Memory & Cognition
2005, 33 (2), 261-269

Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures

AARON S. BENJAMIN
University of Illinois at Urbana-Champaign, Champaign, Illinois

One popular technique in the study of human recognition memory involves the elicitation of *remember* and *know* judgments and the attribution of those judgments to qualitative states of memory retrieval. An alternative view, reviewed here, implicates quantitative, but not qualitative, differences in evidence as the basis for those two judgments. That theory makes two clear and testable predictions: that of criterion shifts in "remembering" and that of isodiscriminability across different response sets. In this experiment, the makeup of the distractor set in a recognition test is shown to influence overall recognition criterion and also rates of "remember" responses. The second portion of the article demonstrates how A' is a poor choice of a measure to test the prediction of isodiscriminability. When this measure is corrected (Equation 7) to make it more consistent with current knowledge about the receiver-operating characteristic in recognition memory, it reveals that there is no difference in discriminability between "remember" and all positive responses.

One of the most fundamental ways in which we use our memory to interact with the world is to make decisions of recognition. If we encounter someone at the gym, our interactions with that person depend on our ability not only to evaluate whether we have ever seen that person before, but also to decide whether we have seen him or her before at that very gym. Thus, every act of recognition involves an evaluation of *general* familiarity or novelty, as well as some degree of *specific* information retrieval about past encounters. Whether these processes are separate or unified and the relative order and time course by which they supply information from memory are matters of current theoretical debate.

The approach that I examine critically here is one in which the relative contributions of the multiple sources of information that feed the recognition apparatus are elicited by asking subjects to probe their own phenomenological states during recognition and to reveal their insights to the experimenter. In particular, subjects are asked to evaluate whether an act of recognizing a stimulus is accompanied by a general sense of familiarity, but one lacking specific details about the prior encounter, or by a feeling of retrieving specific elements from memory of the episode (Gardiner & Richardson-Klavehn, 2000; Tulving, 1985). The former state is called *knowing*, and the latter, *remembering* (hereafter, K and R). This approach utilizes careful instructions to subjects about what their phenomenological sense will be like under each of these conditions; subjects' reports of these states are thus thought to accurately index the two qualitatively different types of processes or information, alluded to above, that subjects can rely on when making recognition decisions.

The use of this technique is straightforward and has a particular face validity that makes it appealing to researchers interested in the decision processes underlying recognition. The results of explorations with R/K judgments have indicated that some manipulations, such as depth of processing, influence R, but not K, judgments (Gardiner, 1988) and that others, such as priming, influence K, but not R, judgments (Rajaram, 1993). Other manipulations affect both judgments in similar (Gardiner, Ramponi, & Richardson-Klavehn, 1999) or opposite (Parkin & Russo, 1993) ways. A good review of this work has been provided by Gardiner and Richardson-Klavehn (2000).

There are numerous theories of recognition that postulate dual contributions to the recognition decision, and these self-reports map well onto some terminological distinctions but poorly onto others. It is not the goal here to evaluate whether models of recognition that are portrayed as having a single process are superior or inferior to those that explicitly possess multiple processes; rather, I will address the specific question of whether the data

Thanks to Elliot Hirshman and John Dunn, who provided helpful commentaries on this manuscript, as well as to Sameer Bawa, Benjamin Johnson, and Matt Rambert, all of whom assisted in running subjects for the experiment reported here. Correspondence should be addressed to A. S. Benjamin, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820 (e-mail: asbenjam@uiuc.edu).

gathered from experiments that elicit such judgments substantiate the claim that the judgments themselves derive from two qualitatively different sources of evidence. The stakes for the resolution of this issue extend well beyond the purely theoretical discussions of the R/K procedure in recognition: Many recent studies of memory in which neuropsychological techniques have been used with elderly (e.g., Basdin & Van der Linden, 2003) or pathological (e.g., Verfaillie, Giovanello, & Keane, 2002) populations, as well as brain imaging studies with fMRI (e.g., Henson, Rugg, Shallice, Josephs, & Dolan, 1999) and ERP (e.g., Trott, Friedman, Ritter, & Fabiani, 1997) have taken the valid relationship between these verbal responses and underlying forms of recognition as a starting point.

An alternate view has been well articulated by Donaldson (1996), Dunn (2004), Hirshman and Master (1997), Postma (1999), and others, and will be reviewed only cursorily here. That explanation follows from a simple extension of the Theory of Signal Detection (TSD), which provides a statistical model for human or nonhuman performance on tasks that require a decision maker to classify an ambiguous stimulus as belonging to one category or another. The version of this theory that applies to recognition memory treats each recognition test probe as an ambiguous signal and the classification task as the decision between a prior exposure during a particular study phase or no such exposure. Because nothing is really new in this world, including recognition test stimuli, all such stimuli elicit from the recognizer some degree of mnemonic evidence. The decision-making aspect of the task is to translate that ambiguous amount of evidence into a recognition decision.

Because unstudied stimuli can be highly familiar by virtue of recent or potent extraexperimental encounters and because studied stimuli can elicit low familiarity—perhaps because of a failure to attend or faithfully encode the stimulus during study—the probability distributions of evidence for studied and unstudied stimuli overlap, sometimes considerably, as is shown in Figure 1. If we assume that these distributions are nonzero throughout the entire evidence scale, there is no amount of evidence—or lack thereof—that is perfectly indicative of whether an item has been studied or not. The subject is thought to confront this problem by setting a recognition criterion—that is, a point on the evidence axis beyond which stimuli will be endorsed as recognized and below which they will be rejected. Variants of this model account well for recognition performance (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff, Van Zandt, & McKoon, 1995; Rotello, Macmillan, & Van Tassel, 2000), and extensions of this model into multidimensional space have been quite successful in describing versions of recognition memory tasks that range in the degree to which specific retrieved information is necessary for accurate performance (Banks, 2000; Rotello, Macmillan, & Reeder, 2004).

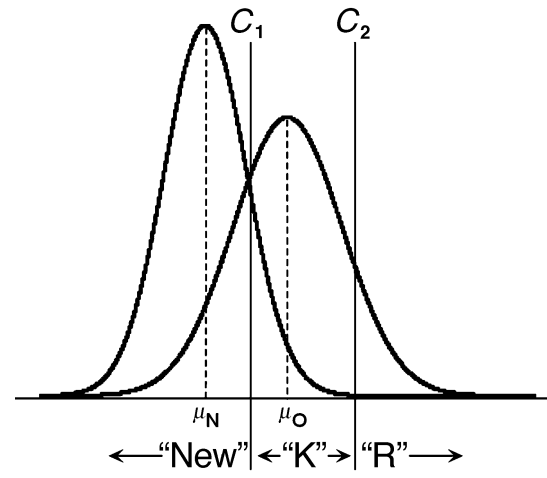

Figure 1. Hypothetical probability distributions of evidence for unstudied (left) and studied (right) items. The means of the distributions are indicated, as are two decision criteria. C_1 is the criterion between responses of *new* and *old*, and C_2 is the criterion between responses of *remember* and *know*.

This model is applied to the data from R/K experiments by assuming that each subject sets two criteria on the evidence axis, one corresponding to a cutoff for R judgments and the other to a cutoff for K judgments. Because K is a normatively poorer basis for recognition by any account, that criterion is assumed to be lower than the one for R. General criterion effects in R/K judgments have been previously observed when different methods for eliciting the judgments have been compared (Eldridge, Sarfatti, & Knowlton, 2002). More critically, Dunn (2004) has recently shown that the TSD models can easily handle those very patterns of R/K data that are touted as evidence for the validity of the assumption relating responses to phenomenological states.

Hirshman and Hentzler (1998; see also Strack & Förster, 1995) tested the core assumption of this model by evaluating how a manipulation of criterion affected R rates. The TSD dual-criterion model implies that such manipulations will shift both criteria, thus affecting R rates in a straightforward manner and K rates in a more complex way that is determined by both by the magnitude and the direction of criterion shift, as well as by the global location of those criteria with respect to the evidence distributions (Hirshman & Master, 1997). In their experiment, they informed subjects prior to the recognition test that previously seen items made up either 30% or 70% of the total test items that they were about to view. (It is important to note that this information was false: The actual rate of studied stimuli on the test was 50% in both conditions.) Because this instruction affects

the estimated a priori rate of targets on the test, the subjects set a more stringent criterion for recognition when they believe the targets to be rarer. This bias approximates the likelihood ratio at the optimal criterion placement,

$$\beta = \frac{f_S(C)}{f_U(C)}, \quad (1)$$

in which $f_D(X)$ represents the height of the function D at criterion location X. When the prior odds of the targets and the distractors are equal, β_{optimal} is unity and C_{optimal} is a function thereof. When the targets are more likely (or are assumed to be more likely, as in Hirshman and Hentzler's experiment), $\beta_{\text{optimal}} < 1$, and C_{optimal} decreases correspondingly. Thus, criteria—both in terms of the evidence axis and the likelihood ratio—shift left under liberal decision standards and right under conservative decision standards.

Hirshman and Hentzler (1998) showed that this manipulation affected the proportion of R responses, as well as the criterion for all positive responses in the experiment. The view that R and K represent distinct operations and consequent phenomenological states cannot readily accommodate this finding. However, their manipulation of response bias was a bit unorthodox. Explicitly telling the subjects about the makeup of the test may have induced the subjects into biased reporting of their phenomenological states in order to appear as though their memories and recollection ability were more accurate than they actually were. This interpretation is supported by an odd datum in their experiment, as has been pointed out by Gardiner, Richardson-Klavehn, and Ramponi (1998): False alarm rates in that experiment were much higher in the liberal criterion condition than are typically seen in experiments on recognition memory, suggesting that the manipulation that they employed may have pushed performance outside the boundaries of what some might consider normal recognition memory conditions. In the present experiment, criterion differences were elicited in a subtler manner and false alarm rates were constrained to a more typical range.

Hirshman and Hentzler (1998) took advantage of the fact that subjects' estimates of the a priori rate of targets on a recognition task influence their criterion placement. Although this effect is usually achieved by actually varying this rate, doing so in a task such as this one is not a viable experimental option. An unequal number of targets makes the comparison of the number of R responses across conditions impossible.

The other major way in which recognition bias can be induced in a recognition task is to manipulate the discriminability of the studied and the unstudied stimuli. Because optimal criterion scales with the distance between the distributions, as revealed by Equation 1, manipulations that increase discriminability via learning also tend to lead to stricter criteria (Hirshman, 1995). However, manipulating discriminability by using a learning variable would be a poor choice for the present investigation. Such manipulations can also affect the rate of Remembering; thus, differences in R responses between conditions can reflect either criterion differences or actual differences in learning and consequent Remembering.

In this article, use was made of the task introduced by Benjamin and Bawa (2004, Experiment 1), in which discriminability is influenced by a manipulation of distractor plausibility, rather than learning. In their task, distractors for recognition tests were drawn either from the same semantic categories as those for the previously studied items or from novel categories. They showed that subjects set a more lenient criterion on a test on which the distractors were semantically unrelated to the targets. In this experiment, that logic was extended to the R/K paradigm, and I examine how manipulations of distractor plausibility affect rates of R responses. Because criterion placement scales with discriminability, the TSD dual-criterion view predicts that R responses should be rarer in the condition in which the distractors are more plausible (and hence, less discriminable), despite the fact that learning and memory for the studied stimuli should be equivalent between conditions.

Although it is difficult to know for sure that those variables that affect yes/no criterion placement will affect R/K criterion placement similarly, several sets of data extant make that assumption plausible. First, Hirshman and Hentzler (1998) revealed concomitant changes in R rates in response to a clear manipulation of overall bias. Second, Stretch and Wixted (1998) showed that confidence criteria fanned out with a manipulation of discriminability, suggesting that confidence criteria were roughly calibrated with likelihood ratios. To the degree that R/K judgments are akin to confidence ratings—the TSD view suggests that they are analogous—those judgments should show a similar effect. In the present case, in which the location of the evidence distributions for the studied items was equivalent between conditions, criteria to the right of the intersection of the distributions would always shift in the same direction in order to maintain a constant likelihood ratio. This was not true to the left of the intersection. However, since criteria tended to be somewhat more conservative than optimal and, thus, lie to the right of the intersection, it was quite likely that the preconditions would be met for both criteria shifting in the same direction.

The basic logic for this prediction is outlined in Figure 2, in which the greater discriminability of studied and unstudied items on tests with less plausible distractors is represented by the fact that the probability distribution for mnemonic evidence yielded by those distractors is further to the left than in the case in which the distractors are more plausible. This figure also shows reasonable criterion placement differences as a function of that manipulation; because those criteria are also further left for the case representing the test with less plau-

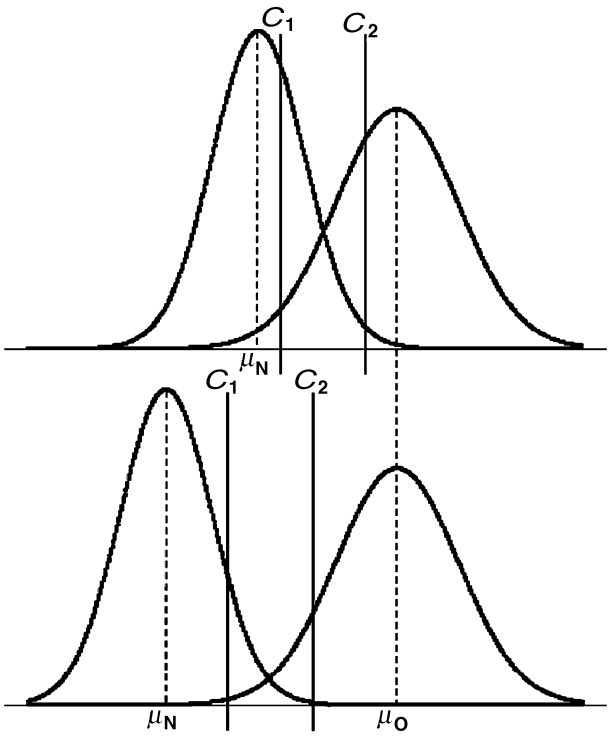

Figure 2. An extension of the theoretical formulation from Figure 1 to an experiment in which distractor plausibility is manipulated. In the top panel, the distractors are similar to studied items, and the distributions overlap to a greater degree than do those in the bottom panel, in which the distractors are relatively dissimilar to studied items. Because decision criteria scale with discriminability, the criteria represented in the top panel are more conservative than those in the bottom panel.

sible distractors, greater rates of *old* and R responses are predicted in that condition.

The TSD model of R/K judgments also makes the prediction that the discriminability of old from new items should be equivalent for remembered and known items. That is, because such judgments are seen as decision phenomena in the TSD model, the placement of the distributions—and thus, discriminability—should not vary with R/K judgments. This prediction has not been confirmed in recent analyses (Gardiner et al., 2002). However, I will demonstrate in a later section of this article how the measure that has been used to assess discriminability in these analyses (A') varies with criterion location and, thus, will argue that it is an unsubstantial test of this prediction.

METHOD

Subjects

Seventy undergraduates at the University of Illinois participated in order to partially fulfill course requirements for an introductory course in psychology. Half of the subjects were quasi-randomly assigned to each condition.

Design

Distractor plausibility was manipulated by drawing distractors from either the same or different semantic categories as those for the studied items. It was manipulated between subjects and was the only independent variable. Hits and false alarms were treated as separate dependent measures in the analysis, as was proportion of R responses within each of those measures.

Materials and Procedure

Each subject studied one of three counterbalanced versions of the study list, each of which contained 10 items from each of 10 semantic categories. Each version of the study list contained items in a unique order that was blocked by category. There were four versions of the recognition test, each of which included a total of 100 items. Half of the subjects received a test with 50 more plausible distractors (drawn from the same semantic categories as those for the study items; see Benjamin & Bawa, 2004), and the other half received a test with 50 less plausible distractors. The other 50 test items had been included during the study phase. The eight versions of the test corresponded to which test condition the subjects were in, which version of the two study lists the subjects had been exposed to, and which half of the items from that study list were included on the test.

The study items were presented in Microsoft Powerpoint on a PC computer at the rate of 4 sec/item, with a 1-sec interstimulus interval. After the study phase, the subjects engaged in a short (~5 min) distractor task in which they completed math problems. After the distractor phase, the subjects were read the instructions for the recognition test and the R/K judgments. The instructions for the judgments were modeled after those described in Gardiner, Ramponi, and Richardson-Klavehn (1998), but without including a *guess* response. The subjects were then given one of the four full pages of test items and were asked to circle previously studied items, as well as the corresponding R/K judgment immediately to the right of that item.

RESULTS AND DISCUSSION

All the results reported here are reliable at the $\alpha = .05$ level, using two-tailed tests, and are summarized in Table 1. The manipulation of distractor plausibility affected false alarm rates [$t(68) = 4.99$], suggesting that the differences between the conditions translated effectively into subjective plausibility. More important, a difference in the hit rates [$t(68) = 3.08$] implied criterion differences between the conditions.

Criterion Shifts in Remembering

The critical test was the comparison of R responses across plausibility conditions. R responses were more common for endorsements of old items in the condition with less plausible distractors [$t(68) = 2.01$]. This result

Table 1
Hit Rates (HR), False Alarm Rates (FAR), and Remember (R) Responses to Old [*p*(R|Old)] and New [*p*(R|New)] Items for Recognition Tests With More Plausible or Less Plausible Distractors

Distractor Condition	HR	FAR	*p*(R|Old)	*p*(R|New)
More plausible	.79	.18	.72	.10
Less plausible	.89	.02	.80	.02

Table 2
Measures of Discrimination (A' and A_R) Across Conditions for Remember (R) and All Positive (R+K) Recognition Responses

	Discrimination Measure and Response Set				
	A'			A_R	
Distractor Plausibility	R		R+K	R	R+K
More plausible	.89	>	.88	.89	.89
Less plausible	.95	<	.97		

Note—A_R could not be estimated for the less plausible distractor condition because of the large number of subjects for whom the false alarm rate was zero. Greater than/less than denote that difference was reliable.

mirrors the finding of Hirshman and Hentzler (1998) and reveals that criterion shifts in recognition change the rate at which subjects claim to Remember studied items. This finding is incompatible with the view that such R judgments reflect the correct report of a phenomenological state of memory that should presumably be uninfluenced by the type of distractors on a recognition test.

Isodiscriminability Across Response Sets

The second critical test of the TSD view is that measures of discrimination should not differ across R responses and all positive recognition responses. Table 2 shows the data from the two conditions reparameterized as A' and reveals a failure of this prediction: For the condition with more plausible distractors, A' is higher for remembered items [$t(34) = 2.21$]. However, for the condition with less plausible distractors, A' is higher for all items (remembered and known) than for remembered items [$t(34) = 5.02$].

Measures of Discrimination in Yes/No Recognition Tasks

If R/K judgments reflect nothing more than difference in criterion—and not something qualitative about the nature of the memory supporting those judgments—measures of memory discriminability that are independent of criterion should be equal for the different judgments (e.g., Donaldson, 1996). It is thus a challenge for this view that A' varies across these difference response sets in the present data. In a thorough and impressive review, Gardiner, Ramponi, and Richardson-Klavehn (2002) showed that this prediction typically fails when discriminability is assessed using A', a measure that has been proposed to be a nonparametric measure of memory discriminability (but see, e.g., Donaldson, 1993; Pastore, Crawley, Berens, & Skelly, 2003).

From a theoretical perspective, the fact that the effects of response set differ across our two experimental conditions is something of a mystery. In the next portion of this article, I will address sources of variance in A' that are artifactual to the measure and will consider whether this result—as well as the findings of Gardiner et al.'s (2002) meta-analysis—are contaminated by problems inherent to A' (see also Dobbins, 2001; Dunn, 2004; Macmillan, Rotello, & Verde, in press).

"Nonparametric" A' and the receiver-operating characteristic in recognition memory. A close examination of A' suggests that the failure of the prediction of isodiscriminability may lie in overly optimistic expectations about the invariance of A' with distributional characteristics. To examine this claim, the nature of the measure must be examined and also what is known about such distributions in recognition memory must be considered. Related problems with A' have been discussed by Macmillan and Creelman (1996), Pastore et al. (2003), and Snodgrass and Corwin (1988).

A' (Grier, 1971) estimates the theoretical quantity A, which corresponds to the area under the function relating hit rates to false alarm rates across the spectrum of all possible decision criteria:

$$A = \int_0^1 R(x)\, dx, \tag{2}$$

in which $R(x)$ is the function relating false alarm to hit rates: HR = R (FAR). This function is called a *receiver-operating characteristic* (ROC), and an example is shown in Figure 3. When discriminability is nil, that function is a straight line along the major diagonal in probability space, and the area under that line is .5. As discriminability increases, the data points fall increasingly above that line, and the area increases correspondingly. Experiments in which multiple criteria are employed, via either confidence ratings or a bias manipulation, allow the estimation of multiple points on that function; the ROC can then be extrapolated more or less accurately depending on the number of data points, their distribution in probability space, and the validity of the theoretical form used to fit those points.

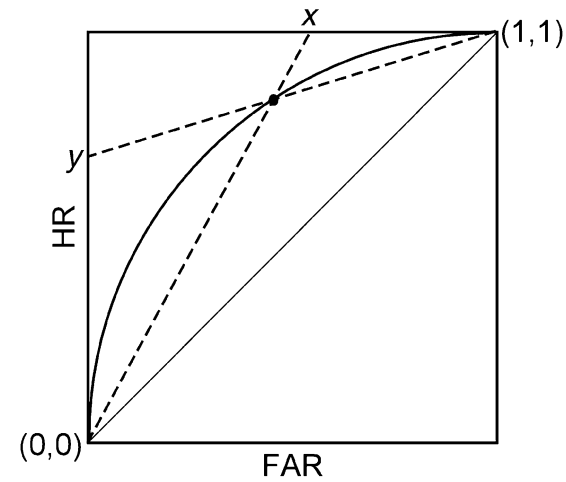

Figure 3. A hypothetical receiver-operating characteristic with lines [(0,0),(x,1)] and [(1,1),(0,y)], indicating the basis for the computation of A'.

A' is one of several measures that can be used to estimate the area under the ROC, and because it is derived from the ROC, rather than from specific assumptions about the form of the probability distributions that underlie the ROC (as do measures such as d' or HR − FAR), it may be more robust than other measures to violations of those assumptions. In addition, it has great appeal because it does not require that the experimenter elicit more than a single hit/false-alarm rate pair. In Figure 3, A' is the average of the area under the line (0,0), (x,1) and the area under the line (0,y), (1,1). This quantity is

$$A' = \frac{1}{4}\left(3 + y - x\right) \tag{3}$$

in terms of the geometry of the probability space. Solving for the coordinates x and y in terms of the experimental statistics yields

$$A' = \frac{1}{4}\left(3 + \frac{\text{HR} - \text{FAR}}{1 - \text{FAR}} - \frac{\text{FAR}}{\text{HR}}\right). \tag{4}$$

These formulations make apparent the fact that A' will vary considerably along the isodiscriminability line, because the area under the two lines will differ as the data point diverges from the negative diagonal (along which HR = 1 − FAR). A' can either under- or overestimate A, depending on where performance lies relative to the inflection point of the ROC.[1] This is a vital issue, because the two postulated criteria in the R/K task differ by definition in their location in ROC space. Thus, criterion variance introduces a source of variability in A' judgments that is independent of discriminability.

Mathematically, it is apparent that A' can vary with constant A and will misestimate that value unless a relatively restricting set of assumptions is met. In that sense, it is not *nonparametric*—different assumptions about the parameters of the underlying evidence distributions affect the accuracy of the estimate. Because the testability of the assumption of invariant discriminability relies on the validity of this measure, it is important to critically examine A' under circumstances that approximate those apparent in normal recognition.

Slope of the *z*ROC in recognition memory. When the ROC is transformed into z-space, underlying Gaussian probability distributions yield linear *z*ROCs. In recognition memory, such linear *z*ROCs are ubiquitous (but see Arndt & Reder, 2002; Rotello et al., 2004; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). Thus, the slope of the *z*ROC and its responsiveness to experimental manipulations has become a critical test of models of recognition (Ratcliff, Sheu, & Gronlund, 1992).

There is ample evidence that the slope of the *z*ROC varies with recognition performance (e.g., Glanzer et al., 1999; Heathcote, 2003; Hirshman & Hostetter, 2000), but it has almost universally been found that the value is less than 1, indicating that the probability distribution for old items has greater variance than the new-item distribution when interpreted in the context of the unequal-variance Gaussian version of TSD (Ratcliff et al., 1992; cf. DeCarlo, 2002). If A' is truly nonparametric, it should be robust to this violation of the equal-variance assumption, but the considerations mentioned above suggest that a closer examination is warranted.

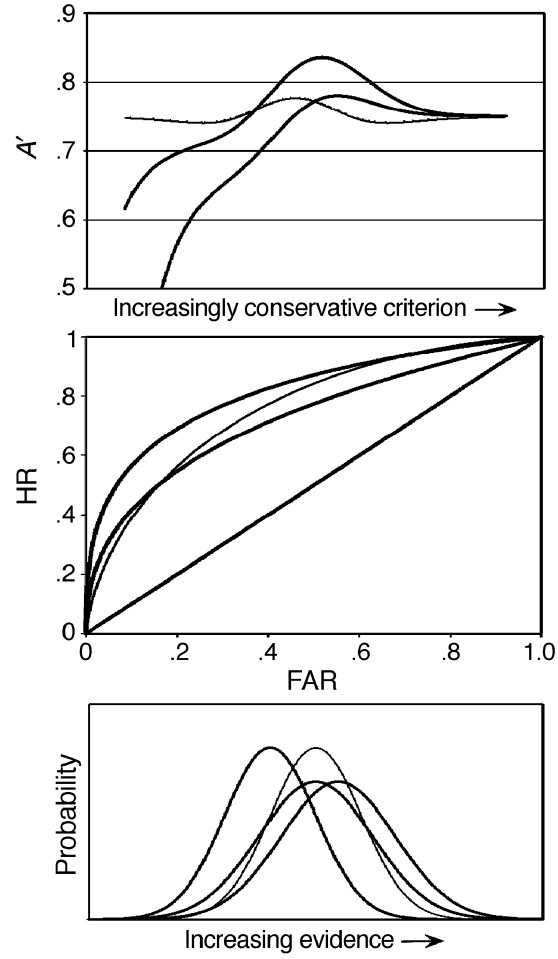

Figure 4. Estimated A' (top panel) and receiver-operating characteristic functions (middle panel) across the decision criterion spectrum for the three old-item probability distributions described in Table 2 and shown in the bottom panel.

Figure 4 depicts estimates of A' (top panel), ROC functions (middle panel), and underlying evidence distributions (bottom panel) for three cases. The thin line represents the case in which the evidence probability distribution for studied items is of the same variance as that for the new items and lies one unit standard deviation to the right. In this case, the ROC is symmetrical, the *z*ROC has a slope of 1 (not shown), and the value of A' varies from .74 to .78 across the spectrum of possible criterion placements. The actual value of A in this case is .76, so it is clear that A' estimates this value faithfully across that range.

The other two distributions, shown in thicker lines, represent distributions that have 33% greater variance.[2] These cases more validly represent what is known about the form of the recognition ROC. The means of these

Table 3
Three Measures of Discrimination (A, A', and A_R) for Noise-Free Data Simulated From Three Probability Distributions of Evidence for Old Items (see Figure 4)

	Discrimination Measure and Accuracy of Measure				
Old-Item Distribution Characteristics	A	A': Range	A': Close Scores*	A_R: Range	A_R: Close Scores
$\mu = 1, \sigma = 1$	.76	.74–.78	1.00	.42–.97	.07
$\mu = 1, \sigma = 1.33$	.73	.08–.78	.07	.71–.80	.99
$\mu = 1.5, \sigma = 1.33$	.82	.61–.84	.25	.80–.87	.99

Note—The new-item probability distribution is fixed at $\mu = 0$ and $\sigma = 1$, and the moments of the old-item distributions are measured in new-item standard deviation units. *This number is the proportion of scores across the criterion spectrum that are within .02 of the actual value of A.

distributions are at 1.0 and 1.5 units to the right of the mean for the new distribution, scaled in terms of the standard deviation of the new distribution. For these cases, the ROC is not symmetrical, and the *z*ROC has a slope equal to the ratio of the standard deviations of the new and the old distributions. The critical aspect of this figure is what happens to A' along the isodiscriminability contour. As can be seen in Table 3, A' varies considerably, and there is only a very small range for which it is accurate (see also Dunn, 2004).

In both of these theoretical cases, A' appears to be a poor measure because of its range, and in neither case do they center about the actual value of A. Also, A' varies nonmonotonically with criterion. The punchline is that *A' can either increase or decrease with criterion, even when discriminability is constant* (although the shapes of the curves suggest that increases are more likely). The strange effects apparent in our experiment—in which R responses reveal apparent greater discriminability than do all positive responses in one condition, but the opposite result obtains in the other—may be attributable to the nonmonotonic form of the A' function shown in the top panel of Figure 4. More important, *A' is a poor estimator of A when the underlying probability distributions are not of equal variance*.

Valid indices of discrimination. Of interest is whether the empirically observed differences in A' reflect differences in accuracy or are artifacts of the measure. The approach to this problem is to use a measure that approximates A more validly than A'. To do this conclusively, however, it is necessary to collect multiple HR/FAR pairs across the criterion spectrum—by using either confidence ratings or manipulations of bias—and to estimate the form of the ROC more accurately. Unfortunately, such a procedure fundamentally changes the task in several ways. First, to estimate a subject's ROC faithfully, the number of test items must be large enough to supply each confidence bin or bias condition with a sufficient number of data points. This necessitates much longer tests or multiple study–test cycles. Second, it is not clear that the means by which subjects make R and K judgments would not be different in a paradigm in which they were asked to make concurrent confidence ratings,[3] thus leading to the concern that results from such a task might not generalize to the bulk of the literature in which R/K judgments have been used. This conjecture is supported by the fact that a recent meta-analysis (Rotello et al., 2004) revealed that the slope of the *z*ROC function in R/K experiments is closer to 1 than the traditional estimates obtained from standard recognition experiments. Nonetheless, this issue will be solved definitively only when better estimates of A are obtained in R/K experiments.

Another approach is to augment empirically obtained data with reasonable assumptions based on the form of the recognition ROC. This technique allows the user to extrapolate a sensible ROC from a single data point without resorting to the somewhat odd mechanics of A'. The cost is that, if the assumptions about the ROC are incorrect, the estimates can be misleading. Of course, this is the same cost that one pays in using any summary measure of discrimination; the goal is thus to try to pick less controversial assumptions. The growing literature in which the form of the ROC in recognition has been investigated, reviewed only briefly above, provides a logical starting point for this process.

Remember that the goal is to estimate the form of $R(x)$, as indicated in Equation 2. The complicated form of the ROC makes that function difficult to estimate (and difficult to integrate, as well). If one assumes that the underlying distributions are Gaussian, A can be estimated from the *z*ROC, either in terms of the parameters of the old-item distribution (μ,σ) or in terms of the y-intercept and slope of the *z*ROC (y_0, m):

$$A_z = \Phi\left(\frac{\mu_s}{\sqrt{1+\sigma_s^2}}\right) = \Phi\left(\frac{y_0}{\sqrt{1+m^2}}\right), \qquad (5)$$

in which Φ indicates the cumulative Gaussian transform.

The large number of demonstrations that the *z*ROC is linear and very few reports to the contrary, and these only under somewhat peculiar experimental conditions (Arndt & Reder, 2002; Yonelinas et al., 1996), indicate that the assumption of Gaussian evidence distributions should be uncontroversial.

Next, one has to know something about the slope of the *z*ROC. This is a trickier domain, and no one choice will be correct for all situations. In addition, it is impossible to evaluate the correctness of a particular choice in the absence of multiple data points from which to estimate a line. However, one can make a choice that is more consonant with knowledge about the ROC than that embodied in the computation of A'. For this experiment, I selected the value of .75, as used in the example laid out

in Figure 4 and justified in note 2. Given this value, only one parameter remains, and it can be estimated from the data. Because the *z*ROC is linear,

$$y_0 = Z(\text{HR}) - mZ(\text{FAR}), \quad (6)$$

in which m is assumed by fiat to be .75 in this case. Combining Equations 5 and 6 yields a model for the estimation of A, assuming a *z*ROC with a slope of .75:

$$A_{\text{R}} = \Phi\left(\frac{Z(\text{HR}) - .75[Z(\text{FAR})]}{1.25}\right), \quad (7)$$

in which the R subscript indicates that this value is based on assumptions relevant to recognition memory.

How does this measure fare for the simulated data presented in Figure 4? It should do well, because these data derive from the very model that underlies the corrections implemented in Equation 7. And it does: Table 3 shows that the range of (arbitrarily) accurate estimates is considerably smaller than A' and that the proportion of estimates within a reasonable range of A is considerably higher. Not surprisingly, this measure fails for Condition 1, in which the distributions are of equal variance.

Table 2 shows what happens when this measure is used to evaluate data from the more-plausible distractor condition in the present experiment. The difference in A' that was apparent between the two response sets does not obtain when A_{R} is used. Thus, when a measure is used that does not have the undesirable qualities of A' under conditions in which the probability distributions are likely unequal, the predicted result of the signal detection view obtains (see Dunn, 2004, Argument 2, for a similar discussion of problems with A').

This result must be qualified, of course. First and foremost, I am attempting to defend a null hypothesis, which is a dicey adventure in the best of circumstances. Second, to do so, a measure has been created that makes several simplifying assumptions about the processes underlying recognition—most critically, that the variance of the signal distribution is ~1.33. This method is unarguably inferior to techniques that estimate the ROC on the basis of multiple data points, and the result will ultimately be qualified by what researchers obtain using those procedures. Nonetheless, it is as justifiable, if not more so, than the use of A', which has dominated the literature in which the questions posed here have been evaluated. Finally, even though this same result obtains with a range of estimates of variance for the signal distribution (~0.4 to ~0.8 in this case), the claim that this slope is closer to 1 (Rotello et al., 2004) challenges this conclusion. Nonetheless, it is clearly worthwhile to evaluate this question of invariant discriminability more deeply, rather than to defer to the findings for which A' has been used.

CONCLUSIONS

The question of what sources of evidence inform recognition decisions and how they do so is central to the development of refined models of recognition memory. The central claim of this article is that the use of subjective judgments as unequivocal indicators of particular underlying processes is an approach that needs careful scrutiny before those judgments are taken as proxies for those processes. A simple extension of TSD can explain many of the apparent dissociations evident in R/K judgments (Hirshman & Hentzler, 1998) and makes two additional predictions. First, R judgments should vary with shifting decision criteria. This prediction was confirmed by Hirshman and Master (1997) and was replicated here in a novel paradigm. Second, measures of memory discriminability that are independent of decision criteria should be equivalent for R judgments and all positive recognition judgments. This prediction has been invalidated when A' has been used as the measure of discriminability (Gardiner et al., 2002), but questions have been raised here about the appropriateness of that measure, and it has been shown that an alternative related measure (A_{R}) does not show that effect.

REFERENCES

Algarabel, S., Gotor, A., & Pitarque, A. (2003). Remember, know, confidence and the mirror effect. *European Journal of Cognitive Psychology*, **15**, 589-605.

Arndt, J., & Reder, L. M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 830-842.

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

Basdin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology*, **17**, 14-24.

Benjamin, A. S., & Bawa, S. (2004). Manipulations of distractor plausibility induce an asymmetric criterion shift in recognition. *Journal of Memory & Language*, **51**, 159-172.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, **109**, 710-721.

Dobbins, I. G. (2001). The systematic discrepancy between A' for overall recognition and remembering: A dual-process account. *Psychonomic Bulletin & Review*, **8**, 587-599.

Dobbins, I. G., Kroll, N. E., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember–know analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1306-1315.

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society*, **31**, 271-274.

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.

Eldridge, L. L., Sarfatti, S., & Knowlton, B. J. (2002). The effect of testing procedures on remember–know judgments. *Psychonomic Bulletin & Review*, **9**, 139-145.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, **16**, 309-313.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness & Cognition*, **7**, 1-26.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1999). Response deadline and subjective awareness in recognition memory. *Consciousness & Cognition*, **8**, 484-496.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory*, **10**, 83-98.

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). New York: Oxford University Press.

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

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 500-513.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75**, 424-429.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1210-1230.

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

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 302-313.

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

Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test recognition memory: The relationship between presentation duration and slope. *Memory & Cognition*, **28**, 161-166.

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

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, **3**, 164-170.

Macmillan, N. A., Rotello, C. M., & Verde, M. F. (in press). On the importance of models in interpreting remember–know experiments: Comments on Gardiner et al.'s (2002) meta-analysis. *Memory*.

Parkin, A. J., & Russo, R. (1993). On the origin of functional differences in recollective experience. *Memory*, **1**, 231-237.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" *A′* and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, **10**, 556-569.

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. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518-535.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General*, **124**, 352-374.

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**, 558-616.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory & Language*, **43**, 67-88.

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

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.

Trott, C. T., Friedman, D., Ritter, W., & Fabiani, M. (1997). Item and source memory: Differential age effects revealed by event-related potentials. *NeuroReport*, **8**, 3373-3378.

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

Verfaillie, M., Giovanello, K. S., & Keane, M. M. (2002). Recognition memory in amnesia: Effects of relaxing response criteria. *Cognitive, Affective, & Behavioral Neuroscience*, **1**, 3-9.

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.

NOTES

1. This point will be on the negative diagonal when the underlying distributions are of equal variance.

2. This value represents a slope of the *z*ROC of .75. That value was selected because it is the average of 16 conditions in analogous experiments discussed in a recent paper by Heathcote (2003, Experiments 1 and 2). Although .80 is often chosen as a representative value (from the original report by Ratcliff et al., 1992), this value of .75 is probably more appropriate for our data set, because the slope tends to decrease with increasing accuracy, at least under some conditions.

3. There are several experiments (e.g., Algarabel, Gotor, & Pitarque, 2003; Dobbins, Kroll, & Liu, 1998; Donaldson, 1996; Rajaram, 1993) in which both confidence and elicit R/K judgments have been evaluated, but only one in which an ROC analysis has been reported (Dobbins et al., 1998), and that report gave between-subjects ROCs, which are likely to be fundamentally different from ones based on individual subjects (cf. Heathcote, 2003).

(Manuscript received April 13, 2004;
revision accepted for publication June 22, 2004.)