



Remember–know models as decision strategies in two experimental paradigms

Caren M. Rotello ^{*}, Neil A. Macmillan

Department of Psychology, Box 37710, University of Massachusetts, Amherst, MA 01003-7710, USA

Received 2 March 2006; revision received 25 July 2006

Available online 18 September 2006

Abstract

In the remember–know paradigm, subjects report the subjective basis for their “old” response to a memory probe to be either recollection of specific details (“remembering”) or familiarity (“knowing”). The response rates for these judgments are often taken as direct measures of underlying processes, but this process-pure account is implausible in view of the known effects of experimental paradigm. Here, we explore two such paradigms: the *remember-first* method, in which a remember response is solicited first, followed by “new” or “know” for nonremembered items; and the *trinary* paradigm, in which a single response adjudicates among the “remember,” “know,” and “new” alternatives. We expand these paradigms to include rating responses, allowing us to distinguish among a set of quantitative models. Subjects’ decision rules, inferred from the models providing the best fit, varied. In both experiments, however, a one-dimensional strength model described a majority (48 of 70) of subjects and the apparently natural process-pure model was rarely supported (4 subjects). The remember–know paradigm is commonly justified by the intuition that it recruits two processes, but in the present experiments “remember” and “know” responses most often depend on a single strength variable.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Remember–know; Memory models; Recollection and familiarity; Recognition memory

In the remember–know recognition memory experiment, the decision that a tested item has been studied is elaborated by describing the prior presentation as explicitly “remembered” or simply familiar (“known”). Remember–know judgments (Tulving, 1985) have been used in hundreds of recognition memory experiments, usually with the goal of measuring the relative contributions of the underlying recollective and familiarity-based processes. Indeed, many of these experiments have

observed dissociations of remember and know responses across experimental conditions (for a review see Gardiner & Richardson-Klavehn, 2000). These dissociations have been taken as evidence of the process purity of the judgments: “remember” responses reflect recollection, and “know” responses reflect familiarity in the absence of recollection.

This interpretation of the data may be correct but it is not self-evident, and in designing the experiments reported in this paper we adopted an agnostic view toward the question of what remembers and knows mean. To make progress toward an answer, we assumed that responses in recognition memory experiments are

^{*} Corresponding author. Fax: +1 413 545 0996.

E-mail address: caren@psych.umass.edu (C.M. Rotello).

based on a partition of a decision space. This is not a controversial position: the simplest old–new recognition task has long been modeled as a binary division of a single strength dimension and analyzed using signal detection theory (SDT: Banks, 1970; Lockhart & Murdock, 1970). If ratings are added to the task, the strength dimension is divided by multiple criterion points into regions corresponding to the responses. These regions are usually taken to reflect levels of response bias, a conclusion that relies on the constancy of sensitivity across different ratings (Macmillan & Creelman, 2005).

In our approach, existing models of the remember–know paradigm are described as partitions of a two-dimensional decision space. This perspective was first adopted by Tulving (1985), and his proposal is shown in Fig. 1. The underlying familiarity and recollection dimensions are labeled “semantic information” and “episodic information,” and a criterion curve divides events leading to “old” and “new” decisions. An item may be deemed Old if it evokes high values on either dimension [points (a,z) and (c,x)] or moderate values on both [point (b,y)]. To describe existing models in this space requires three modest modifications: (1) specification of the form of the old–new bound; (2) specification of a bound for making the remember–know decision, an idea that is only implicit in Tulving; and (3) interpretation of the axes, which have been labeled in different ways by different theorists. Taking a neutral stance on this last question, we call the axes simply x (semantic, in Tulving’s proposal) and y (episodic). (We have reversed his axes to be consistent with past work.)

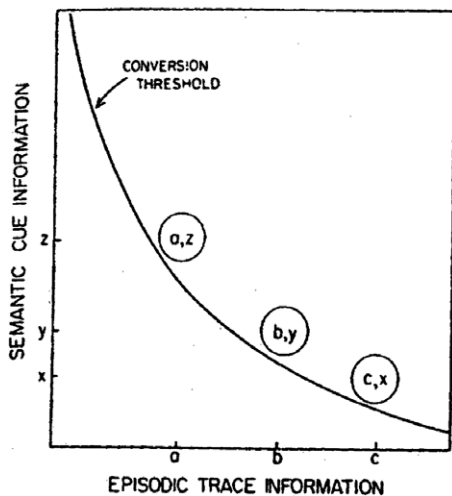


Fig. 1. Schematic diagram of the synergistic ephory model of retrieval. A two-dimensional memory space. Events differ in the episodic and semantic strength, and those above the curved bound are deemed “old.” Items stronger in episodic strength tend to be remembered and those stronger in semantic strength are known. From Tulving (1985), with permission from the author and the publisher.

The plan of the article is as follows: first, existing models for the basic (nonrating) remember–know task are summarized. We then argue that these models can be better distinguished if they are elaborated to include ratings. The type of ratings required depends on details of the experimental paradigm, and we present models for several designs. Two experiments, reported next, employ designs (*remember-first* and *trinary*) for which distinct models seem appropriate, and we test the degree to which the decision rule is chosen to fit the experimental paradigm. The pattern of outcomes for individual subjects leads us to conclude that experimental design has a modest influence on strategy in remember–know decision-making.

Models without ratings

The process-pure model

The most straightforward interpretation of the remember–know task is offered by the *process-pure model*, which assigns responses as directly as possible to the two presumed underlying processes of familiarity and recollection. A sufficiently high value on the y dimension leads to a “remember” response; if y is less than this criterial value but x exceeds a familiarity criterion, a “know” response is made. If neither x nor y is large enough, the response is “new.” The partition of the decision space corresponding to this rule is shown in Fig. 2A.

The process-pure model is often implicitly assumed in empirical work: the rate of “recollection” is estimated from remember responses, the contribution of “familiarity” from know responses corrected for the opportunity to make them, i.e.,

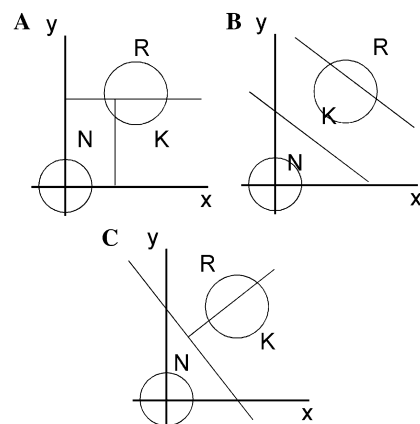


Fig. 2. Decision space for the remember–know task without ratings: (A) process-pure model, (B) one-dimensional model, (C) sum-difference model (STREAK).

$$P(\text{recollection}) = P(\text{“remember”}),$$

$$P(\text{familiarity}) = P(\text{“know”})/[1 - P(\text{“remember”})]. \quad (1)$$

The model assumes that a remember response is always made if there is sufficient y -type information, whereas a know response is a secondary option.

To develop predictions from this model, SDT assumptions are commonly used. In Fig. 2, the distributions resulting from Old and New items are shown as circles (equal-probability contours) that are the locus of points a common distance from the distribution mean. Because the Old and New distributions are assumed to be normal, they extend infinitely and overlap. The model has been elaborated in three different ways. Reder et al. (2000; Diana, Reder, Arndt, & Park, 2006) derived Eq. (1) from a process model, *Source of Activation Confusion (SAC)*, in which episodic and semantic information are separately encoded in a network. Murdock (2006) grafted Eq. (1) onto the *TODAM* model, identifying the x axis as item strength and the y axis as associative strength. Yonelinas (2001) has assumed that the remember process has a threshold character, leading to the prediction by his *dual-process model* that remember false alarms have a “nonmemorial” source.

The one-dimensional model

Although subjects seem to have little difficulty in following remember–know instructions, it is possible that what they believe to be qualitatively distinct kinds of memory are merely two levels of strength that differ only quantitatively. The *one-dimensional model* (Donaldson, 1996) captures this possibility. Whereas Donaldson did not specify the nature of the strength dimension, Wixted and Stretch (2004) suggested that it is simply the sum of x and y (familiarity and recollection, in their treatment). A decision variable that represents such a sum corresponds to a line of positive slope (if the sum is unweighted, the slope of the line is 1; see Macmillan & Rotello, 2006). As Fig. 2B shows, the decision space is divided into three regions, “remember” being assigned to values above the remember criterion, “know” to values between the remember and old criteria, and “new” to values below the old criterion. The model is formally identical to one for a 3-category rating experiment (“sure yes,” “unsure,” “sure no”). Experiments that have compared rating (sure-unsure) responses with remember–know judgments have sometimes been interpreted as failing to support that isomorphism (e.g., Rajaram, 1993), but Dunn (2004) has shown that both types of judgment are consistent with the one-dimensional model if response criteria are allowed to differ.

An early test of the one-dimensional model (Donaldson, 1996; Gardiner & Gregg, 1997) assumed that the variances of the Old and New distributions were the

same, so that sensitivity should be identical whether estimated from the old or the remember criterion. Meta-analyses (Dunn, 2004; Rotello, Macmillan, & Reeder, 2004; Macmillan, Rotello, & Verde, 2005) have revealed that this prediction is approximately correct. This good news for the model in isolation is bad news for connecting it with the larger recognition memory literature, in which the variance of the Old distribution is almost always found to be greater than that of the New distribution (Ratcliff, Sheu, & Gronlund, 1992).

The sum–difference model (STREAK)

In STREAK (Rotello et al., 2004) x is labeled global strength, y specific strength, and memory judgments are based on a sum or difference of these values. A judgment of “old” is more likely when the sum $y + x$ is large, because greater values of either aspect of strength justify that decision. A “remember” response is more likely when the relative contribution of specific strength ($y - x$) is large; it is the balance of the two strength components, not the absolute quantity of either, that determines the response. Neither judgment is process-pure. The resulting partition of the decision space is shown in Fig. 2C: the sum of x and y increases toward the upper right corner, whereas the difference between y and x increases toward the upper left corner.

Comparing the models

Saturated models, plausible parameter values, and the nature of underlying dimensions

The three models just described are conceptually distinct, particularly in their proposals about the “remember” response: is it based on the y variable alone, the sum of y and x , or the difference between y and x ? But the process-pure model and STREAK are saturated, with 4 parameters to describe 4 independent data points; the one-dimensional model has 3 parameters in its simplest version, but if unequal variances are allowed or if the remember criterion is variable (as proposed by Wixted & Stretch, 2004) it is saturated as well. Comparing models by seeking the “best fit” is not possible, because in general all models fit perfectly.

Another possibility is to examine the parameter estimates of the models to see if they are reasonable. Murdock (2006) argued that parameter values estimated from the process-pure model made sense for four data sets, but Macmillan and Rotello (2006) showed that STREAK’s parameters were equally plausible. Even the 3-parameter one-dimensional model, which was applied by Dunn (2004) to these same data sets, tells a convincing story.

Theorists have chosen different terms to describe the variables being encoded; can we determine whether item and associative memory strength, as in Murdock (2006),

is a superior pair of labels to semantic and episodic strength (Reder et al., 2000), familiarity and recollective strength (Wixted & Stretch, 2004), or global and specific strength (Rotello et al., 2004)? Two hurdles arise: First, as we have seen, simply examining parameter values tends not to discriminate among these terms. Second and more fundamentally, any of the decision rules can be combined with any of the axis labels, so even if the decision models could be discriminated the implications for verbal description of the space are nil.

Paradigm dependence

There are multiple procedures for eliciting remember-know judgments, and some of these seem to map naturally onto particular decision models. The most common method has been to first ask for an old-new decision, then collect remember-know judgments for words that are judged “old.” This experimental design is most consistent with the sum-difference decision bounds in Fig. 2C (Rotello et al., 2004): test probes are “old” if they fall above the old-new bound, and if so they are compared with the remember-know bound. In contrast, the most natural decision sequence under the process-pure view (Fig. 2A) is to first decide if a test probe is remembered (does the event falls above the criterion for remembering on the y -axis?). If not, the second stage of judgment determines whether the probe is new or known by comparing its strength with the know criterion on the x -axis. The one-dimensional model seems equally adept at handling either type of decision.

The first experiment reported here uses the relatively rare remember-first paradigm; the second adopts the *ternary* paradigm, in which a single decision is made among the three alternatives “remember,” “know,” and “new.” In choosing these designs we built on our previous experiments with the popular old-first design (Rotello et al., 2004) and hoped to find a match between the paradigms and the contrasting models.

Rating experiments

Any approach to comparing the models must circumvent the problem of saturation. One solution is to collect data in multiple conditions in which some parameters can be presumed to be constant. For example, Rotello, Macmillan, Hicks, and Hautus (in press) manipulated response bias between conditions, with no effect on sensitivity. Cary and Reder (2003) varied list strength, list length, and other variables, again increasing the number of data cells relative to the number of parameters in the SAC model.

In this paper we adopt a second strategy, the analysis of confidence ratings. Ratings designs allow the construction of receiver operating characteristic (ROC) curves, which can be powerful in distinguishing alternative models. Typically, ROC curves in recognition memory tasks are based on subjects' confidence ratings that

each test probe had been studied; these old-new ROCs are usually best described with an unequal-variance Gaussian model (Wixted, in press). For the remember-know experiment other types of ratings, corresponding to the many combinations of model and paradigm, are possible. Each such design requires a model to describe it, and we spell out these rating models next.

Models for rating paradigms

Old-first paradigm

The old-first paradigm is the dominant method in the remember-know literature, and we did not collect new data in this form for the present study. Either of the two required judgments (old-new and remember-know) can be converted to a rating, and both types of ratings have been used.

For old-new ratings followed by a binary remember-know judgment, STREAK and the one-dimensional model postulate a series of decision bounds parallel to the old-new bound in Figs. 2B and C and make essentially identical predictions about the form of ROC.¹ Rotello et al. (2004, Exp. 1) verified these predictions, but ROCs based on only the “know” responses were better described by the one-dimensional model than by STREAK. Similarly, Rotello et al. (in press, Exp. 1) found STREAK's fit to be inferior to that of the one-dimensional and dual-process models. In these tests, it appears that the independence assumption (graphically, the orthogonality of the old-new and remember-know bounds in Fig. 2C) fails. No process-pure ratings model has been developed for the old-first task, nor is it obvious from an examination of Fig. 2A whether the remember criterion, the know criterion, or both should change with confidence.

For remember-know ratings preceded by a binary old-new judgment, STREAK postulates a series of decision bounds orthogonal to the old-new bound (Fig. 2C), whereas the one-dimensional model's additional criteria are necessarily parallel to its old-new bound (Fig. 2B). In spite of this difference in the assumed structure of remember-know ratings, the two existing comparisons (Rotello et al., 2004, Exp. 2; Rotello et al., in press, Exp. 2) show that both models provide excellent accounts of the data and are superior to dual-process accounts. Again, no process-pure model has been developed.

¹ Quantitative predictions for the one-dimensional model can be found in Rotello et al. (in press) and for STREAK can be derived from the nonrating expressions in Appendix B of Rotello et al. (2004)

Remember-first paradigm

As noted earlier, the remember-first design is intuitively more compatible with the process-pure model (Fig. 2A), and a handful of experiments have included ratings in this paradigm. For example, Yonelinas and Jacoby (1995, Exp. 3) (see also Yonelinas, 2001, Exps. 2a–2c) asked their subjects to first decide if a test probe was remembered; if it was not, they were then asked to rate their confidence that the word was new or known along a 6-point scale. We use this method in Experiment 1 and evaluate the ability of a process-pure model, the one-dimensional model and a version of STREAK to describe the data.

The models (specified mathematically in Appendix A) are illustrated in Fig. 3. In all models, the Old and New distributions are circular bivariate normal with means of (d_x, d_y) and $(0,0)$ and standard deviations of l and s , respectively. In the process-pure model (Fig. 3A) a decision bound on the y -axis divides “remember” from “not remember” responses and an orthogonal series of criteria on the x -axis define the new-know ratings. The one-dimensional model inserts the new-know ratings between the old criterion and the remember criterion (Fig. 3B). A third model (STREAK Patio, Fig. 3C) restricts remember responses to a corner of the decision space, as in the original version of STREAK.

The dual-process model (Yonelinas, 2001) is not generally identical to the process-pure model shown in Fig. 3A because it assumes that remembering is a high-threshold rather than a continuous process. However, in the absence of confidence ratings on the remember response, the dual-process model cannot be distinguished from the process-pure representation.

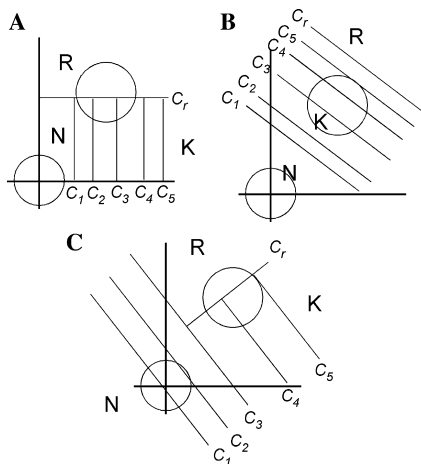


Fig. 3. Decision space for the remember-first task with ratings (Experiment 1): (A) process-pure, (B) one-dimensional, (C) STREAK Patio.

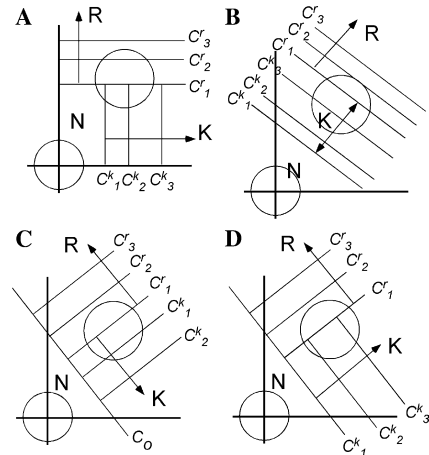


Fig. 4. Decision space for the trinary (RKN) task with ratings (Experiment 2): (A) process-pure, (B) one-dimensional, (C) STREAK Parallel, (D) STREAK Parquet.

Trinary paradigm

A number of experiments have used the trinary remember-know-new decision task (e.g., Parkin & Russo, 1993; Hicks & Marsh, 1999). The trinary paradigm is logically neutral, allowing the subject maximal flexibility to devise a response strategy. Unlike the old-first and remember-first paradigm, this task does not lend itself to ratings between two responses; instead, it seems natural to ask for confidence judgments of the responses themselves. In Experiment 2, subjects followed up remember and know responses by rating their confidence on a 3-point scale.

Our rating analyses for this task build on the same models as those for the remember-first task. Fig. 4A shows the process-pure model with its orthogonal remember and know criteria. The one-dimensional model (Fig. 4B) inserts both subsets of ratings along its single decision axis. The remaining two models are based on STREAK in assuming remember ratings to be based on different amounts of the variable $y - x$. In the STREAK Parallel model (Fig. 4C) the know ratings are based on this same variable, so that increasing confidence in a “know” response indicates increasing rejection of remembering. In the STREAK Parquet model, on the other hand (Fig. 4D), know ratings are based on $y + x$ and increasing confidence in knowing indicates increasing rejection of the “new” alternative.

Experiment 1—the “remember-first” paradigm

Experiment 1 employed a two-part judgment task like that used by (Yonelinas & Jacoby (1995, Exp. 2); see also Yonelinas, 2001, Exps. 2A–2C). For each test

word, subjects first decided whether they remembered any details of its appearance on the study list, such as its position on the list, the word that was studied before or after it, or their emotional reaction to the word. For words that were not remembered, subjects decided whether the word was new or known (sure it was studied, despite the absence of recollected details). They made this second decision on a 6-point scale that was anchored with Sure New at one end and Sure Know at the other. These “know” ROCs were then fit with all four models for the remember-first task (Fig. 3); because the judgment task corresponds naturally to the process-pure bounds, that model was expected to provide the best fit to the data.

Method

Participants

Twenty-three undergraduate students from the University of Massachusetts participated in the experiment for course credit. All participants were native speakers of English. One participant who did not follow the instructions was excluded from the analyses.

Stimuli

One hundred and twenty nouns were selected from the MRC Psycholinguistic Database (Coltheart, 1981). The words were divided equally into 2 lists that were closely matched for number of syllables (mean = 1.4), length (mean = 5.12 letters), and frequency (Kucera & Francis, 1967; mean = 98, standard deviation = 113). Fourteen 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 studied one list of 60 words, and the other half studied the other list. The words were presented in a different random order for each subject, along with 4 practice items. Each word was presented in the center of the computer screen for 1250 ms, with a 750 ms inter-stimulus interval. Three primacy and 3 recency items were also included in the study list. Participants were instructed to study the words carefully in preparation for an upcoming memory test.

After the study list, participants were given instructions about the nature of the memory test. They were told that they would see a series of words, some of which were Old and others were New. Standard definitions of remembering and knowing were provided on a paper handout (Rajaram, 1993); the remember-know distinction was reinforced through verbal descriptions. For each test probe, subjects first judged whether or not they *remembered* the word. They responded “remembered” or “not remembered” by pressing “z” or “/” key on the computer keyboard, respectively. If they claimed to remember, there were no additional questions about that probe. If they failed to remember a test probe, they were asked to decide whether that was because it was a New word or because they *knew* they studied it. This new-know response was made on a 6-point rating scale, ranging from “(1) sure new” to “(6) sure know.” Participants were asked to distribute their new-know responses across all 6 ratings.

These instructions were supplemented by a brief practice task that included 8 words (half Old; half New). During the practice phase, participants were encouraged to ask questions and to explain their “remember” and “know” responses to the experimenter.

Following the practice phase, participants completed the recognition memory test. Sixty Old and 60 New words were presented in random order. The experimenter was not present in the room during the test phase of the study.

Results

Table 1, which presents the overall response proportions, shows that 62% of Old items and 18% of New items were “remembered,” whereas 20% in each category were “known.” We later compare these values with those from the literature (and from Experiment 2). Of the models we are evaluating, only the equal-variance one-dimensional model is unsaturated and can be tested in this (nonrating) form. The prediction that overall memory sensitivity should be equal whether measured for “remember” or “old” responses (Donaldson, 1996) is confirmed: d is 1.22 for the “remember” criterion and 1.18 for the “old” criterion.

The know-new rating data can be used to construct a “know” ROC curve, as shown in Fig. 5. The first point

Table 1
Overall response proportions from both experiments. Standard errors are in parentheses

Experiment	Old items			New items		
	“remember”	“know”	“new”	“remember”	“know”	“new”
1 ^a	0.62 (.22)	0.20 (.16)	0.19 (.10)	0.18 (.15)	0.20 (.11)	0.62 (.19)
2	0.42 (.03)	0.30 (.02)	0.28 (.02)	0.11 (.02)	0.20 (.02)	0.69 (.03)

^a A know response was defined as a rating of 4–6 on the new-know response scale, a new response as a rating of 1–3.

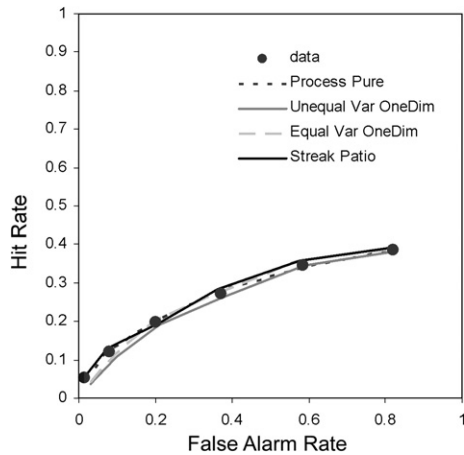


Fig. 5. “Know” ROCs from Experiment 1. The superimposed functions are predicted by the various models, each generated with maximum-likelihood fits to all of the data simultaneously.

on this ROC is based on the highest-confidence “know” judgment (rating 6) to Old (y -axis) and New items (x -axis); the second point reflects “know” judgments for ratings 5 and 6, and so forth. The ROC does not rise to the point (1, 1) because some of the test probes were judged to be remembered; this happened more often for Old than New items. Each model makes a proposal about the dimension that is partitioned by the ratings, as Fig. 3 makes clear. For the process-pure model, criteria along the x -axis produce this partition, whereas for the other contenders the criteria are spaced along a decision axis that combines x and y . Fig. 5 includes curves corresponding to the predictions of the four models fit to these group data (see Appendix A for details), and it is clear that each of them provides a satisfactory description of the results.

A more detailed comparison can be obtained from fitting the models to individual subjects’ data. Because the models have different numbers of free parameters, we used Akaike’s Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC, Schwarz, 1978) to compare the goodness-of-fits (smaller

values of these statistics indicate better fit). The average values of these statistics, given in Table 2, are quite similar for the four models, confirming the visual impression gained from the ROCs. For individual subjects, however, there is a clear pattern: by both measures, the one-dimensional model was supported most often (13 or 17 of 22 subjects) and the process-pure model was supported least (2 or 1 subjects).

The differences in AIC and BIC values are small, so that one may reasonably wonder whether we are overfitting (or at least, overinterpreting) our data. Burnham and Anderson (2002) suggest two ways of comparing the fit across models: (1) using Akaike weights, which indicate the weight of the evidence supporting model i out of a set of N models, and (2) using *evidence ratios*, which compare two particular Akaike weights (one for model i and one for model j) using simple division. The Akaike weight of a model is an estimate of the posterior probability of that model given the data, relative to the posterior probabilities of the other models under consideration. Thus, these weights sum to 1 over the N models, and $1/N$ is the expected weight for N equally well-fitting models; larger Akaike weights indicate better support for the model. Evidence ratios reflect the odds of one model over another, so larger evidence ratios also indicate greater support for model i .

Table 2 provides mean Akaike weights for each model, when it was the winning model. These weights are all between .44 and .55, showing that the best-fitting model for each subject received some support (compared to the .25 value if all models were equal), to about the same degree for all models. Table 2 also displays evidence ratios for the winning model over the next-best model of a different type (i.e., we did not compare the equal- and unequal-variance versions of the one-dimensional model). When the one-dimensional model (in either form) provided the best fit, it resulted in an average evidence ratio of 3.82 over the next-best model of another type; STREAK Patio had a lower average winning evidence ratio, 2.05. The process-pure model fared slightly worse, with an average winning evidence ratio of only 1.60.

Table 2

Mean AIC and BIC values for fits of the models to individual subjects’ data in Experiment 1, the number of subjects best fit by each model, and related AIC-based statistics

Model	Process-pure	STREAK Patio	One-dimensional	
			Unequal-variance	Equal-variance
Mean AIC	348.04	348.52	349.42	350.04
No. of subjects for whom AIC is lowest	2	7	5	8
Mean Akaike weight, when the model was the winner	.50	.55	.44	.53
Evidence ratio over next-best model of different type	1.60	2.05	3.82	
Mean BIC	373.12	373.60	371.72	369.56
No. of subjects for whom BIC is lowest	1	4	2	15

Table 3
Average parameter values for four models fit to individual subjects' data from Experiment 1

Model	Parameters								
	d_x	d_y	s	C_1	C_2	C_3	C_4	C_5	C_r
Process-pure	0.56 (.48)	1.54 (.87)	1.07 (.51)	-0.70 (.45)	0.07 (.47)	0.78 (.78)	1.40 (.93)	4.43 (5.04)	1.23 (.94)
STREAK Patio	1.69 (.98)	0.62 (.36)	0.93 (.33)	-0.77 (.60)	-0.21 (.54)	0.18 (.47)	0.99 (.41)	3.22 (2.04)	0.92 (1.02)
One-dimensional, Unequal variance ^a	1.18 (.56)		0.75 (.24)	-0.65 (.46)	-0.15 (.36)	0.20 (.36)	0.48 (.36)	0.69 (.42)	0.82 (.56)
One-dimensional, Equal-variance	1.33 (.72)		1.0 (fixed)	-0.75 (.47)	-0.11 (.46)	0.34 (.51)	0.65 (.48)	0.88 (.51)	1.02 (.62)

Standard errors are in parentheses.

Note: The process-pure and STREAK models have 9 free parameters, the one-dimensional unequal-variance model has 8, and the one-dimensional equal-variance model has 7. There are 12 independent frequencies in the data matrix (see Appendix A).

^a The one-dimensional model has a single sensitivity parameter, d .

The models can also be compared by considering the estimated parameter values, which are given in Table 3. The models make quite different statements about the data; for example, the one-dimensional model infers a lower sensitivity than the others, which in turn differ in the relative sensitivity they claim for the x and y dimensions. The one-dimensional model also makes more reasonable claims about the placement of C_5 ; particularly large values occurred in the two-dimensional models whenever subjects failed to report “know” responses with highest confidence (rating 6). In the one-dimensional model, the absence of “know-6” responses simply compressed the region between C_5 and C_r onto essentially the same value.

Finally, we examined the parameter estimates for the two versions of the one-dimensional model for consistency. For each subject, we computed the difference in AIC values (unequal variance minus equal variance) and correlated those differences with the slope that was estimated by the unequal-variance version of the model. When the estimated slope is far from 1, a properly discriminating AIC statistic should be more likely to point to the unequal-variance model finding the better fit; estimated slope that is close to 1 should result in the AIC statistics indicating that the equal-variance model provided the better description of the data. Positive AIC differences indicate that the equal-variance model fit better; negative values indicate that the unequal-variance model fit better. Thus, the expected correlation is positive: negative AIC differences should occur when the slope is small (far less than 1), and positive differences should occur when the slope is large (close to 1). Indeed, the observed correlation is +0.47, $p < .05$. We also considered the mean slope estimated with the unequal-variance model as a function of whether the equal-variance model or unequal-variance model was selected: using AIC to select the model, these values were 1.10 and 0.79, Welch's $t(17) = 2.45$, $p < .05$; using BIC, these values were 0.94 and 0.81, $t(20) = 0.76$. When deciding between the competing equal- and unequal-variance

assumptions of the one-dimensional model, the information statistics tend to select the equal-variance version when variances are equal and the unequal-variance version when they are not, a reassuringly predictable result.

Discussion

Similar remember-first experiments (Yonelinas & Jacoby, 1995, Exp. 3; Yonelinas, 2001, Exps. 2A–2C) have obtained similar “old” response rates but much lower “remember” response rates than the present study (about .36 vs .62 for Old items and about .03 vs .18 for New items, averaging over studies). Rotello, Macmillan, Reeder, and Wong (2005) have shown that differences in “remember” response rates are easily obtained with changes in remember-know instructions: “remember” judgments are reduced when subjects are encouraged to be conservative about what justified remembering, as was done in the Yonelinas studies.

Within the framework of our decision-space/model-fitting strategy, we reached two complementary conclusions, one expected and the other surprising. First, subjects adopted a variety of interpretations of the remember-know instructions. The remember-know response dichotomy is entirely subjective, leaving plenty of room for interpretational differences. Second, however, the most popular choice was the one-dimensional model in which qualitative differences in types of memory of the sort usually postulated by employers of the remember-know paradigm play no role. Although surprising, this outcome serves as a caution that subjective reports do not directly reveal underlying processes.

Experiment 2—the trinary paradigm

In Experiment 2, participants made remember-know-new decisions for each word; “remember” and “know” judgments were followed by 3-point rating scales appropriate for the judgment. The trinary task is

agnostic about the sequence of judgments (ON + RK seems as plausible as R + KN), so the resulting data should be particularly diagnostic about subjects' preferred underlying decision bounds.

Method

Participants

Subjects were 51 students from the same population as Experiment 1; none had participated in that study. The data from three participants were dropped because of a failure to use the full range of responses, and the reported data are based on the remaining 48 subjects. All participants were fluent speakers and readers of English.

Stimuli

The stimuli were the same as in Experiment 1.

Procedure

The study phase was identical to that in Experiment 1. Following the study, participants received standard remember-know instructions, as in Experiment 1, plus the following instructions:

For each word, you will be asked to say whether you 'remember' seeing it, whether you 'know' you saw it, or whether you believe it is new. If you can recall any details from your original experience of reading this word earlier in the experiment, you 'remember' the word—in this case, press the RED key. If the word feels familiar even though you do not remember any details, you 'know' you saw it—in this case, press the BLUE key. If you do not think you saw the word earlier, press the YELLOW key for 'new.' When you remember or know a word, you will be asked to rate your confidence for that judgment on a 3-point scale.

The "z" key was marked with a red sticker, the "/" key was marked with a blue sticker, and the space bar was marked with a yellow sticker. Following these instructions, participants were tested on 8 practice words (4 Old; 4 New) to verify their understanding of the task. Each test word was shown centered on the computer screen along with the words "remember," "know," and "new" and a reminder of the button to be pressed for each response. If the participant responded "remember," the test word remained on the screen and they were asked to "Rate the number of details you remember about reading this word." Below this instruction was printed a three-point scale: (1) "few details," (2) "some details," and (3) "lots of details." If the participant responded "know," they were told to "Rate your feeling of knowing about reading this word." Below this instruction was printed a three-point scale: (1) "weak feeling of knowing," (2) "moderate feeling of knowing,"

and (3) "strong feeling of knowing." If the participant responded with "new," no scale was presented. After a delay of 750 ms, the next test word was presented. Subjects were encouraged to ask questions during this practice phase, or at any other point during the experiment.

Finally, the test instructions were repeated, and the 60 studied and 60 new words were presented in random order in the center of the screen. The test procedure was identical to the practice phase.

Results

The overall response proportions are shown in Table 1. Old–new discrimination ($d' = 1.08$) was similar to that observed in Experiment 1 ($d' = 1.18$), but subjects distributed their responses more evenly across the response options (42% "remember" responses to Old items, compared to 62% in the remember-first task). The response proportions were typical of the trinary task: across 47 conditions in the literature, the mean response rates were 44% "remember" and 29% "know" for Old items (compared to 42 and 30% here), and 6% "remember" and 18% "know" for New items (compared to 11 and 20% here). As in Experiment 1, we fit the equal-variance version of the one-dimensional model to these response proportions. That model's prediction that sensitivity should be the same at the "old" ($d' = 1.08$) and "remember" ($d' = 1.02$) criteria was upheld.

The confidence rating data were used to generate remember and know ROCs, which are shown in Fig. 6 for the group data. The left point in each curve reflects the highest confidence responses (i.e., "strong know" or "remember lots of details"); each point to the right relaxes the decision criterion. The curves do not reach (1,1); instead, the vector sum of the upper remember point and upper know point equals (F, H), the point representing the overall false-alarm and hit rates.

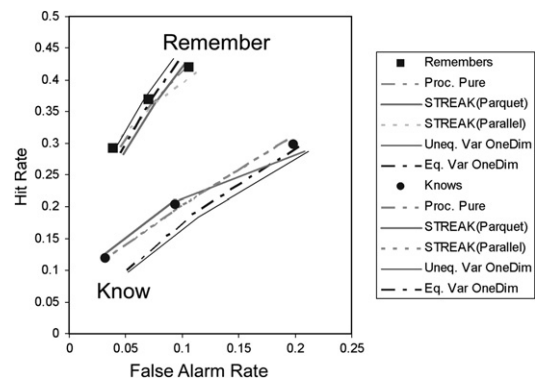


Fig. 6. "Remember" and "know" ROCs from Experiment 2 group data. The superimposed functions are predicted by the various models, each generated with maximum-likelihood fits to all of the data simultaneously.

The five models described earlier were fit to these rating data; each makes different assumptions about how the ratings of remember and know responses are generated, as is shown in Fig. 4 (see Appendix B for details). The resulting theoretical ROCs are superimposed on the observed data, and all of the models do a reasonable job. Although the group data suggest that the one-dimensional model is less successful than the others, it is well known that average data may tell a story that is inconsistent with individuals' performance (e.g., Maddox, 1999), and that turned out to be the case here.

We also fit each subject's individual data with the models and compared their goodness-of-fit using AIC and BIC. The results, shown in Table 4, emphasize the very small differences in the models' ability to fit these data. Nonetheless, data from 73% of the subjects were best fit with the one-dimensional model by the AIC criterion, 94% according to BIC; the other models each described a portion of the remaining subjects.

As in Experiment 1, we computed Akaike weights for each model, when it was the winning model (see Table 4). These weights are all between .34 and .46, showing that the best-fitting model for each subject received some support (compared to the .20 value if all models were equal), to about the same degree for all models. Table 4 also displays evidence ratios for the winning model over the next-best model of a different type. When the one-dimensional model (in either form) provided the best fit, it resulted in an average evidence ratio of 3.42 over the next-best model of another type. Although AIC indicates that STREAK Parallel provided the best account of 7 subjects' data, the evidence ratio of this model over the next-best was a paltry 1.22. The process-pure and STREAK Parquet models each accounted for only a few subjects' data, yielding average winning evidence ratios of 1.58 and 1.08, respectively.

We again compared AIC statistics and parameter values for the equal- and unequal-variance versions of the one-dimensional model. When AIC indicates that the best-fitting model is the equal-variance version, the z ROC slope estimated by the unequal-variance fit should be close to 1.0 (the value assumed in the equal-

variance model). In contrast, when the unequal-variance model is selected by AIC, then the slope estimate should be relatively far from 1.0. The correlation between the difference in AIC values (unequal variance minus equal variance) and the slope estimated by the unequal-variance model was -0.31 ($p < .05$). This correlation is negative because the unequal-variance model typically estimated a slope greater than 1 (see Table 5), whereas in Experiment 1 it typically estimated a slope *less* than 1.

Average model parameters are shown for each model in Table 5. The models again differ in the interpretation given to the data; it is particularly notable that the estimates of s vary widely both between models and (as indicated by the large standard errors) between subjects. As a result, the key data are the number of subjects best described by each model (Table 4), and the conclusion we tentatively reach is that subjects in the trinary paradigm may adopt any of several decision strategies, but most often rely on the one-dimensional model.

Discussion

The trinary task may appear to be a purer remember-know paradigm than others because the three alternatives have equal status, but our data confirm those of Hicks and Marsh (1999), who found that it induced a liberal bias toward the remember response. We also replicate their finding that the task produced the same sensitivity as others. An implication of these results is that as long as performance is summarized by an arguably bias-free measure (e.g., d) no task need be left behind; conversely, summarizing the data in unprocessed terms (e.g., the remember hit rate) can lead to conflicting conclusions when different paradigms are used.

The addition of ratings to the remember response permitted the construction of the remember ROCs plotted in Fig. 6. These functions clearly show that the remember response is graded, a finding observed in the individual data of all but one subject and also reported by Rotello et al. (2005, in press). These data present problems for the dual-process model (Yonelinas, 2001), in which "rememberers" are imagined to arise from

Table 4

Mean AIC and BIC values for fits of the models to individual subjects' data in Experiment 2, the number of subjects best fit by each model, and related AIC-based statistics

Model	Process-pure	STREAK		One-dimensional	
		Parallel	Parquet	Unequal-variance	Equal-variance
Mean AIC	299.28	299.88	299.54	300.10	300.43
No. of subjects for whom AIC is lowest	2	7	4	8	27
Mean Akaike weight, when the model was the winner	0.38	0.36	0.34	0.39	0.46
Evidence ratio over next-best model of different type	1.58	1.22	1.08	3.42	
Mean BIC	324.40	324.97	324.63	322.40	319.94
No. of subjects for whom BIC is lowest	1	1	1	5	40

Table 5
Model parameter values for Experiment 2

Model	Parameters										
	d_x	d_y	s	C_o	C_1^f	C_2^f	C_3^f	C_1^r	C_2^r	C_3^r	
Process-pure	1.19 (2.08)	0.77 (2.14)	2.17 (6.49)	—	1.19 (2.00)	1.92 (2.32)	3.64 (4.17)	1.46 (.65)	1.67 (.61)	2.52 (1.93)	
STREAK Parallel	0.61 (0.75)	0.83 (0.69)	0.58 (0.25)	0.31 (0.41)	0.70 (.97)	1.68 (1.55)	—	0.09 (.70)	-0.12 (.75)	-0.85 (2.12)	
STREAK Parquet	0.89 (0.67)	1.48 (1.16)	1.11 (0.69)	—	0.62 (0.83)	1.41 (1.06)	3.42 (4.35)	0.14 (.72)	-0.10 (.76)	-0.88 (2.37)	
One-dim., Unequal variance ^a	1.29 (0.74)	1.70 (3.43)	1.70 (3.43)	—	0.59 (.61)	0.99 (.65)	1.24 (.61)	1.47 (.57)	1.67 (.55)	2.16 (0.98)	
One-dim., Equal-variance	1.26 (0.68)	—	1.0 (fixed)	—	0.59 (.60)	0.94 (.60)	1.19 (.56)	1.49 (.59)	1.68 (.57)	2.24 (1.28)	

Standard errors are in parentheses.

Note: The process-pure and STREAK models have 9 free parameters, the one-dimensional unequal-variance model has 8, and the one-dimensional equal-variance model has 7. There are 12 independent frequencies in the data matrix (see Appendix B).

^a The one-dimensional model has a single sensitivity parameter, d .

a threshold mechanism. In Experiment 1 we could not distinguish this model from the continuous version of the process-pure model because no ratings were applied to the “remember” response, but in this experiment the threshold variant clearly fails.

These conclusions are based on the group data; what can be said about individuals? The key is the distribution of best-fitting models: Although the differences in AIC and BIC values were small, the one-dimensional model provided the best description for 73 (AIC) to 93 (BIC) percent of our subjects, and the various models of remember-know judgments appear to be empirically distinguishable. Each of the other models had adherents, however, and just as in the remember-first task the average data reflect a mixture of strategies. As always, drawing detailed conclusions from group data is risky.

The trinary rating task contains a potential ambiguity: when a “remember” response is given with high confidence, is the subject implicitly contrasting this judgment with the alternative “know” or the alternative “new”? Our models take different stands on this issue. In the process-pure model and one-dimensional models, higher remember ratings reflect increasing confidence that the item was not New. In all models, higher remember ratings correspond to greater distance from the “know” region, but in the one-dimensional and STREAK Parallel models these ratings are algebraically opposite to increasing know ratings whereas in the process-pure and STREAK Parquet models such ratings are independent of the level of know rating. To the extent our data support the one-dimensional model, therefore, higher remember ratings are contrasted with both “known” and “new” items. Our conclusion that the one-dimensional model is commonly used implies that there is usually no conflict between the two contrasts.

General discussion

Our starting point in this paper was Tulving (1985) original proposal, illustrated in Fig. 1, that remember and know responses draw on two aspects of memory strength. Using SDT techniques, we expanded this framework to include explicit distributions of strength, and decision bounds to convert observed strength values into responses. One set of such bounds, our two-dimensional strength models, closely resembled those implied by Tulving. A second set, the process-pure model, assigned responses to values of one or the other of the strength values. A third set, the one-dimensional models, reduced the two strengths into a single, summed value. To distinguish these possibilities we added rating responses to the standard remember-know design; in each of two experiments we employed a specific variant of the standard paradigm. The data

lead to both substantive and methodological conclusions.

Substantive conclusions about remember–know judgments

We have reached three substantive conclusions, all of them in some degree of conflict with the conventional wisdom of the field. First, not all subjects adopted the same decision rules. Second, there was little effect of paradigm. Third (and most important), one-dimensional strength models were the most successful in accounting for the results. We discuss these conclusions in order.

Individual differences

Although the one-dimensional model provided the best fit for the most subjects, other models better described the performance of a substantial minority in these experiments. Whatever led to the adoption of these distinct strategies, it was not under our control and unless it was, implausibly, a fixed personality trait, it must have been under the control of the subject. We have not shown that an individual can change decision rules, but the most parsimonious explanation of the variability in models' success that we observe across individuals is that there is also potential variability over time within an individual.

Limited effect of paradigm

To our surprise, variation in experimental paradigm did not greatly affect the distribution of decision rules for our subjects. In past experiments, alterations in format have sometimes affected the pattern of data; for example, Hicks and Marsh (1999) found a different pattern of data for the trinary paradigm and the old–first paradigm (an old–new decision followed, for items called old, by a remember–know judgment). However, they were able to characterize the change as one of response bias. When SDT models are fit to data, changes in response bias influence only parameter values, not goodness-of-fit, and there is in fact a large response-bias difference between the data of our two experiments (see Table 1). We further expected that the remember–first design in Experiment 1 would push subjects toward a process-pure model, but here our intuitions clearly failed us.

Dominance of the one-dimensional model

The success of the one-dimensional model adds another notch to an already well-ornamented belt. In our studies of response-bias effects (Rotello et al., in press), and of memory for emotional materials (Dougal & Rotello, in press), the one-dimensional model was the most successful tested; Wixted and Stretch (2004) and Dunn (2004) have cited many aspects of remember–know data that are well-described by this model. We

implemented Wixted and Stretch's version of the model, in which the single strength dimension arises from the sum of x - and y -strengths, so the model's dominance does not contradict the idea that two processes underlie remember–know judgments.

A strong implication of the modeling outcomes, however, is that remembers and knows cannot be directly identified with separate processes. Even within a paradigm designed to favor a process-pure strategy (Experiment 1), few subjects adopted it. Variants of the process-pure rule are widely assumed by researchers, in two forms: (1) the dual-process model (Yonelinas, 2001), in which the model is modified to make recollection a threshold process; and (2) a model-free approach in which remembers and knows are taken to directly reflect distinct memory processes (Gardiner & Richardson-Klavehn, 2000). The former approach has been strongly urged by its proponents, but increasing evidence weighs against it (Wixted, in press). The latter interpretation appears self-evident to many researchers, but simplicity does not trump accuracy. In our view, decision processes complicate all experiments in the field of memory, and understanding them is a prerequisite to understanding memory itself. This observation leads to the methodological morals to be drawn from this investigation.

Methodological conclusions about collecting and analyzing remember–know data

Our findings have implications for the value of models, the use of confidence ratings, and the averaging of data across subjects.

Models

Signal detection theory provides an obvious precedent for a shift from reified interpretation of responses to one that is more model-based. It is now commonly known that “yes” responses in a detection task cannot be taken as direct evidence that a “detect” event rather than a “nondetect” event has occurred. Rather, an adjustable response criterion is customarily assumed to affect the categorization decision, and a bias-free sensitivity measure is calculated to characterize performance. The models for the remember–know paradigm that we have considered are a bit more complex, having multiple adjustable decision bounds rather than one, but in other respects the situation is the same: conclusions about underlying processes are based on models, and the sensitivity parameter (or parameters) is most often of primary interest. It is only by application of models to data that we can interpret the remember–know responses at all.

Ratings

Our ability to detect the variety of models being applied came entirely from the use of rating paradigms.

Without ratings or other methods, such as explicit response-bias manipulations (Rotello et al., in press) that increase the number of degrees of freedom, different decision strategies cannot be diagnosed. Proponents of competing theories will almost always be able to account for data to which they can apply a saturated model (that is, one in which there are as many free parameters as degrees of freedom), and some of the inconclusive nature of the debate about remember–know judgments may arise from the (explicit or implicit) application of such models.

Individual-subject analyses

Finally, the common assumption that all subjects in an experiment are doing the same thing is likely to be wrong in the remember–know paradigm. Individual differences are generally disguised by averaging, and when simple models are involved such averaging can be benign (Macmillan & Kaplan, 1985). But in more complex cases averaging can distort the results; recall that in both experiments the best-fitting model for the average data was discrepant (by at least one goodness-of-fit measure) from the model that described the greatest number of subjects. In the typical remember–know study, testable data for individual subjects require no more trials than are usually collected when the data are to be averaged, so there is little cost and much gain from conducting analyses at this level.

Our conclusions about the choice of model by subjects prejudice one final analytic issue: model mimicry. Perhaps the one-dimensional model does a good job of mimicking its competitors, so that in the face of inevitable variability it wins more often than it should. Such an effect could account for the dominance of this model, although it does not appear able to explain the wide distribution of “best” models across subjects. A useful next step, which we plan to pursue, is to explore the mimicry abilities of the present models, and of others that have been developed for the remember–know paradigm. Meanwhile, we strongly caution researchers against the use of remember–know judgments as a measure of recollection or familiarity: the weight of the evidence indicates that, for the vast majority of subjects in a variety of tasks, remember and know responses differ only quantitatively, not qualitatively.

Acknowledgments

This research was supported by a grant from the National Institutes of Health (R01 MH60274) to C.M.R. and N.A.M. We are grateful to Ruthanna Gordon and Mungchen Wong for collecting the data, and to John Dunn, Andy Yonelinas, and an anonymous reader for comments that were consistently helpful and sometimes insightful.

Appendix A. Ratings models for the remember-first paradigm

All models depend on the two-dimensional decision space illustrated in Fig. 3. The bivariate normal distribution due to New items is represented as a circle centered at (0,0) and has a standard deviation of s on both axes. The distribution due to Old items, which is displaced by d_x on the x -axis and d_y on the y -axis, has standard deviations of 1.

In the remember-first task, subjects first make a binary remember/not remember (rem/~rem) judgment. If the decision is ~rem, a rating response is made on a 6-point scale from new to know. The resulting data matrix has the form:

	Remember	~Remember				
		6 = Know	5	4	3	2
Old						
New						

A.1. Process-pure model

In the process-pure model (Fig. 3A), the horizontal line divides rem from ~rem; the orthogonal distance from the mean of the New distribution to this line is C_r . Five criteria divide the ~rem region into new-to-know rating categories; their orthogonal distances from the mean of the New distribution are $C_1, C_2, C_3, C_4,$ and C_5 . Denoting the normal integral by Φ , the response rates to lures are:

$$\begin{aligned}
 P(\text{rem}|\text{New}) &= \Phi\left(-\frac{C_r}{s}\right) \\
 P(6|\text{New}) &= \Phi\left(\frac{C_r}{s}\right)\Phi\left(-\frac{C_5}{s}\right) \\
 P(i|\text{New}) &= \Phi\left(\frac{C_r}{s}\right)\left[\Phi\left(\frac{C_i}{s}\right) - \Phi\left(\frac{C_{i-1}}{s}\right)\right], \quad i = 2 \dots 5 \\
 P(1|\text{New}) &= \Phi\left(\frac{C_r}{s}\right)\Phi\left(\frac{C_1}{s}\right)
 \end{aligned} \tag{A.1}$$

For targets:

$$\begin{aligned}
 P(\text{rem}|\text{Old}) &= \Phi(d_y - C_r) \\
 P(6|\text{Old}) &= \Phi(C_r - d_y)\Phi(d_x - C_5) \\
 P(i|\text{Old}) &= \Phi(C_r - d_y)[\Phi(C_i - d_x) - \Phi(C_{i-1} - d_x)], \quad i = 2 \dots 5 \\
 P(1|\text{Old}) &= \Phi(C_r - d_y)\Phi(C_1 - d_x)
 \end{aligned} \tag{A.2}$$

A.2. One-dimensional model

The model is shown in the two-dimensional space in Fig. 3B, but only the single decision dimension need be considered. On this dimension New items have mean 0 and standard deviation s , Old items mean d and standard deviation 1. Six criteria partition the dimension, the highest (C_r) dividing rem from ~rem, the others (C_1 to C_5) defining new to know rating categories. For lures:

$$\begin{aligned}
 P(\text{rem}|\text{New}) &= \Phi\left(-\frac{C_r}{s}\right) \\
 P(6|\text{New}) &= \Phi\left(\frac{C_r}{s}\right) - \Phi\left(\frac{C_5}{s}\right) \\
 P(i|\text{New}) &= \Phi\left(\frac{C_i}{s}\right) - \Phi\left(\frac{C_{i-1}}{s}\right), \quad i = 2 \dots 5 \\
 P(1|\text{New}) &= \Phi\left(\frac{C_1}{s}\right)
 \end{aligned} \tag{A.3}$$

For targets:

$$\begin{aligned}
 P(\text{rem}|\text{Old}) &= \Phi(d - C_r) \\
 P(6|\text{Old}) &= \Phi(C_r - d) - \Phi(C_5 - d) \\
 P(i|\text{Old}) &= \Phi(C_i - d) - \Phi(C_{i-1} - d), \quad i = 2 \dots 5 \\
 P(1|\text{Old}) &= \Phi(C_1 - d)
 \end{aligned} \tag{A.4}$$

A.3. STREAK Patio

In the STREAK Patio model (Fig. 3C), the partition into remember, know, and new responses is just as in the original STREAK model for the old-first task (Rotello et al., 2004, Fig. 4). For the remember-first task, a “remember” response is made for points in the upper left corner. If an item is not remembered, then a rating from “6” (sure “know”) to “1” (sure “new”) is made. The 5 decision bounds dividing these ratings are parallel: the middle (old–new) bound is C_3 from the origin, C_4 and C_5 are higher in the space and extend only to the remember bound, whereas C_1 and C_2 are lower in the space and extend infinitely.

A.3.1. Lures

The remember rate is exactly as in the original STREAK model (Rotello et al., 2004, Eq. B3). The notation is simplified if the axes are rotated to be parallel to the decision bounds and the criteria are measured as orthogonal distances of these bounds from the origin. Then the remember rate is

$$P(\text{rem}|\text{New}) = \Phi\left(-\frac{C_3}{s}\right)\Phi\left(-\frac{C_r}{s}\right) \tag{A.5}$$

For the three “know” responses in the ratings:

$$\begin{aligned}
 P(6|\text{New}) &= \Phi\left(\frac{C_r}{s}\right)\Phi\left(-\frac{C_5}{s}\right). \\
 P(i|\text{New}) &= \Phi\left(\frac{C_r}{s}\right)\left[\Phi\left(\frac{C_i}{s}\right) - \Phi\left(\frac{C_{i-1}}{s}\right)\right], \quad i = 4, 5
 \end{aligned} \tag{A.6}$$

Criteria 1 and 2 operate just like the old–new criterion, so proportions for the lowest 3 response categories are simply written:

$$\begin{aligned}
 P(1|\text{New}) &= \Phi\left(\frac{C_1}{s}\right), \text{ and} \\
 P(i|\text{New}) &= \Phi\left(\frac{C_i}{s}\right) - \Phi\left(\frac{C_{i-1}}{s}\right), \quad i = 2, 3.
 \end{aligned} \tag{A.7}$$

A.3.2. Targets

The equations for targets are analogous to those for lures, except that the mean of the Old distribution is (d_x, d_y) instead of $(0, 0)$ and the standard deviation is 1 rather than s . Continuing to refer to rotated axes, we denote the mean of the Old dis-

tribution in terms of the axis that determines whether an item is remembered or not, and the axis that defines the old–new classification: (d_o, d_r) , where d_r is the mean distance along the axis that defines the rem/~rem decision and d_o is the mean distance along the axis that defines old–new. The distances d_r and d_o are related to d_x and d_y by

$$d_o = \frac{2d_x d_y}{\sqrt{d_x^2 + d_y^2}} \text{ and } d_r = \frac{d_y^2 - d_x^2}{\sqrt{d_x^2 + d_y^2}}. \tag{A.8}$$

Note that the Pythagorean distance D between the means of the two distributions is (as it must be) the same for either set of axes:

$$D = \sqrt{d_o^2 + d_r^2} = \sqrt{d_x^2 + d_y^2}. \tag{A.9}$$

Thus:

$$\begin{aligned}
 P(\text{rem}|\text{Old}) &= \Phi(d_o - C_3)\Phi(d_r - C_r) \\
 P(6|\text{Old}) &= \Phi(C_r - d_r)\Phi(d_o - C_5). \\
 P(i|\text{Old}) &= \Phi(C_r - d_r)[\Phi(C_i - d_o) - \Phi(C_{i-1} - d_o)], \quad i = 4, 5 \\
 P(1|\text{Old}) &= \Phi(C_1 - d_o), \text{ and} \\
 P(i|\text{Old}) &= \Phi(C_i - d_o) - \Phi(C_{i-1} - d_o), \quad i = 2, 3.
 \end{aligned} \tag{A.10}$$

Appendix B. Ratings models for the trinary paradigm

All models depend on the same two-dimensional decision space assumed for the remember-first task, but with the different decision bounds shown in Fig. 4. In this task, subjects first make 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. The resulting data matrix has the form:

	New	Know			Remember		
		1	2	3	1	2	3
Old							
New							

B.1. Process pure

The process-pure model for the trinary task is displayed in Fig. 4A. It is very similar to the process-pure model for the remember-first task (Eqs. A.1 and A.2), except that there are more remember criteria (C_1^r to C_3^r in order of decreasing willingness to say “remember”) and fewer know–new criteria (C_1^k to C_3^k in order of decreasing willingness to say “know”). Thus for lures:

$$\begin{aligned}
 P(\text{rem}, 3|\text{New}) &= \Phi\left(-\frac{C_3^r}{s}\right) \\
 P(\text{rem}, i|\text{New}) &= \Phi\left(\frac{C_{i+1}^r}{s}\right) - \Phi\left(\frac{C_i^r}{s}\right), \quad i = 1, 2 \\
 P(\text{know}, 3|\text{New}) &= \Phi\left(\frac{C_1^k}{s}\right)\Phi\left(-\frac{C_3^k}{s}\right) \\
 P(\text{know}, i|\text{New}) &= \Phi\left(\frac{C_1^r}{s}\right)\left[\Phi\left(\frac{C_{i+1}^k}{s}\right) - \Phi\left(\frac{C_i^k}{s}\right)\right], \quad i = 1, 2 \\
 P(\text{new}|\text{New}) &= \Phi\left(\frac{C_1^r}{s}\right)\Phi\left(\frac{C_1^k}{s}\right)
 \end{aligned} \tag{B.1}$$

For targets:

$$\begin{aligned}
 P(\text{rem}, 3|\text{Old}) &= \Phi(d_y - C_3^r) \\
 P(\text{rem}, i|\text{Old}) &= \Phi(C_{i+1}^r - d_y) - \Phi(C_i^r - d_y), \quad i = 1, 2 \\
 P(\text{know}, 3|\text{Old}) &= \Phi(C_1^r - d_y)\Phi(d_x - C_3^k) \\
 P(\text{know}, i|\text{Old}) &= \Phi(C_1^r - d_y)[\Phi(C_{i+1}^k - d_x) - \Phi(C_i^k - d_x)], \quad i = 1, 2 \\
 P(\text{new}|\text{Old}) &= \Phi(C_1^r - d_y)\Phi(C_1^k - d_x)
 \end{aligned}
 \tag{B.2}$$

B.2. One-dimensional model

The one-dimensional model for this task (Fig. 4B) is exactly the same as the one-dimensional model for the remember-first task (Fig. 3B), with the exception that the response categories are now relabeled. Instead of a single remember response and 6 ratings from know to new, there are now 3 remember responses, 3 knows, and a new.

B.3. STREAK Parallel

The STREAK Parallel model (Fig. 4C) is exactly the same as the model for the old/new binary, remember-know rating task described in Rotello et al. (2004). It assumes that the highest confidence “know” response corresponds to the region farthest from the “remember” regions, that is, reflects high confidence that the judgment should not be “remember.”

B.4. STREAK Parquet

Alternatively, the STREAK Parquet model (Fig. 4D) proposes that high confidence “know” responses are those farthest from the “new” region. Know criteria are thus parallel to the old–new boundary and orthogonal to the remember boundaries. The expressions for lures are:

$$\begin{aligned}
 P(\text{rem}, 3|\text{New}) &= \Phi\left(-\frac{C_1^k}{s}\right)\Phi\left(-\frac{C_3^r}{s}\right) \\
 P(\text{rem}, i|\text{New}) &= \Phi\left(-\frac{C_1^k}{s}\right)\left[\Phi\left(\frac{C_{i+1}^r}{s}\right) - \Phi\left(\frac{C_i^r}{s}\right)\right], \quad i = 1, 2 \\
 P(\text{know}, 3|\text{New}) &= \Phi\left(-\frac{C_3^k}{s}\right)\Phi\left(\frac{C_1^r}{s}\right) \\
 P(\text{know}, i|\text{New}) &= \left[\Phi\left(\frac{C_{i+1}^r}{s}\right) - \Phi\left(\frac{C_i^r}{s}\right)\right]\Phi\left(\frac{C_1^r}{s}\right), \quad i = 1, 2 \\
 P(\text{new}|\text{New}) &= \Phi\left(\frac{C_1^k}{s}\right).
 \end{aligned}
 \tag{B.3}$$

The expressions for targets are simplified by partitioning the distance between the means of the distributions into the distance along the remember-know axis, d_r , and the distance along the old–new axis, d_o (see Eq. A.8).

$$\begin{aligned}
 P(\text{rem}, 3|\text{Old}) &= \Phi(d_o - C_1^r)\Phi(d_r - C_3^r) \\
 P(\text{rem}, i|\text{Old}) &= \Phi(d_o - C_1^r)[\Phi(d_r - C_{i+1}^r) - \Phi(d_r - C_i^r)], \quad i = 1, 2 \\
 P(\text{know}, 3|\text{Old}) &= \Phi(d_o - C_3^k)\Phi(d_r - C_1^r) \\
 P(\text{know}, i|\text{Old}) &= [\Phi(d_o - C_{i+1}^k) - \Phi(d_o - C_i^k)]\Phi(d_r - C_1^r), \quad i = 1, 2 \\
 P(\text{new}|\text{Old}) &= \Phi(C_1^k - d_o).
 \end{aligned}
 \tag{B.4}$$

References

- Akaike, H. (1973). Information theory as 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.
- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81–99.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.
- Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*, 231–248.
- Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.
- Diana, R., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: a review of arguments in favor of a dual process account. *Psychological Bulletin & Review*, *13*, 1–21.
- Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–533.
- Dougal, S., & Rotello, C. M. (in press). Remembering emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*.
- Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, *111*, 524–542.
- Gardiner, J. M., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review*, *4*, 474–479.
- 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.
- Hicks, J. L., & Marsh, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, *6*, 117–122.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*, 185–199.
- Macmillan, N. A., & Rotello, C. M. (2006). Deciding about decision models of remember and know judgments: A reply to Murdock (2006). *Psychological Review*, *113*, 657–665.
- Macmillan, N. A., Rotello, C. M., & Verde, M. F. (2005). On the importance of models in interpreting remember-know experiments: Comments on Gardiner et al.'s (2002) meta-analysis. *Memory*, *13*, 607–621.
- Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*, 354–375.

- Murdock, B. B. (2006). Decision-making models of remember/know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review*, *113*, 648–656.
- Parkin, A. J., & Russo, R. (1993). On the origin of functional differences in recollective experience. *Memory*, *1*, 231–237.
- Rajaram, S. (1993). Remembering and knowing: two means of access to the personal past. *Memory and 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.
- Reeder, 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, and 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.
- Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (in press). Interpreting the effects of response bias on remember-know judgments using signal-detection and threshold models. *Memory & Cognition*.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
- Wixted, J. T. (in press). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*.
- Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*, 616–641.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12.
- 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., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: effects of size congruency. *Journal of Memory and Language*, *34*, 622–643.