

## On the importance of models in interpreting remember–know experiments: Comments on Gardiner et al.’s (2002) meta-analysis

Neil A. Macmillan, Caren M. Rotello, and Michael F. Verde

*University of Massachusetts, Amherst, MA, USA*

From a meta-analysis of recognition experiments using the remember–know–guess paradigm, Gardiner, Ramponi, and Richardson-Klavehn (2002) reported two findings that they viewed as evidence against the one-dimensional model for that paradigm: (1) Memory strength increased when know responses were added to remember responses, decreasing when guess responses were also included. (2) The accuracy of guess responses was correlated with the location of the old–new criterion in the one-dimensional model for the paradigm, implying that guesses were influenced by decision processes. We question both findings. The first result is contradicted by a signal-detection (SDT) analysis, which shows that both know and guess responses reduced estimated memory strength. The discrepancy results from the properties of  $A'$ , the measure of accuracy used by Gardiner et al., which we argue is flawed. The second result follows directly from the one-dimensional model, in which accuracy and response criteria are fixed. The authors’ reasons for rejecting the one-dimensional model are thus not persuasive, but it can nonetheless be rejected because ROC curves implied by the data are inconsistent with ROCs derived from ratings experiments. A two-dimensional SDT model (Rotello, Macmillan, & Reeder, 2004) accounts for both sets of data. The analysis illustrates the importance of models in interpreting remember–know data.

Recognising a test item as one that was presented previously in a memory experiment can be accompanied by either of two subjective experiences: remembering, in which the earlier presentation is recollected, or knowing, a nonspecific feeling of familiarity. In the remember–know paradigm (Tulving, 1985), items that are recognised as old must be placed in one of these two categories. Remembering and knowing are said to reflect separate memorial processes: auto-noetic vs noetic, episodic vs semantic (both distinctions drawn by Tulving), unconscious vs conscious (Jacoby, Yonelinas, & Jennings, 1997), or fluent vs distinctive (Rajaram, 1996). In this paper we focus on a variant of this procedure in which reports of a

third kind of subjective experience, guessing, are also allowed.

In a recent meta-analysis, Gardiner, Ramponi, and Richardson-Klavehn (2002) evaluated findings from 86 remember–know–guess experiments. Four aspects of their analysis concern us here. First, following an analogous remember–know survey by Donaldson (1996), they employed a one-dimensional signal-detection approach that focused on the accuracy measure  $A'$ . Second, using that measure, they came to the conclusion that the inclusion of “know” responses in addition to “remembers” increased memory accuracy. Third, they found that the inclusion of “guess” responses in addition to “remembers” and

---

Correspondence should be addressed to Neil A. Macmillan, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003, USA. Email: nam@psych.umass.edu

We thank John Gardiner and an anonymous reviewer for helpful comments on an earlier draft, and particularly thank Drs Gardiner and Cristina Ramponi for providing details concerning their remember–know–guess database. A paper based on some of this material was presented at the 44th meeting of the Psychonomic Society, Vancouver, Canada, in November 2003. The work was supported by a grant (R01 MH60274) from NIH to CMR and NAM.

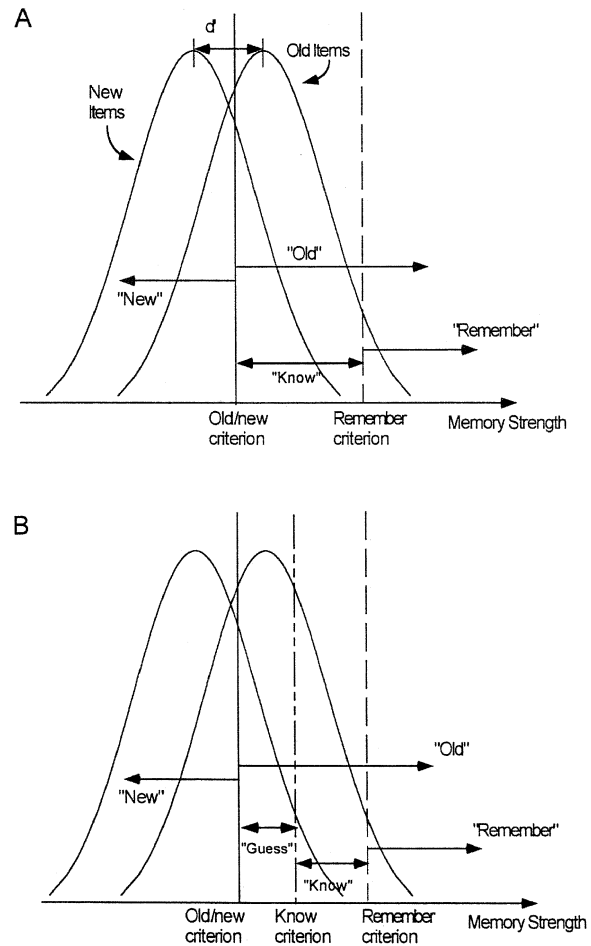
“knows” lowered accuracy to approximately the level observed for “rememberers” alone. Fourth, they found that “guess” rather than “know” accuracy was correlated with the overall response criterion. Their calculations led them to the conclusion that a one-dimensional model, in which remember, know, and guess responses reflect levels of a single evidence continuum, was not adequate to describe the data.

We argue that this analysis makes strong and demonstrably incorrect assumptions about the underlying memory representation. We show that  $A'$  does not have the advantages usually claimed for it, and that the apparent improvement in accuracy obtained by adding “knows” to “rememberers” is an artefact of using this index. The correlational pattern that led to Gardiner et al.’s conclusion about decision processes is shown to derive from a method for calculating sensitivity that is not justified by any current model; the method guarantees the observed pattern, which is therefore not instructive about decision processes.

We thus find nothing in Gardiner et al.’s report that undermines the one-dimensional model. That model can be shown to be inadequate, however, when a wider range of recognition memory results are considered, and we briefly summarise an alternative model that is superior to the one-dimensional model in this respect. Our larger point is that conclusions about the representation and processes underlying remember-know judgements cannot be justified without a rigorous model, rigorously applied.

## THE ONE-DIMENSIONAL MODEL

Somewhat surprisingly, the dominant quantitative model of the remember-know paradigm does not postulate distinct processes. Donaldson (1996), Hirshman and Master (1997), and Inoue and Bellezza (1998) proposed instead that “remember” and “know” responses correspond to different regions on a single strength continuum. The representation draws on signal detection theory (SDT), as shown in Figure 1a. Items have variable strength, the average being greater for Old items than for New. Two criteria are used to partition the strength axis: a high value that is required for a “remember” response and a lower value that is adequate for an “old” response. Events in the region between the two criteria lead to “know” responses.



**Figure 1.** (a) The one-dimensional SDT model for the remember-know procedure. Old and New items differ in average strength. Participants use two criteria to divide the strength axis, responding “remember” for values above the upper criterion, “know” for values between the criteria, and “new” for values below the lower criterion. (b) The same model extended to the remember-know-guess procedure. A third criterion is added to allow for a “guess” region.

The model has three parameters:  $d'$ , the distance between the Old and New means; and the two criteria. The experiment generates four independent data values: the hit and false-alarm rates (“old” responses to Old and New items) and the remember hit and remember false-alarm rates (“remember” responses to Old and New items). One degree of freedom remains to test whether the model provides a good fit to data.

Donaldson (1996) evaluated the model on a set of 80 published experimental conditions. His strategy was to ask whether sensitivity remained constant when it was evaluated at the remember-know criterion (from remember hit and remember false-alarm rates) and the old-new criterion (from

ordinary hit and false-alarm rates). Using the sensitivity measure  $A'$  as an alternative to  $d'$  he found no reliable difference between these two estimates, in support of the model; but Gardiner and Gregg (1997) determined that the difference ( $A'$  was .86 for “old” and .83 for “remember”) was significant when evaluated by a sign test.

Gardiner et al.’s (2002) meta-analysis applied the same approach to experiments using a slightly more complex remember-know-guess design. In the one-dimensional model, “guess” responses are assumed to reflect values of strength below the know criterion but higher than a third, old-new cutoff (see Figure 1b). The model can be tested in the same way as for the conventional remember-know design, but now there are six independent data values (hits and false-alarms corresponding to each of the three criteria) and four parameters.

Gardiner et al.’s results are summarised in Table 1. The left-hand columns give the “remember”, “know”, and “guess” rates for Old and New items. The “remember” rates are viewed by the one-dimensional model as hit and false-alarm rates at the most severe criterion. To obtain hit and false-alarm rates for the know criterion, “know” responses are added to the “remember” responses; these totals are given in the R+K column. To obtain hit and false-alarm rates at the lowest criterion, “guess” responses are added as well; these are given in the R+K+G column.

The last row of the table gives the results of the Gardiner et al. analysis. Adding “know” responses to “remember” increases  $A'$  slightly, and adding “guess” responses as well decreases it to approximately the level found for “remember.”<sup>1</sup> Gardiner et al. argued that these changes are inconsistent with the one-dimensional model. (Gardiner et al. also calculated  $A'$  from the “know” and “guess” responses alone. Because these values cannot be interpreted as accuracy estimates we refer to them as *sham A'*; we discuss this index later in the paper.)

<sup>1</sup>The values for Lure response rates in Table 1 differ slightly from those reported by Gardiner et al., because of an editing error. The response proportions were correctly reported for each study, but the means and  $A'$  values were not updated as preliminary data in Experiments 10 and 11 were finalized. Thus, the values we report for  $A'$  also differ slightly from Gardiner et al.’s numbers. The conclusions were unchanged by these corrections, with the one exception that the mean difference in  $A'$  between remember and remember + know + guess responding, which was a (one-tailed) significant .012 in the original report, is reduced to an insignificant .007.

**TABLE 1**  
Remember, Know, Guess, and combined response rates for targets and lures

	<i>R</i>	<i>K</i>	<i>G</i>	<i>R+K</i>	<i>R+K+G</i>
Targets	.338	.198	.093	.535	.629
Lures	.030	.077	.103	.106	.209
$A'$	.787	[.684]	[.492]	.810	.794

Mean values for the database from Gardiner et al., 2002. R = Remember, K = Know, G = Guess. Brackets indicate sham values that are not true estimates of sensitivity. Values differ slightly from those in Gardiner et al.—see footnote 1.

### **$A'$ VERSUS $d'$ IN THE ONE-DIMENSIONAL MODEL**

An important part of Gardiner et al.’s (2002) survey (and that of Donaldson before them) is the choice of  $A'$  as a measure of accuracy. Figure 1 displays normal distributions, consistent with a common assumption in applying SDT, and according to this model the appropriate measure of sensitivity is  $d'$ . Does it matter which of these variables is used?

To find out, we calculated  $d'$  from the standard formula

$$d' = z(\text{hit rate}) - z(\text{false-alarm rate}) \quad (1)$$

for each of the three criteria for each entry in Gardiner et al.’s database. The formula is indeterminate for arguments of 0 or 1, and we deleted six cases with values of 0 from the analysis. (Alternative treatments of those cases did not change our conclusions.) The results are shown in Table 2, which includes  $A'$  values from Table 1 for comparison. It can be seen that moving from the remember criterion to the know criterion *decreases*  $d'$  [ $t(79) = 2.09, p = .0397, SE_{diff} = 0.245$ ], even though the opposite result occurred for  $A'$ . Both changes are small but statistically reliable. The same effect arose in Donaldson’s earlier analysis: In the studies he surveyed,  $d'$  was 1.71 for the old criterion and 1.80 for the remember criterion,

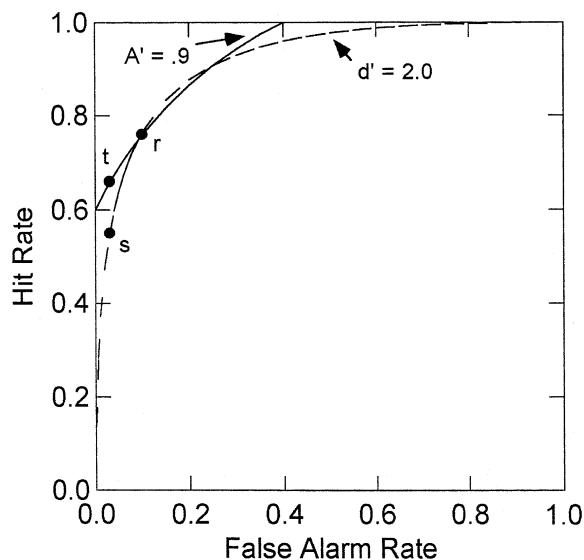
**TABLE 2**  
Estimates of sensitivity in the R-K-G database

<i>Statistic</i>	<i>R</i>	<i>R+K</i>	<i>R+K+G</i>
$A'$	0.787	0.810	0.794
$d'$	1.438	1.381	1.236
$A_z$	0.845	0.839	0.812

R = remember, K = know, G = guess.

reversing the order obtained with  $A'$ . Why should two statistics that each purport to measure accuracy lead to conflicting conclusions? Is it possible to choose between them in a principled way?

Researchers who use  $A'$  often defend it as being “nonparametric”, that is, making no assumptions. We return to this question later, but for now it is sufficient to note that  $A'$ ,  $d'$ , and every other potential sensitivity measure make a prediction about how hit and false-alarm rates will change if response bias varies and sensitivity does not. That is, every index has an implied ROC curve. Figure 2 illustrates this point. Suppose a participant in a remember-know experiment obtains “old” hit and false alarm rates corresponding to point  $r$  on the curve, where  $A' = .9$  and  $d' = 2$ . Then if the normal version of the one-dimensional model is correct, the “remember” hit and false-alarm rates will occur at a point lower on a curve of constant  $d'$ , say point  $s$ . If, on the other hand, accuracy as measured by  $A'$  remains fixed, the “remember” hit and false-alarm rates will fall lower than  $r$  on a curve of constant  $A'$ , say point  $t$ . The two statistics  $A'$  and  $d'$  are equally assumption laden—they assume a particular pattern of hit and false-alarm rates when accuracy is constant—and a choice cannot be made between them on grounds of economy.



**Figure 2.** ROC curves implied by different measures of sensitivity. The  $d'$  ROC is the set of possible (false-alarm, hit) points that can occur if  $d' = 2$ , and the  $A'$  ROC is the set of possible (false-alarm, hit) points that can occur if  $A' = .9$ . The point  $r$  lies on both curves. If accuracy is constant in terms of  $d'$  a participant with a more conservative criterion would operate at point  $s$ , whereas if accuracy is constant in terms of  $A'$  such a participant would operate at point  $t$ .

## THE GARDINER ET AL. DATA IN ROC SPACE

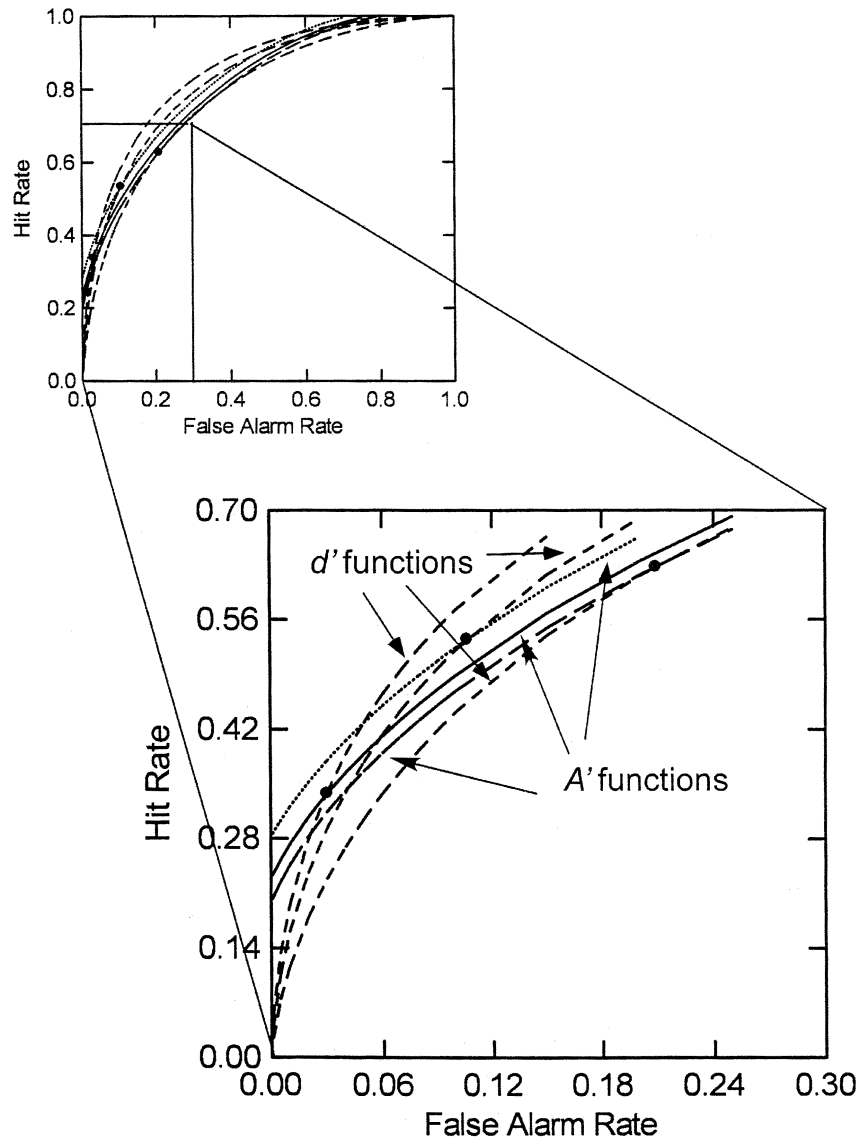
We can now see how Gardiner et al., and Donaldson before them, obtained their contradictory results. In Figure 3, three points are plotted in ROC space, one for the remember hit and false-alarm rates, one for the remember plus know proportions, and the third for the remember plus know plus guess proportions. Three  $d'$  ROCs are drawn, one connecting points with the same  $d'$  as the remember point, one for the remember plus know point, and one for the remember plus know plus guess point. Three  $A'$  ROCs are also drawn, in the same way.<sup>2</sup>

For remember and know responses, the pattern in Figure 3 is similar to that found by Donaldson (1996) and by Gardiner and Gregg (1997) for the original remember-know database. Recall that they too found  $A'$  to either stay constant or be higher for “old” responses, whereas  $d'$  was higher for “remember” responses. The figure shows that there is no contradiction in these comparisons. The  $A'$  and  $d'$  curves are most different near the edges of the space, and data from remember-know experiments invariably contain at least one point in this region, because the “remember” responses can only be a fraction of the “old” responses and “remember” false-alarm rates are always low. The parameter values that lead to the problem we have been discussing are not idiosyncratic or unlucky, but are inherent in the remember-know design.<sup>3</sup>

What about the guess responses? In terms of  $A'$ , accuracy declined slightly when “guess” responses were included, to a level (.794) similar to remember  $A'$  (.787). In  $d'$  terms the guess point lies on a curve of much lower sensitivity (1.236) than the other points (1.438 and 1.381).

<sup>2</sup>The values of  $d'$  (1.47, 1.34, and 1.14) and  $A'$  (.807, .820, and .799) represented by the curves are slightly different from those in Tables 1 and 2. The curves are for values corresponding to average hit and false-alarm rates, whereas the tables give averages of the values of  $d'$  and  $A'$  that are computed from individual studies.

<sup>3</sup>Dobbins (2001) also noted that  $A'$  tended to be greater for remember than for remember plus know responses, and characterized this result as the effect of the rigid form of the implied ROC for  $A'$ .



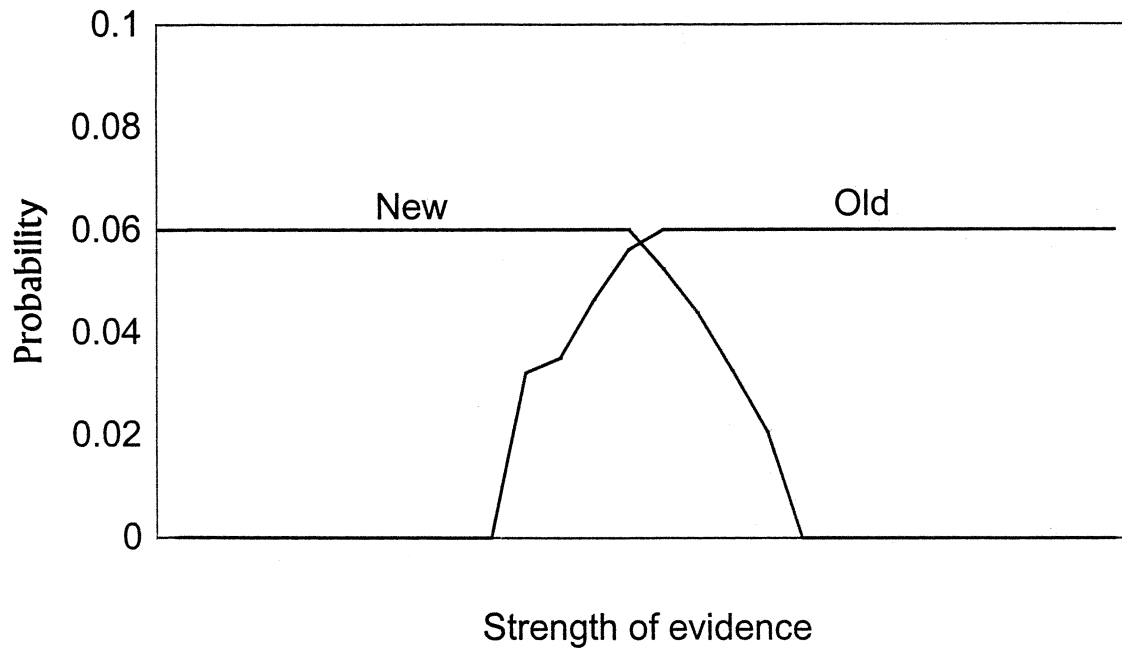
**Figure 3.** Average data from the Gardiner et al. (2002) meta-analysis plotted in ROC space. The leftmost point is obtained from “remember” responses, the middle point from the sum of “remember” and “know” responses, and the rightmost point from the sum of “remember”, “know”, and “guess” responses. The curves implied by  $A'$  show that  $A'$  is largest at the “know” criterion. The curves implied by  $d'$  show that  $d'$  is largest at the “remember” criterion and smallest at the “guess” criterion.

### THE TROUBLE WITH $A'$

Looking at remember-know data alone does not help us in deciding whether the assumptions underlying  $A'$  or  $d'$  are more appropriate, and thus whether adding know responses increases or decreases memory accuracy. Advocates of  $A'$  cite two advantages for it: That it is nonparametric, and that it displays less statistical bias than  $d'$ . Neither of these claims is true.

### Assumptions underlying $A'$

We have just seen that  $A'$  has an implied ROC, and is thus consistent with a limited set of underlying distributions. What kind of distributions are implied by  $A'$ ? Macmillan and Creelman (1996) showed that the form of these distributions changes with accuracy level. At low levels, the implied ROCs are similar to those of an SDT model with logistic distributions, which are similar to Gaus-



**Figure 4.** Underlying distributions consistent with  $A'$ . Note that low values of strength can only be produced by New items and high values only by Old items. This is characteristic of moderate to high values of  $A'$ .

sians. At higher levels, they are similar to those predicted by threshold models; a representation consistent with  $A' = .9$  is shown in Figure 4. The shape of the underlying distributions cannot be uniquely determined from the implied ROC, but an important aspect of the distributions *can* be inferred. Notice that low-strength values arise only from New items and high-strength values only from Old items, with intermediate values possible for either. This is a characteristic of threshold models, and except for unusual designs (Rotello, Macmillan, & Van Tassel, 2000; Yonelinas, 1997), such models have not been supported in recognition memory research.

We have established, then, that  $A'$  (1) is not, contrary to its publicity, nonparametric; and (2) predicts a threshold-like ROC shape that is not consistent with data in recognition (or in other fields). In addition, (3) it predicts symmetric ROCs. We shall see later in the paper that this too is a serious limitation.

This litany of flaws is not new (for a recent discussion, see Pastore, Crawley, Berens, & Skelly, 2003), but the  $A'$  statistic has nonetheless seduced many researchers. The rationale for using it can be expressed in the following syllogism: Area under the ROC is a nonparametric measure of sensitivity (Green, 1964);  $A'$  estimates the area under the ROC; therefore,  $A'$  is a nonparametric

measure of sensitivity. The fallacy in this reasoning is the second premise. In fact,  $A'$  estimates the area under an ROC of a particular shape, and this shape is not what is observed when full ROCs are collected.

What Green (1964) showed was that the area under the *full* ROC is a nonparametric estimate of sensitivity (in particular, that it equals the proportion correct obtained by an unbiased observer in the corresponding two-alternative forced-choice design), so the most desirable area measure is one based on a multi-point ROC. If there is only a single point, the best one can do is to estimate the area under an ROC that has a known shape and goes through that point. For normal distributions, this measure is called  $A_z$ ; it is a monotonic transformation of  $d'$ , and its values for the current data sets are given in Table 2. The pattern of accuracies for  $A_z$  is, of course, the same as for  $d'$ .

### Statistical bias of $A'$ and $d'$

The second purported advantage of  $A'$ , cited by Gardiner et al. (2002), is that its estimates suffer less from statistical bias than those of  $d'$ . This claim is due to Donaldson (1993), who noted that both  $A'$  and  $d'$  extrapolate from single ROC points

by assuming underlying distributions of equal variance. If this assumption is wrong, these estimates will be systematically biased.

To discuss statistical bias, we adopt a standard convention in which the estimate of a parameter  $p$  is written as  $\hat{p}$ . In this notation, the functions Donaldson evaluated are:

$$\begin{aligned} &(\hat{A}' - A_z)/A_z \text{ and} \\ &(\hat{d}' - d_a)/d_a . \end{aligned} \tag{2}$$

That is, he compared observed  $A'$  to true  $A_z$ , the area under the binormal ROC; and observed  $d'$  to true  $d_a$ , a generalisation of  $d'$  to the unequal-variance case. For seven slopes ( $s$ ), four values of  $d_a$ , and nine criteria, he found that  $A'$  was less biased in 86% of the cases.

However,  $A'$  and  $d'$  are on different scales:  $d'$  is a distance measure, and for a fairer comparison we need to transform it into an area measure like  $A'$ . A monotonic transformation of  $d'$  is  $A_z$ , the area under the unit-slope (equal-variance) ROC. Our measures of statistical bias, then, were

$$\frac{\hat{A}' - (A_z | s)}{(A_z | s)}$$

and (3)

$$\frac{(\hat{A}_z | s = 1) - (A_z | s)}{(A_z | s)} .$$

That is, we compared observed  $A'$  and observed  $A_z$ , both assuming unit ROC slope, to true  $A_z$  based on a variety of slopes. We used five slopes, six values of  $A_z$ , and nine criteria. The result, which was very different from Donaldson's, is summarised in Table 3. Ignoring the equal-variance case (which is guaranteed to favour  $A_z$ ),  $A_z$  is superior to  $A'$  in 52% of cases for low  $d'$ , 62% of cases for moderate  $d'$ , and 90% of cases for high  $d'$ . Figure 5 shows the regions of ROC space in which each measure is superior.

We conclude, then, that neither of the putative advantages of  $A'$  (being nonparametric and being

less statistically biased) holds. On the bias issue,  $d'$  (or its area transform) appears superior. On the matter of assumptions, neither measure is assumption free, and the threshold entailments of  $A'$  argue against it. About the remember-know-guess literature, we conclude (consistent with the detection-theoretic statistics in Table 2) that adding "know" responses to "remembers" lowers sensitivity slightly, and adding "guesses" as well lowers it more substantially. Gardiner et al.'s (2002) methods uncovered neither of these results, but nonetheless led them to a conclusion with which we agree: The one-dimensional model, as portrayed in Figure 1b, does not account for the pattern of sensitivities found in remember-know-guess experiments.

### DECISION FACTORS IN KNOWING AND GUESSING

The second major conclusion reached by Gardiner et al. (2002) was that "guess" responses are influenced by decision processes. Their supporting argument has several steps. First, for each subjective category (remember, know, and guess) calculate the proportion of correct responses (hit rate) and incorrect responses (false-alarm rate) using that category. These data are plotted in ROC space (Figure 2 in Gardiner et al.) to reveal characteristics of the categories, and the hit and false-alarm rates are combined to estimate the accuracy associated with each category. Second, determine whether each category might be influenced by decision processes by correlating these sensitivities with the overall (old/new) response criterion. We first consider the estimation of sensitivity by category, then the correlational findings.

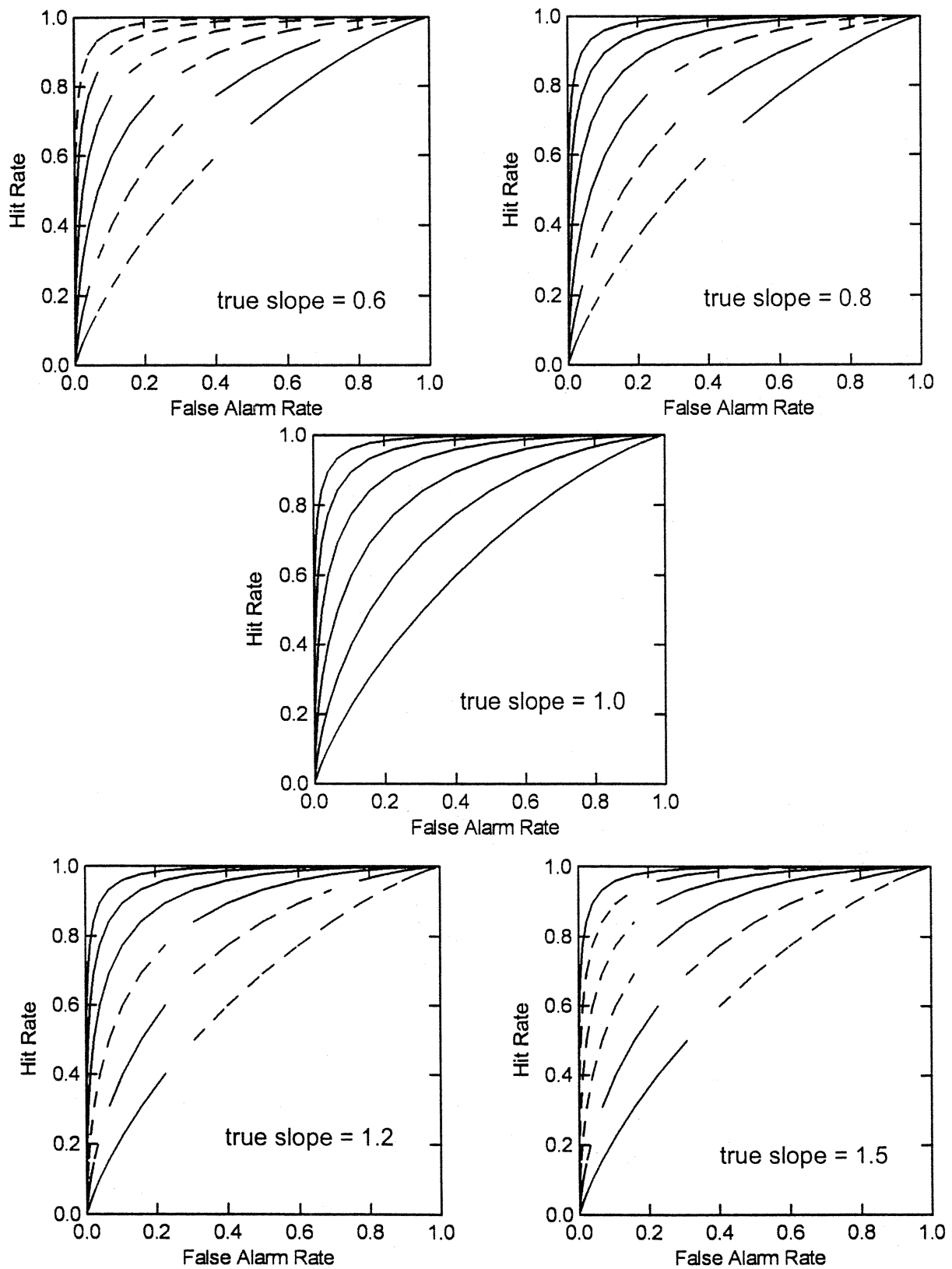
#### The accuracy of subjective categories

In our earlier discussion of sensitivity estimates from the remember-know-guess database,  $A'$  and  $d'$  were always estimated from hit and false-alarm proportions represented as tail areas in Figure 1: Areas above the remember criterion are "remember" responses, those above the know criterion are "remembers" plus "knows", and those above the guess (or old-new) criterion are "remembers" plus "knows" plus "guesses". Equation 1 implies that only such areas can be used to find  $d'$ , and the situation for  $A'$  is analogous.

**TABLE 3**

Percentage of cases for which  $A_z$  is a more accurate estimator than  $A'$

<i>ROC slope</i>	$d' = 0.5 \text{ or } 1.0$	$d' = 1.5 \text{ or } 2.0$	$d' = 2.5 \text{ or } 3.0$
0.6	52.4	50.0	73.8
0.8	52.4	76.2	100
1.0	100	100	100
1.2	52.4	72.6	100
1.5	52.4	50.0	85.7



**Figure 5.** ROC curves shaded according to whether  $A_z$  (solid lines) or  $A'$  (dashed lines) provide a less biased estimate of true  $A_z$ . Both estimators assume unit slope, whereas true slope varies across panels.

However, Gardiner et al. (2002), following Donaldson (1996), also calculate hits and false alarms for “know” and “guess” responses alone. As can also be seen in Figure 1, these proportions correspond to areas *between* criteria. The values of these proportions in themselves are of interest, but when the formulas for  $A'$  and  $d'$  are applied to them the result is not a measure of sensitivity according to the one-dimensional model. It is to make this distinction salient that we have referred to the estimates obtained in this way as *sham*  $A'$  and *sham*  $d'$ ; we also use the term *category-specific accuracy* to describe them.

Gardiner et al.’s findings about category-specific accuracy are quite striking. “Remember” false-alarm rates were always less than .10, hit rates ranged as high as .75, and average  $A'$  was .787. “Know” hit rates were in every case higher than false-alarm rates; average *sham*  $A'$  was .684, and the authors concluded that “know as well as remember responses revealed a consistent power to discriminate between targets and lures” (p. 90). For “guess” responses, there were as many cases of below-chance (hit rate less than false-alarm rate) as above-chance performance. Average *sham*  $A'$  was .492, and Gardiner et al. comment that, as this is less than .5, “the  $A'$  ... for guess responses did not exceed chance” (p. 91).<sup>4</sup>

In the one-dimensional model, however, this pattern is to be expected in the absence of any fluctuations in sensitivity. Consider the hypothetical representation in Figure 6a, which approximately captures the average values in the data base:  $d' = 1.5$ , and the three criteria are equally spaced, with  $c$  (the criterion location relative to the halfway point between the means) equal to 1.5, 0.75, and 0 for “remember”, “know”, and “guess” responses. Using the remember criterion, the hit and false-alarm rates are .227 and .012; using the know criterion, they are .5 and .067; and using the guess criterion they are .773 and .227. The correct value of  $d'$  can be obtained at any of the criteria: For remember,  $d' = z(.227) - z(.012) = 1.5$ ; for know,  $d' = z(.5) - z(.067) = 1.5$ ; and for guess,  $d' = z(.773) - z(.227) = 1.5$ .

What values will category-specific  $d'$  have? For “know” responses alone the hit rate is .5 – .227 = .273, the false-alarm rate .067 – .012 = .055, and

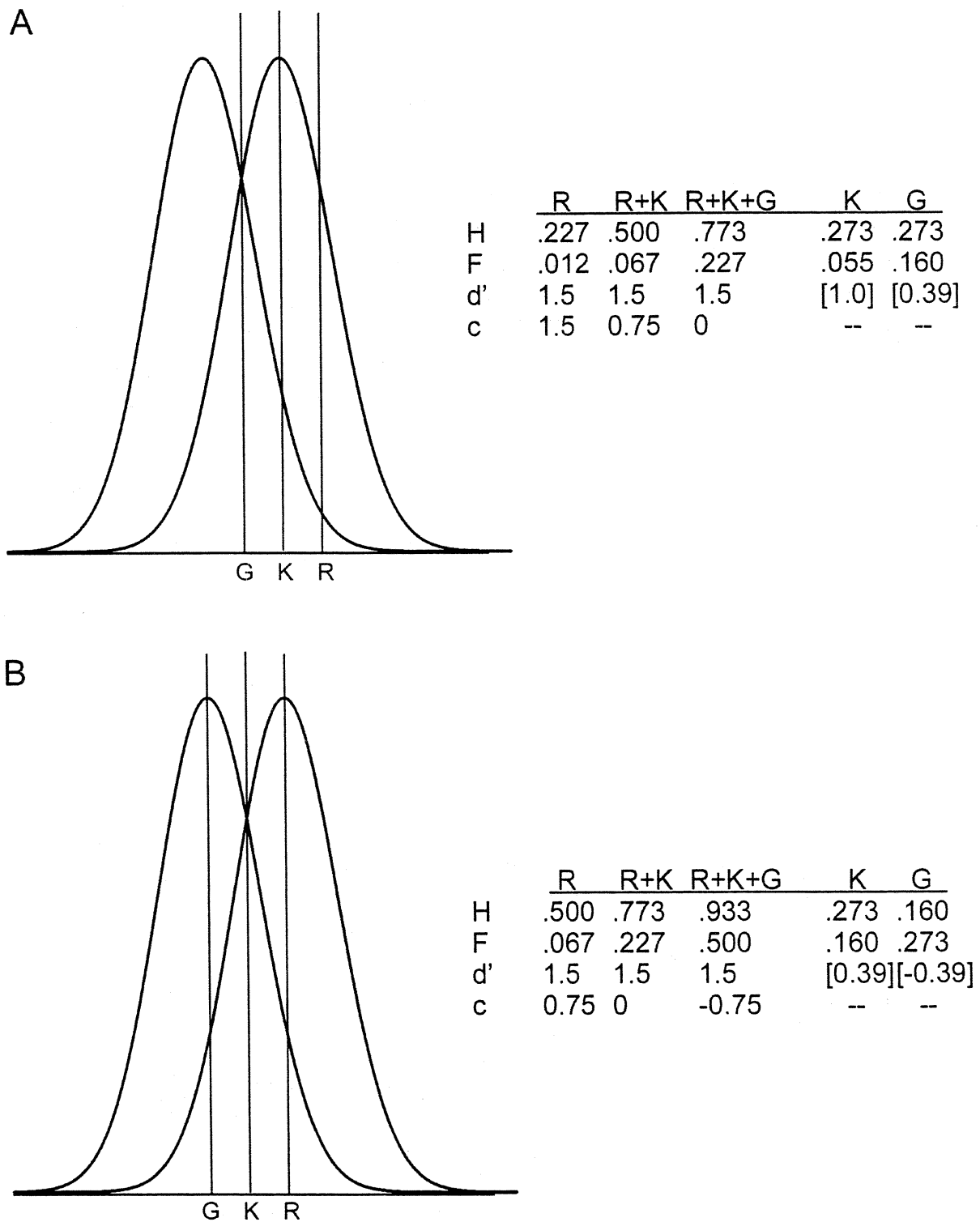
*sham*  $d' = 1.0$ ; a similar procedure for “guess” responses alone results in *sham*  $d' = 0.39$ . But clearly “sensitivity” has not really dropped for “know” and “guess” responses: According to the one-dimensional model, sensitivity is a constant. Application of Equation 1 to category-specific hit and false-alarm rates does not yield an estimate of  $d'$ .

Figure 6b is identical to 6a except that all criteria are lowered by 0.75 units. Repeating the earlier calculation with the remember criterion, the hit and false-alarm rates are .5 and .067; with the know criterion, they are .773 and .227; and using the guess criterion they are .933 and .5. As before, the correct value of  $d'$  can be obtained at any of the criteria (see the figure). The *sham*  $d'$  value for the know category is now  $z(.273) - z(.160) = 0.39$ , and for the guess category  $z(.160) - z(.273) = -0.39$ . This negative value is worthy of comment: In the one-dimensional model, categories below the midpoint on the strength axis will always produce a false-alarm rate that is higher than the hit rate. As in Figure 6a, however, sensitivity has not decreased, much less become negative. This example will be useful in the next section.

There is no objection, of course, to finding the category-specific hit and false-alarm rates, but to use them to estimate accuracy (and point to low values of accuracy as a substantive finding) requires a model. The one-dimensional model is not it, but would other possibilities justify this approach? For example, the dual-process model of Yonelinas (2001) has been applied to remember-know data. Items may be remembered with a high-threshold (false-alarm free) process; failing that, the participant consults a continuous strength dimension that is divided into categories as in the one-dimensional model. Although the model has not been applied to the remember-know-guess design, it is clear that the strength axis would have to be divided into at least three regions (depending on whether “remembers” ever resulted from it rather than the high-threshold process), and the pattern just demonstrated for the one-dimensional model would again arise.

In our own model, STREAK (Rotello, Macmillan, & Reeder, 2004), old vs new judgements are based on a weighted sum of specific and general information whereas remember vs know judgements rely on a weighted difference of the same components. Although the model postulates two dimensions, it resembles the one-dimensional model in dividing an underlying space into regions

<sup>4</sup>Gardiner et al. (2002) provide the formula for  $A'$  that applies when the hit rate  $H$  is greater than or equal to the false-alarm rate  $F$ . When  $H < F$ , the alternative formula presented by Aaronson and Watts (1987) was used:  
 $A' = .5 - (F - H)(1 + F - H) / [4F(1 - H)]$ .



**Figure 6.** Hypothetical criterion locations for the one-dimensional model. (a) The  $d'$  of 1.5 can be estimated by using areas to the right of any of the criteria, but if only regions between criteria are considered, lower sham values of accuracy [in brackets] are found. (b) Same representation but with more lenient criterion locations. Category-specific accuracy is again lower than true accuracy; in the guess region, which lies below the equal-bias point, the category-specific hit rate is less than the false-alarm rate, leading to negative sham  $d'$ . True  $d'$ , which must be estimated from tail areas of the distributions, is always 1.5.

corresponding to each possible response. The two sensitivity parameters (for specific and general information) are again constants, and the various category-specific hit and false-alarm rates cannot be used directly to estimate them. To our knowledge, then, there is no model that assigns theoretical meaning to accuracy estimates based solely on “know” or “guess” responses.

**The correlation between category-specific accuracy and old-new criterion**

Gardiner et al. asked whether category-specific accuracy (as measured by sham  $A'$ ) was correlated with the overall, old-new decision criterion (as measured by  $B''_D$ ).<sup>5</sup> Their Figure 3 shows that the correlation is slightly negative for “remember” responses ( $r = -.18, ns$ ), slightly positive for “knows” ( $r = .14, ns$ ), but substantial for “guesses” ( $r = .62$ ). This result is interesting because Donaldson (1996), in a similar analysis, concluded that “knows” but not “remembers” displayed such a correlation. Gardiner et al.’s findings suggested to them that it was the “guesses” hidden in those “knows” that were to blame in the pure remember-know paradigm studied by Donaldson. For the remember-know-guess design, they concluded that “most, if not all, of the decision processes were captured by the guess responses” (p. 95). They find the correlations between response criterion and accuracy (for the “guess” response in their survey and the “know” response in Donaldson’s) to be worrisome: “Varying response criteria is likely to invalidate the responses as measures of auto-noetic and noetic awareness” (p. 95).

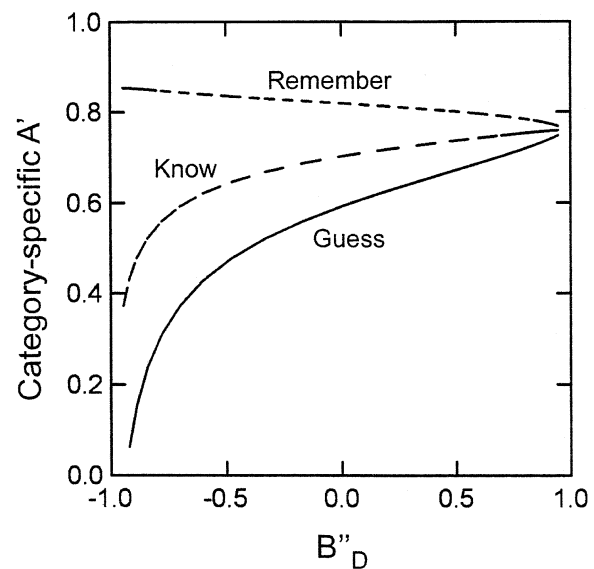
A little more detail about the effect observed by Gardiner et al. will be useful. Values of  $B''_D$  have a theoretical minimum of  $-1$  (most lenient) and a maximum of  $+1$  (most conservative), and the studies in Gardiner et al.’s database span almost the entire range. However, examination of the scatterplot (Gardiner et al., Figure 3) shows that below-chance values of  $A'$  predominate when the criterion is below 0, above-chance values when it is positive. The correlation results from this systematic difference.

It turns out that just such a pattern is predicted by the one-dimensional model, which has fixed

<sup>5</sup> Although often treated as nonparametric, the measure  $B''_D$  has been shown (Macmillan & Creelman, 1996) to estimate the criterion location in a signal-detection model with logistic distributions.

sensitivity and stable criteria. We have already found (Figure 6b) that lenient criteria lead to below-chance values of category-specific sensitivity (negative for  $d'$ , below .5 for  $A'$ ) and conservative values lead to above-chance values (Figure 6a). According to the one-dimensional model, the observed correlation *must* occur.

Figure 7 shows the values of category-specific  $A'$  as a function of  $B''_D$ , assuming that  $d' = 1.5$  and that the three criteria are equally spaced, with  $c$  values 0.75 apart. The curves capture the general trend in the Gardiner et al. scatterplot, and demonstrate that their data are consistent with the one-dimensional model. (The data in Gardiner et al.’s Figure 3 show variability that is lacking here, because a range of parameter values are represented there. However, other selections of sensitivity and bias spacing lead to patterns similar to Figure 7.) Note that remember  $A'$ , a true, tail-area-based measure, is not quite constant even though  $d'$  is. This effect, similar to that found by Gardiner et al., results entirely from the choice of  $A'$  rather than  $d'$  as an accuracy index. The only implication of the correlational analysis for decision processes, then, is that guesses correspond on average to lower levels of strength than remembers or knows; in some studies, the old-new criterion falls below the equal-bias point, in other studies above it.



**Figure 7.** Predicted dependence of category-specific  $A'$  on old-new criterion location according to the one-dimensional model for the remember-know-guess paradigm. Equal variance normal distributions are assumed,  $d'$  is set to 1.5 (the approximate average in the database), and the three criteria are equally spaced, 0.75 units apart.

Gardiner et al. make the more general point that response criteria can be manipulated within studies, and this is certainly true. Does this invalidate the remember-know method, as the authors fear? In our view, it merely invalidates the idea that any of the categories used in these experiments has a threshold character; if instead “remember” and “know” responses are based on one or more kinds of strength (as in the one-dimensional model and in STREAK, but not in the dual-process model), the ability to adjust criteria in response to instructions and other constraints is to be expected. In STREAK, there are two separate criteria, for old-new and for remember-know judgements, and participants can systematically and separately adjust these criteria.

We conclude, then, that the correlational pattern reported by Gardiner et al. is easily explained and sheds little light on memorial processes. However, the differences in accuracy associated with different criterion placements do require explanation, and we have found that the nature of this pattern depends on whether  $A'$  or  $d'$  is used to assess it (Table 2). We have criticised  $A'$  on internal grounds, but there is also a strong positive advantage to the signal-detection approach: SDT models have been very successful in the broader field of recognition memory. We discuss these findings next.

## ITEM-RECOGNITION RATING EXPERIMENTS AND THEIR IMPLICATIONS

### Support for the Gaussian model

The one-dimensional model derives from the more general set of models of recognition memory which assume that judgements are based on memory strength, a variable that combines various kinds of evidence that a test item appeared on a study list (Morrell, Gaitan, & Wixted, 2002). A natural framework for such models is signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 1991), and SDT strength models have been popular ever since they were first proposed early in the cognitive era (Banks, 1970; Egan, 1958; Green & Moses, 1966; Lockhart & Murdock, 1970). The current generation of “global” memory models is more explicit about the type of evidence that contributes to the strength axis, and the premises of SDT are supported by

their infrastructure (especially TODAM: Murdock, 1982).

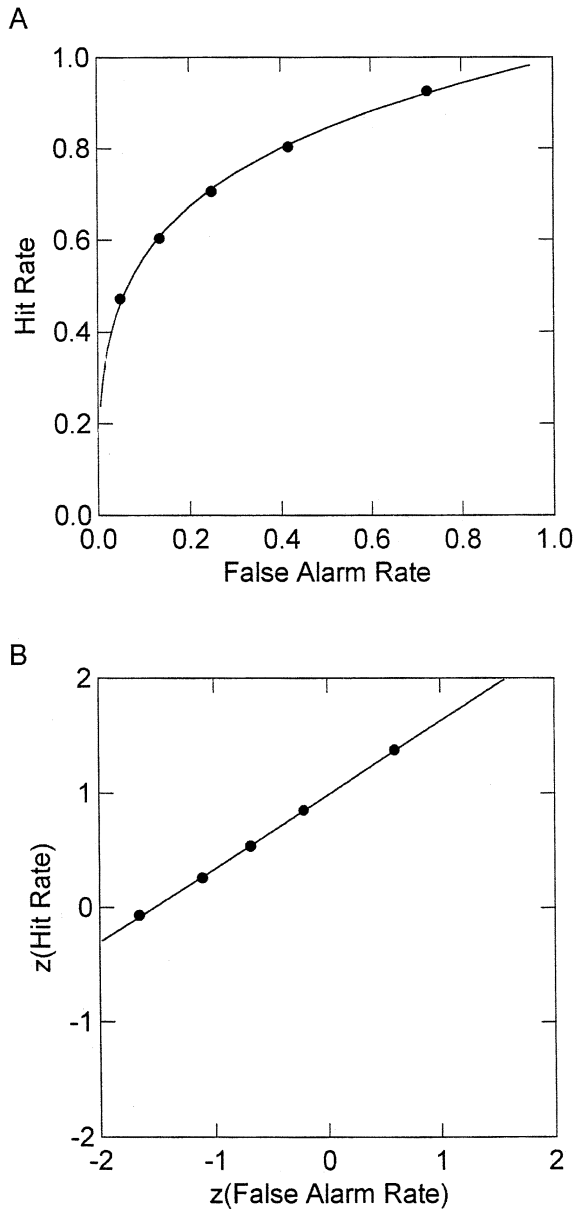
Support for the normal-distribution assumption of these models comes from ROC curves collected by asking participants to provide confidence ratings for their recognition judgements. In this paradigm, one ROC point is obtained by considering only “old” responses at the highest level of confidence, a second by considering the highest two levels together, and so forth. This paradigm has been heavily used in recognition. The one-dimensional model we have been discussing says that shifting from a remember-know to an old-new criterion is like including lower confidence levels, so data from these confidence rating experiments are critical to evaluating the representation proposed by the model.

Except for a few experiments with specialised designs (e.g., associative recognition), the data from rating experiments are uniformly in agreement with the Gaussian model. Figure 8 shows a ROC obtained from such a recognition memory rating experiment (Rotello et al., 2000, Exp. 1), in two forms: Panel (a) plots hits against false-alarms, and panel (b) plots the  $z$ -scores of these proportions. If the underlying distributions are normal, the  $z$ -ROC is a straight line; the data in Figure 8, like most ROCs in the recognition memory literature, are convincingly linear.

### Support for the unequal-variance model

Additional information can be obtained from the slope  $s$  of the line. If this  $s$  equals 1, the standard deviations of the underlying Old and New distributions are the same; if not,  $s$  provides an estimate of their ratio. In the literature (and in Figure 8),  $s$  is consistently less than 1, implying that the standard deviation corresponding to the New items' distribution is smaller than that of the Old items. (See Heathcote, 2003, for recent examples.)

Returning now to our analysis of the remember-know data, recall that our estimate of  $d'$  was greater for the remember point than the know point, and greater for the know than the guess point. Graphically, the statistic  $d'$  is the vertical distance in  $z$ -space between a point and the chance line ( $z_H = z_F$ ), and this distance is greater for points in the left side of the space (more conservative responses). Thus the pattern that appears in the data (Gardiner et al.'s survey as well as Donaldson's) is consistent with a



**Figure 8.** ROC curves from a recognition memory experiment (Rotello et al., 2000, Experiment 1). (a) Probability of correctly recognising an Old item (hit) vs probability of incorrectly identifying a New item as a target (false-alarm). The distinct points were obtained by a rating procedure. (b) The same ROC on  $z$ -coordinates. The curve is well described by a straight line, as is expected if underlying distributions are normal. The slope of the line estimates the ratio of the standard deviations of the underlying lure and target distributions. (Reprinted by permission.)

recognition memory ROC that has a slope less than 1, as expected based on previous item recognition studies not using the remember-know paradigm.

However, the agreement between the data and conventional ROCs disappears when the two

paradigms are compared quantitatively. For each condition in Gardiner et al.'s meta-analysis, we calculated the ROC slope implied by looking at the remember and the know point. The average of these two-point slopes was 0.976. Although this is less than 1, it is much larger than the value of about 0.8 found in the item-recognition literature (Ratcliff, McKoon, & Tindall, 1994). We also calculated the slope implied by including the guess responses, that is, the slope derived from the remember plus know responses and the remember plus know plus guess responses. This slope was 0.642, substantially too *low* to be consistent with conventional ROC data. Each of these results is difficult for the one-dimensional model to accommodate.

**Inconsistency between rating and remember-know data**

What can we conclude, then, about the one-dimensional model for the remember-know and remember-know-guess paradigms?

For the *remember-know-guess* paradigm the data are unambiguously inconsistent with the model: Including “guess” responses with “remembers” and “knows” lowers apparent sensitivity substantially, and allowing unequal variances to resolve the discrepancy requires an unusually small ROC slope. It appears likely that when given a button to use when they feel they are guessing, participants use it at least some of the time for just that purpose.

For the *remember-know* paradigm the pattern of accuracy estimates deviates from the predictions of an equal-variance SDT model only slightly: Statistical significance arises from the large  $N$ s in these meta-analyses. Considering only the remember-know literature, the small deviations are encouraging, and other writers have found this outcome consistent with the one-dimensional model (Dunn, 2004; Wixted & Stretch, in press). Further support for the one-dimensional model has come from reanalyses of experiments in which “remember” and “know” responses are shown to “dissociate”, that is, to behave differently when independent variables are manipulated. Such results have often been interpreted as implying two processes rather than one, and Gardiner et al. seem to take this view. However, Dunn (2004) has shown that the one-dimensional model can easily account for such dissociations if the variables in question influence

more than one parameter, and Rotello et al. (2004) draw a similar conclusion.

The result that is fatal to the model, in our view, is the discrepancy between the equal-variance assumption of the one-dimensional model and the unequal-variance assumption supported by rating data. A normal-distribution representation provides a good description of remember-know experiments, but it is not the same as the representation underlying rating experiments, and both cannot be correct. Gardiner et al. (2002) summarised their analysis by saying that existing results "... do not support a quantitative trace strength model according to which ... [remember, know, and guess] responses merely reflect different response criteria" (p. 83). We have argued that their criticisms of the one-dimensional model are unpersuasive, but we do agree with this conclusion on the quite different grounds that the model is inconsistent with the results of a large class of rating experiments.

### SOME POSITIVE CONCLUSIONS

To our knowledge, the only successful account of both remember-know and item-recognition ratings data is provided by STREAK (Rotello et al., 2004), a two-dimensional SDT model. In developing STREAK we were guided not only by this important conflict in the data, but also by a generally acknowledged shortcoming of the one-dimensional model: The idea that "remember" and "know" responses really tap different levels of confidence on a single strength dimension is at odds with experimental intuition and non-quantitative theorising, and indeed with the rationale for the remember-know distinction itself. STREAK postulates two continuous sources of memorial information, general and specific, and proposes that old-new decisions are based on a weighted sum of these variables, whereas remember-know judgements are based on a weighted difference. Both criteria can be adjusted continuously, consistent with experiments in which response bias effects occur.

Other accounts of remember-know data will no doubt be proposed. Our more general point is that any satisfying account must be consistent with the complete range of recognition memory findings, and also with the subjective experiences of those doing the remembering.

A final conclusion concerns the role of models in interpreting remember-know experiments.

Attempts to understand such data without a model, which have dominated the literature, have led to influential conclusions about the processes underlying these subjective categories. Many of these conclusions do not withstand rigorous analysis. When a quantitative model is being evaluated, qualitative tests are in addition unnecessary. The present report illustrates the power and surprise value of a quantitative approach.

Manuscript received 13 March 2003

Manuscript accepted 22 April 2004

PrEview proof published online 1 October 2004

### REFERENCES

- Aaronson, D., & Watts, B. (1987). Extensions of Grier's computational formulas for  $A'$  and  $B''$  to below-chance performance. *Psychological Bulletin*, *102*, 439-442.
- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81-99.
- Dobbins, I. G. (2001). The systematic discrepancy between  $A'$  for overall recognition and remembering: A dual-process account. *Psychonomic Bulletin & Review*, *8*, 587-599.
- 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.
- Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Technical Note AFCRC-TN-58-51). Bloomington, IN: Indiana University Hearing and Communication Laboratory.
- 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., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory*, *10*, 83-98.
- Green, D. M. (1964). General prediction relating yes-no and forced-choice results. *Journal of the Acoustical Society of America*, *36*, 1042 (Abstract).
- Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, *66*, 228-234.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Heathcote, A. (2003). Item recognition memory and the ROC. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1210-1230.
- Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition*, *25*, 345-351.

- Inoue, C., & Bellezza, F. S. (1998). The detection model of recognition using know and remember judgments. *Memory & Cognition*, *26*, 299–308.
- Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness: Carnegie Mellon Symposia on Cognition* (pp. 13–47). Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- 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. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
- 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.
- Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1095–1110.
- Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.
- Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. (2003). Nonparametric  $A'$  and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*, 556–569.
- Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 365–377.
- Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763–785.
- Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, *43*, 67–88.
- Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional detection theory model. *Psychological Review*, *111*, 588–616.
- Tulving, E. (1985). Memory and consciousness. *Canadian Journal of Psychology*, *26*, 1–12.
- Wixted, J. T., & Stretch, V. (in press). In defense of the signal-detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*.
- Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*, 747–763.
- Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*, 361–379.