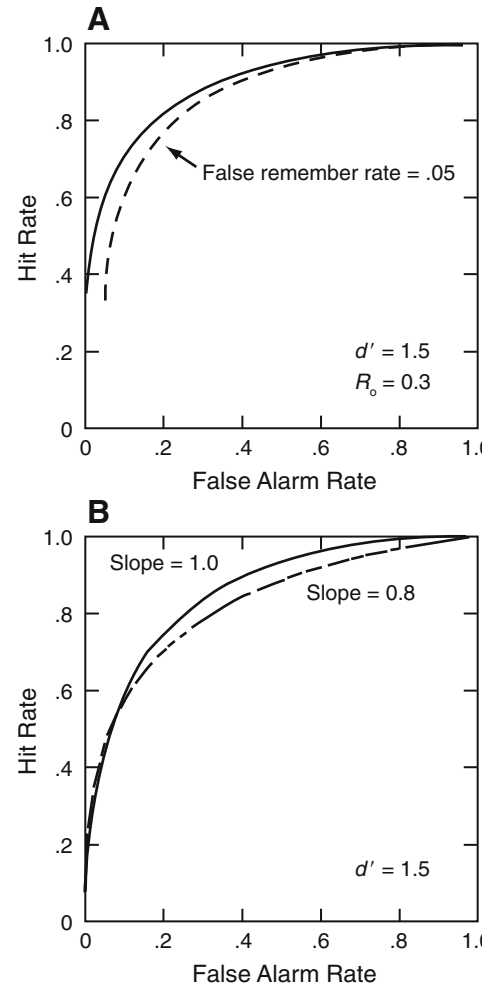

Figure 2. Old–new ROCs for three models. Highest-confidence responses occur at the left of the space; moving right along each theoretical function cumulates over lower levels of confidence. (A) Dual-process model assuming $d' = 1.5$ and $R_o = 0.3$. The dashed curve assumes a nonzero false remember rate; the solid curve does not. (B) One-dimensional model or STREAK assuming $d' = 1.5$. The solid curve assumes equal variance; the dashed curve assumes unequal variance ($s = 0.8$).

and concluded that the rating and two-point ROC slopes were approximately equal. However, 7 of the 10 within-subjects experiments they discussed collected memory judgments in an atypical fashion. In 4 experiments (from Stretch & Wixted, 1998), participants made speeded old–new decisions, then remember–know judgments, and finally rated their confidence from "guess" to "sure"; which response was being rated may have been ambiguous to the participants. In 3 experiments (from Yonelinas, 2001), the participants pressed the R key to indicate they remembered a test item, or else they rated their confidence that it had been studied; no "know" responses were actually collected. On average, these 7 atypical studies resulted in similar rating and two-point ROC slopes. The remaining 3 experiments summarized by Wixted and Stretch (2004) used a more standard design: The participants first made old–new confidence ratings and then remember–know judgments. The two-point slope in each of these experiments was greater than the rating ROC slope, consistent with Rotello et al.'s (2004) and Dunn's (2004) meta-analyses. None of the experiments summarized by Wixted and Stretch compared the ROC slopes statistically. The present experiments allow a within-subjects evaluation of the slopes using the standard design, in which old–new confidence ratings are followed by remember–know judgments.

The Two-Dimensional Signal Detection Model: STREAK

STREAK (Rotello et al., 2004) postulates a two-dimensional memory representation in which items differ in both specific and global strength. As illustrated in Figure 1C, both Old and New items lead to bivariate normal distributions in this space, the standard deviation of New items being about 0.8 that of Old items. The diagonal decision boundaries in Figure 1C imply that old–new judgments depend on a weighted sum of specific and global strength and the remember–know distinction on a weighted difference. In a rating experiment, multiple old–new decision boundaries are parallel to each other and perpendicular to this sum axis. The old–new ROC predicted by STREAK is similar to that generated by the one-dimensional model, but the unequal standard deviations of the New and Old distributions imply an asymmetric curve (Figure 2B). On *z*-coordinates, the ROC is a straight line, with slope equal to the ratio of the standard deviations.

The consequence of a response-bias manipulation is to shift the old–new boundary or boundaries, moving points along the ROC as in the one-dimensional model. This shift perforce changes the rate of "remember" responding, but the orthogonal relation between the decision bounds for old–new and remember–know judgments allows for an interesting internal test: The proportion of "old" judgments that are categorized as "remember" is predicted to be constant when old–new response bias changes, and this is true for both hits ("old" responses to targets) and false alarms ("old" responses to lures). We call this relation the *response-ratio invariant*.

STREAK takes no position on the two-point slope issue that is so critical for the one-dimensional model. Because remember–know judgments are based on different information than old–new judgments, STREAK neither predicts nor is troubled by equality between rating and two-point slopes.

The Present Experiments

In two experiments, we attempted to manipulate participants' tendency to identify test items as "old." In both experiments, items called "old" were then categorized as remembered or known. To take advantage of the distinct predictions of the various models, rating responses were used to construct ROC curves; in Experiment 1, the ratings were applied to old–new judgments, whereas in Experiment 2 they were applied to remember–know judgments.

EXPERIMENT 1

Experiment 1 was designed as a replication of the Hirshman and Henzler (1998) study mentioned earlier. We told some participants that 70% of the recognition test items had been studied and others that 30% of the items had been studied. (In both cases, half of the test items were actually old.) The *70% Old* condition was intended to produce a liberal bias in which more items would be called "old" than in the second, conservative bias condition. The major improvement in our version of the experiment was the collection of confidence ratings on the old–new judgments; the ROCs thus generated allow stronger tests of the competing models. The dual-process models predict that response bias should affect the familiarity process and not recollection. The one-dimensional model, but not STREAK, predicts that two-point slopes obtained from the remember–know judgments should equal those obtained from the rating procedure. STREAK predicts that the instructional manipulation of response bias should affect only the old–new criterion and should produce a response-ratio invariant.

Method

Participants. Forty-eight volunteers agreed to participate in exchange for extra credit in psychology courses at Louisiana State University. An equal number of participants were randomly assigned to each of the testing conditions, as described in the Procedure. Two participants in the *30% Old* condition were excluded from the analysis because their performance was below chance (A_g, the area under the ROC, equaled .41 in both cases). The participants were tested individually in sessions that lasted approximately 30 min.

Materials. We chose 170 words of varying frequency from the Kučera and Francis (1967) corpus. Ten words were chosen randomly and fixed as primacy and recency buffers in the study list, leaving 160 words to be presented for the recognition test. Two sets of 80 words were constructed randomly from these remaining items. For counterbalancing purposes, half of the participants in each test instruction condition received the first set of 80 as studied words, whereas the remaining participants received the other set. The orders of both the study and test lists were randomized for each participant.

Procedure. All participants were told that they would be studying a list of words for an unspecified memory test to be given later. Ninety words were then presented for 1.5 sec each in the center of

Table 1
Data From Experiment 1: Proportions of Positive Recognition Responses and Standard Errors to Targets and Lures by Subjective Experience and Test Condition, and ROC Slopes Obtained From Ratings and From Remember–Know (R/K) Data

	70% Old		*30% Old*	
Test Condition	*P*	*SE*	*P*	*SE*
Responses to Targets				
"Remember"	.53	.04	.40	.04
"Know"	.27	.03	.22	.03
Total "old"	.80	.02	.62	.03
"Remember" | hit	.66	.04	.65	.04
Responses to Lures				
"Remember"	.20	.02	.09	.03
"Know"	.25	.03	.13	.023
Total "old"	.45	.03	.22	.03
"Remember" | false alarm	.43	.05	.33	.06
*z*ROC Slopes				
From old–new ratings	0.79	.07	0.77	.08
From R/K (two-point)	1.14	.12	0.83	.08

the computer monitor. Each study word was preceded by a blank screen (250-msec duration), a fixation stimulus (***; 500 msec), and a second blank screen (250 msec). The first 5 and last 5 words in the study list were presented as primacy and recency buffers and were not tested. The other 80 studied words served as targets in the recognition test.

Following study, the participants were given instructions for the recognition test. They were told that they would be making an old–new (studied–unstudied) decision and rating their confidence for each test item. For words called "old," regardless of the level of confidence, the participants were told they would also make a remember–know judgment. NEW and OLD adhesive labels were affixed to the F and J keys of the computer keyboard, respectively. The numbers 1 (*guessing*), 2 (*somewhat sure*), and 3 (*very sure*) on the computer keyboard were used to indicate confidence. Remember–know responses were made with the M and B keys, which were labeled with *R* and *K*, respectively.

The remember–know instructions were identical to those used by Hicks and Marsh (1999). In short, test items that were accompanied by retrieval of contextual detail surrounding initial exposure were to be labeled "remember," and items that evoked a feeling of familiarity in the absence of recollective detail were to be labeled "know." Finally, participants in the *30% Old* condition were told that 30% of the items on the recognition test (3 out of 10) were studied items, and that the remaining items were new. Participants in the *70% Old* condition were told that 70% of the test items (7 out of 10) had been studied.

After the experimenter confirmed that the participants understood these test instructions, the 160 test items (80 targets and 80 lures) were presented individually in the center of the monitor. The phrase "NEW or OLD?" appeared under each test word as it was presented. Following the old–new judgment, the participants rated their confidence in that decision. For each item called "old," a remember–know judgment was requested.

Results

The recognition probabilities for Old and New items for each test instruction condition, given in Table 1, confirm that the bias manipulation had its intended effect: For both targets and lures, the proportion of items called "old" was greater in the *70% Old* than the *30% Old* condition [targets, .80 vs. .62, $t(44) = 4.54$, $p < .001$; lures, .45 vs. .22, $t(44) = 5.16$, $p < .001$]. This bias was significant for "remember" responses [targets, .53 vs. .40, $t(44) = 2.29$, $p < .05$; lures, .20 vs. .09, $t(44) = 3.06$, $p < .01$] and for "know" responses estimated with the independence correction {Yonelinas & Jacoby, 1995; P(independent know) = P("know") / [1 − P("remember")]} [targets, .57 vs. .37, $t(44) = 3.45$, $p < .01$; lures, .31 vs. .14, $t(44) = 4.23$, $p < .001$].

The ROCs are shown in Figure 3; the points on an ROC are simply response proportions at different levels of bias (as inferred from the rating responses). The form of the old–new ROC is consistent with the underlying normal distributions of memory strength assumed by the SDT models (that is, the curves were approximately linear when plotted on *z*-coordinates). All points appear to lie on a single function. The ROC slopes were in the range expected from the item-recognition literature (Ratcliff, Sheu, & Gronlund, 1992): The means across participants were 0.77 and 0.79, which did not differ from one another [$t(44) = 0.16$].

Model-Based Assessments

The competing models have been previously tested either by examining remember and know proportions or by examining ROCs generated from ratings, but not by considering both simultaneously. The experimental data consist of nine possible frequencies (responses of "remember" or "know" for ratings 4–6 but just "old" for ratings 1–3) for each of the two stimulus classes. To test the models against this full data set, it was necessary to spell out certain assumptions of the dual-process and one-dimensional models, and we developed two versions of each. STREAK is explicit about how the two aspects of the data combine, although it has not previously been fit to complete data sets of this form. A description of each model tested follows; the exact equations for predicted proportions are given in the Appendix.

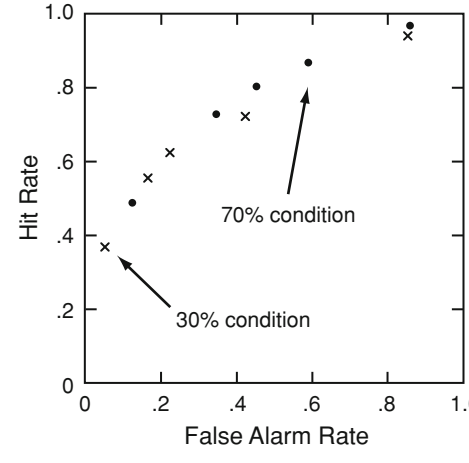

Figure 3. Old–new ROC data from Experiment 1. Highest-confidence judgments are plotted as the leftmost point in each condition, and successive points cumulate over other levels of confidence.

For the dual-process model, we assumed that true recollection operated on a fraction R_o of the target trials and false recollection on ε of all trials. The familiarity process was therefore assumed to be effective on the remainder of the trials ($1 - R_o - \varepsilon$ of target trials and $1 - \varepsilon$ of lure trials). In the *standard* version of the model, remember responses were predicted to occur only at the highest confidence level. The predicted proportion of "remember" responses in other cells of the design was set equal to .01 times the probability of recollection ($R_o + \varepsilon$ for Old items; ε only for New items).[2] In the *extended* version of the model, we allowed "remember" responses to be distributed over all levels of confidence that an item was old (rating 4, 5, or 6). The standard dual-process model has two sensitivity parameters (R_o, d'), five criteria, and ε, a measure of false remembering responses. The extended version adds two parameters that describe how the "remember" responses are distributed over the three ratings.

We also tested two implementations of the one-dimensional model, a *fixed* version, in which all decision criteria were fixed, and a *variable* version proposed by Wixted and Stretch (2004). In the variable model, the old–new rating criteria are fixed, but the location of the remember–know boundary is sampled from a normal distribution on each trial. The fixed version of the one-dimensional model has eight parameters: sensitivity, six criteria, and the slope of the *z*ROC. The variable criterion model adds one parameter: the standard deviation of the remember–know criterion distribution.

Finally, we tested the original version of STREAK. The parameters estimated were six decision criteria (five for old–new confidence and one for remember–know), two sensitivity parameters (d_x and d_y), and the slope of the *z*ROC (s).

All of the models were fit using maximum likelihood estimation. The dual-process and one-dimensional models were implemented in Excel's Solver module; for STREAK, a separate program was written that used a variant of downhill SIMPLEX to find parameters. Resulting parameter estimates (including locations of the old–new and remember–know criteria, but not of those for intermediate ratings) are shown in Table 2. We consider each model separately and then provide a statistical comparison of fits across all models.

Dual-process models. A key test of all models is that the sensitivity parameters should remain constant under our manipulation while bias parameters change.[3] Within the dual-process models, the response-bias manipulation should not affect the rate of recollection R_o, the accuracy of the familiarity process d', or the remember false alarm rate ε, but should instead affect the criterion locations within the familiarity process (Yonelinas, Regehr, & Jacoby, 1995). As can be seen in Table 2, estimates of

Table 2
Mean Best-Fitting Parameter Estimates and Standard Errors for STREAK, the One-Dimensional Model, and the Dual-Process Model for Experiment 1

	70% Old		*30% Old*	
Model Parameters	*M*	*SE*	*M*	*SE*
STREAK				
d_x	0.51	.05	0.57	.05
d_y	1.24	.16	1.69	.28
C_o	0.06	.07	0.73	.10
C_r	0.08	.11	−0.28	.10
One-dimensional model with fixed criteria				
d'	1.47	.28	1.10	.13
K_o	0.26	.10	0.93	.12
K_r	0.14	1.09	0.84	.57
Slope s	0.84	.10	0.71	.09
One-dimensional model with variable remember–know criterion				
d'_1	1.35	.19	1.21	.11
K_o	0.18	.10	0.83	.10
K_r	0.61	.13	1.28	.17
Total Old variance	1.18	.17	1.03	.20
$1/s$ (strength)	0.89	.08	0.47	.14
$1/t$ (criterion)	0.36	.07	0.21	.13
Standard dual-process model				
R_o	0.33	.03	0.32	.03
d'	0.54	.11	0.49	.12
ε	0.20	.03	0.09	.03
Extended dual-process model				
R_o	0.33	.03	0.32	.03
d'	0.54	.11	0.49	.12
ε	0.20	.03	0.09	.03
$r_6 = (1 - r_5 - r_4)$	0.74	.03	0.79	.03
r_5	0.23	.03	0.19	.03

Table 3
AIC and BIC Statistics for the Data From Experiments 1 and 2, Averaged Over Individual Participants' Fits

				One-Dimensional		Dual-Process	
Experiment	Condition	Measure	STREAK	Fixed Criteria	Variable Remember Criterion	Standard	Extended
1	*30% Old*	AIC	604.4	507.2	512.1	562.0	519.8
		BIC	632.1	531.8	539.8	586.6	550.5
	70% Old	AIC	634.9	548.1	541.4	637.3	554.4
		BIC	662.5	572.7	569.1	661.9	585.2
2	Conservative	AIC	270.6	N/A	269.1	N/A	N/A
		BIC	257.4	N/A	257.9	N/A	N/A
	Liberal	AIC	345.3	N/A	344.6	N/A	N/A
		BIC	332.1	N/A	333.4	N/A	N/A

Note—Lower values indicate better fit.

both R_o and d' increased slightly in the *70% Old* condition in comparison with the *30% Old* condition, but neither change was statistically reliable for either version of the model [R_o, .33 vs. .32, $t(44) = 0.38$; d' for both models, 0.54 vs. 0.49, $t(44) = 0.30$]. The difference in false remembering was reliable, however [for both versions of the model, ε: .20 vs. .09, $t(44) = 3.05$, $p < .01$].

The dual-process model also assumes that decision criteria change between conditions. For both versions of the model, the criterion that determines Confidence Level 2 showed the biggest (and only reliable) change over conditions [for both models: .05 vs. .41, $t(44) = 3.71$, $p = .001$].

One-dimensional SDT model. As shown in Table 2, both versions of the one-dimensional model interpret the data to reveal a substantially more liberal old–new decision criterion K_o in the *70% Old* condition than in the *30% Old* condition [fixed model, 0.26 vs. 0.93, $t(44) = 4.25$, $p < .001$; variable criterion model, 0.18 vs. 0.83, $t(44) = 4.53$, $p < .001$]. Both models also indicate a liberal shift in the remember–know criterion K_r in the *70% Old* condition, but the difference was only reliable in the variable criterion model fit [variable criterion, 0.61 vs. 1.28, $t(44) = 3.13$, $p < .01$; fixed model, 0.14 vs. 0.84, $t(44) = 0.55$]. Sensitivity did not differ between the two conditions [variable criterion, $t(44) = 0.62$; fixed, $t(44) = 1.21$], nor did the slope of the *z*ROC [variable model, $t(44) = 0.72$; fixed, $t(44) = 0.97$]. The variable criterion model allows the variance of the Old distribution to be partitioned into a strength variance ($1/s^2$) and criterion variance ($1/t^2$). The strength standard deviation ($1/s$) was reliably larger in the *70% Old* condition [0.89 vs. 0.47, $t(44) = 2.60$, $p = .05$], but the criterion standard deviation ($1/t$) was unchanged [0.36 vs. 0.21, $t(44) = 1.07$, $p = .29$].

This pattern of results is substantially as expected according to the model: Bias parameters respond to the instructional manipulation, whereas sensitivity parameters remain constant. An outcome worth noting is that K_r moves about as much as K_o under response bias changes (104% as great a change, according to the fixed model; 82%, according to the variable model.) This effect is not predicted by the model, but is not inconsistent with it.

A second test of the fixed one-dimensional model is the comparison of two-point *z*ROC slopes obtained from the "remember" and "old" responses and the *z*ROC slopes resulting from the recognition rating data. According to the model, these two slopes should be equal, but (as can be seen in Table 1) the two-point slopes are greater than the recognition slopes. The increase is reliable in the 70% condition [*70% Old*, 1.14 vs. 0.79, $t(21) = 2.39$, $p < .05$; *30% Old*, 0.83 vs. 0.77, $t(19) = 0.53$].[4] In this respect, the data do not entirely support this model.

STREAK. Like the one-dimensional model, STREAK has two main decision criteria that could be affected by instructions: an old–new criterion and a remember–know bound. The model expects instructions aimed at the old–new criterion to affect only that criterion but, like the one-dimensional model, does not prohibit a change in the remember–know criterion. The numerical parameter values reveal that C_o shifts more between biasing conditions than C_r (0.67 units vs. 0.36), but both parameter changes are statistically reliable [C_o, $t(44) = 5.80$, $p < .001$; C_r, $t(44) = 2.45$, $p < .02$]. The sensitivity and slope parameters did not change significantly between conditions [d_x, $t(44) = 0.90$; d_y, $t(44) = 1.43$; slope, $t(44) = 0.70$].

A second test of the STREAK model, also based on the independence of the two decision criteria, is the predicted response-ratio invariant: Given that only the old–new criterion is influenced by the instructions, the proportion of "old" judgments that are followed by "remember" responses should be unaffected by old–new response bias. This invariant held for Old items: P("remember" | hit) was .64 in the *30% Old* condition and .65 in the *70% Old* condition [$t(46) = 0.278$]. For new items, the corresponding response rates were .33 and .43, which are also not significantly different [$t(46) = 1.36$, $p = .179$]. This test provides some support for STREAK, but its power is low: We had about a 40% chance of detecting a medium effect ($\eta^2 = .5$) and about a 10% chance of detecting a small effect ($\eta^2 = .2$).

Overall fits of models. Another way of evaluating the models is to compare their relative fit using likelihood measures that adjust for differences in degrees of freedom. To that end, AIC and BIC statistics (both discussed

Table 4
Proportion of "Remember" Judgments for Words Judged "Old" in Old–New Bias Experiments That Collected Remember–Know Judgments

		Studied Words, $P(R \mid H)$		Lures, $P(R \mid F)$	
Experiment	Condition	Conservative	Liberal	Conservative	Liberal
Strack & Förster (1995), Experiment 1	High-frequency words	.48	.43	.22	.08
	Low-frequency words	.58	.60	.20	.21
Strack & Förster (1995), Experiment 2		.70	.39	.17	.17
Gardiner, Richardson-Klavehn, & Ramponi (1997)	High-frequency words	.44	.47	.09	.10
	Low-frequency words	.54	.56	.12	.09
Hirshman & Henzler (1998)	Slow study	.45	.45	.28	.24
	Fast study	.31	.33	.28	.24
Postma (1999)		.58	.64	.40	.20
Mean		.51	.48	.21	.16

Note—$P(R \mid H)$, P("remember" | hit); $P(R \mid F)$, P("remember" | false alarm).

by Myung & Pitt, 2002) were computed for all of the models for each participant and are reported in Table 3. According to AIC, the one-dimensional model provided the best description of the data for 19 of the 22 participants (12 with the variable criterion model; 7 with the fixed criterion version). The remaining 3 participants were best fit with the dual-process model (2 with the standard version; 1 with the extended model). According to BIC, 19 participants' data were best described by a version of the one-dimensional model (9 with variable criterion; 10 with fixed criterion); 3 were best fit with the standard version of the dual-process model. In the *70% Old* condition, AIC indicated that the best fit was provided by some form of the one-dimensional model for 19 of the 24 participants (13 variable criterion; 6 fixed criterion); the remaining 5 participants were best fit by the extended dual-process model. According to BIC, the one-dimensional model provided the best fit for all 24 participants (12 variable criterion; 12 fixed). The differences between the AIC and BIC conclusions occur because BIC penalizes models more severely for additional parameters, and the extended dual-process model has the most parameters.

The actual values of these fit measures also indicate that the one-dimensional model provided the best overall description of the data (see Table 3). The improvement in fit provided by the extended dual-process model is also apparent in these values: The average AIC and BIC statistics are smaller in the extended dual-process model than in the standard model. The discrepancy is especially noticeable in the 70% condition. Of course, the advantage conferred by the extended model is to soften the high-threshold nature of the recollection process by allowing "remember" responses to be distributed across ratings. In this way, the improvement in fit occurs because the extended dual-process model is more like the one-dimensional model.

Discussion

The manipulation of participants' beliefs about the proportion of studied items on the test list was expected to influence response bias, but not accuracy. All of the models agree that response criteria were affected, but they draw different pictures of the effect. And none of the models is completely successful.

For our implementations of the dual-process model, the criteria in the familiarity process do change, as expected, in response to instructions. The model does better in the *30% Old* condition: The false remember parameter ε takes on a lower value, and the overall fit is moderately good. In the *70% Old* condition, ε takes on a rather high value (.20). "Remember" responses were observed at several levels of old–new confidence, contrary to the predictions of the standard model. The extended model does allow for this result, and therefore provides a better fit on the average (see Table 3).[5] However, this benefit comes at the price of obscuring the threshold nature of remembering, a fundamental assumption of the dual-process approach.

Both signal detection models conclude that sensitivity is the same in the two bias conditions and that the old–new criteria differ appropriately; in this respect, they are successful. Both also find that the remember–know criterion shifts (by about the same amount as the old–new criteria for the one-dimensional model and to a smaller degree by STREAK). Although this effect does not contradict the models, nothing in the instructions requires it.

The one-dimensional model produced the best quantitative fits. The signature prediction of the fixed version of the model—that the slope of the two-point *z*ROC obtained from remember–know judgments should equal the slope of the old–new rating *z*ROC—failed in the *70% Old* condition. Wixted and Stretch (2004) proposed the variable criterion version of the one-dimensional model to address exactly this slope discrepancy, and it does appear to be a factor in fitting the data. For the 21 participants whose data were best fit by the variable criterion model (according to BIC), 17 (81%) had two-point remember–know *z*ROC slopes that were steeper than their recognition slopes. In contrast, only 8 of the 22 participants best described by the fixed version of the one-dimensional model (36%) had overly steep two-point slopes.

STREAK's unique prediction, that the proportion of "old" judgments categorized as remembered is invariant between conditions, is confirmed for both Old and

New items. Although the finding for New items must be tempered by the low power of the test, it confirms a result found repeatedly in the literature. Table 4 summarizes data on the response-ratio invariant from past experiments that have manipulated old–new response bias in the remember–know paradigm. Although there is notable variability, statistically, there is no difference in the average response rates for Old or New items [Old, .51 vs. .48, $t(8) = 0.75$; New, .21 vs. .16, $t(5) = 1.59$, $p = .17$].

STREAK's fits were the poorest, and an examination of the data suggests an explanation. Although the model does not allow the remember–know criterion to move when the old–new criteria shift, it is easy to imagine that participants' remember–know responses do depend on the rating they have just given. Indeed, the correlation between "old" and "remember" response rates over confidence levels was .95 in the *30% Old* condition and .96 in the *70% Old* condition. If modified to allow the remember criterion to shift with the old–new rating, the model would certainly yield a better likelihood value. However, without some constraint on criterion placement, the number of free parameters would increase. (See Rotello et al., 2004, for additional discussion of this issue.)

There is a commonality in the application of the three models. All postulate that one of several processes is influenced by old–new response bias, but when the models are applied, one discovers that multiple components are affected. The dual-process model found changes in the distribution of responses due to familiarity (as expected), and also in the distribution of responses due to false recollection (at best, a surprise). The signal detection models inferred constant sensitivity and found changes in both old–new criteria (as expected) and the remember–know criterion (as neither expected nor forbidden). In STREAK, there was also an interaction between the old–new rating and the remember–know response that contributed to the model's inferior fit.

EXPERIMENT 2

Experiment 1 manipulated response bias by changing participants' expectations about the proportion of studied items on the test list, and old–new confidence ratings were collected. Both of these manipulations operated on the same response scale, and there is evidence that they interacted. In Experiment 2, we again manipulated old–new response bias, but collected ratings on the remember–know response, separating the two decisions.

Response bias was controlled by asking some participants to guess "old" and others to guess "new" when they were uncertain about a test probe's status. For each item called "old," they rated their degree of remembering or knowing on a 6-point scale (a strategy first used in Rotello et al., 2004, Experiment 2). This design does not allow us to calculate the slope of the old–new *z*ROC, but there are two benefits: (1) the biased old–new responses are not confounded with the ratings, and (2) the signal detection models and the dual-process models make different predictions about the form of the ROCs that result from the remember–know ratings (Rotello et al., 2004). As in Experiment 1, all models predict that only response bias changes across conditions; in addition, STREAK predicts the response-ratio invariant.

Method

Participants. Forty-two undergraduate students at the University of Massachusetts participated in Experiment 1 for extra credit in their psychology courses. All were native English speakers. Data were discarded for 1 participant who did not follow the instructions. Each of the remaining 41 participants was randomly assigned to one of the two conditions: 21 to the conservative group and 20 to the liberal group.

Stimuli. We selected 120 English nouns from the MRC Psycholinguistic Database (Coltheart, 1981). The words were divided equally into two lists that were closely matched on number of syllables ($M = 1.5$), number of letters ($M = 5.2$), and linguistic frequency (Kučera & Francis, 1967; $M = 158$, $SD = 113$). Ten additional words were drawn from the same pool to serve as practice, primacy, and recency items.

Procedure. The experiment consisted of a study phase, a practice phase, and a final test phase. In the study phase, half of the participants in each condition studied one list of 60 words, and half studied the other list. The words were presented in random order along with 4 practice items; a primacy item and a recency item were also included. Each word was presented in the center of a computer screen for 1,250 msec, with a 750-msec interval between words. The participants were instructed to study the words carefully in preparation for an upcoming memory test.

Following the study phase, the participants were given instructions about the nature of the memory test. They were told that they would see a series of words, some that were old and others that were new. For each word, the participants first judged whether or not they recognized it from the study phase, using the binary response alternatives "old" and "new." The participants in the conservative condition were instructed to respond "new" when they were unsure about having studied a word ("because people are sometimes overconfident in their recognition memory"), whereas participants in the liberal condition were instructed to respond "old" when they were unsure ("because people are sometimes less confident in their recognition memory than they should be").

Whenever participants called a word "old," they were also prompted to make a remember–know judgment on a 6-point scale. The endpoints of this scale were labeled in terms of the standard definitions of remembering and knowing developed by Rajaram (1993): "1) *remember specific aspects of the experience*" and "6) *know it feels very familiar, but nothing specific*." The intermediate ratings on this scale were described (but not specifically labeled) in terms of a range of recognition experience that is less detailed than specific remembering but more detailed than simply knowing. The participants were asked to "try to use the full range of this scale to reflect the way your feeling of recognition varies from word to word."

These instructions were supplemented by a brief practice task that included four old practice words and four new words. During the practice phase, the participants were encouraged to ask questions and to explain their "remember" and "know" responses to the experimenter.

The test itself included all 60 old words and 60 new words, presented in random order. The experimenter was not present in the room during the actual testing.

Results

Response proportions. Table 5 presents the mean values of the response proportions for the conservative and liberal conditions. For this purpose (and for the estimation of model

Table 5
Data From Experiment 2: Proportions of Positive Recognition Responses and Standard Errors to Targets and Lures by Subjective Experience and Test Condition, and ROC Slopes Obtained From Ratings and From Remember–Know (R/K) Data

	Conservative		Liberal	
Test Condition	*P*	*SE*	*P*	*SE*
Responses to Targets				
"Old"	.60	.04	.77	.03
"Remember" | "old"	.41	.04	.56	.04
"Remember" | hit	.66	.04	.72	.03
Responses to Lures				
"Old"	.16	.02	.38	.03
"Remember" | "old"	.05	.01	.12	.02
"Remember" | false alarm	.31	.05	.30	.04
*z*ROC Slopes				
From old–new ratings	N/A		N/A	
From R/K (two-point)	0.82	.15	0.89	.12

Note—*P*("remember" | *x*) values treat ratings of 1–3 as "remember," 4–6 as "know."

parameters reported below), ratings of 1–3 on the remember–know scale were counted as "remember" responses, and ratings of 4–6 were counted as "know" responses.[6]

Our biasing instructions were effective: The participants in the conservative condition were less willing to call test items "old" than the participants in the liberal condition, exhibiting significantly lower hit rates [.60 vs. .77, $t(39) = 3.78$, $p < .01$] and false alarm rates [.16 vs. .38, $t(39) = 5.33$, $p < .001$]. The participants in the conservative condition also produced fewer "remember" responses to both Old [.41 vs. .56, $t(39) = 2.63$; $p = .05$] and New [.05 vs. .12, $t(39) = 2.62$, $p = .05$] items, but this difference was proportional to the differences in the hit and false alarm rates. The proportion of "old" responses that were followed by "remember" judgments was not reliably different across conditions for either Old [$t(39) = 1.08$] or New [$t(39) = 0.13$] items. That is, the response-ratio invariant held, as predicted by STREAK.

Model-based assessments. The dual-process model cannot be fit to the full set of data: It can describe old–new or know–new rating tasks, but because "remember" and "know" responses are controlled by different processes, it cannot describe remember–know ratings. The model does predict the form of the ROCs, and we evaluate its success at this task in the next section.

The competing signal detection models were fit to the data for each individual participant. As for Experiment 1, a key test of the models is that their sensitivity parameters should remain constant under our manipulation while their bias parameters change. The resulting parameter values (shown in Table 6) broadly confirm this prediction.

The one-dimensional model for this experiment is straightforward: The lowest criterion divides "new" from "old" responses, and five higher criteria distinguish subcategories of the "old" region ranging from "know" to "remember." (See the Appendix for details.) Sensitivity, measured in units of the New distribution standard deviation d'_1, did not differ across conditions [1.53 vs. 1.82, $t(39) = 0.78$]. The standard deviation ratio also did not differ [0.87 vs. 0.75, $t(39) = 0.72$], and its magnitude was consistent with values observed in the recognition literature (Ratcliff et al., 1992). In contrast, both the old–new (K_o) and remember–know (K_r) decision criteria were more liberally placed in the liberal condition [K_r, 1.30 vs. 1.88, $t(39) = 3.20$, $p < .01$; K_o, 0.33 vs. 1.07, $t(39) = 5.41$, $p < .001$]. The irrelevant criterion K_r thus moved 78% as far as the relevant criterion (K_o).

For STREAK, the estimated parameter values support the predictions for this experiment. The sensitivity parameters (d_x and d_y) were not reliably different in the conservative and liberal conditions [for d_x, $t(39) = 0.52$; for d_y, $t(39) = 0.30$], indicating that biased test instructions did not influence the quality of the memories. The old–new criterion (C_o) was significantly higher (more conservative) in the conservative condition than in the liberal condition [0.89 vs. 0.22, $t(39) = 2.18$, $p < .05$]. In contrast, the remember–know criterion (C_r) was little changed [0.48 vs. 0.56, $t(39) = 0.56$]; it moved 12% as far as C_o.

Remember–know rating ROCs. The ROCs based on "remember" responses are shown in Figure 4 for the conservative (circles) and liberal (squares) conditions. These ROC points are based on ratings of remembering for both Old and New items, cumulating from strongest "remember" decisions to weakest "remember" (strongest "know"). Unlike typical ROC curves, these ROCs do not reach (1, 1): The highest point is determined by the hit and false alarm rates.

Note first that these data immediately challenge the dual-process model's assumption that remember and know judgments are based on different underlying processes: It is not clear that a dual-process participant would find instructions to judge variation between remembering and knowing sensible. In addition, the dual-process model assumes that "remember" responses are based on a high-threshold recollection process, so that a single (low) remember false

Table 6
Mean Best-Fitting Parameter Estimates and Standard Errors for STREAK and the One-Dimensional Model for Experiment 2

	Conservative		Liberal	
Model Parameters	*M*	*SE*	*M*	*SE*
STREAK				
d_x	0.73	0.22	0.61	0.06
d_y	1.36	0.26	1.27	0.10
C_o	0.89	0.29	0.22	0.07
C_r	0.48	0.12	0.56	0.09
s	0.80	0.19	0.65	0.08
One-dimensional				
d'_1	1.53	0.20	1.82	0.33
K_o	1.07	0.09	0.33	0.10
K_r	1.88	0.14	1.30	0.12
s	0.87	0.12	0.75	0.09

Note—In the unequal-variance one-dimensional model (when $s \neq 1$), sensitivity may be measured in units of the standard deviation of either the Old or New distributions. We used the New standard deviation, so we report d'_1; in this way, the sensitivity values are comparable across models.

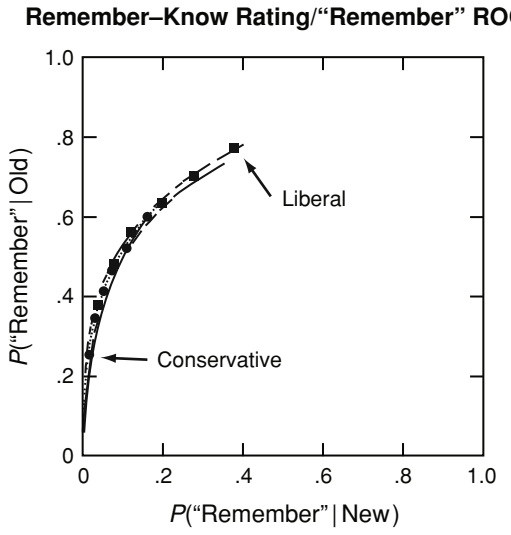

Figure 4. Remember–know rating / "remember" ROC data from Experiment 2. *P*("remember"|Old) is plotted against *P*("remember"|New). Circles represent the conservative condition and squares represent the liberal condition. The superimposed curves were generated with the one- and two-dimensional models; all of the curves fall in the same region of space. One-dimensional fits are the upper dashed curve (for liberal) and the dotted curve (for conservative); two-dimensional fits are the lower dashed curve (for liberal) and the solid curve (for conservative).

alarm rate should be observed, but the data in Figure 4 show a wide range of remember false alarm rates.

In contrast, these data present no challenge to the SDT models: Both the one-dimensional model and STREAK postulate remember–know criteria that partition a decision axis so that more highly rated "remember" responses are allocated to stronger memories. These models also do a good quantitative job of accounting for the ratings from this experiment, as can be seen in Figure 4 from the close correspondence between the observed points and the model predictions. Because memory sensitivity did not differ reliably across conditions, the predicted functions look very similar for the two bias instructions. However, the ROC differences produced by changing levels of C_o or K_o (and K_r) are evident in the narrower range of values predicted (and observed) in the conservative condition in comparison with the liberal condition.

The two SDT models make predictions that cannot be distinguished for this set of data. The mean AIC statistics for STREAK and the one-dimensional model are essentially identical in both conditions of this experiment (see Table 3); the same is true for the BIC statistics. According to AIC, the number of participants whose data were best described by the one-dimensional model was 17 of 21 in the conservative condition and 15 of 20 in the liberal. According to BIC, 18 participants' data in the conservative condition were better described by STREAK, as were the data for 15 of the participants in the liberal condition. These reversals reflect the extremely small differences in fit between the two models and suggest that the models are performing similarly.

Discussion

The participants claimed to recognize fewer items in the conservative condition than in the liberal condition. Application of the signal detection models indicated that memory sensitivity did not change with instructions intended to bias "old" decisions, but old–new bias became more liberal when the participants were asked to guess "old" whenever they were uncertain that an item had been studied. According to the two-dimensional model, remember–know bias was unaffected by old–new bias: The number of "remember" responses increased in the liberal condition relative to the conservative condition, but only because the rate of "old" judgments increased. The rate of remembering given "old" decisions was unchanged. The one-dimensional model can accommodate this response-ratio invariant by assuming that K_r became more liberal in the liberal condition. No dual-process model of this task has been presented, perhaps because it does not provide a natural description of the remember–know rating task.

GENERAL DISCUSSION

Our goal has been to understand the locus of response-bias effects in remember–know experiments. In two studies, we used instructions to manipulate participants' willingness to call a test item "old." In both cases, "remember" responses were also affected: More conservative old–new responding resulted in fewer "remember" responses. We collected ratings of old–new confidence (Experiment 1) or of the degree of remembering versus knowing (Experiment 2), allowing us to construct ROC curves whose shape was of theoretical interest.

In order to locate the mechanism for the observed effects, we have relied on quantitative models of the remember–know paradigm, and it is worth reviewing the rationale and logic underlying this strategy. All models interpret bias effects as shifts in response criteria, but each proposes a different basis for the "remember" response. In the dual-process model, this response arises from recollection and is thus process-pure. In the one-dimensional model—especially the Wixted and Stretch (2004) version tested in Experiment 1—"remember" responses are based on the *sum* of recollective strength and familiarity. In STREAK, it is the *difference* between these two strengths that mediates remembering. The relative success of the models is relevant to discriminating among these opposing theoretical statements.

Conclusions Arising From Dual-Process Models

The first step in applying this modeling strategy is to determine which accounts are plausible. Examining several different aspects of the data, we have concluded that the dual-process model does not provide a good description. For example, the parameter that accounts for false remembering, ε, changes with response bias rather than remaining constant and small in value. Experiments can be designed so as to minimize false remember responses (and, therefore, ε). For example, participants can be encouraged to only say "remember" when they can describe the nature

of the remembered detail to the experimenter (Yonelinas et al., 1996). Alternatively, the response options can be "remember" or a rating of the item's familiarity (Yonelinas, 2001). These designs limit the remember false alarm rates, which improves the apparent fit of the dual-process model. However, most experiments in the literature have not placed such strict requirements on the nature of "remember" responses, and in the more common designs (old–new followed by remember–know judgments), the data are not favorable to the dual-process model (Rotello, Macmillan, Reeder, & Wong, 2005). This failure is important because of the wide application of dual-process models in the literature.

More quantitatively, goodness of fit was inferior in Experiment 1: Only a few participants' data were best described by our implementation of the dual-process model, even when the model was extended so that "remember" responses could appear at multiple confidence levels.

Conclusions Arising From Signal Detection Models

The SDT models provide a good overall description of these results: Our instructional manipulations led to changes in criteria but not in sensitivity. Two aspects of the data that are troubling for dual-process models are congenial within the SDT approach: "Remember" responses can occur at a variety of old–new confidence levels, and ratings of remembering and knowing can be made on a continuous scale (as in Experiment 2). On goodness-of-fit measures, the fixed one-dimensional model provided the best fit for the majority of participants in Experiment 1, and fared about as well as STREAK with the data from Experiment 2.

The data do indicate that, if these models are correct, the decision rules adopted by our participants are not those that are most natural in model terms. For Experiment 1, the one-dimensional model concludes that the remember–know criterion shifted by about as much as the old–new criterion, and STREAK concluded that it shifted by about half as much. In addition, STREAK detected a correlation between rating response and the proportion of "remember" responses in Experiment 1 that, while reflecting an understandable strategy, led to inferior fits. In Experiment 2, the one-dimensional model found changes in the remember–know criterion of about three quarters the magnitude of the old–new shift, whereas, according to STREAK, the remember–know criterion moved only slightly. There is no reason participants could not adopt these decision strategies, but our findings lead us to have greater confidence in the models' ability to separate sensitivity from bias than to describe the decision rules in detail.

Can we choose between the one- and two-dimensional versions of the SDT model? Remember that these models have different, even contradictory views on the nature of the "remember" response, so deciding between them is of some importance. Unfortunately, the answer appears to depend on the task. In Experiment 1, ratings of old–new confidence were correlated with "remember" responses, which caused STREAK some difficulty and resulted in an inferior overall fit. In Experiment 2, ratings of the remember–know continuum provided data that were equally well described by the one- and two-dimensional models. Further complicating the theoretical picture, each model has a "signature" prediction, and these data slightly favor STREAK over the one-dimensional model. The fixed one-dimensional model predicts that the slope of the two-point *z*ROC equals that of the ratings ROC. Past experiments suggest that this is false: In recognition confidence rating experiments, the *z*ROC slope averages approximately 0.8 (e.g., Ratcliff et al., 1992), whereas two-point *z*ROC slopes have a mean of about 1.0 (Rotello et al., 2004). The data from Experiment 1 allow us to compare *z*ROC slopes for recognition confidence ratings and for remember–know judgments for the same participants: In both bias conditions, the two-point slopes were greater than the recognition slopes; significantly so in the *70% Old* condition.

STREAK's unique prediction is the response-ratio invariant: When old–new response bias changes, the rate of remembering changes proportionally so that the same fraction of hits and false alarms are followed by remember judgments under both conservative and liberal old–new bias. This invariant was observed between conditions in both experiments, and for both Old and New words. Neither form of the one-dimensional model predicts the response-ratio invariant, though the result can be accommodated by allowing both decision criteria to move.

Implications for Remember–Know Research

The remember–know paradigm has been extensively used to explore for dissociations, experimental manipulations that affect the rates of "remember" and "know" responding differently. Typically, if one rate changes and the other stays the same, or if the two rates change in opposite directions, a dissociation is inferred and the independent variable in question is thought to influence only one of two processes, or to influence both in opposite ways.

How might the presence of response bias muddy such conclusions? Signal detection models allow us to distinguish between dissociation-like effects that are due to changes in sensitivity and those that are due to response bias. Much discussion of dissociations tacitly assumes that they are of the former type, so that a change in experimental conditions changes the relative efficacy of the two underlying processes. The dual-process model offers support for this process-pure view, at least with regard to "remember" responses. To see how application of SDT models could alter such reasoning, let us examine two cases in which response patterns shift while sensitivity is fixed.

First, the remember and know rates may change in opposite directions but leave the overall "old" and "new" rates unchanged; both SDT models interpret this as a change in the remember–know criterion. Second, "old" and "remember" rates [that is, hit and false alarm rates as well as *P*("remember"|Old) and *P*("remember"|New)] may change in the same direction, but proportionately;

STREAK assigns responsibility for this result to a change in the old–new criterion, and the one-dimensional models assume a change in both criteria. A qualitative approach might describe the first pattern and not the second as a dissociation, but in neither case has the sensitivity of an underlying process changed. Rather, according to the SDT models, both outcomes reflect adjustments in decision strategy. Without an interpretive model, this important distinction between sensitivity and bias cannot be made.

REFERENCES

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, **13**, 1-21.

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

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

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

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. (1997). On reporting recollective experiences and "direct access to memory systems." *Psychological Science*, **8**, 391-394.

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., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science*, **9**, 61-65.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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

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

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

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In H. Pashler & J. [T.] Wixted (Eds.), *Stevens' Handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 429-460). New York: Wiley.

Postma, A. (1999). The influence of decision criteria upon remembering and knowing in recognition memory. *Acta Psychologica*, **103**, 65-76.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, **21**, 89-102.

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

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 computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 294-320.

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.

Strack, F., & Förster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, **6**, 352-358.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.

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

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. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, **25**, 747-763.

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., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness & Cognition*, **5**, 418-441.

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory & Language*, **34**, 622-643.

Yonelinas, A. P., Regehr, G., & Jacoby, L. L. (1995). Incorporating response bias in a dual-process theory of memory. *Journal of Memory & Language*, **34**, 821-835.

NOTES

1. Donaldson's proposal is often called "the signal detection model." We avoid this term because all the models we consider here incorporate detection-theoretic assumptions, differing in other characteristics of the assumed underlying processes.

2. Other implementations of the dual process are possible; our conclusions are limited to the versions we tested.

3. Note that response bias was manipulated between participants. Because model fitting was done at the individual level, both response bias and sensitivity parameters were free to vary in all models.

4. Sample sizes were reduced by the few participants who did not make any remember false alarms. For those participants, two-point *z*ROC slopes cannot be calculated. Another concern was raised by a reviewer: If the most conservative points on the *z*ROC show a steeper slope than the overall data, then the remember–know data will necessarily show a steeper slope than the old–new data. To evaluate this explanation of the slope difference, we compared the two-point slopes based on remember–know data to two-point *z*ROC slopes calculated from the two highest-confidence "old" ratings. The two-point remember–know slopes were significantly steeper than the two-point recognition slopes in the 70% condition [1.14 vs. 0.85, $t(21) = 2.08$, $p = .05$] and were marginally steeper in the 30% condition [0.83 vs. 0.64, $t(19) = 1.86$, $p = .078$].

5. The extended model improved the fit for 18 of the 22 participants in the *30% Old* condition, and for 21 of the 24 in the *70% Old* condition.

6. We also considered a different division of the ratings, in which 1–5 counted as "remember" and 6 as "know." Estimates of the "remember" response proportions and the C_r parameter in STREAK differed numerically in that analysis, but none of the patterns of significance or conclusions changed.

APPENDIX
Models for Remember–Know Rating Experiments

The three types of models that we evaluated against these data have all been tested before, but it was necessary to elaborate the dual-process and one-dimensional models to apply them to the present experiments, especially Experiment 1. We attempted to extend the fundamental assumptions of the models in natural ways, but faced some choices about how to do this.

DUAL-PROCESS MODEL

Yonelinas and his colleagues have often tested the dual-process model against ROC curves in standard old–new recognition memory experiments (e.g., Yonelinas, 1994, 1997). Yonelinas (2001) theorized that the intercept of such a curve should equal the proportion of "remember" responses to Old items (corrected for "remember" responses to New items) in a remember–know experiment, and tested this hypothesis in an experiment in which the response options were "remember" plus a rating scale from "know" to "new." The paradigm used in the present experiments is less restrictive, allowing "remember" responses for any rating in the "old" region.

One strategy for testing the model is to fit an old–new ROC with the constraint that the intercept equal $P(\text{"remember"} \mid \text{Old}) - P(\text{"remember"} \mid \text{New})$. Another strategy would be to fit the old–new ROC and also to fit the "remember" responses in the highest-confidence rating. We rejected both of these approaches because they do not try to account for the distribution of "remember" and "know" responses across all ratings. Instead, we fit the distribution of "remember," "know," and "new" responses to each rating category. Following Yonelinas (2001), we assumed that "remember" responses should all be assigned the highest level of confidence, so that the predicted frequencies of "remember" responses for ratings 5 and 4 is zero. This *standard model* captures at least the spirit of the dual-process hypothesis, and we tested it (with predicted "remember" proportions of .01, rather than 0, to allow calculation of likelihoods). In the following equations for this *standard* model, R_o is the true remember rate, ε is the false remember rate, d is the mean of the Old distribution, C_1 to C_5 are the criterion locations (C_1 being the most liberal), and Φ is the cumulative normal distribution function.

Targets

$$
\begin{aligned}
&P(\text{Rem \& 6}) = .98\,(R_o + \varepsilon)\\
&P(\text{Rem \& 5}) = .01\,(R_o + \varepsilon)\\
&P(\text{Rem \& 4}) = .01\,(R_o + \varepsilon)\\
&P(\text{Know \& 6}) = (1 - R_o)\,\Phi(-C_5 + d)\\
&P(\text{Know \& 5}) = (1 - R_o)\,[\Phi(C_5 - d) - \Phi(C_4 - d)]\\
&P(\text{Know \& 4}) = (1 - R_o)\,[\Phi(C_4 - d) - \Phi(C_3 - d)]\\
&P(3) = (1 - R_o)\,[\Phi(C_3 - d) - \Phi(C_2 - d)]\\
&P(2) = (1 - R_o)\,[\Phi(C_2 - d) - \Phi(C_1 - d)]\\
&P(1) = (1 - R_o)\,\Phi(C_1 - d)
\end{aligned}
\tag{A1}
$$

Lures

$$
\begin{aligned}
&P(\text{Rem \& 6}) = .98\,\varepsilon\\
&P(\text{Rem \& 5}) = .01\,\varepsilon\\
&P(\text{Rem \& 4}) = .01\,\varepsilon\\
&P(\text{Know \& 6}) = (1 - \varepsilon)\,\Phi(-C_5)\\
&P(\text{Know \& 5}) = (1 - \varepsilon)\,[\Phi(C_5) - \Phi(C_4)]\\
&P(\text{Know \& 4}) = (1 - \varepsilon)\,[\Phi(C_4) - \Phi(C_3)]\\
&P(3) = (1 - \varepsilon)\,[\Phi(C_3) - \Phi(C_2)]\\
&P(2) = (1 - \varepsilon)\,[\Phi(C_2) - \Phi(C_1)]\\
&P(1) = (1 - \varepsilon)\,\Phi(C_1)
\end{aligned}
\tag{A2}
$$

A less restrictive model allows "remember" responses to be distributed freely across ratings 6, 5, and 4. Such a modification gives the model more flexibility at the expense of making it more similar to the detection-theoretic models. This *extended model* replaces the fixed distribution of remember responses (.98, .01, .01) with free parameters that sum to 1; thus, the standard model we implemented is a specific case of the extended model.

Targets

$$
\begin{aligned}
&P(\text{Rem \& 6}) = (R_o + \varepsilon)(1 - r_5 - r_4)\\
&P(\text{Rem \& 5}) = (R_o + \varepsilon)\,r_5\\
&P(\text{Rem \& 4}) = (R_o + \varepsilon)\,r_4
\end{aligned}
\tag{A3}
$$

Lures

$$P(\text{Rem \& 6}) = \varepsilon\,(1 - r_5 - r_4)$$
$$P(\text{Rem \& 5}) = \varepsilon\, r_5$$
$$P(\text{Rem \& 4}) = \varepsilon\, r_4 \tag{A4}$$

ONE-DIMENSIONAL MODEL

In the one-dimensional model for 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 C_1 to C_5 for the old–new judgment, and the remember–know criterion C_r could fall within any rating in the "old" region. The *fixed* version of this model predicts that if the remember–know criterion falls in the region corresponding to rating i, then both "remember" and "know" responses can occur for that rating, only "remember" responses can occur for ratings greater than i, and only "know" responses can occur for ratings less than i. (As for the dual-process model, predicted values of 0 were replaced with .01, and the remaining cells adjusted slightly so that the predicted proportions summed to 1.) The New distribution has $M = 0$ and $SD = 1$, and the Old distribution has $M = d$ and $SD = 1/s$. The equations depend on the location of C_r; we spell out only the case in which C_r falls in Rating Category 5 (the most common case in our experiments).

Targets

$$P(\text{Rem \& 6}) = .98\,\Phi[-s(C_5 - d)]$$
$$P(\text{Rem \& 5}) = .98\,\{\Phi[s(C_5 - d)] - \Phi[s(C_r - d)]\}$$
$$P(\text{Rem \& 4}) = .01$$
$$P(\text{Know \& 6}) = .01$$
$$P(\text{Know \& 5}) = .98\,\{\Phi[s(C_r - d)] - \Phi[s(C_4 - d)]\}$$
$$P(\text{Know \& 4}) = .98\,\{\Phi[s(C_4 - d)] - \Phi[s(C_3 - d)]\}$$
$$P(3) = .98\,\{\Phi[s(C_3 - d)] - \Phi[s(C_2 - d)]\}$$
$$P(2) = .98\,\{\Phi[s(C_2 - d)] - \Phi[s(C_1 - d)]\}$$
$$P(1) = .98\,\Phi[s(C_1 - d)] \tag{A5}$$

Lures

$$P(\text{Rem \& 6}) = .98\,\Phi(-C_5)$$
$$P(\text{Rem \& 5}) = .98\,[\Phi(C_5) - \Phi(C_r)]$$
$$P(\text{Rem \& 4}) = .01$$
$$P(\text{Know \& 6}) = .01$$
$$P(\text{Know \& 5}) = .98\,[\Phi(C_r) - \Phi(C_4)]$$
$$P(\text{Know \& 4}) = .98\,[\Phi(C_4) - \Phi(C_3)]$$
$$P(3) = .98\,[\Phi(C_3) - \Phi(C_2)]$$
$$P(2) = .98\,[\Phi(C_2) - \Phi(C_1)]$$
$$P(1) = .98\,\Phi(C_1) \tag{A6}$$

The *variable* version of the one-dimensional model, proposed by Wixted and Stretch (2004), permits the remember–know criterion to vary from trial to trial and thus allows for both "remember" and "know" responses at every rating. The five old–new criteria are constant, as in the fixed version, but the remember–know criterion is now a random variable C that has an *average* location of C_r and a standard deviation of $1/t$. The model says that a rating of i arises when strength x is between C_{i-1} and C_i and the remember–know decision is based on whether x is above or below the current value of C. The process governing the location of C is independent of that determining the strength x, so these variables are appropriately represented as orthogonal dimensions, as shown in Figure A1.

Ratings depend only on strength, and the C_i therefore divide the space into vertical strips. Remember judgments require that $x > C$—that is, $C - x < 0$—so in the "old" region (ratings 4, 5, and 6) "remember" and "know" responses should be separated by the diagonal unit-slope line $C = x$. For simplicity of calculation, however, we use a stepping-stone decision rule that approximates this (but is slightly more conservative with regard to saying "remember"): If strength is between C_i and C_{i+1}, a "remember" response is given if the remember–know criterion is lower than C_i.

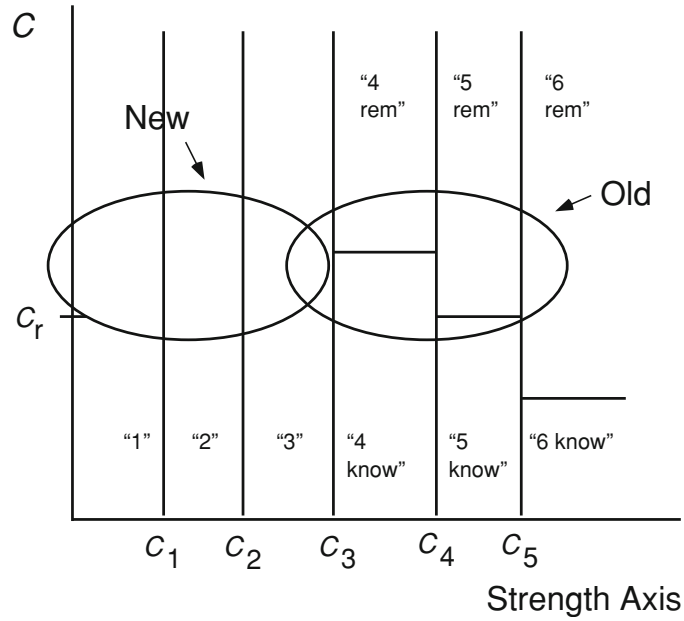

Figure A1. A decision space for the variable criterion version of the one-dimensional model for the remember–know rating design. Old and New distributions are represented as ellipses. The horizontal axis represents memory strength and is divided by five criteria (C_1 to C_5) into six vertical strips that correspond to the rating response. The vertical axis represents the location of the remember–know criterion. "Remember" and "know" responses are given only following ratings of 4, 5, or 6; the location of the criterion is set so that the likelihood of saying "remember" increases with confidence. C_r is the mean location of the remember–know criterion.

Targets

$$
\begin{aligned}
&P(\text{Rem \& 6}) = \Phi[s(d - C_5)]\,\Phi[t(C_5 - C_r)]\\
&P(\text{Know \& 6}) = \Phi[s(d - C_5)]\,\Phi[t(C_4 - C_5)]\\
&P(\text{Rem \& 5}) = \{\Phi[s(C_5 - d)] - \Phi[s(C_4 - d)]\}\,\Phi[t(C_4 - C_r)]\\
&P(\text{Know \& 5}) = \{\Phi[s(C_5 - d)] - \Phi[s(C_4 - d)]\}\,\Phi[t(C_r - C_4)]\\
&P(\text{Rem \& 4}) = \{\Phi[s(C_4 - d)] - \Phi[s(C_3 - d)]\}\,\Phi[t(C_3 - C_r)]\\
&P(\text{Know \& 4}) = \{\Phi[s(C_4 - d)] - \Phi[s(C_3 - d)]\}\,\Phi[t(C_r - C_3)]\\
&P(3) = \Phi[s(C_3 - d)] - \Phi[s(C_2 - d)]\\
&P(2) = \Phi[s(C_2 - d)] - \Phi[s(C_1 - d)]\\
&P(1) = \Phi[s(C_1 - d)]
\end{aligned}
\tag{A7}
$$

Lures

$$
\begin{aligned}
&P(\text{Rem \& 6}) = \Phi(-C_5)\,\Phi[t(C_5 - C_r)]\\
&P(\text{Know \& 6}) = \Phi(-C_5)\,\Phi[t(C_r - C_5)]\\
&P(\text{Rem \& 5}) = [\Phi(C_5) - \Phi(C_4)]\,\Phi[t(C_4 - C_r)]\\
&P(\text{Know \& 5}) = [\Phi(C_5) - \Phi(C_4)]\,\Phi[t(C_r - C_4)]\\
&P(\text{Rem \& 4}) = [\Phi(C_4) - \Phi(C_3)]\,\Phi[t(C_3 - C_r)]\\
&P(\text{Know \& 4}) = [\Phi(C_4) - \Phi(C_3)]\,\Phi[t(C_r - C_3)]\\
&P(3) = \Phi(C_3) - \Phi(C_2)\\
&P(2) = \Phi(C_2) - \Phi(C_1)\\
&P(1) = \Phi(C_1)
\end{aligned}
\tag{A8}
$$

To fit the remember–know rating data collected in Experiment 2, we assumed that the remember–know boundary fell at the middle rating category (i.e., three ratings for remembering and three for knowing). The resulting equations are:

Targets

$$P(6) = \Phi[s(C_6 - d)]$$
$$P(k) = \Phi[s(C_{k+1} - d)] - \Phi[s(C_k - d)], k = 1 \ldots 5$$
$$P(\text{“new”}) = \Phi[s(C_1 - d)] \quad \text{(A9)}$$

Lures

$$P(6) = 1 - \Phi(C_6)$$
$$P(k) = \Phi(C_{k+1}) - \Phi(C_k), k = 1 \ldots 5$$
$$P(\text{“new”}) = \Phi(C_1) \quad \text{(A10)}$$

(Manuscript received September 8, 2004;
revision accepted for publication September 27, 2005.)