Psychonomic Bulletin & Review
2004, 11 (4), 616-641

In defense of the signal detection interpretation of remember/know judgments

JOHN T. WIXTED
University of California, San Diego, La Jolla, California

and

VINCENT STRETCH
University of Southern Mississippi Gulf Coast, Long Beach, Mississippi

Donaldson (1996) argued that remember/know judgments can be conceptualized within a signal detection framework by assuming that they are based on two criteria situated along a strength-of-memory decision axis. According to this model, items that exceed a high criterion receive a remember response, whereas items that only exceed a lower criterion receive a know response. Although a variety of findings have been presented in evidence against this idea, Dunn (2004) recently showed that detection theory is fully compatible with those findings. We present a variety of new results and new analyses that weigh strongly in favor of the detection interpretation. We further show that a dual-process account of recognition memory is compatible with a unidimensional detection model despite the common notion that such a model necessarily assumes a single process. The key assumption of this model is that individual recognition decisions are based on both recollection and familiarity (not on one process or the other).

A running debate in the recognition memory literature pits a long-standing theoretical framework known as signal detection theory against a popular multiple-process theory based on remember and know responses (Donaldson, 1996; Dunn, 2004; Gardiner & Gregg, 1997; Gardiner, Richardson-Klavehn, & Ramponi, 1998; Hirshman, 1998; Hirshman & Henzler, 1998; Hirshman & Master, 1997; Inoue & Bellezza, 1998; Malmberg, Zeelenberg, & Shiffrin, 2004; Xu & Bellezza, 2001; Yonelinas, 2002). The prototypical version of detection theory involves two equal-variance Gaussian distributions (one representing targets and the other representing lures) and one decision criterion situated along a decision axis. A test item that generates a memory signal that exceeds the decision criterion is declared to be *old*; otherwise, it is declared to be *new* (as illustrated in the upper panel of Figure 1). Although the equal-variance detection model is useful for illustrating the qualitative predictions made by signal detection theory, much evidence has accumulated over the years suggesting that a quantitatively more accurate version of the theory is an unequal-variance model in which the standard deviation of the target distribution slightly exceeds that of the lure distribution (as illustrated in the lower panel of Figure 1). An important feature of the standard detection model is that it assumes that recognition decisions are based on a unidimensional *strength-of-memory* variable.

An alternative view holds that recognition memory is based on at least two processes that correspond to distinct subjective states. One process involves recollection of the contextual associations of the original experience, whereas the other process, considered by most to reflect familiarity and by some to reflect semantic memory, does not (Cary & Reder, 2003; Karayianni & Gardiner, 2003; Tulving, 1985; Yonelinas, 2002). Because the subjective states associated with these two memory processes differ, subjects can indicate which process underlies their decision whenever they decide that a test item is *old*. They typically do this by saying "know" if their decision is based on familiarity (or semantic memory) and by saying "remember" if their decision is based on recollection.

The signal detection versus remember/know debate was initiated by Donaldson (1996), who argued that remember and know responses might reflect different degrees of memory strength, instead of qualitatively different memory processes (cf. Knowlton & Squire, 1995). Donaldson suggested that subjects approach the task by adopting two decision criteria, as illustrated in Figure 2. One is situated at a relatively low point on the decision axis (the know criterion), and the other is situated at a relatively high point on the decision axis (the remember criterion). A remember response indicates that the memory strength of the test item exceeds the high remember criterion, whereas a know response indicates that its strength falls above the know criterion but below the remember criterion. If the detection interpretation is correct, then remember/know judgments do not reflect qualitatively different forms of memory, so results based on these judgments cannot be used to effectively investigate dual-process theories of recognition memory.

Correspondence concerning this article should be addressed to J. T. Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093-0109 (e-mail: jwixted@ucsd.edu).

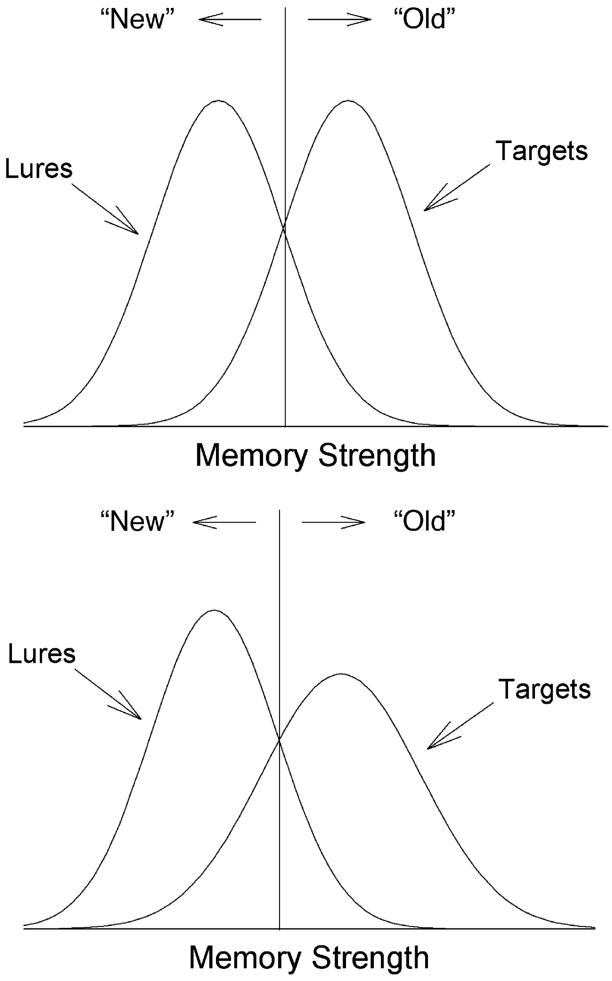

Figure 1. Equal-variance (upper panel) and unequal-variance (lower panel) signal detection models of recognition memory.

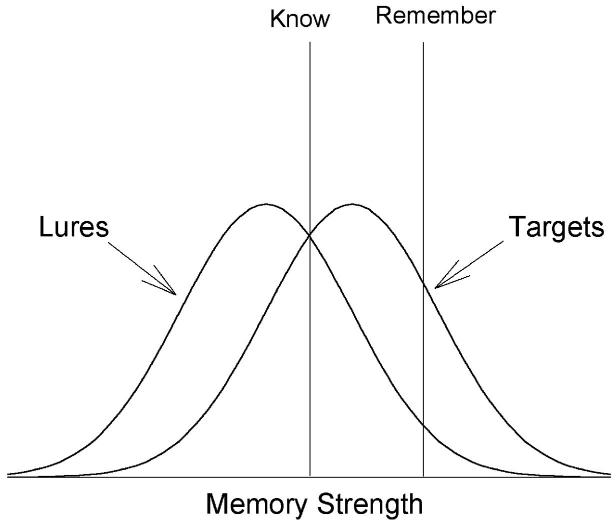

Figure 2. Signal detection interpretation of remember/know judgments.

Much has been written on this topic in the ensuing years, with several researchers suggesting that the detection account has been empirically refuted (e.g., Gardiner, Ramponi, & Richardson-Klavehn, 2002; Yonelinas, 2002). Dunn (2004) recently showed that all the prior results that have been taken as evidence against the detection account are actually compatible with it. In the present article, we report a variety of new findings that directly test (and support) strong predictions about remember/know judgments that are made by signal detection theory.

In what follows, we will use the phrase *dual-process theory* to refer to the widely held view that recognition memory is based on recollection and familiarity (e.g., Mandler, 1980), an idea that we do not dispute. Furthermore, we will use a phrase such as *dual-process remember/know theory* to refer to the idea that remember judgments denote recollection-based responses and that know judgments denote familiarity-based responses (an idea we do dispute). This is important to emphasize at the outset, because it is often assumed that the standard unidimensional detection model of recognition memory is incompatible with a dual-process theory of recognition memory. It is incompatible with some versions of dual process theory, such as the idea that recognition decisions are based either on recollection or on familiarity, but not with others, such as the idea that recognition decisions are based on an aggregate strength variable that consists of a combined recollection and familiarity signal.

If recollection is a binary (i.e., all-or-none) process, as it is sometimes thought to be, it would make sense to assume that subjects respond on the basis of *either* recollection *or* familiarity when making a recognition decision about an individual test item. After all, no amount of familiarity could further increase the maximum level of confidence that accompanies the complete recollection of the encoding event. All dual-process theories of recognition memory assume this either/or decision strategy, and if these theories are right, a unidimensional signal detection model does not apply. However, as will be detailed in a later section of this article, if recollection and familiarity are both continuous variables (so that both can be low, medium, or high in strength), it would make more sense for the subject to combine them into a single memory signal, instead of relying on one process or the other. And if the two processes are combined into an aggregate strength-of-memory variable, the signal detection model illustrated in Figure 1 would apply in a natural way. An important implication of this view is that *old*/*new* recognition memory itself is not process pure, so remember/know judgments cannot be process pure either.

We will begin our empirical inquiry into this matter with a consideration of the relationship between confidence ratings and remember/know judgments, a relationship about which signal detection theory has much to say. This discussion will be based largely on new analyses of data previously reported by Stretch and Wixted (1998). In that study, we asked subjects to supply both remember/know judgments and confidence ratings for every recognition decision, much as Tulving (1985) did in his original remember/know study. A critical issue in our investigation concerns the average confidence and reaction times associated with remember *false alarms*. Whereas remember false alarms are treated by most remember/know theorists as a minor annoyance to be ignored or subtracted out, they actually exhibit reliable and systematic effects that happen to correspond to the predictions of signal detection theory. For example, we show that remember false alarms are made more quickly and with higher confidence than know *hits*, effects that are anticipated by the simplest detection model. We then will conduct a detailed review of various correlational analyses that have been taken to pose a particular challenge to the signal detection interpretation of remember/know judgments. A closer look at the relevant data and the computation of new correlations again suggest just the opposite. Particularly telling in this regard is the fact that hit rates and false alarm rates of all kinds, including remember hit and false alarm rates, are positively correlated across individuals. The positive correlation between remember hit and false alarm rates has not been previously noticed, but it is just what a detection account predicts.

The fact that the detection analysis makes clear predictions about remember false alarms is one important way in which it differs from all dual-process interpretations of remember/know judgments. As Higham and Vokey (2004) have put it, dual-process models "have no mechanism to account for systematic variation in false R [remember] ratings" (p. 715). In fact, only one dual-process theory of remember/know judgments offers any interpretation at all of remember false alarms, and that is the high-threshold/detection model proposed by Yonelinas and his colleagues (e.g., Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998). According to this model, recollection is a threshold process. If recollection occurs, it occasions a remember judgment (if remember/know judgments are required) or a high-confident *old* response (if confidence ratings are required). If recollection fails to occur for a particular item, responding is based on familiarity instead. If the familiarity of the test item exceeds a criterion level, it occasions a know judgment; otherwise, the item is declared to be new. On this view, remember false alarms reflect guesses, not recollection, because, as Yonelinas et al. (1998) put it, "participants cannot truly recollect new items" (p. 330). Similarly, Eldridge, Sarfatti, and Knowlton (2002) stated that remember responses "reflect a qualitatively distinct type of memory in which the specific encoding context is reexperienced. Because episodic recollections are unlikely to occur for unexperienced events, one would expect the false alarm rate to be consistently low" (p. 142). Even so, the remember false alarm rate is almost always greater than zero, so some explanation for the apparent recollection of unexperienced events is needed. According to the threshold model, remember false alarms amount to guesses. That is, for whatever reason, subjects sometimes claim to remember an item even though recollection has not occurred. The fact that remember false alarms are regarded as guesses makes it difficult to derive clear predictions about them, but it is probably fair to say that this way of thinking does not naturally predict a variety of effects that are clearly predicted by signal detection theory.

No other dual-process account of remember false alarms can be identified in the literature, probably because most remember/know theorists would rather eliminate than explain them. It is, after all, awkward to assert that randomly chosen lures (which often are selected from a large pool of possible lures) occasion the recollection of an encoding event that never happened. Then again, as Roediger and McDermott (1995) showed, the false memories produced by the DRM procedure frequently generate remember responses, and subjects often supply recollective details associated with these false memories. Perhaps all recognition tests give rise to false recollection, with the only difference being that the DRM procedure allows one to predict in advance which lure will be the one that is likely to be falsely recollected. Although no one has advanced a false-recollection theory of remember false alarms, it seems natural to consider this possibility as we examine the systematic and heretofore unnoticed regularities associated with remember false alarms.

CONFIDENCE RATINGS AND REMEMBER/KNOW JUDGMENTS

An inescapable implication of the signal detection interpretation of remember/know judgments is that there should be a relationship between remember/know judgments and confidence ratings. Just as subjects can be asked to classify their recognition decisions as remember or know, they also can be asked to indicate a degree of confidence associated with each recognition decision. Figure 3 shows the standard signal detection interpretation of confidence ratings that are, in this example, made on a 5-point scale (with a rating of 1 reflecting *low confidence* and a rating of 5 reflecting *high confidence*). As shown in the figure, these ratings are assumed to be based on additional decision criteria arrayed along the memory strength axis. The *old*/*new* decision criterion is now labeled 1_{Old}, and any item with a strength that exceeds that value receives an *old* decision. If the strength of the test item falls between the 1_{Old} and the 2_{Old} criteria, the *old* decision is made with a confidence rating of 1. If the strength of the test item instead falls between the 2_{Old} and the 3_{Old} criteria, the *old* decision is made with a confidence rating of 2 (and so on). The higher the memory strength of the test item, the higher the confidence rating. It is also typically assumed that the farther the strength falls above the *old*/*new* criterion, the *faster* the response will be. Thus, high-confident responses should be made more quickly than low-confident responses, a result that is commonly observed (e.g., Ratcliff & Murdock, 1976). The question of interest is how these observations bear on the signal detection interpretation of remember/know judgments.

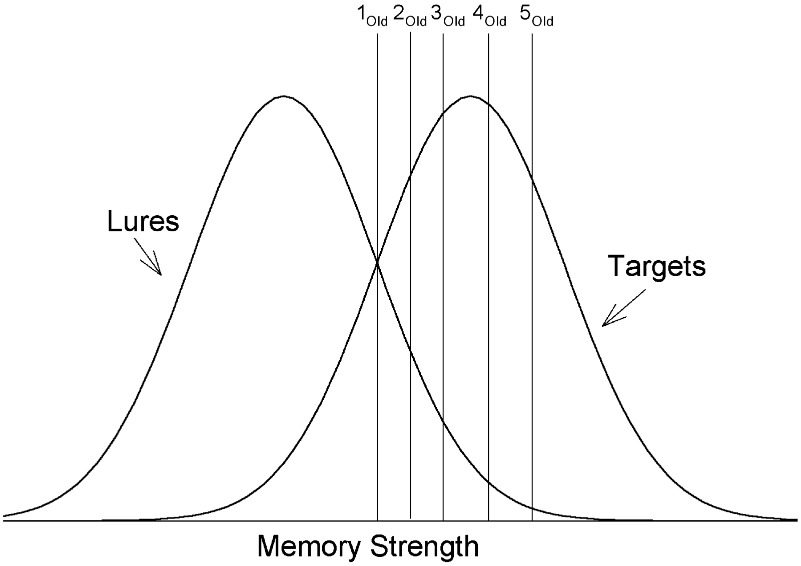

Figure 3. Signal detection interpretation of confidence ratings made on a 5-point scale (only criteria for *old* responses are shown, though a similar set of criteria could be shown for *new* responses as well).

Confidence and Reaction Times Associated With Remember/Know Judgments

Figure 4 presents the signal detection model with the confidence and remember/know criteria placed on the same memory strength axis (i.e., Figure 4 is the combination of Figures 2 and 3). The leftmost criterion is the *old*/*new* decision criterion, and it is now labeled with both a “1_{Old}” and a “K” (meaning that any item that falls above that criterion but below the remember criterion receives an *old* decision that is classified as a know response).

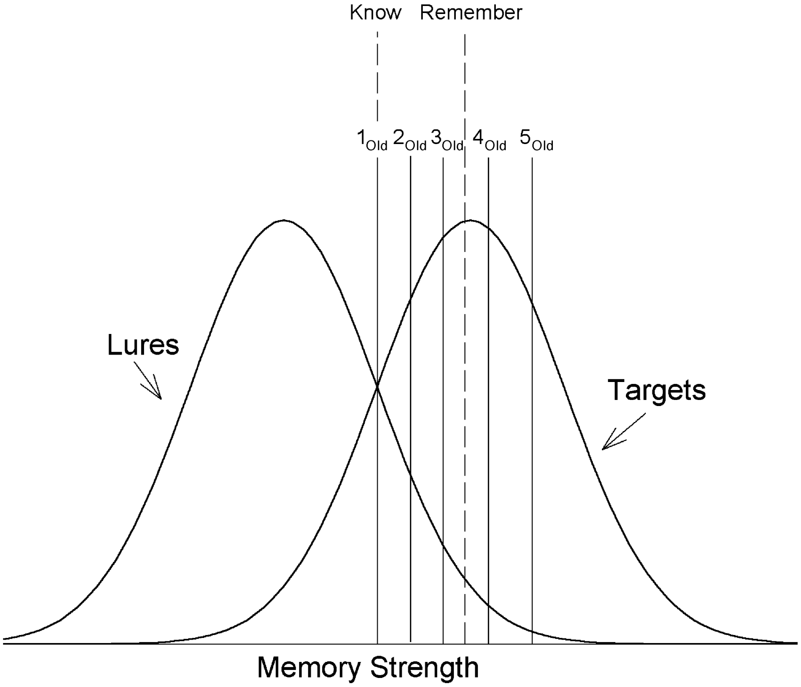

Figure 4. Signal detection interpretation of confidence ratings and remember/know judgments.

From this figure, it is easy to see that remember responses should be made with relatively high confidence (and relatively quickly), as compared with know responses.

It is well known that remember responses to *targets* are made with higher confidence than are know responses (Tulving, 1985), and this makes sense from both the dual-process point of view, according to which remember responses reflect recollection, and the detection point of view, according to which remember responses reflect strong memories. From the dual-process point of view, if one recollects the prior episodic encounter with a test stimulus, one ought to be highly confident when reporting that the test stimulus was, indeed, encountered on a prior occasion. From the signal detection point of view, the remember criterion is situated at a relatively high point on the memory strength axis, so the strong memories that exceed that criterion would also exceed a relatively high confidence criterion.

Signal detection theory goes beyond this and further predicts that remember *false alarms* will be made with greater confidence than know false alarms. This prediction follows from the fact that remember false alarms are made to lures with a strength that happens to exceed the (high) remember criterion. This is a key consideration. Conspicuously missing from most remember/know studies is a discussion of what remember false alarms might represent (the notable exception being Yonelinas and colleagues, as was mentioned earlier). Most often, it is simply noted that the remember false alarm rate is low, and the theoretical interpretation is based exclusively on the remember hit rates (as Higham & Vokey, 2004, also observed). This is unfortunate given that remember false alarms may hold the key to revealing what a remember response actually represents. After all, the subject does not know that the test item is a lure when he or she claims to remember it. Something about that lure presumably prompts a remember response, and it is worth considering what that something might be. According to the detection account, items that give rise to remember responses are high in strength (and far from the *old*/*new* decision criterion). That being the case, remember false alarms, like remember hits, not only should be made with high confidence but also should be made more quickly than their know counterparts.

To investigate this issue, we reanalyzed data reported by Stretch and Wixted (1998). In their Experiment 1, subjects first studied a list of 48 words and then completed a recognition test in which they were presented with 48 targets randomly intermixed with 48 lures. On the recognition test, the subjects were first asked to make an *old*/*new* judgment and to do so as quickly as they could without sacrificing accuracy. For any item that received an *old* decision, the subjects were then asked to make a remember/know judgment. Finally, all the items were given a confidence rating on a scale of 1 to 5 (*low* to *high*). Table 1 presents the average confidence and average reaction times (RTs) associated with remember and know judgments. It is important to emphasize that the RT data correspond to the initial *old*/*new* decision that was *subsequently* classified as a remember or know response. Thus, the remember/know instructions, which typically instruct subjects to first decide whether or not recollection occurred and to then classify their response as know only if recollection failed, would presumably not influence the timing of the initial *old*/*new* decision. For that decision, the subjects were instructed to respond quickly without sacrificing accuracy. This is a two-step remember/know procedure, as opposed to the one-step procedure in which subjects classify items as remember, know or new. Eldridge et al. (2002) argued that the two-step procedure, which is the kind of procedure originally used by Tulving (1985), is more likely to induce responding based on recollection (for remember responses) and familiarity (for know responses) than is the one-step procedure (cf. Hicks & Marsh, 1999).

Table 1
Average Confidence Ratings and Average Reaction Times (in milliseconds) for Remember and Know Judgments to Targets and Lures From Experiment 1 of Stretch and Wixted (1998)

	Confidence		Reaction Time	
	Remember	Know	Remember	Know
Targets	4.64	3.03	718	797
Lures	3.82	2.80	734	812

As uniquely predicted by the detection account, remember hits *and* remember false alarms were made more quickly and with higher confidence than their know counterparts (Table 1). With regard to targets, the mean confidence in remember responses (4.64) significantly exceeded mean confidence in know responses [3.03; $t(26) = 25.3$]. More interesting, mean confidence in remember responses to *lures* (3.82) significantly exceeded mean confidence in know responses to lures [2.80; $t(26) = 8.29$]. Similar trends are observed in the RT data. With regard to targets, the mean RT for remember responses is significantly less than the mean RT for know responses [$t(26) = -4.09$], and the same is true for the lures [$t(26) = -2.43$].

Perhaps the most counterintuitive and compelling prediction made by signal detection theory is that remember responses to *lures* should be made more quickly and with higher confidence than know responses to *targets*. These predictions follow from the fact that lures that happen to exceed the remember criterion will have greater strength than targets that do not. It is hard to imagine a dual-process account predicting such a curious outcome. However, as predicted by detection theory, the mean RT of remember responses to lures was significantly faster than the mean RT of know responses to targets [$t(26) = -2.72$], and the mean confidence in remember responses to lures significantly exceeded the mean confidence in know responses to targets [$t(26) = 8.29$].

Stretch and Wixted (1998) reported three additional experiments in which remember/know judgments and confidence ratings were collected in the same manner as that in

Experiment 1. The confidence results observed in Experiment 1 were replicated in each (e.g., in every case, the average confidence in remember false alarms significantly exceeded the average confidence in know hits), but significant effects were not always obtained for the RT measure. In Experiment 2, mean RTs were ordered in the same manner as in Experiment 1, but only the difference between remember responses to targets (mean RT = 717 msec) and know responses to targets (mean RT = 765 msec) was significant [$t(26) = -3.81$]. A marginally significant effect in the same direction was observed for responses to lures [remember mean RT = 756 msec, know mean RT = 793 msec, $t(26) = -1.68$, $p = .10$], and although similar trends were observed in Experiments 3 and 4, the differences were small and did not approach statistical significance. Thus, both the confidence data and the RT data support the detection account, but the RT results are somewhat less compelling, because significant effects are not always obtained.

Although the predicted RT effects did not always materialize, these tests offered real opportunities for the detection account to fail. This is worth emphasizing because some have argued that signal detection theory offers little more than a post hoc description of empirical results because it can adjust itself to any outcome (e.g., Gardiner et al., 2002). But this is not the case. Had remember false alarms been made with lower confidence than know false alarms, or had they been made more slowly than know false alarms, the detection account would have been hard to sustain. As it turns out, the required results are generally obtained.

The standard dual-process interpretations of remember/know judgments do not seem to offer a compelling explanation for these findings. For example, the idea that remember *false alarms* should be made quickly or with high confidence does not follow from the general notion that remember responses reflect recollection (because lures should not be recollected in the first place). If remember false alarms are construed as guesses, as the high-threshold account assumes (e.g., Yonelinas et al., 1996; Yonelinas et al., 1998), they need not be made rapidly or with high confidence. In fact, it is easy to imagine that guesses would be made with low confidence, contrary to what has been observed. If remember false alarms are, instead, construed as instances of false recollection, the high confidence associated with those responses makes sense. However, one might expect recollection-based hits and false alarms to be made significantly more slowly than familiarity-based know responses. That is, dual-process theories of recognition memory typically hold that familiarity-based responses are made quickly, whereas recollection-based responses are made more slowly (Mandler, 1980; Yonelinas, 2002). Thus, if know responses to targets were made more quickly than remember responses to targets, this would provide converging evidence that remember/know judgments correspond to recollection and familiarity. But in Experiments 1 and 2 of Stretch and Wixted (1998) at least, the opposite result was observed.

Some dual-process remember/know models do seem to predict that recollection-based remember responses should be made more quickly than familiarity-based know responses. Cary and Reder (2003), for example, proposed that familiarity-based responses are made only after an initial attempt at recollection fails, even for *old/new* decisions that do not involve remember/know judgments. It is not clear how to reconcile a model such as this with prior evidence suggesting that familiarity-based responding is typically fast, but the model is consistent with the RT data for remember and know hits presented here. This model does not predict the occurrence of remember false alarms, because as is typically the case, the model assumes that recollection does not occur for items that were never presented. However, if we modify this model to allow for the possibility of false recollection (e.g., due to associative activation of some of the lures), it becomes compatible with the results discussed thus far. That is, remember responses are faster than know responses because subjects always attempt to respond on the basis of recollection before they respond on the basis of familiarity, and remember false alarms are made with higher confidence than know hits because those false alarms involve a (false) recollective experience.

Distribution of Confidence Ratings for Remember/Know Judgments

The predictions considered above were concerned with average confidence ratings for remember and know judgments, but the model shown in Figure 4 makes a clear prediction about the distribution of confidence ratings for remember/know judgments as well. The remember criterion in this example is placed between the 3_{Old} and 4_{Old} confidence criteria. A subject whose various decision criteria are arranged in this manner should provide confidence ratings of 1, 2, or 3 for know responses and 3, 4, or 5 for remember responses. That is, if the confidence criteria and the remember/know criteria remain rigid with respect to each other throughout a recognition test, confidence ratings should overlap by, at most, one value. In this example, some know responses and some remember responses would receive confidence ratings of 3, but no know response would receive a confidence rating of 4 or 5, and no remember response would receive confidence rating of 1 or 2.

Slight item-to-item variability in the placement of the remember criterion with respect to the confidence criteria would expand the confidence rating overlap to some degree. For example, if the remember criterion were placed between the 3_{Old} and the 4_{Old} confidence criteria for some items and between the 4_{Old} and the 5_{Old} confidence criteria for others, remember and know responses would share two confidence ratings (3 and 4). Somewhat more variability in the placement of the remember criterion would mean that three confidence ratings are shared (2, 3, and 4), and even more variability would mean that 4 or 5 confidence ratings would be shared. If all the subjects showed extensive overlap in the confidence ratings associated with remember and know judgments, it would be hard to find any evidence in those ratings that would favor the signal detection account (i.e., the variability would obscure any detection process that might exist).

Each of the 36 subjects in Stretch and Wixted's (1998) Experiment 1 studied two lists of 48 words, and they completed a recognition test after each. Thus, their data yielded 72 opportunities to investigate the relationship between the confidence ratings and the remember/know judgments at the individual subject, individual test level. In 3 of those cases, no assessment was possible, because the subject supplied only remember responses. Interestingly, in all 3 of these cases, the confidence ratings associated with remember responses spanned the range from 3 to 5. Such responding is entirely consistent with a detection account in that it suggests that the memory strength of items exceeding the remember criterion is graded (thereby accounting for the different levels of confidence). If remember responses denoted mental time travel (i.e., all-or-none recollection), one might expect to see such responses associated only with the highest confidence rating. Instead, multiple confidence levels are observed, as has been known since Tulving (1985) first introduced the remember/know distinction. Rotello, Macmillan, and Reeder (2004) have provided additional convincing evidence that remember responses are graded.

Of the remaining 69 protocols, 25 (36%) conformed to the pattern predicted by the model shown in Figure 4, in that remember and know responses overlapped by, at most, one confidence rating. A representative example is presented in the upper panel of Table 2 (and the complete data set for all 36 subjects is presented in Appendix A). Combining *old* responses to targets and lures, this subject made 18 remember responses with a confidence rating of 5, 9 with a confidence rating of 4, and none with a confidence rating of 3 or less. By contrast, this subject's know responses received confidence ratings of 2, 3, or 4 only. A pattern such as this would be observed if the remember criterion was placed between the 4_{Old} and the 5_{Old} confidence criteria and remained fixed with respect to the confidence criteria throughout the recognition test.

In an additional 22 of the 69 response protocols (32%), confidence ratings to remember and know responses overlapped by two values, due to a single response. A representative example is presented in the middle panel of Table 2. This subject's performance can, for the most part, be represented by a model in which the remember criterion is placed between the 3_{Old} and the 4_{Old} confidence criteria, but some variability in the remember criterion would have to be assumed in order to accommodate the one know response that received a confidence rating of 4.

Table 2
Distribution of Confidence Ratings Associated With Remember and Know Judgments for Three Representative Subjects From Experiment 1 of Stretch and Wixted (1998)

	Confidence Rating				
Judgment	1	2	3	4	5
Remember	0	0	0	9	18
Know	0	4	7	5	0
Remember	0	0	3	7	31
Know	0	1	10	1	0
Remember	0	1	6	7	11
Know	0	2	10	2	0

The remaining 22 protocols (32%) corresponded to a detection model that assumes somewhat greater variability in the location of the remember criterion. These protocols exhibited overlap in three or more of the confidence ratings. A representative protocol exhibiting this pattern is presented in the bottom panel of Table 2, which exhibits overlap in confidence ratings of 2, 3, and 4. A protocol such as this would be produced if the remember criterion varied in placement throughout the recognition test between the 2_{Old} and 3_{Old} confidence criteria and between the 4_{Old} and 5_{Old} confidence criteria.

That the remember criterion might exhibit item-to-item variability hardly seems like a radical proposition. In fact, Macmillan and Creelman (1991) addressed the issue of criterion variability in their well-known signal detection text and concluded that it is almost certainly the case that such effects occur. Ordinarily, there is no way to tell whether the decision criteria vary from trial to trial. With two sets of criteria, however, we can determine whether they vary with respect to each other. Although there is no way to tell which set of criteria vary, or whether they both do, it seems reasonable to suppose that the placement of the remember criterion (the definition of which was learned mere minutes before taking the recognition test) would exhibit more item-to-item variability than the placement of the confidence criteria would (which subjects have understood for years). In fact, researchers often go to considerable lengths to ensure that subjects understand the meaning of remember and know judgments. Yonelinas (2001), for example, stated the following: "To ensure that subjects understood the test instructions, they were asked to describe the remember–know distinction back to the experimenter, and the instructions were repeated if the participant appeared to have misunderstood the distinction" (p. 363). Similarly, Rajaram (1993) stated that "after reading these instructions, each subject was asked to explain to the experimenter how she/he would make the 'remember' and 'know' judgments on the basis of the instructions provided. If they were confused about the distinction, the experimenter clarified the instructions further before the test phase began" (p. 92). No such efforts were made to ensure that the subjects understood the distinction between varying degrees of confidence, presumably because the subjects were (quite reasonably) assumed to already understand the distinction very well.

By assuming some variability in the placement of the remember criterion, these data are easily reconciled with a detection account, but no other model would seem to offer an explanation for the overall pattern that was observed. The general notion that remember responses reflect all-or-none recollection of the original encoding event naturally predicts that they would be made with high confidence and only to targets. Instead, remember responses are made with varying degrees of confidence and are made to both targets and lures. Allowing for the possibility that some remember responses are guesses would account for the presence of remember false alarms, and it might be the case that subjects express varying degrees of confidence when they make remember guesses. However, this way of thinking does not seem to predict that confidence in remember false alarms would be so high, on average, or that those responses would exhibit such a small amount of overlap with the confidence ratings for know responses.

The idea that remember false alarms reflect false recollections instead of guesses seems more compatible with results such as these. If remember false alarms reflect false recollection, the high confidence associated with those responses makes sense. Such recollection does not appear to be an all-or-none process, because as with targets, remember responses to lures are associated with varying degrees of confidence (even though confidence is high, on average). Although an all-or-none model has trouble accounting for results such as these, Cary and Reder's (2003) dual-process remember/know model does not, because they assume that recollection is based on a continuously distributed variable. If recollective strength falls below a threshold, recollection does not occur. If it falls above that threshold, a recollective experience does occur. Although they discuss recollection in categorical terms (as either happening or not), it is but a small step to assume that the further above the threshold the recollective strength falls, the higher one's confidence will be. This model does not incorporate any provision for remember false alarms, but once again, adding a false recollection component would seem to bring it into line with the graded remember responses that occur to lures as well as to targets.

Thus, a dual-process remember/know model that is compatible with the results considered thus far would hold that recollection is a continuously distributed variable (not an all-or-none variable) and that it is fallible in the sense that recollection can occur for lures as well as for targets. But if the recollective process is thought to involve a continuously distributed and fallible variable, it seems odd to assume that responding is based *either* on recollection *or* on familiarity, which is an idea that the remember/know methodology depends on. If there are two independent, continuously distributed variables, both of which are useful but imperfect indicators of prior occurrence, it makes more sense to combine them in some way, rather than responding on the basis of one or the other (a point we will consider in more detail later). In fact, it seems odd that subjects would not learn that this is true on the basis of their everyday experiences. Still, if one is willing to assume (1) that *old*/*new* recognition memory is process pure even though recollection and familiarity are continuously distributed, fallible variables, (2) that subjects attempt recollection before responding on the basis of familiarity (thereby accounting for relatively slow know responses), and (3) that remember false alarms reflect false recollection (thereby accounting for their high confidence), the results described to this point are compatible with a dual-process interpretation of remember/know judgments.

Although such a model is compatible with the results considered to this point, it is worth noting that other findings in the literature pose a problem for it. Presumably, the false recollection that accounts for remember false alarms would arise because some of the lures are similar to or are associatively activated by the targets that appeared on the list. As noted by Wixted and Squire (2004), however, the remember false alarm rate for amnesics is often very high. In two recent studies (Manns, Hopkins, Reed, Kitchener, & Squire, 2003; Yonelinas et al., 2002), for example, the remember false alarm rate for amnesic patients was 16%. Schacter, Curran, Galluccio, Milberg, and Bates (1996) also described an amnesic patient with a very high remember false alarm rate. It is difficult to imagine that patients who have great difficulty forming true memories would be so adept at forming false ones, so results such as these weigh against a process-pure, false recollection account of remember false alarms. A more likely interpretation is the one offered by a detection model: Amnesics, knowing that their memories are poor, set a more lenient remember decision criterion on the memory strength axis, so that a greater proportion of familiar lures exceeds that criterion (cf. Curran, Schacter, Norman, & Galluccio, 1997).

Receiver-Operating Characteristic Analyses

It has been known for many years that an equal-variance detection model, such as the one shown in the upper panel of Figure 1, does not apply to human recognition memory. Instead, every relevant analysis suggests that the standard deviation of the target distribution is somewhat greater than that of the lure distribution, as is shown in the lower panel of Figure 1 (e.g., Ratcliff, Sheu, & Gronlund, 1992). The issue of the relative variance of the target and lure distributions is typically addressed by examining the properties of the receiver-operating characteristic (ROC). An ROC is simply a plot of the hit rate versus the false alarm rate for different levels of bias. A typical ROC is obtained by asking subjects to supply confidence ratings for their recognition memory decisions. Several pairs of hit and false alarm rates can then be computed by cumulating from different points on the confidence scale (beginning with the most confident responses). Thus, a high-confident hit and false alarm rate pair is obtained by computing the proportion of targets and proportion of lures that receive a high-confident *old* response. Another hit and false alarm rate pair is obtained by computing the proportion of targets and the proportion of lures that receive either a high- or a medium-confidence *old* response (and so on for the remaining confidence levels). The easiest way to analyze an ROC is to convert the hit and false alarm rates to *z*-scores and then plot *z*(Hit) versus z(FA). If a Gaussian signal detection model underlies performance, the plot should be accurately described by a straight line. Moreover, the slope of that line provides an estimate of the ratio of the standard deviation of the lure distribution to the standard deviation of the target distribution ($\sigma_{\text{Lure}}/\sigma_{\text{Target}}$). If an equal-variance model applies, the slope should be 1, but if the standard deviation of the target distribution exceeds that of the lure distribution, the slope should be less than 1. Previous meta-analyses of confidence-based ROC data generally have shown that the slope of the best-fitting line is, on average, approximately 0.80 (Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff et al., 1992). Thus, the standard deviation of the target distribution is typically about 1.25 (i.e., 1/0.80) times that of the lure distribution.

Remember/know ROC meta-analyses. If remember and know decisions reflect different placements of decision criteria along a strength-of-memory axis, and if z-ROCs are constructed from those responses, one would expect the slope of that ROC to be less than 1 (and in the vicinity of 0.80). The procedure for constructing such ROCs would be identical to the procedure outlined above for confidence ratings, except that only two pairs of hit and false alarm rates (i.e., two points on the ROC) would be computed. One point would consist of hit and false alarm rates based on remember responses only, and another would consist of hit and false alarm rates based on remember and know responses combined.

Rotello et al. (2004) analyzed many two-point remember/know ROCs and found that the slope estimate was reliably greater than the expected 0.80, with the mean and median of 373 independent estimates found to be 1.01 and 0.92, respectively. Dunn (2004) reported essentially the same result, using a different method. Specifically, across 400 remember/know conditions in previous studies, Dunn found that d' estimates based on remember hit and false alarm rates generally matched d'estimates based on overall hit and false alarm rates. This would happen only if an equal-variance model applied, so that the slopes of the remember/know ROCs were, on average, close to 1.0. Whereas Dunn took the equivalent d' estimates to be consistent with a signal detection interpretation of remember/know judgments, Rotello et al. took that same result to weigh against it. This result is consistent with the equal-variance detection model shown in the upper panel of Figure 1, but the quantitatively more accurate version of the theory (according to all prior analyses, anyway) is the unequal-variance model shown in the lower panel. Thus, the fact that remember/know ROCs suggest an equal-variance model, although appearing to support the detection account, appears to pose a slight problem for it. In fact, largely on the basis of this result (namely, the remember/know ROC slope should have been close to 0.80, but it was close to 1.0 instead), Rotello et al. rejected the unidimensional detection model and advanced a multidimensional detection model instead.

However, both Dunn's (2004) meta-analysis and Rotello et al.'s (2004) meta-analysis were based on a review of procedurally diverse studies, many of which differed greatly from the kind of studies that generated the expectation that the ROC slope should be close to 0.80. An ROC slope of 0.80 is not an immutable constant; instead, the slope is a function of the experimental variables (e.g., list length) used in the recognition study (Heathcote, 2003). The slopes are usually less than 1 and, given the procedures typically used, tend to average about 0.80. But the procedures used in many remember/know studies are often quite unlike those used in standard recognition memory studies. To take one of many examples, some of the remember/know studies have investigated recognition memory for English and Polish folksongs (Gardiner & Radomski, 1999). The remember/know ROC slopes for these studies were quite large (greater than 1 in every case), and it is simply not known what the confidence-based ROC slopes would have been for this procedure. Perhaps they would have been large as well. This problem cannot be addressed merely by increasing the size of the remember/know database that is analyzed. No matter how large the database is, one cannot necessarily assume that the average remember/know ROC slope should be 0.80. Instead, one can assume only that the average remember/know ROC slope should be similar to the average confidence-based ROC slope, whatever that might be. A more direct way to test whether remember/know judgments reflect different placements of decision criteria on a memory strength axis is to construct a confidence-based ROC and then determine whether or not the ROC data generated by remember/know judgments fall on the same line when the same procedure is used. They should if the detection account is correct. We consider this very kind of analysis next.

Direct comparisons of remember/know and confidence-based ROCs. Figure 5 shows the ROCs constructed by pooling the confidence data over subjects for each of the four experiments reported by Stretch and Wixted (1998). The points on the ROC were obtained by computing hit and false alarm rates for items receiving a response that exceeded each confidence criterion (e.g., 5_{Old}, 4_{Old}, 3_{Old}, and so on) and then taking the z-transform of each value. The results are typical in that the points generally trace out a straight line, although some evidence of an inverted-U is apparent for Experiment 1 (see Heathcote, 2003, for a discussion about interpretations of nonlinear z-ROCs).

Of particular interest is the question of where the remember hit and false alarm rates and the remember + know hit and false alarm rates fall. As shown in Figure 5, they fall almost exactly where they should according to the confidence-based ROC. That the remember + know point falls on the ROC is not surprising, because it represents the overall hit and false alarm rate, as one of the confidence points does, too. More remarkable (if the detection interpretation is incorrect, that is) is the fact that

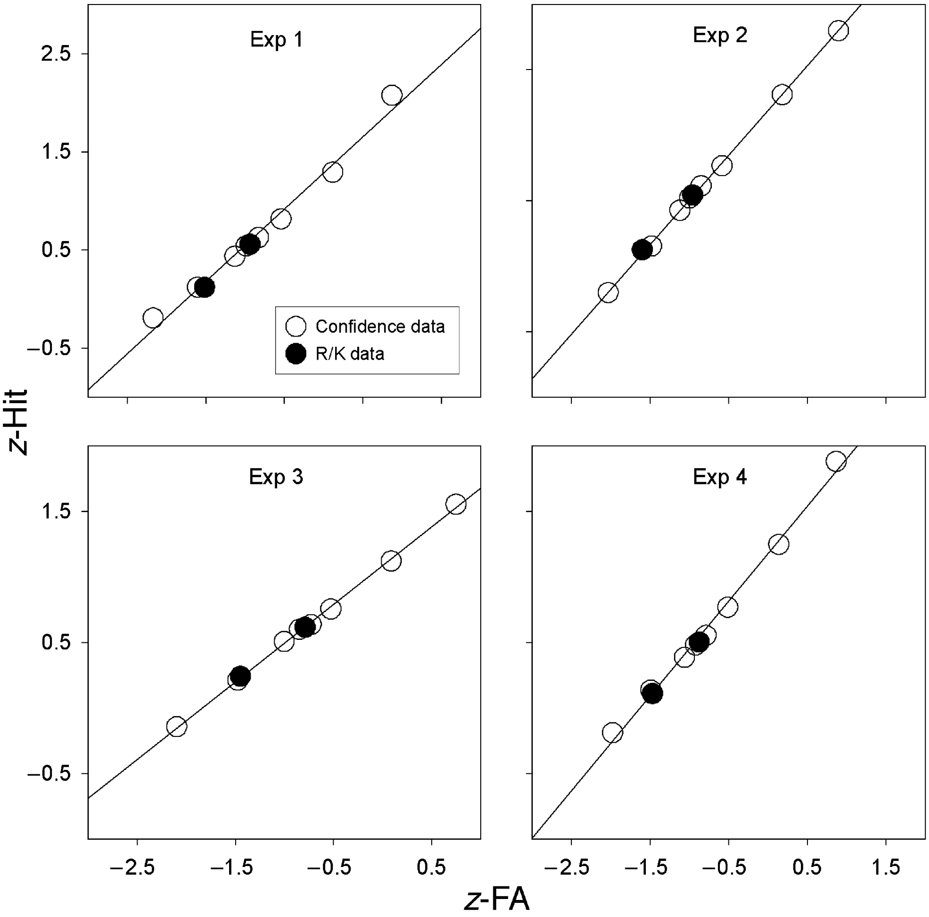

Figure 5. Confidence-based (open circles) and remember/know-based (filled circles) receiver-operating characteristic data plotted in *z* space for Experiments 1–4 of Stretch and Wixted (1998). The lines show the least squares fit of the confidence data.

the point representing the remember hit and false alarm rates falls so close to the confidence-based ROC in every case.

Table 3 shows the slope of the four ROCs based on the confidence ratings and based on the remember/know judgments. Across the four experiments, the average slope of the confidence-based *z*-ROC is 0.68, which, as usual, suggests an unequal-variance detection model. This value is less than the typical value of .80, but it is well within the range commonly observed in prior studies. The average slope of the remember/know-based ROCs is 0.66, which is not substantially different from the confidence-based slope estimate. The next set of values in Table 3 is from a similar study reported by Yonelinas (2001). In the full- and divided-attention conditions of Experiments 1, 2a, 2b, and 2c, confidence ratings and remember/know judgments were taken after each test item (as in the study by Stretch & Wixted, 1998). Yonelinas (2001) did not perform ROC slope analyses, but the data needed to do so were presented in Table 1 of that article. Averaged across the eight conditions, the confidence and the remember/know ROC slopes were very similar (0.75 and 0.73, respectively) and were suggestive of the typical unequal-variance model. The next set of values in Table 3 shows the confidence-based and remember/know-based ROC slopes for two similar experiments reported by Rotello et al. (2004). The mean confidence and the remember/know ROC slopes were reported by those authors to be 0.86 and 0.94, respectively. Again, an unequal-variance detection model is implied in both cases, although the remember/know ROC in this case yielded a slope that is somewhat larger than the confidence-based ROC. Over the three studies represented in Table 3, the mean of the mean slopes is 0.76 for the confidence-based ROCs and 0.78 for the remember/know-based ROCs.

In studies such as these, in which confidence ratings and remember/know judgments were taken for each recognition test item, the slope estimates not only should be equal on average, they also should be highly correlated from experiment to experiment. The estimates are, after all, theo-

Table 3
Confidence-Based and Remember/Know-Based ROC Slopes for Experiments in Which Subjects Made Confidence and Remember/Know Judgments After Each Item

	ROC Slope	
	Confidence Ratings	R/K Judgments
Stretch and Wixted (1998)		
Experiment 1	0.74	0.76
Experiment 2	0.68	0.65
Experiment 3	0.59	0.57
Experiment 4	0.72	0.66
Mean	0.68	0.66
Yonelinas (2001)		
Experiment 1		
Full	0.63	0.73
Divided	0.76	0.81
Experiment 2A		
Full	0.64	0.60
Divided	0.72	0.65
Experiment 2B		
Full	0.79	0.73
Divided	0.84	0.82
Experiment 2C		
Full	0.74	0.70
Divided	0.85	0.80
Mean	0.75	0.73
Rotello, Macmillan, and Reeder (2004)		
Experiment 1	0.81	0.85
Experiment 2	0.91	1.03
Mean	0.86	0.94
Mean of Means	0.76	0.78

retically determined by *exactly the same* pair of underlying distributions in each experiment. For the 14 experiments shown in Table 3 (4 from Stretch & Wixted, 1998, 8 from Yonelinas, 2001, and 2 from Rotello et al., 2004), the correlation between the two slope estimates was .85, which is highly significant ($p < .001$).

An advantage of the studies discussed above is that they allowed a comparison between confidence-based and remember/know slope estimates while holding the experimental procedure constant. This was accomplished by asking subjects for both confidence ratings and remember/know judgments after each test item. It could be argued that a better approach would be to use a between-subjects design. Although this would add an additional element of variability, it would eliminate any chance that remember/know judgments might somehow contaminate confidence ratings (or vice versa). Gardiner and Java (1990) conducted one such study. In their Experiment 2, subjects studied words and nonwords for a subsequent recognition test, during which they also supplied remember/know judgments. In their Experiment 3, the procedure was the same, except that the subjects supplied confidence ratings instead (sure vs. unsure). Table 4 shows the slopes of the two-point ROCs from these two experiments. Although the values are rather variable, the average slope for the confidence-based ROC across the two studies was 0.85, whereas the average slope for the remember/know-based ROC was 0.84. Again, the results in both cases point to the standard unequal-variance detection model.

Rajaram (1993) also conducted a study in which confidence ratings (sure vs. unsure) and remember/know judgments were taken in two different experiments that otherwise used the same procedure. In both experiments, repetition primes or unrelated primes were presented prior to the presentation of the recognition test items, and the results of an ROC analysis performed on these data are also shown in Table 4. Averaging across the two experiments, the slopes of the confidence-based and remember/know ROCs were identical (0.65) and, yet again, were consistent with the standard unequal-variance detection model. In Experiment 3 of Yonelinas (2001), one group of subjects studied words under deep and shallow encoding

Table 4
Confidence-Based and Remember/Know-Based ROC Slopes for Experiments in Which One Group of Subjects Supplied Confidence Judgments and Another Group of Subjects Supplied Remember/Know Judgments

	ROC Slope	
	Confidence Ratings	R/K Judgments
Gardiner and Java (1990)		
Words (Experiments 2 and 3)	0.94	0.60
Nonwords (Experiments 2 and 3)	0.75	1.08
Mean	0.85	0.84
Rajaram (1993)		
Repetition (Experiments 3 and 4)	0.74	0.68
Unrelated (Experiments 3 and 4)	0.56	0.62
Mean	0.65	0.65
Yonelinas (2001)		
Deep (Experiment 3)	0.65	0.72
Shallow (Experiment 3)	0.78	0.84
Mean	0.72	0.78
Mean of Means	0.74	0.76

Note—The slope estimates from Gardiner and Java (1990) and Rajaram (1993) were taken from the database described by Rotello, Macmillan, and Reeder (2004).

conditions and provided remember/know judgments, and a different group of subjects was exposed to the same procedure, except that they supplied confidence ratings. On average, the slope of the confidence-based ROC was 0.72, and the slope of the remember/know-based ROC was 0.78. Across all three studies shown in Table 4, the mean of the mean slopes for the confidence-based ROCs was 0.74, whereas the corresponding value for the remember/know-based ROCs was 0.76. Thus, with the procedure equated, the results shown in Tables 3 and 4 suggest that the slopes are similar and not far from the expected value of 0.80 whether the ROC is constructed using confidence ratings or remember/know judgments.

Although most of the evidence suggests that the slopes are the same, it may be premature to dismiss the possibility that the remember/know ROC slope is greater than the confidence-based ROC, because there is at least one study that showed the slopes to be quite different. Rajaram, Hamilton, and Bolton (2002) conducted a study that was similar to the one conducted by Gardiner and Java (1990). In each of two experiments, subjects studied lists consisting of a mixture of words and nonwords in two different sessions. In the first session, remember/know judgments were taken after each recognition decision, and in the second (1 week later), confidence judgments were taken. The order in which the judgments were taken was not counterbalanced, out of a concern that asking for confidence judgments first would contaminate remember/know judgments. As Rajaram et al. put it, "Remember–know instructions are more elaborate than confidence judgments, and care was taken to ensure that the participants fully grasped the nature of, and the distinction between, these experiential states. Therefore, this testing order was selected to ensure that confidence judgments did not contaminate the remember–know judgments with a carryover effect to the second session" (p. 230). This complicates a direct comparison, because carryover effects might have occurred in the other direction as well, but the results are worth considering anyway.

The remember/know and sure/unsure hit rates were estimated from Figures 1 and 2 of Rajaram et al. (2002) in order to compute ROC slope estimates (the corresponding false alarm rates were obtained from the text). In both experiments, the remember/know-based slopes (0.97 and 1.20 in Experiments 1 and 2, respectively) far exceeded the confidence-based slopes (0.55 and 0.66, respectively).

What accounts for these slope differences? The remember hit rates in these experiments were well below the *sure* hit rates, as if a very high remember criterion was used. That being the case, a detection account would anticipate that the remember false alarm rates would be correspondingly lower than the *sure* false alarm rates. But the remember false alarm rates were, from the detection point of view, unexpectedly high (and that accounts for the much larger remember/know ROC slopes). This is best illustrated by considering the nonword data from their Experiment 1. The remember hit rate (.15) was much lower than the *sure* hit rate (.32), so the remember false alarm rate should have been lower than the *sure* false alarm rate of .01 as well. Instead, it was higher (.02). The only way for the detection model to cope with this result is to assume that remember/know judgments were based on a different set of distributions than the confidence judgments. The results of this study, coupled with the remember/know meta-analyses reported by Rotello et al. (2004) and Dunn (2004), would appear to leave open the possibility that the remember/know ROC slope is greater than the confidence-based ROC slope, despite of the data summarized in Tables 3 and 4. It is, after all, possible that had confidence ratings been taken in the many remember/know studies that were included in the meta-analyses, the average confidence-based ROC slope would have been close to 0.80.

What would the implications be if the ROC slope difference turns out to be real? As was indicated in the previous section, the evidence suggests that the location of the remember criterion exhibits item-to-item variability with respect to the confidence criteria. In the absence of such variability, the slope of the confidence-based ROC should be exactly the same as the slope of the remember/know ROC. However, if the remember criterion varies from item to item, the slope of the ROC would increase accordingly. This is not an intuitively obvious result, but the Monte Carlo simulations described in Appendix B show that variability in the location of the remember criterion serves to increase the slope of the ROC. The evidence reviewed in the prior section clearly indicates that such variability exists. That is, in most of the response protocols in Table A1 of Appendix A, there is more overlap in the confidence ratings associated with remember and know judgments than can be explained by the fixed-criterion model illustrated in Figure 4. In a few of the protocols, the overlap was considerable. If a unidimensional detection model is assumed, results such as these require the assumption of criterion variability. If that variability is due mainly to item-to-item variations in the location of the remember criterion (perhaps because the distinction between recollection and familiarity is not trivially easy for subjects to grasp), the slope of the remember/know ROC would increase accordingly.

It is important to keep in mind that the tests described above offered every opportunity for detection theory to fail in a spectacular way. If remember responses reflect recollection, an unusually low remember false alarm rate would not have been a shocking result (because not recollecting any lures is the expected outcome from the dual-process point of view). The farther the remember false alarm rate falls below its expected value relative to the confidence-based ROC, the *lower* the remember/know ROC slope will be. As such, even an ROC slope as low as 0.25 would not pose the slightest problem for this theory. But the remember false alarm rate is accurately predicted by the confidence-based ROC, and the slight departure from the expected value that may exist (i.e., the remember false alarm rate may be slightly too high) does not seem to be in the direction that a recollection account of remember responses would predict. By contrast, that possible deviation is easily understood in terms of criterion variability.

CORRELATIONAL ANALYSES

Another front in the dual-process versus signal detection interpretations of remember/know judgments involves correlations between various behavioral measures (across individuals, across conditions within a study, or across studies). In the past, some correlational analyses have related a hit rate (or a false alarm rate) to a derived measure such as B'', which is the bias measure that corresponds to the supposedly nonparametric discriminability measure A' (Donaldson, 1996; Gardiner et al., 2002). We avoid that strategy here because B'' is a theoretically peculiar measure (e.g., Macmillan & Creelman, 1996) and because its computational formula involves hit and false alarm rates. That being the case, the measures being correlated (e.g., false alarm rate vs. B'') are not statistically independent, so the theoretical significance of any correlation that might be observed is hard to interpret. Instead, we focus mainly on correlations between raw hit and false alarm rates, which are independent.

Standard signal detection considerations lead to the prediction that the hit rate and the false alarm rate will be positively correlated across individual subjects and negatively correlated across experimental conditions. These predictions apply to the overall hit and false alarm rates, *as well as to the remember hit and false alarm rates*. Although not widely appreciated, these effects are typically observed even though there is no reason at all to expect that they would be given a dual-process interpretation of remember/know judgments. In addition, across experiments, the remember, the know, and (when the option is included) the guess false alarm rates should all be highly correlated with each other, and they are.

Correlations Across Individuals

Previously reported correlations. Dobbins, Khoe, Yonelinas, and Kroll (2000) conducted an individual-subject correlational analysis that was said to pose a particular challenge for signal detection theory, but a closer look at their argument and a reanalysis of their data suggest that the opposite may be true. Figure 6 is a reproduction of Figure 1 from Dobbins et al., and it shows the basis for a specific prediction that they tested. The figure shows possible criterion settings for conservative, neutral, and liberal observers. Liberal observers have their decision criteria set more to the left than conservative observers, and as the remember criterion shifts to the left, a larger portion of the target distribution falls to the right of it. Thus, the remember hit rate (as well as the overall hit rate) should be higher for more liberal subjects. Similar considerations apply to the false alarm rates. That is, liberal observers should have a higher remember false alarm rate and a higher overall false alarm rate than do conservative observers. Thus, if the model depicted in Figure 6 is accurate, the *remember hit rate* and the *overall false alarm rate* should be positively correlated. Both should be high for liberal subjects, and both should be low for conservative subjects. No such correlation is predicted by the dual-process model, because according to that account, remember hits reflect the products of a recollection process that is unrelated to the processes that underlie false alarms.

Across individual observers, Dobbins et al. (2000) found no correlation between the remember hit rate and the overall false alarm rate. A reanalysis of the data reported by Stretch and Wixted (1998) yields the same result.

As Dunn (2004) has observed, the remember and the know criteria may shift across individuals in tandem or they may not, and neither outcome would confirm or contradict the detection interpretation of remember/know judgments. The nonsignificant correlation between the remember hit rate and the overall false alarm rate suggests that a subject who interprets the instructions for a remember response to mean that the remember criterion should be set at a higher than average point on the memory strength axis does not also tend to interpret the instructions for a know response to mean that the know criterion should be set at a higher than average point on the memory strength axis. No detection theory principle requires that it be otherwise.

Then again, it is true that detection theory does not make a definite prediction about the correlation in question, and it would have offered an easily understood explanation for the correlation had it materialized. By contrast, the dual-process interpretation of remember/know judgments specifically predicts the absence of this correlation, because responses based on recollection (i.e., Remember hits) should not be related to the nonrecollection processes that give rise to false alarms. Thus, while the data are not incompatible with a detection account (contrary to what the authors argued), it does seem fair to say that the nonexistent correlation is a point in favor of the dual-process view.

Dobbins et al. (2000) also tested the predicted positive correlation between the overall hit rate and the overall false alarm rate. This prediction is not based on a consideration of the *relationship* between the remember and know criteria but, instead, arises from a consideration of the know criterion itself (the location of which determines the overall hit and false alarm rates). A positive correlation between hit and false alarm rates is predicted only if subjects differ in the degree of bias evident in their response protocols (as illustrated in Figure 6). In practice, individuals do vary considerably in the degree to which they are biased, so much so that a strong positive correlation between overall hit and false alarm rates is, indeed, typically observed. It is true of all three experiments reported by Dobbins et al. and is true of the data reported by Stretch and Wixted (1998) as well. The relevant correlational statistics are presented in the first column of Table 5. Dobbins et al. argued that although detection theory is consistent with the presence of this

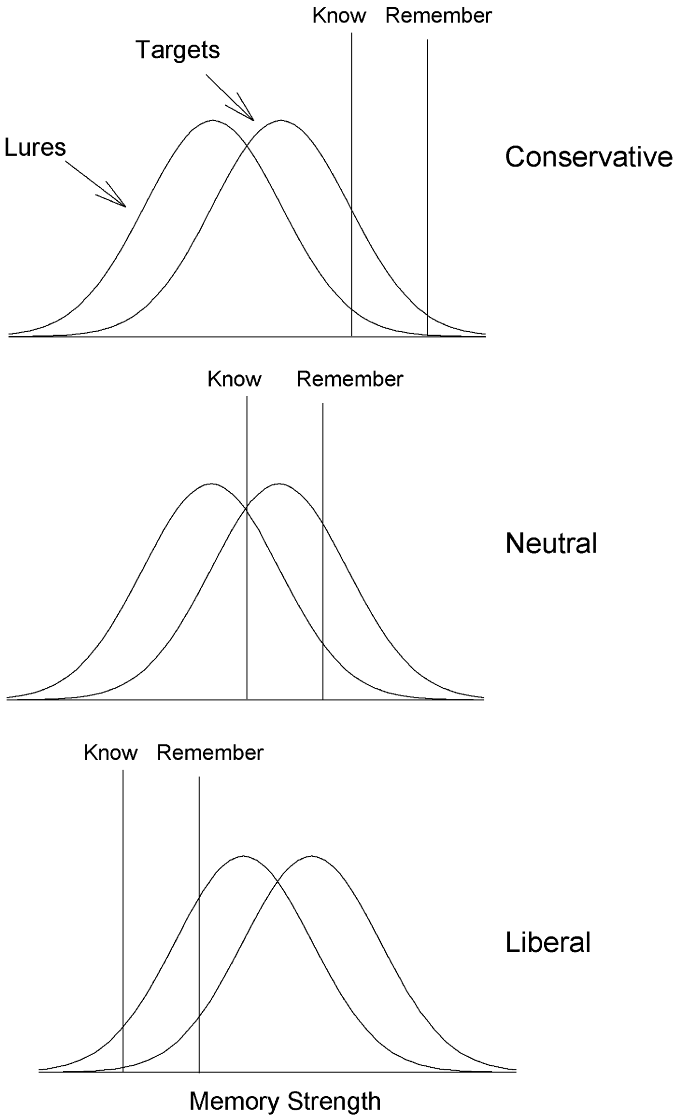

Figure 6. Signal detection interpretation of conservative, neutral, and liberal remember/know observers.

correlation, it is inconsistent with the nonsignificant correlation between the remember hit rate and the overall false alarm rate, discussed above. By contrast, both of these outcomes were shown by Dobbins et al. to match the predictions of their dual-process account, which assumes a detection process for know responses only.

Newly computed correlations. Two additional correlations are especially relevant to the detection account of remember/know judgments. One involves the correlation between the remember hit and false alarm rates, and the other involves the correlation between the know hit and false alarm rates. Neither of these correlations was reported by Dobbins et al. (2000), but the authors of that study provided us with their data, and we will analyze them next.

If conditions are such that variations in the placement of the know criterion yield a significant correlation between the overall hit and false alarm rates (and they are), then, if remember responses are based on a high criterion, it is reasonable to expect the same to be true of the effects of variations in the placement of the remember criterion on remember hit and false alarm rates. Subjects who place the remember criterion high on the memory strength axis ought to have a low remember hit rate and a low remember false alarm rate. The opposite should be true of subjects who place the remember criterion low on the memory strength axis.

Hirshman and Henzler (1998) showed that when subjects are induced to shift the remember criterion across conditions by means of instructions, the remember hit and false alarm rates covary. Results such as these suggest that remember judgments are, indeed, based on a movable decision criterion, as Donaldson (1996) originally proposed. Although this evidence seems decisive to

Table 5
Individual Subject Correlations between Overall Hit and False Alarm Rates (Hit vs. FA), Remember Hit and False Alarm Rates ($\mathrm{R_{Hit}}$ vs. $\mathrm{R_{FA}}$), Know Hit and False Alarm Rates (K_{Hit} vs. K_{FA}), and Corrected Remember Hit and False Alarm Rates (R vs. R_{FA}) From Two Remember/Know Studies (Dobbins et al., 2000, and Stretch & Wixted, 1998)

	Hit vs. FA	R_{Hit} vs. R_{FA}	K_{Hit} vs. K_{FA}	R vs. R_{FA}
Dobbins, Khoe, Yonelinas, and Kroll(2002)				
Experiment 1	.558*	.514*	.743*	.272
Experiment 2	.829*	.747*	.884*	.481*
Experiment 3	.445*	.526*	.698*	.449*
Stretch and Wixted (1998)				
Experiment 1	.400*	.468*	.733*	.317**
Experiment 2	.563*	.386*	.626*	.242
Experiment 3	.511*	.475*	.566*	.386*
Experiment 4	.487*	.427*	.441*	.265

*$p < .05$. **$p < .06$.

us, Hirshman and Henzler's study has been criticized for having induced subjects to make remember responses on the basis of something other than recollection, thereby accounting for the high remember false alarm rates observed in the more liberal conditions (Gardiner et al., 2002). Perhaps, but the same kind of correlation between remember hit and false alarm rates may arise quite naturally without inducing a criterion shift of any kind. Individual subjects already differ in how liberal or conservative they are, so the remember hit rate may correlate with the remember false alarm rate for that reason alone.

As shown in the second column of Table 5, the remember hit rate is, indeed, strongly (and significantly) correlated with the remember false alarm rate in all three experiments reported by Dobbins et al. (2000) and in all four experiments reported by Stretch and Wixted (1998). This positive correlation does not seem to have resulted from a few subjects who happened to have high remember hit and false alarm rates (and who might be suspected of having misunderstood the instructions). When the correlation was run on the 10 subjects with the lowest remember false alarm rates (ranging from .01 to .03) from Experiment 1 of Stretch and Wixted, it was still positive ($r = .65$, $p < .05$). Given the robustness of this correlation, it seems likely to be evident in most of the remember/know studies that have been performed to date. Thus, in the future, authors should specifically test for the presence of this correlation and then grapple with its theoretical significance.

It also follows from all of this that the know hit rate should correlate positively with the know false alarm rate. That is, if the remember and the know criteria are uncorrelated across individuals (as the evidence suggests is the case), we may, for the moment, think of the remember criterion as being fixed and ask what the effect would be on the know hit and false alarm rates as the location of the know criterion varies across individuals. The know hit rate corresponds to the area under the target distribution *between* the know and the remember criteria, and the know false alarm rate corresponds to the area under the lure distribution between the know and the remember criteria. As the know criterion shifts to the left, both areas will increase. Thus, the know hit and false alarm rates should vary together across individuals. As shown in the third column of Table 5, this correlation is highly significant in every case.

The crucial point to make here is that both know responses and remember responses exhibit clear criterion effects. That is, as the relevant criterion changes, hit rates (whether remember or know) *and* false alarm rates change together, a point that has not been previously appreciated. What are we to make of this correlation if the dual-process interpretation of remember/know judgments applies? According to the threshold model advanced by Yonelinas and his colleagues (Yonelines et al., 1998), familiarity-based know responses arise from a signal detection process, so the criterion effects observed for know responses are explained in essentially the same manner as the one we are advocating. Recollection-based remember responses, however, theoretically arise from a threshold process (i.e., the subject either does or does not recollect the item). In addition, subjects are assumed to sometimes use the remember option when they are simply guessing, which is why remember false alarms are occasionally observed. According to this way of thinking, remember criterion effects are evident because subjects who guess a lot would have high remember hit and false alarm rates, whereas those who are reluctant to guess would have lower remember hit and false alarm rates.

Except for its high-threshold flavor, the guessing explanation of criterion effects is not unlike the explanation offered by signal detection theory. That is, variations in the rate of remember guessing across subjects have essentially the same effect as variations in the location of the remember criterion across subjects. A complication for the high-threshold account is that if all remember false alarms are guesses, it seems odd that these responses are made quickly and with high confidence (as was shown earlier). It is much easier to imagine that a guess would be made with low confidence precisely because it is a guess. Still, if we accept the possibility that remember false alarms reflect high-confident guesses, we can use the high-threshold correction-for-guessing formula to obtain a pure estimate of recollection. Once corrected, the remember hit rate (now, theoretically, a pure measure of recollection) would not be expected to still correlate with the remember false alarm rate (a pure measure of the rate of guessing).

To investigate this, the remember hit rate for individual subjects in the three experiments reported by Dobbins et al. (2000) and in the four experiments reported by Stretch and Wixted (1998) were corrected, using the following standard formula:

$$R = \left(R_{\mathrm{Hit}} - R_{\mathrm{FA}}\right) / \left(1 - R_{\mathrm{FA}}\right),$$

where R represents the corrected remember hit rate for an individual subject, R_{Hit} represents the observed remember hit rate for that subject and R_{FA} represents the remember false alarm rate. This estimate of true recollection is based on standard high-threshold assumptions, according to which the remember hit rate is based on a combination of true recollection and guessing, whereas the remember false alarm rate is based on guessing only (Yonelinas et al., 1996; Yonelinas et al., 1998).

According to the high-threshold model, there is no reason to expect that the probability of true recollection would be correlated with the probability of making a remember response on the basis of a guess. According to signal detection theory, the above correction formula would not necessarily be expected to eliminate the positive correlation (because the formula is not based on detection theory considerations).

Once each subject's score was corrected in this way, we correlated R with R_{FA}. As shown in the fourth column of Table 5, a positive correlation was still evident in every case, and it was statistically significant (or marginally so) in four of seven cases. These results offer compelling evidence for the signal detection interpretation of remember/know judgments, and they pose another difficult challenge for the idea that remember responses provide a pure measure of recollection. Why would the probability of guessing (if that is what a remember false alarm represents) correlate so strongly with the probability of pure recollection?

The positive correlation between remember hit and false alarm rates also raises problems for the idea that remember false alarms reflect false occurrences of all-or-none recollection. Why would subjects who have a high degree of recollection also be prone to the formation of false memories? Such an outcome is possible, of course, but it would be just as easy to imagine that the opposite would occur. Perhaps, then, recollection is a continuously distributed variable, so that a criterion is needed to decide whether enough recollective detail has been retrieved to warrant a remember response. In that case, a high remember false alarm rate would not indicate that the subject was especially prone to the formation of false memories. Instead, it would suggest a liberal setting for the remember criterion. This explanation is the same as that offered by the simplest detection account, except that it assumes that *old*/*new* recognition decisions are process pure even though recollection and familiarity are continuously distributed and imperfect indicators of prior occurrence. In that case, however, it would be inefficient to respond on the basis of either recollection or familiarity when making an *old*/*new* recognition decision (although Sherman, Atri, Hasselmo, Stern, & Howard, 2003, have proposed just such a model). Still, such a model is consistent with the observed correlations between hit and false alarm rates.

The Remember Mirror Effect

The signal-detection–based predictions discussed above arise because individual subjects exhibit different degrees of bias (with some being quite liberal and others being quite conservative). Across individuals, those differences in response bias entail a positive correlation between the hit rate and the false alarm rate. Across conditions that differ in strength, however, the opposite effect is usually observed (Glanzer, Adams, Iverson, & Kim, 1993). When the distance between the target and the lure distributions changes, the average criterion changes accordingly, and this gives rise to the well-known mirror effect (which entails a *negative* correlation between hit and false alarm rates) across experimental conditions. The generic detection-based interpretation of the mirror effect is illustrated in Figure 7. This tendency for the criterion to shift as a function of d' is not especially pronounced across subjects, which is why differences across subjects in the level of bias they exhibit yields a positive correlation between hit and false alarm rates. But wherever subjects happen to place the decision criterion in a weak memory condition, they tend to shift it farther to the right on the decision axis in a strong memory condition. That average shift is what is illustrated in Figure 7.

The mirror effect is a phenomenon that theoretically reflects a change in the location of the *old*/*new* (i.e., know) criterion as a function of strength. If remember responses are criterion based as well, one might expect to see a remember mirror effect for the same reason. After all, the remember criterion is, theoretically, just another criterion situated on the memory strength axis, but it is not otherwise fundamentally different from the know criterion.

Figure 8 illustrates the prediction of interest. The upper panel illustrates a weak (low d') condition, and the lower panel illustrates a strong (high d') condition. As before, the know criterion is placed at a point farther to the right on the memory strength axis in the strong condition, as compared with the weak condition, thereby giving rise to the standard mirror effect. If the remember criterion shifts across conditions in the same way, one would expect to see a remember mirror effect as well. This is not a necessary prediction of detection theory, but it is a natural one.

Keep in mind that this figure represents average criterion placements. Data reviewed earlier already suggest that individual subjects do not place the remember and the know criteria on the memory strength axis in correlated fashion. But wherever subjects place their remember and know criteria, they may place them higher on the memory strength axis in a strong condition than in a weak condition. If so, an overall mirror effect, as well as a remember mirror effect, ought to be observed.

A survey of many relevant remember/know studies suggests that the remember mirror effect is as universal as the standard mirror effect. Yonelinas (2002) reviewed a wide range of remember/know studies in which strength was manipulated in a way that bears on the predicted remember mirror effect. To observe a strength-based mirror effect, strength must be manipulated across lists. If strength is manipulated within a list (e.g., half the items receive deep processing and half shallow processing), the

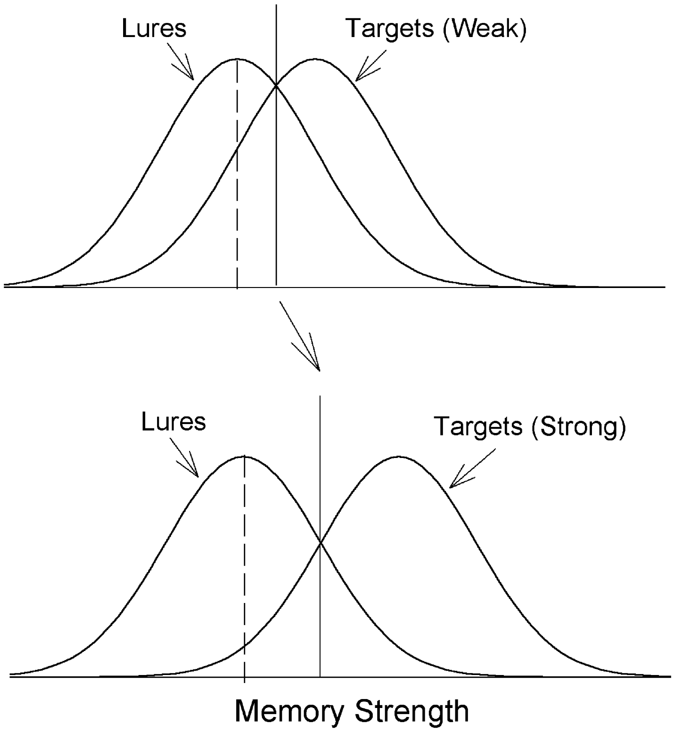

Figure 7. Signal detection interpretation of the mirror effect.

ensuing recognition test typically involves a mixture of the strong and weak targets randomly intermixed with lures. As such, only one false alarm rate is procedurally possible. If strength is manipulated across lists (e.g., one list receives deep processing and the other shallow processing), by contrast, independent false alarm rates can be obtained after each list. This allows one to see whether a higher remember hit rate is associated with a lower remember false alarm rate.

Of the remember/know experiments reviewed by Yonelinas (2002), strength was manipulated across lists by manipulating levels of processing in two studies, by dividing attention during encoding in three studies, by manipulating study duration in four studies, and by manipulating retention interval in eight studies. Out of these 17 cases, 13 show clear evidence of a remember mirror effect, 2 show no change in the remember false alarm rate across conditions, and 2 show an effect in the wrong direction. A representative study was reported by Gregg and Gardiner (1994), in which they manipulated levels of processing across lists. Not surprisingly, accuracy was much higher in the deep condition than in the shallow condition (the d' values were 2.13 and 0.97, respectively), and an overall mirror effect was observed. That is, the overall hit rate was higher in the deep condition than in the shallow condition (.88 vs. .65, respectively), and the overall false alarm rate was lower (.17 vs. .28, respectively). Of more interest for present purposes was the fact that a mirror effect was observed for the remember responses as well. That is, the remember hit rate was higher in the deep condition than in the shallow condition (.61 and .31, respectively), and the remember false alarm rate was lower (.01 and .05, respectively). Higham and Vokey (2004) also noted systematic effects of such variables as word frequency on the remember false alarm rate: Low-frequency words yield a higher remember hit rate and a lower remember false alarm rate than do high-frequency words.

The remember mirror effect is readily understood in terms of signal detection theory (as illustrated in Figure 8), but what are its implications if a remember response is taken to reflect recollection? It is not clear how lures can be recollected in the first place, and it is hard to imagine that, even if they can be falsely recollected (e.g., because they were associatively activated during study), more lures would be recollected in the *shallow* encoding condition than in the deep encoding condition. The data are most easily interpreted in terms of a model that assumes that the remember criterion shifts across strength conditions in a manner similar to the way in which the *old*/*new* criterion is already known to shift. Thus, these findings add to the growing body of evidence suggesting that a decision criterion is involved in remember responses (Hirshman & Henzler, 1998; Rotello et al., 2004). Once one allows for this possibility, the idea that *old*/*new* recognition decisions are process pure becomes more difficult to sustain. Why, for example, would a subject ignore the fact that an item is quite familiar just because a moderate amount of recollective detail associ-

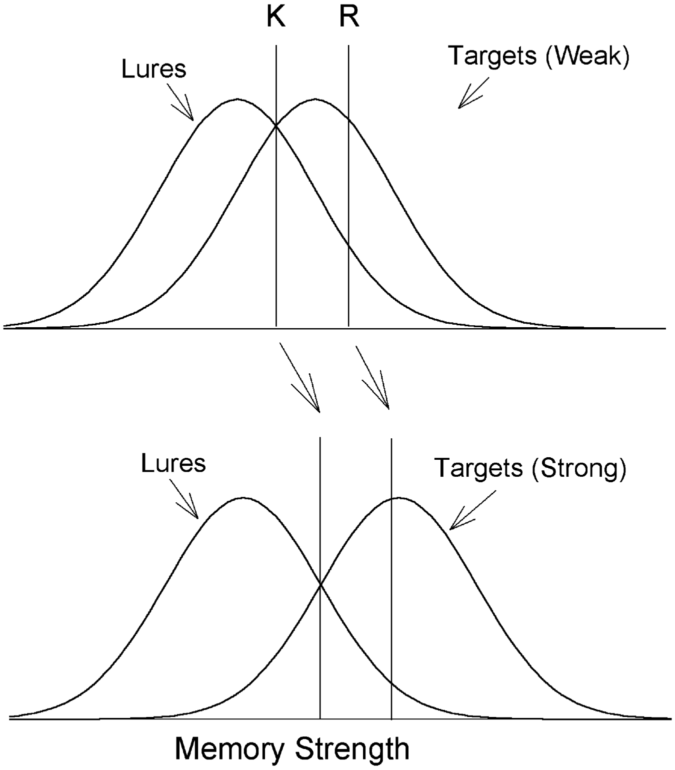

Figure 8. Signal detection interpretation of the remember mirror effect. K, know; R, remember.

ated with the item has been retrieved? A moderate amount of recollection coupled with a reasonably high level of familiarity would point strongly to the prior occurrence of the item on the list (more strongly than would either process considered alone). It seems reasonable to suppose that subjects would learn that this is true on the basis of a lifetime of experience making recognition decisions.

The Correlation Between Remember and Know (and Guess) False Alarm Rates

The predictions discussed above were concerned with the relationship between hit rates and false alarm rates. However, detection theory leads to the expectation that there will be orderly relations among various false alarm rates as well. In the studies just mentioned, for example, both the overall false alarm rate (a component of the traditional mirror effect) and the remember false alarm rate (a component of the newly identified remember mirror effect) tended to decrease when *d′* increased, and vice versa. Thus, averaged across subjects, the remember and the know criteria appeared to shift in tandem across conditions (approximately). If so, it should also be the case that the remember false alarm rate and the know false alarm rate will be positively correlated across conditions.[1] When the criteria move to the left in tandem, both the area under the lure distribution to the right of the remember criterion (which corresponds to the remember false alarm rate) and the area of the lure distribution *between* the know and the remember criteria (which corresponds to the know false alarm rate) increase. This holds true so long as the know criterion does not shift so far left that it falls below the mean of the lure distribution (in which case, the overall false alarm would exceed 50%, which is very rare). Across the 17 studies that were analyzed above for evidence of a remember mirror effect, it was, indeed, typical to find that the remember and the know false alarm rates were positively correlated across conditions. In Gregg and Gardiner's (1994) study, for example, the know false alarm rates in the deep and the shallow conditions were .16 and .23, respectively, and the corresponding remember false alarm rates were .01 and .05.

Interestingly enough, this positive correlation between the remember false alarm rate and the know false alarm rate is not only evident across conditions that vary in strength within studies; it is also evident across studies that differ in many respects (type of experimental manipulation, subject population, etc.). Donaldson (1996) reviewed 80 remember/know studies and Gardiner et al. (2002) reviewed 86 remember/know/guess studies, and the relevant false alarm rates were highly correlated in each meta-analysis. The correlation between the remember and know false alarm rates in Donaldson's meta-analysis was .662 (r^2 = .387). For Gardiner et al.'s (2002) meta-analysis, it was .531 (r^2 = .282). Both of these correlations were highly significant. In fact, all of the relevant false alarm rates were highly correlated in Gardiner et al.'s (2002) data set. The correlation between the know false alarm rate and the guess false alarm rate across the 86 studies was .489 ($r^2 = .239$), and the correlation between the remember false alarm rate and the guess false alarm rate was .593 ($r^2 = .351$).

These false alarm rate correlations all make sense if one assumes that the average remember and know (and guess, where applicable) criteria shift more or less in tandem across studies.[2] How a dual-process interpretation of remember/know judgments might explain these correlations is not entirely clear. If remember false alarms reflect false recollections, why would the rate of false memory formation in a given study correlate with the rate of incorrectly *guessing*? If remember false alarm rates reflect guesses rather than false memories, why would subjects continue to make remember false alarms even when a guess option is provided for them?

A MULTIPROCESS SIGNAL DETECTION MODEL

Remember/know theorists argue that recognition memory is based on at least two processes, recollection and familiarity (Mandler, 1980). Signal detection theorists usually argue that recognition memory is based on a unidimensional strength-of-memory variable. The fact that a unidimensional axis is assumed to underlie recognition decisions might create the impression that a detection model is necessarily a single-process model. This, however, is untrue. Recollection may be a graded phenomenon (as familiarity is), not an all-or-none phenomenon (e.g., Kelley & Wixted, 2001; Qin, Raye, Johnson, & Mitchell, 1999; Rotello et al., 2004; Sherman et al., 2003). If so, the idea that responding is based *either* on recollection *or* on familiarity seems peculiar. If partial recollective information is available, it would make sense for the subject to combine that information with whatever level of familiarity is available to arrive at a sense of prior occurrence (Kelley & Wixted, 2001). And that combined memory signal would be what a detection model denotes as *memory strength*. For a subject to instead consider either partial recollection or familiarity when making a decision about a particular test item would be like a juror who considers either one piece of evidence (e.g., eyewitness testimony) or another piece of evidence (e.g., fingerprints), but never both together, when assessing the strength of evidence pointing to the defendant's guilt. It can be done, but it would be odd nonetheless.

Prior debates in the remember/know literature have been concerned with such issues as whether the memory processes that underlie remember and know judgments are mutually exclusive or independent. Some argue that memories are encoded in such a way that they will later be retrieved on the basis of one process or the other (Gardiner, 1988). Others argue that the two processes are not mutually exclusive, so that if the decision for a particular item had not been made on the basis of recollection, there is some chance that it would have been based on familiarity (Yonelinas et al., 1996). Although the mutually exclusive versus independent ways of thinking differ in important ways, they agree on a key point that we dispute. Specifically, they agree that for a particular test item, subjects base their decisions either on one process or on the other. What follows is a concrete example of a model that instead assumes that both recollection and familiarity contribute to *old*/*new* recognition decisions at the level of the individual test item.

A Unidimensional Dual-Process Signal Detection Model

Basing recognition decisions on a combined recollection/familiarity signal is entirely compatible with the standard unidimensional signal detection model depicted in Figures 1–4. To make this more concrete, consider a multiprocess model that assumes that recognition memory decisions are made on the basis of a strength variable that is equal to the sum of a baseline strength (e.g., preexperimental familiarity) and experimentally added strength. The added strength may come in the form of increased familiarity, increased recollective strength, or both. Assume that baseline memory strength is distributed normally with a mean (μ_{Lure}) of 100 units and a standard deviation (σ_{Lure}) of 15 units (so that $\sigma^2_{\text{Lure}} = 225$). Further assume that when an item appears on a list, baseline strength is incremented by a familiarity value drawn from a rectangular distribution with a range from 0 to 15 (mean, μ_{Fam}, = 7.5) and by a recollective value drawn from a rectangular distribution with a range from 0 to 30 (mean, μ_{Ret}, = 15). A rectangular distribution is assumed for the sake of simplicity (e.g., it ensures no negative increments). The variance of a rectangular distribution is $(a - b)^2/12$, where a and b represent the upper and lower limits of the range, respectively. Thus, the variance of the familiarity strength distribution, σ^2_{Fam}, is $(15 - 0)^2/12$, or 18.75, and the variance of the recollective strength distribution, σ^2_{Ret}, is $(30 - 0)^2/12$, or 75.

Under this scenario, the mean strength of the target distribution will be

$$\begin{aligned}\mu_{\text{Target}} &= \mu_{\text{Lure}} + \mu_{\text{Fam}} + \mu_{\text{Ret}} \\ &= 100 + 7.5 + 15 = 122.5,\end{aligned}$$

and, assuming that the baseline, familiarity, and retrieval values are independent, the variance of the target distribution will be

$$\begin{aligned}\sigma^2_{\text{Target}} &= \sigma^2_{\text{Lure}} + \sigma^2_{\text{Fam}} + \sigma^2_{\text{Ret}} \\ &= 225 + 18.75 + 75 = 318.75.\end{aligned}$$

Thus, whereas the standard deviation of the lure distribution is 15, the standard deviation of the target distribution would be $\sqrt{318.75} = 17.85$. The slope of the ROC that would be generated by this model is $\sigma_{\text{Lure}}/\sigma_{\text{Target}}$, which works out to 0.84. Monte Carlo simulations based on this simple model yield z-ROC data that are very nearly linear and that have a slope of about 0.84. A summation rule (i.e., memory strength = familiarity + recollection) is assumed merely because it is the simplest combination rule that subjects might rely on, but other combination rules might turn out to be more appropriate.

According to this model, most items that exceed the remember criterion would be associated with considerable recollection. That is, in most cases, one reason why the item is high enough in strength to exceed the remember criterion is that the recollective component is strong. Thus, it would still be reasonable to interpret remember responses to critical lures in the DRM paradigm (for example) as evidence in favor of false recollection. However, because there are two sources of memory strength (recollection and familiarity), this would not be a necessary interpretation, because it is possible for some items to fall above the remember criterion mainly on the basis of one or the other process. An especially familiar lure, for example, might be given a remember response even though no recollective strength is involved. Thus, on this view, remember responses are not process pure; instead, they denote items that are high in strength because either the recollective component or the familiarity component (or both) is high in strength.

Similarly, items that receive a know response are not process pure, because the associated memories are not, in this model, devoid of recollective detail. Instead, a know response denotes an item that is low in strength because neither the recollective component nor the familiarity component is strong. Even so, some recollective detail may be present. In agreement with this notion, Hicks, Marsh, and Ritschel (2002) showed that subjects often make accurate source judgments (which depend on recollection) for items that receive a know response.

In light of these considerations, it should be clear that to advocate a detection interpretation of remember/know judgments is not necessarily to call into question a dual-process interpretation of recognition memory; instead, what is called into question is the idea that remember responses reflect a relatively pure measure of recollection and that know responses reflect a relatively pure measure of an independent familiarity (or semantic memory) process. More generally and more interestingly, the very idea that *old/new* recognition memory itself is process pure is called into question. If recognition memory is not process pure, the remember/know methodology may not be the most profitable way to investigate the familiarity and recollective processes that underlie recognition. However, other methods, such as those once used by Mandler (1980), still seem well suited to the task, because those methods do not rely on the accuracy of subjective reports (reports that cannot be especially accurate if memory itself is not process pure). To take just one example, Mandler found that the degree to which a list is semantically organized predicts recognition performance after a long retention interval but not after a short retention interval. Because semantic organization influences recollection, but not familiarity, results such as these suggest that recognition performance following a short retention interval has a large familiarity component, whereas performance following a long delay is determined largely by recollection. Mandler interpreted results such as these to mean that the familiarity component fades rapidly with time, whereas the recollection component is more durable. The important point to make here is that Mandler's methods of investigating dual-process theory of recognition memory are compatible with the dual-process unidimensional signal detection model outlined above.

Multidimensional Signal Detection Models

Several multidimensional detection models have been advanced in recent years, and it is worth considering how those models relate to the dual-process unidimensional detection account of remember/know judgments that we have discussed up to this point. Banks (2000) proposed a multidimensional detection model that assumes that the relevant decision variable depends on what the subject is asked about the prior status of the test item. For example, imagine that subjects *see* one list of words and then *hear* another list. These lists could be followed by a standard *old/new* recognition test ("did this word appear on either of the prior lists?") and then, for items that are declared to be old, by a source memory test ("was the item seen or heard?"). These are two quite different questions about the prior status of the test item.

In the initial *old/new* recognition task, any source information that is recollected (e.g., "I remember hearing this word") would add to evidence that the item was previously encountered. So would any nonsource information that is recollected (e.g., "I clearly recollect thinking about the ocean when this word was presented"), and so would any familiarity that the item generates. That is, as we see it, the relevant decision variable would be *memory strength*, and that strength would be composed of both recollection and familiarity.

If, after declaring an item to be old, subjects are now asked to indicate the source of the test item (e.g., "was this item on the seen list or the heard list?"), the decision variable would need to change. No longer would it suffice to base one's decision on a summed strength variable composed of recollective detail (whether auditory or visual) and familiarity. Whether previously seen or previously heard, the items are equally strong (on average), so relying on memory strength would be of no use. Instead, to answer the source question, the nature of the recollective detail becomes critical, and the familiarity of the test item becomes irrelevant. As such, the subject might now attend only to source-relevant recollective detail, and the decision variable might change from aggregate strength to, say, the difference between seen and heard recollective detail. If the amount of seen detail that is recollected exceeds the amount of heard detail that is recollected, the source decision would be "seen"; otherwise, it would be "heard." Note that this is a different decision variable than the memory strength variable that applies when the question concerns whether or not the item was encountered on a previous list. The work reported by Banks (2000) suggests that the decision variable does change when the task (*old*/*new* recognition vs. source identification) changes.

For a remember/know task, the subject's job is simply to indicate the basis for the *old*/*new* decision that was just made, not to answer a new question pertaining to the prior status of the test item. As such, subjects need not switch to a new decision variable to get the "right" answer, although they could. Our unidimensional model assumes that they do not and that, instead, they set two decision criteria along a strength-of-memory axis and respond accordingly. A newly proposed multidimensional model of remember/know judgments assumes otherwise.

Rotello et al. (2004) proposed a multidimensional account that assumes that memories vary along two dimensions, global and specific strength, which are somewhat analogous to familiarity and recollection. To make an *old*/*new* decision, subjects sum the two components of strength and then declare the item to be old if that sum exceeds a decision criterion. This is exactly like our assumption that subjects sum familiarity and recollective strength and then use that summed value to decide whether or not the item appeared on the list (cf. Kelley & Wixted, 2001). However, whereas we assume that subjects simply set two decision criteria to make remember/know judgments, they assume that subjects base their remember/know judgments on a different decision variable. In fact, their account is much like the one described above for making source judgments, in that they assume that subjects compute the *difference* between the specific and the global strengths of the test item when trying to decide whether the item was remembered or known. If that difference exceeds a criterion, subjects declare the item to have been remembered; otherwise, it is declared to be known.

As with the unidimensional model that we are advocating, this model assumes that two decision criteria are involved. Unlike the unidimensional account, it also assumes that two *decision variables* are involved, one being the sum of global and specific memory strength (G + S) and the other being the difference between global and specific memory strength (G − S). Adding that assumption increases the range of outcomes that can be accommodated, as compared with the unidimensional model. For example, it is easy for the multidimensional model to account for the differing slopes of the confidence-based and remember/know ROCs discussed earlier, and this was the main reason why Rotello et al. (2004) rejected the unidimensional model in favor of their new account. However, the increased flexibility of this model comes at a price. None of the predictions considered earlier concerning confidence and RTs associated with remember and know judgments are compelled by the multidimensional model. For example, an item for which summed strength (G + S) is high would generate a high-confident *old* response, just as in our unidimensional model. However, unlike in the unidimensional model, that same item would not necessarily be more likely to generate a remember response, because the difference between G and S is as likely to be positive as it is to be negative. Thus, the compelling relationship between confidence ratings and remember/know judgments does not follow naturally from this model. A simpler unidimensional model with criterion variability, which retains all of the predictions about confidence and RTs discussed earlier and also accommodates the slight differences in confidence-based and remember/know ROCs that may exist, seems like the most parsimonious account.

CONCLUSION

If there is one key point to emphasize, it is that remember/know theorists have not come to grips with remember false alarms or know false alarms (as Higham & Vokey, 2004, noted as well). With regard to the former, it is often simply noted that the remember false alarm rate in a particular study was low (e.g., Gardiner, Gregg, Mashru, & Thaman, 2001). Low or not, the remember false alarm rate correlates positively with the remember hit rate across individuals who differ in bias, and it correlates negatively with the remember hit rate across conditions that differ in strength. In other words, clear *criterion effects* are evident for both remember and know responses. Moreover, remember responses to *lures* are made more quickly and with higher confidence than know responses to *targets*, as if remember responses are made to high-strength items regardless of their list status. The range of findings that are easily understood in terms of signal detection theory and that pose a dilemma for dual-process accounts of remember/know judgments is summarized in Table 6. These systematic effects should be explained, not swept away by the observation that the remember false alarm rate is low. The absolute magnitude of the remember false alarm rate is not particularly relevant; what is relevant is that remember false alarms exhibit systematic effects, and those effects just happen to be the ones that are anticipated by the prevailing signal detection model of recognition memory.

To their credit, Yonelinas and colleagues have taken the important step of specifying a theory of false alarms

Table 6
Summary of Some Empirical Findings That Are Consistent With the Signal Detection Interpretation of Remember/Know Judgments

1. Remember false alarms are made with higher confidence than know hits.
2. Remember false alarms are made more quickly than know hits.
3. The overlap between remember and know confidence ratings is low.
4. Confidence-based ROCs closely predict remember/know ROCs.
5. Remember hit and false alarm rates are correlated across subjects.
6. Remember, know, and guess false alarm rates are correlated across studies.
7. Amnesics have a high remember false alarm rate.

within the remember/know paradigm (e.g., Yonelinas, 2002; Yonelinas et al., 1996; Yonelinas et al., 1998). In their view, remember false alarms are guesses (not false memories, for example), and know false alarms arise when lures are familiar enough to exceed a decision criterion (i.e., they have proposed a detection account of know false alarms). However, to construe remember false alarms as guesses encounters some difficulties (e.g., remember false alarms are made with higher confidence than know hits, and the remember false alarm rate correlates with the remember hit rate even when the latter has been corrected for guessing, using the standard high-threshold formula), but at least a coherent theory has been articulated. Generally speaking, however, dual-process remember/know theorists are silent about the nature and meaning of remember and know false alarms.

If remember false alarms do not reflect guessing, they may reflect false recollection instead. However, recollection is generally thought to be a slow process, relative to familiarity, but remember false alarms are made quickly—more quickly than know hits. In addition, the remember false alarm rate correlates with the remember hit rate across individuals, and that result does not seem to fall naturally out of the idea that remember false alarms reflect false recollections. Across studies, the remember false alarm rate correlates highly with the know and guess false alarm rates. If remember false alarms reflect false recollections, it seems odd that the rate of false recollection would correlate so highly with the rate of guessing. Finally, amnesics often have high remember false alarm rates. It does not seem likely that individuals who have difficulty forming true memories would be especially prone to forming false ones.

Although no one has proposed it, a dual-process remember/know model that can be reconciled with most of the evidence reviewed here would hold that recollection and familiarity are both continuous and imperfect decision variables, but *old*/*new* recognition decisions are process pure anyway. Both processes would be imperfect because lures, as well as targets, would give rise to varying degrees of recollection and familiarity (false recollection in the case of lures). A decision criterion would need to be set for each process, and this would account for the evident criterion effects for both remember and know judgments. Like the model described by Cary and Reder (2003), this model would also need to assume that, for each test item, subjects first attempt to make an *old*/*new* recognition decision based on recollection and then resort to familiarity only if that attempt fails (thereby accounting for the fact that remember decisions are made more quickly than know decisions). Although this model accounts for some of the basic facts, it does not clearly predict that remember false alarms would be made with higher confidence than know hits or that the confidence ratings for remember and know judgments would exhibit such a small degree of overlap.

The findings reported here are most easily reconciled with the simple notion that *old*/*new* recognition decisions are based on memory strength, which may involve a combination of recollection and familiarity, and that subjects set two decision criteria to make remember and know judgments (as Donaldson, 1996, proposed). On this view, remember false alarms should be made quickly and with high confidence, and the correlation between remember hit and false alarm rates across individuals is to be expected. As a group, amnesics should have a high remember false alarm rate for the same reason that they tend to have a high overall false alarm rate. Specifically, knowing that their memories are poor, they set a lower decision criterion so that at least some of the targets will exceed it. The price they pay for setting a lower criterion is that a high proportion of lures exceed it as well (cf. Curran et al., 1997).

According to the detection account we advocate, remember responses are associated with items that are relatively high in both recollection and familiarity, and know responses are associated with items that are relatively low in both processes. Various dissociations that have been observed for remember and know judgments may seem to suggest otherwise, but, as has been thoroughly explained elsewhere (e.g., Donaldson, 1996; Dunn, 2004; Inoue & Bellezza, 1998), behavioral dissociations between remember/know judgments do not provide compelling evidence that they tap into different processes. Such dissociations turn out to be entirely compatible with a detection account.

The same holds true for dissociations in brain activity. For example, using fMRI methodology, Eldridge, Knowlton, Furmanski, Bookheimer, and Engel (2000) found that activity in the hippocampus was higher, relative to baseline, only for hits that were accompanied by a remember response. Hits that were accompanied by a know response yielded activity comparable to that observed during quiet baseline periods. Because some have argued that the hippocampus is involved mainly in recollection (not familiarity), results such as these appear to validate the assumption that remember responses are process-pure indicators of recollective success.

A simpler possibility is that the increased brain activity associated with remember responses reflects the higher level of activity that is associated with the retrieval of strong memories. The retrieval of weaker memories—those that give rise to a know response—may not have yielded detectable hippocampal activity because that activity was measured with respect to a baseline period, one that was free of experimenter-imposed activity but that may have involved a considerable amount of encoding and retrieval nonetheless. In this regard, Stark and Squire (2001) found that activity in the medial temporal lobe was higher during quiet rest periods (which is the baseline used by Eldridge et al., 2000) than during mind-numbing tasks such as deciding whether or not numbers were odd or even. On the basis of results such as these, they concluded that "rest is apparently an active condition associated with significant cognitive activity" (p. 12765). In fact, activity during "rest" may be as high as that associated with the retrieval of the relatively weak memories that give rise to know responses. From this point of view, the results reported by Eldridge et al. (2000) imply that the retrieval of strong memories results in strong hippocampal activity, whereas the retrieval of weak memories results in weaker hippocampal activity—the kind of weak activity that is evident during mentally active rest periods. Significant hippocampal activity might have been observed even in the know condition of Eldridge et al.'s (2000) study, had they used a baseline task like the digit odd/even task.

The overall contribution of the present research is not merely to suggest that a procedure that is widely used to investigate recollection and familiarity in recognition memory is flawed. Our inquiry into the remember/know procedure also led us to offer a more positive theoretical contribution as well. Specifically, we suggest that both recollection and familiarity contribute to the recognition decisions associated with individual items, because both processes are continuous and imperfect indicators of prior occurrence (cf. Kelley & Wixted, 2001). Under such conditions, it seems odd to respond on the basis of one process or the other. Instead, it would be much more efficient to take both processes into consideration. If both processes do contribute to individual recognition decisions, *old/new* recognition memory itself is not process pure. This contrasts with the widely held view that individual recognition decisions are based either on one process or on the other. If recollection and familiarity are both continuous processes and if both contribute to recognition decisions, the two venerable views of recognition memory, dual-process theory and signal detection theory, can be easily reconciled.

REFERENCES

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

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory & Language*, **49**, 231–248.

Curran, T., Schacter, D. L., Norman, K. A., & Galluccio, L. (1997). False recognition after a right frontal lobe infarction: Memory for general and specific information. *Neuropsychologia*, **35**, 1035-1049.

Dobbins, I. G., Khoe, W., Yonelinas, A. P., & Kroll, N. E. A. (2000). Predicting individual false alarm rates and signal detection theory: A role for remembering. *Memory & Cognition*, **28**, 1347-1356.

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., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience*, **3**, 1149-1152.

Eldridge, L. L., Sarfatti, S., & Knowlton, B. J. (2002). The effect of testing procedure 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., & 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., Gregg, V. H., Mashru, R., & Thaman, M. (2001). Impact of encoding depth on awareness of perceptual effects in recognition memory. *Memory & Cognition*, **29**, 433-440.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, **18**, 23-30.

Gardiner, J. M., & Radomski, E. (1999). Awareness of recognition memory for Polish and English folk songs in Polish and English folk. *Memory*, **7**, 461-470.

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., & 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., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, **100**, 546-567.

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.

Gregg, V. H., & Gardiner, J. M. (1994). Recognition memory and awareness: A large effect of study–test modalities on "know" responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology*, **6**, 131-147.

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

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

Hicks, J. L., Marsh, R. L., & Ritschel, L. (2002). The role of recollection and partial information in source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 503-508.

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *Quarterly Journal of Experimental Psychology*, **57**, 714-744.

Hirshman, E. (1998). On the utility of the signal detection model of the remember–know paradigm. *Consciousness & Cognition*, **7**, 103-107.

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

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.

Karayianni, I., & Gardiner, J. M. (2003). Transferring voice effects in recognition memory from remembering to knowing. *Memory & Cognition*, **31**, 1052-1059.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 701-722.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 699-710.

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.

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004).Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 540-549.

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

Manns, J. R., Hopkins, R. O., Reed, J. R., Kitchener, E. G., &

Squire, L. R. (2003). Recognition memory and the human hippocampus. *Neuron*, **38**, 127-133.

Qin, J., Raye, C. L., Johnson, M. K., & Mitchell, K. J. (1999). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1110-1115.

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

Rajaram, S., Hamilton, M., & Bolton, A. (2002). Distinguishing states of awareness from confidence during retrieval: Evidence from amnesia. *Cognitive, Affective, & Behavioral Neuroscience*, **2**, 227-235.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, **83**, 190-214.

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

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.

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

Schacter, D. L., Curran, T., Galluccio, L., Milberg, W. P., & Bates, J. F. (1996). False recognition and the right frontal lobe: A case study. *Neuropsychologia*, **34**, 793-808.

Sherman, S. J., Atri, A., Hasselmo, M. E., Stern, C. E. & Howard, M. W. (2003). Scopolamine impairs human recognition memory: Data and modeling. *Behavioral Neuroscience*, **117**, 526-539.

Stark, C. E., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences*, **98**, 12760-12766.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1379-1396.

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

Wixted, J. T., & Squire, L. R. (2004). Recall and recognition are equally impaired in patients with selective hippocampal damage. *Cognitive, Affective, & Behavioral Neuroscience*, **4**, 58-66.

Xu, M., & Bellezza, F. S. (2001). A comparison of the multimemory and detection theories of know and remember recognition judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1197-1210.

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

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

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

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember–know, process dissociation, and receiver operating characteristic data. *Neuropsychology*, **12**, 323-339.

Yonelinas, A. P., Kroll, N. E. A., Quamme, J. R., Lazzara, M. M., Sauve, M., Widaman, K. F., & Knight, R. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, **5**, 1236-1241.

NOTES

[1] These false alarm rates are not independent. A remember false alarm rate of 100%, for example, necessarily implies a know false alarm rate of 0%. But this lack of independence tends to induce an artifactual negative correlation between the various false alarm rates, not the predicted (and observed) positive correlation.

[2] Procedural and subject population differences lead to large differences in the degree of bias evident across experiments. As was discussed earlier in connection with individual subject correlations, differences in bias tend to introduce a positive correlation between hit and false alarm rates. As such, mirror effects that might be clearly observed across strength conditions within individual studies (which hold variables other than strength constant) would be obscured by differences in bias evidence across different studies. Such cross-experiment bias differences would not cancel out the expected positive correlation between remember and know false alarm rates; instead, it would serve only to enhance it.

(*Continued on next page*)

APPENDIX A

Table A1
Distribution of Confidence Ratings Associated With Remember (R) and Know (K) Judgments for Each Subject in Experiment 1 of Stretch and Wixted (1998)

Note—All of the responses are "old" responses, and the data are collapsed across hits and false alarms. Entries in boldface type correspond to the predictions of the detection model illustrated in Figure 4.

		Confidence Rating (Weak)					Confidence Rating (Strong)				
Subject	Judgment	1	2	3	4	5	1	2	3	4	5
1	R	**0**	**0**	**3**	**5**	**10**	0	0	3	3	19
	K	**0**	**0**	**1**	**0**	**0**	0	0	0	0	0
2	R	0	0	1	3	3	**0**	**0**	**0**	**9**	**18**
	K	0	3	7	7	0	**0**	**4**	**7**	**5**	**0**
3	R	0	1	6	7	11	0	8	2	4	30
	K	0	2	10	2	0	2	3	3	0	0
4	R	3	2	5	5	16	0	0	1	2	30
	K	1	3	8	4	0	0	3	7	1	0
5	R	0	0	1	3	11	**0**	**0**	**0**	**0**	**13**
	K	1	5	16	6	0	**0**	**3**	**2**	**5**	**0**
6	R	0	0	2	5	20	0	0	3	9	35
	K	0	0	0	0	0	0	0	0	0	0
7	R	0	0	2	3	22	0	0	2	0	37
	K	3	9	10	4	1	0	1	5	4	0
8	R	**0**	**0**	**0**	**2**	**17**	**0**	**0**	**0**	**3**	**31**
	K	**0**	**2**	**4**	**1**	**0**	**0**	**2**	**4**	**1**	**0**
9	R	**0**	**0**	**0**	**6**	**4**	0	0	9	12	24
	K	**1**	**0**	**7**	**0**	**0**	0	1	3	1	0
10	R	0	0	5	4	17	0	0	3	2	26
	K	0	6	7	1	1	0	1	7	6	0
11	R	**0**	**0**	**0**	**0**	**7**	0	1	3	2	7
	K	**0**	**6**	**4**	**1**	**0**	0	5	7	1	0
12	R	0	0	4	3	14	0	0	1	3	35
	K	0	2	5	6	0	0	0	7	4	1
13	R	**0**	**0**	**0**	**4**	**12**	0	0	1	5	34
	K	**1**	**3**	**11**	**6**	**0**	0	0	2	1	0
14	R	**0**	**0**	**0**	**1**	**27**	**0**	**0**	**0**	**3**	**19**
	K	**0**	**0**	**4**	**1**	**0**	**0**	**0**	**1**	**1**	**0**
15	R	**0**	**0**	**0**	**6**	**6**	**0**	**0**	**0**	**5**	**6**
	K	**0**	**0**	**8**	**3**	**0**	**0**	**0**	**2**	**4**	**0**
16	R	0	0	2	4	7	0	0	1	2	12
	K	5	12	3	5	0	1	6	1	1	0
17	R	0	0	1	3	2	0	2	1	9	10
	K	8	4	10	1	0	5	3	16	4	0
18	R	0	1	8	8	23	0	1	10	7	28
	K	0	0	12	3	1	0	0	4	4	2
19	R	0	0	6	10	18	0	0	1	9	25
	K	2	3	5	1	0	1	2	7	2	0
20	R	0	0	6	12	10	**0**	**0**	**3**	**5**	**9**
	K	0	5	3	4	0	**0**	**1**	**5**	**0**	**0**
21	R	0	0	1	6	32	**0**	**0**	**0**	**1**	**39**
	K	0	0	2	1	0	**0**	**1**	**0**	**0**	**0**
22	R	0	1	10	4	17	1	1	5	13	28
	K	0	1	1	0	0	0	3	5	0	0
23	R	**0**	**0**	**0**	**0**	**11**	**0**	**0**	**0**	**2**	**21**
	K	**0**	**0**	**4**	**8**	**0**	**0**	**2**	**2**	**5**	**0**
24	R	0	1	11	7	9	**0**	**0**	**6**	**12**	**12**
	K	0	9	7	3	0	**0**	**12**	**3**	**0**	**0**
25	R	**0**	**0**	**0**	**0**	**19**	**0**	**0**	**0**	**1**	**23**
	K	**0**	**2**	**5**	**3**	**0**	**0**	**0**	**1**	**3**	**0**
26	R	**0**	**0**	**0**	**0**	**21**	**0**	**0**	**0**	**0**	**24**
	K	**0**	**0**	**5**	**4**	**0**	**0**	**0**	**2**	**5**	**0**
27	R	0	0	1	1	12	0	0	1	13	14
	K	7	7	7	3	0	5	10	6	3	0
28	R	**0**	**0**	**0**	**1**	**5**	0	0	1	10	14
	K	**0**	**0**	**13**	**4**	**0**	0	1	19	3	0
29	R	0	0	3	7	22	0	0	3	7	31
	K	0	2	7	4	0	0	1	10	1	0

APPENDIX A (Continued)

Table A1 (Continued)

		Confidence Rating (Weak)					Confidence Rating (Strong)				
Subject	Judgment	1	2	3	4	5	1	2	3	4	5
30	R	0	0	7	4	7	**0**	**0**	**0**	**8**	**19**
	K	0	7	9	1	0	**0**	**1**	**7**	**0**	**0**
31	R	0	0	1	4	12	0	0	3	9	20
	K	0	3	19	13	0	0	0	4	9	0
32	R	0	2	12	9	32	0	1	0	4	26
	K	0	3	4	1	0	0	1	8	3	0
33	R	0	0	0	1	18	0	0	0	1	15
	K	0	1	6	0	1	0	8	5	0	1
34	R	**0**	**0**	**1**	**0**	**11**	**0**	**0**	**2**	**1**	**4**
	K	**0**	**1**	**2**	**0**	**0**	**1**	**4**	**0**	**0**	**0**
35	R	0	1	2	1	36	0	1	1	0	43
	K	0	0	1	0	0	0	0	1	0	0
36	R	2	2	8	3	15	2	3	5	6	10
	K	2	5	17	2	0	1	3	13	1	0

APPENDIX B

The idea that the confidence-based ROC and the remember/know-based ROC should have the same slope is focused on the implicit assumption that the confidence criteria and the remember/know criteria remain fixed with respect to each other throughout the recognition test. This is, of course, the simplest assumption, but the evidence reviewed above suggests that the location of the remember criterion exhibits variability with respect to the confidence criteria, and we make the assumption that this occurs mainly because of item-to-item variability in the location of the remember criterion (not the confidence criteria). To determine what the ramifications of such item-to-item variability might be, we performed a simple Monte Carlo simulation experiment. For this simulation, the mean of the target distribution was 1.65 standard deviation units above the mean of the lure distribution, and the standard deviation of the target distribution was 1.25 times that of the lure distribution. Confidence criteria were placed 0.93, 0.98, 1.10, 1.60, and 2.10 standard deviations above the mean of the lure distribution, and these remained fixed throughout (i.e., no variability in the confidence criteria was assumed). The know criterion was always placed at 0.93, but the placement of the remember criterion varied from item to item. More specifically, the value for the location of the remember criterion was drawn from a normal distribution with a mean of 1.7 and a standard deviation of σ_{Remember} (the value of which was varied across simulations). All of these values were chosen because, when set that way, the simulation yielded average data much like those reported by Stretch and Wixted (1998).

Each run of the simulation involved randomly drawing 5,000 values from the target and lure distributions and then assigning a response (*old*/*new*), a confidence rating (1–5), and a remember/know judgment depending on the value of the selected item. For example, if a value drawn from the target distribution happened to be 1.8, it would be declared *old* with a confidence rating of 4 and would be given a remember judgment. ROC slopes were then computed on the basis of the confidence-based hit and false alarm rates and the remember/know hit and false alarm rates. The whole process was repeated 10 times.

For the first 10 runs of the simulation, σ_{Remember} was set to 0. Thus, the confidence criteria and the remember/know criteria were fixed throughout the course of the simulated recognition test. In this case, both the confidence-based and the remember/know ROCs should have an average slope of 0.80. In fact, the average remember/know and confidence ROC slopes were quite close to the expected value (0.795 and 0.797, respectively). For the next 10 runs, σ_{Remember} was set to 0.20. This means that the location of the remember criterion varied about its mean of 1.7 from trial to trial but all other criteria were fixed. A restriction was included that the location of the remember criterion must exceed that of the know criterion, but it was otherwise free to vary. For all 10 runs of this simulation, the remember/know ROC slope exceeded the confidence-based ROC slope. The mean values were 0.834 and 0.789, respectively. For the last 10 runs, σ_{Remember} was set to 0.40. Now, the location of the remember criterion varied considerably from trial to trial. As before, the remember/know ROC slope exceeded the confidence-based ROC slope in all 10 cases, and the mean values were 0.907 and 0.803, respectively. Thus, as the variability of the remember criterion increases relative to the confidence criteria, the estimated slope of the ROC increases.

The distribution of confidence ratings for remember and know judgments suggested the presence of item-to-item criterion variability. Although this could have arisen either because of variability in the confidence criteria or because of variability in the remember criterion, we speculated that it was likely to be due to the latter (given that the definition of a remember response was only recently learned). The fact that the slope of the remember/know ROC is sometimes greater than the confidence-based ROC is consistent with this idea.

(Manuscript received October 21, 2003;
revision accepted for publication June 2, 2004.)