Journal of Experimental Psychology: General
2002, Vol. 131, No. 2, 241–254

0096-3445/02/$5.00 DOI: 10.1037//0096-3445.131.2.241

Separating Sensitivity From Response Bias: Implications of Comparisons of Yes–No and Forced-Choice Tests for Models and Measures of Recognition Memory

Neal E. A. Kroll and Andrew P. Yonelinas
University of California, Davis

Ian G. Dobbins
Harvard Medical School

Christina M. Frederick
University of California, Berkeley

A fundamental challenge to psychological research is the measurement of cognitive processes uncontaminated by response strategies resulting from different testing procedures. Test-free estimates of ability are vital when comparing the performance of different groups or different conditions. The current study applied several sets of measurement models to both forced-choice and yes–no recognition memory tests and concluded that the traditional signal-detection model resulted in distorted estimates of accuracy. Two-factor models were necessary to separate memory sensitivity from response bias. These models indicated that (a) memory accuracy did not differ across the tests and (b) the tests relied on the same underlying memory processes. The results illustrate the pitfalls of using a single-component model to measure accuracy in tasks that reflect 2 or more underlying processes.

One consistent measurement problem faced by psychologists, regardless of subdiscipline, is the separation of discriminative abilities from response or decision strategies, which can seriously contaminate or distort estimates of those abilities. This widespread problem has led to the development of statistical decision models whose primary aim is to effectively remove the contribution of strategic response biases or guessing strategies from estimates of discriminative abilities, allowing the effective comparisons of different observers, groups, neuropsychological populations, or experimental conditions on a particular perceptual or cognitive discrimination skill. These scoring methods have become so commonplace that they are often not viewed as models of discrimination processes but rather as simple “corrections” for guessing. However, to the extent that these models are inappropriate for the cognitive processes being measured, serious errors can arise when comparing individuals, conditions, or groups that have adopted different decision strategies. The consequences of such errors could range from as minor as clouding the interpretation of a small experimental project to as serious as misinterpreting the quality of radiological services at different clinical institutions.

One hallmark of a successful decision model is that it yields similar estimates of accuracy across different testing formats for a given observer. More specifically, if the model’s assumptions regarding the nature of the underlying information and the decision process applied to that information are viable, then it should not matter whether the observer is tested using the sequential presentation of targets and lures (yes–no procedure) or using a simultaneous array of items in which only one is the target (forced-choice procedure). In either case, the estimate of accuracy should be highly similar.

In the present study, we took a closer look at this problem from within the domain of human recognition memory. In particular, we were interested in systematically contrasting forced-choice (FC) and yes–no (YN) recognition accuracy estimates within individuals. This interest was based, in part, on a current controversy regarding the relative ease of the testing formats and the relative performance of memory-impaired individuals across the two test types. Below, we briefly discuss the current debate regarding recognition performance across the two test formats and describe the most commonly used index of accuracy for across-test comparison, the signal-detection theory estimate, d'. Following this, we then show how d' yields consistently discordant estimates across the testing formats, present two modifications of this basic model that potentially eliminate the discrepancy, and discuss the relative merits and practical appeal of each. Our point is not purely methodological; on the contrary, our argument is that the method used to estimate sensitivity, independent of response biases, depends on the theoretical model of the processes underlying sensitivity. Through this article, we aim to shed new light on both the methodological and the theoretical issues.

Signal-detection theory is currently being used in a myriad of applications, including visual detection, psychoacoustics, learning,

Neal E. A. Kroll and Andrew P. Yonelinas, Department of Psychology, University of California, Davis; Ian G. Dobbins, Nuclear Magnetic Resonance Center, Massachusetts General Hospital, Harvard Medical School; Christina M. Frederick, Department of Psychology, University of California, Berkeley.

This work was supported by National Institute of Mental Health Grant MH59352-01. We are grateful to Anne E. Bower for her assistance in data collection.

Correspondence concerning this article should be addressed to Neal E. A. Kroll, Department of Psychology, University of California, One Shields Avenue, Davis, California 95616. E-mail: neakroll@ucdavis.edu

weather forecasting, neurophysiology/neuronal response characterization, attention, vigilance, and many others (e.g., see Macmillan & Creelman, 1991; Swets, 1964). The particular deficiency of this decision model and the suggested solutions explored in this article are specific to the testing of memory. However, we hope that by our (a) demonstrating a particular deficit and its ramifications and (b) showing how one develops and chooses among potential modifications of the decision model so that the resulting decision model better represents the psychological model being tested, psychologists in various subfields may be motivated (a) to test whether the unmodified signal-detection model is appropriate for their own psychological specialty, and, if it's not, (b) to suggest model developments and modifications appropriate to their subfield. Many aspects of the problems and the solution we develop here are discussed in Macmillan and Creelman (1991; especially chap. 5). However, it has been our experience in observing ourselves, our students, and our colleagues that many of the problems and potential solutions illustrated in their book are not fully appreciated or understood until they are worked through in a specific, detailed example. In addition, the specific misfit between signal-detection theory and recognition memory performance has also been noted previously, as we discuss below, but it appears that a very important implication that arises from this misfit has not been fully realized.

The particular problem we wish to address here concerns the contradictory evidence regarding the relative ease, and resulting accuracy, of FC recognition tests as compared to YN recognition tests. Most find it intuitively obvious that recognition accuracy on FC recognition tests should be superior to that on YN tests for the same stimuli. For example, in a pilot study leading up to the present experiments, 20 participants who had just completed a series of both types of tests were asked on which they thought they had performed better. A large majority of the participants (i.e., 17 out of 20) reported that they thought they had done better on the FC test than on the YN test. Yet, by the now-typical d' measure of performance, all 20 were better on the YN test.

Not only experimental participants but also experts in the field believe that FC should result in more accurate recognition, and some have argued that differences in performance arise because the two types of tests rely on different memory processes. For example, Deffenbacher, Leu, and Brown (1981) suggested that participants should be more accurate in an FC test because they can simply choose the item with the closest match to a memory trace, whereas in the YN test they should be more prone to error because they are likely to endorse lures that have similarities to memory traces. Others have argued that because test items are presented simultaneously in the FC test, direct comparisons of familiarity should be particularly useful in this test, whereas because items are presented one at a time in the YN test, participants might be more likely to make use of a recollection process whereby they retrieve qualitative information about the previous study events (e.g., Aggleton & Shaw, 1996; Nolde, Johnson, & Raye, 1998).

Some researchers have hypothesized that patients with medial temporal lobe damage have severe deficits in recollection that severely handicap their recall test performance, while they may maintain much of their ability to make recognition memory judgments on the basis of priming (e.g., Hirst et al., 1986) or familiarity (e.g., Huppert & Piercy, 1976, 1978; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998). If this is true, the above speculation suggests these patients might do relatively better on FC tests than on the YN tests. Indeed, there is some support for this position. For example, Freed, Corkin, and Cohen (1987) found that the amnesic patient H.M. was more likely to show a recognition deficit when tested using a YN procedure than when tested using an FC procedure. Similarly, Aggleton and Shaw (1996) found that amnesiacs with restricted hippocampal lesions exhibited near-normal performance on FC recognition memory tests for words and pictures. However, other studies have failed to find noticeable differences in amnesics' memory deficits in the two types of recognition tests (Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Reed, Hamann, Stefanacci, & Squire, 1997).

The assumed differences between the tests have led some to propose that the discrepancies observed across studies result from one laboratory having used the FC test, whereas the other used the YN test (e.g., Griggs & Keen, 1977; Holmgren, 1974). The ability to compare across laboratory practices is especially critical when comparing results from most human laboratory research, which tend to rely heavily on YN testing methodology, with either neuropsychological (e.g., Warrington & James, 1967) or animal research (e.g., Zola-Morgan & Squire, 1992), which rely more heavily on the FC testing methodology. Clearly understanding the correspondences across disciplines requires an understanding of the correspondences across test methodologies.

Macmillan and Creelman (1991, chap. 5) explain the different decision spaces of the YN and FC tasks and demonstrate that both signal-detection theory and choice theory agree exactly on the mathematical relation between the two designs.[1] But they also remark that the empirical results often do not "respect this unanimity" (Macmillan & Creelman, 1991, p. 130).

Unfortunately, the empirical relationship between these two types of recognition memory tests is poorly understood because there have been very few studies that directly compare performance across the two procedures. Further, the results of the existing studies are not entirely consistent. Deffenbacher et al. (1981) examined recognition memory for faces and reported that FC performance was better than YN performance. However, other studies have found the two tests to be equivalent, including an early set of experiments examining memory for nonsense syllables (Green & Moses, 1966) and two more recent experiments examining memory for common words (e.g., Khoe et al., 2000; Yonelinas, Hockley, & Murdock, 1992). Moreover, Khoe et al. (2000) used the remember–know procedure in a recognition test for words in which participants were required to report whether their recognition judgments were on the basis of recollection or on the basis of familiarity in the absence of recollection. There was no

[1] For the YN test, d' is calculated as $z_{\text{hit rate}} - z_{\text{false-alarm rate}}$, whereas in the FC test d' is calculated as $(1/\sqrt{2})\,(z_{\text{proportion correct}} - z_{\text{proportion incorrect}})$. The FC equation includes the $1/\sqrt{2}$ constant because the participant's decision is based on the distribution of memory strength differences between pairs of stimuli, one drawn from the target distribution and the other from the lure distribution. When two normally distributed variables are added or subtracted, the new variable is still Gaussian, but with a variance of 2, because the variance adds (Macmillan & Creelman, 1991, p. 124). It is likely this relationship is, at least partially, the cause of the perception of the relative accuracy of the FC task. That is, a person's perception of the relative accuracy is most certainly more related to the percentage correct (which is almost always higher in FC than in YN) than to the d' measure.

evidence that the contribution of recollection or familiarity differed in the two types of tests.

The continued interest in potential accuracy differences—and perhaps in the types of memory processes involved—across the two recognition memory tests led us to compare these tests directly in the current series of experiments. Given there were already several experiments showing the equivalence of the tests with verbal materials (e.g., Green & Moses, 1966; Khoe et al., 2000; Yonelinas et al., 1992), and given that many of the current suggestions that the two tests involve different processing have used pictorial stimuli (e.g., Aggleton & Shaw, 1996; Freed et al., 1987; Nolde et al., 1998), we examined memory for pictorial stimuli in the current study.

To compare performance across YN and FC tests, it is necessary to derive accuracy measures that are on the same scale and are independent of criterion bias. Although FC scores are not influenced by participant criterion bias,[2] scores on YN tests vary depending on the participant's tendency to respond positively. In the current experiment we used the d' measure from signal-detection theory as an accuracy index (e.g., Macmillan & Creelman, 1991; Swets, 1986). This method has been used extensively in the recognition memory literature and in several previous studies that have compared YN and FC performance (e.g., Green & Moses, 1966; Khoe et al., 2000; Yonelinas et al., 1992). This method makes a number of critical assumptions about how recognition decisions are made. We will hold our discussion of those assumptions until after the presentation of the results from Experiment 1.

Experiment 1 compared recognition memory performance on a YN test and a two-alternative FC test. Results indicated that the relationship between the performances on the two tests was influenced by the response criterion the participant adopted in the YN test. This result suggested that the index used to measure memory accuracy (i.e., d') might have provided an unreliable measure of accuracy. To test this hypothesis, we asked participants in Experiments 2A and 2B to make confidence judgments about their recognition decisions. This allowed us to examine the relationship between the two tests as a function of the response criterion adopted during the YN test. The d' measure of accuracy was found to lead to either a significant advantage or a significant disadvantage on YN compared with FC performance, depending on the participants' response criterion, verifying that d' did not provide a reliable measure of accuracy. However, two different modifications of signal-detection theories of recognition memory (i.e., the dual-process and the unequal-variance signal-detection models) were found to provide good accounts of performance and indicated that YN and FC accuracy did not differ. Finally, in Experiment 3, participants were required to make remember–know judgments at time of test to determine whether the YN and FC recognition tests relied differentially on recollection. The results of Experiment 3 indicated that overall accuracy on the FC and YN tests was comparable and that recollection contributed a similar degree to the performance on the two types of recognition tests.

Experiment 1

Method

Participants and materials. Twenty-four undergraduates at the University of California, Davis, participated in exchange for partial credit in an introductory psychology course. A total of 560 pictures served as the study and test stimuli. These pictures were selected from diverse sources, including CorelDraw Photos, scanned pictures of postcards and magazine pictures, and photos taken by us. They included pictures of nature scenes, buildings, animals, and people in national costume or in various activities such as sports or vocations. Pictures were presented on an IBM-compatible computer on a monitor set at a 640 × 480 pixel resolution. The pictures themselves were set to a pixel resolution of 432 × 548. The resulting pictures that appeared on the screen were 11.5 × 16.5 cm.

Design and procedure. The 560 pictures were randomly divided into four equal sets of 140 pictures each. Two of these sets were used in the study phase, of which one set would be assigned to the FC test and the other to the YN test. Pictures from the remaining two sets were used as the lures in the two tests. The assignment of these sets of pictures to study sets versus lure sets and to FC test conditions versus YN test conditions was counterbalanced across participants.

Prior to the study phase, participants were instructed that they would see a large number of pictures and that their memory for these pictures would be tested immediately after the study phase. During the study phase, each of the 280 pictures was presented for 2 s, centered on the screen. The pictures fated for the FC test and those fated for the YN test were interleaved in a quasi-random order.

Immediately following the study phase, participants began the test phase. During the test phase, FC and YN test trials were again interleaved in quasi-random order. Pictures were presented at the same size and resolution as during the study phase. On YN trials, one picture was presented at the top of the screen, centered on the left–right axis. The participant's task was to indicate whether the current picture had been presented during the study phase ("old"; denoted by *O* on the keyboard) or had not been ("new"; denoted by *N* on the keyboard). On FC trials, two pictures were presented at the top of the screen, one on the far left and the other on the far right. On half of the FC trials, the left-most picture was the studied picture (the target), and on the other half, the left-most picture was the new picture (the lure). The participant's task was to indicate whether the picture on the right half (*Q* on the keyboard) or the left half (*Z* on the keyboard) of the presentation screen had been previously studied. On each type of test trial, the appropriate key-response assignments were shown at the bottom of the screen. Participants were instructed to proceed at their own pace in both test conditions.

Unannounced to the participant, the first 90 test trials (30 FC, 30 YN-old, and 30 YN-new trials) were actually used as a practice test. The target pictures for these practice trials were the first 30 and last 30 pictures presented during the study phase. At the conclusion of the practice test

[2] In fact, FC performance can be influenced by a participant's bias (Macmillan & Creelman, 1996, chap. 5), but this is a very different type of bias than that influencing YN performance. In a YN recognition test, the bias concerns the participant's willingness to endorse a stimulus as being old. By contrast, in an FC test, the participant *must* endorse one of the stimuli on each trial, and thus, the bias, if any, represents a participant's tendency to choose one of the positions (e.g., the left stimulus) more often than the other. Macmillan and Creelman (1991) discussed a few of the situations in which this may be of interest, and, of course, one should check for a bias in the FC performance to be certain that there is not some unintended aspect of the experiment (e.g., response or position preference) affecting the data. If one wishes a measurement of the position bias, the formula for FC d' is calculated as $(1/\sqrt{2})$ $(z_{\text{proportion correct at Position 1}} - z_{\text{proportion incorrect at Position 2}})$. Although both types of bias are frequently referred to as *response bias*, they result from very different causes and have very different effects on performance measures, as we discuss in the text. Hence, we will refer to the bias observed in a YN test as *criterion bias* and the bias observed in an FC test as *position bias*.

session, participants were given feedback on their performance level. Three goals were achieved by this practice session. The first goal was simply to provide participants with practice for this task and to familiarize them with the two testing methods. Second, this practice session functioned as a filled-retention interval for the rest of the pictures tested during the final phase of the experiment. Finally, this practice phase served to eliminate those pictures that might be associated with either primacy or recency effects.

Following a brief discussion of the participant's performance during the practice test session, participants were tested on the 330 critical test trials (i.e., 110 FC, 110 YN-old, and 110 YN-new trials interleaved in a quasi-random sequence).

Results and Discussion

The average scores in the YN and FC tests are presented in Table 1. The proportion of hits and the proportion of false alarms in the YN test were used to obtain accuracy (i.e., d') and criterion bias (i.e., C) scores for each participant. In addition, the proportions of correct responses on the FC test were used to obtain each participant's d' score for the FC test.[3] A comparison of performance on the two tests indicated that participants had significantly higher d' scores on the YN test ($M = 1.81$, $SE = 0.12$) than on the FC test ($M = 1.47$, $SE = 0.11$), $t(23) = 2.93$, $SE = 0.12$. The consistency of this relationship is illustrated in Figure 1, in which each participant's d' on the FC test is plotted as a function of that participant's d' on the YN test. The finding that YN performance was greater than FC performance is surprising given that previous studies either reported no difference between YN and FC accuracy or reported an advantage for FC over YN accuracy.

One possible interpretation of these results is that participants relied to a great extent on relative familiarity on the FC test, whereas on the YN test they relied more on a recollection process (e.g., Aggleton & Shaw, 1996; Nolde et al., 1998). If one assumes that recollection is more accurate than familiarity, then accuracy should be higher for the YN than the FC tests. Another explanation, however, is that recognition accuracy was not different for the two types of tests, but rather the d' measure we used to compare the two tasks provided distorted indices of accuracy. This latter explanation was suggested by the results of a subsequent analysis we conducted.

To further examine the discrepancy between YN and FC accuracy, we plotted the difference in d' on the two tests as a function of the participants' response criterion on the YN test. Figure 2 presents each participant's YN test advantage ($d'_{\mathrm{YN}} - d'_{\mathrm{FC}}$) as a function of that participant's YN response criterion. There was a significant positive correlation, $R^2(28) = .43$, $p < .01$, between response criterion and the YN–FC accuracy advantage. Thus, the participants who adopted the most strict YN response criterion exhibited the largest advantage on the YN test as compared with the FC test. In fact, the regression function passed very close to the (0,0) intercept, indicating that when YN performance was not biased toward either "yes" or "no" responses, the d' values for the two tests were similar. Moreover, for the participants who adopted the most lenient response criterion (i.e., the bottom left points in Figure 2), there was a tendency to perform worse on the YN than the FC tests.

The finding that the accuracy measures were related to response criterion suggests that d' may not have provided a pure measure of accuracy, and thus it may have obscured the relationship between the two tests. There is, in fact, a growing body of empirical evidence indicating that d' does not provide a reliable measure of recognition memory accuracy (e.g., see Yonelinas, 1999). This evidence comes primarily from the study of receiver operating characteristics (ROCs). In an ROC experiment, participants are typically required to make confidence judgments on a YN recognition test. Hits are then plotted against false alarms as a function of response confidence. If the signal-detection model that underlies d' is correct, then the ROC should be curvilinear and symmetrical along the negative diagonal, such as the functions plotted in Figure 3. The predicted ROC is curved because the model assumes Gaussian familiarity distributions, and it is symmetrical because it assumes that old and new item familiarity distributions have equal variance. Recognition memory ROCs, however, are almost never symmetrical (e.g., Donaldson & Murdock, 1968; Egan, 1958; Glanzer & Adams, 1990; Ratcliff, Sheu & Gronlund, 1992; Yonelinas, 1994); rather, they are asymmetrical as illustrated in Figure 4. This means that the model underlying the d' measure of accuracy is inappropriate. Thus, if this model is used it will provide distorted measures of memory accuracy. Note that the asymmetrical ROCs also rule out other models of recognition, such as those underlying A' measures of accuracy, because they also predict symmetrical ROCs (see Macmillan & Creelman, 1996).

To account for the asymmetric ROCs, it is necessary to modify the signal-detection model. Two such modifications have been proposed and have been found to provide accurate and very similar accounts of standard recognition ROCs. One modification is the dual-process signal-detection model (e.g., Yonelinas, 1994), which includes a recollection process in addition to the familiarity process. Because recollection is assumed to increase the hit rate with little or no increase in false-alarm rate, this pushes up the left side of the ROC, causing it to become asymmetrical. An alternative modification is the unequal-variance signal-detection model (e.g., Macmillan & Creelman, 1991; Swets, 1986), which includes an additional factor that influences the variance of the old-item familiarity distribution. As the variance of the old-item distribution increases, the ROC becomes more asymmetrical. (It should be noted that when Macmillan & Creelman, 1991, discuss the unequal-variance model, they set the variance of the old items to 1.0 and calculate the relative size of the variance of the new items.)

The asymmetry of recognition memory ROCs implies that measures of d', calculated on the basis of the initial signal-detection model, should vary with response criterion. That is, as response criterion becomes more strict, the d' values will increase. As an illustration, if d' was used to estimate accuracy at the three points seen in Figure 3, it would suggest that accuracy differed: The d'

[3] The d' and C values for FC performance were also evaluated with the formula provided in Footnote 2. The corresponding values for the average proportion of correct responses in the left position and for the average proportion of error responses in the right position are presented in Table 1. None of the participants showed a large position bias, and there was no consistent pattern of position bias. It may be worthwhile to reiterate here that bias in the YN test refers to a participant's tendency to accept a stimulus as old. In the FC test, bias refers to the tendency to prefer a particular location. Thus, our finding virtually no bias on the FC test simply means that participants did not have a tendency to choose more left-sided pictures than right-sided pictures or vice versa.

Table 1
Mean Proportion and Standard Error of Hits and False Alarms (FAs) in the Yes–No (YN) Tests and of Left Correct and Right Incorrect Responses in the Forced-Choice (FC) Tests

	Experiment 1				Experiment 2A[a]				Experiment 2B[a]				Experiment 3[b]			
	Hits		FA		Hits		FA		Hits		FA		Hits		FA	
Test format	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
YN	.716	.017	.141	.020	.698	.022	.145	.015	.571	.021	.250	.017	.641	.025	.123	.015
FC	.832	.018	.164	.021	.855	.016	.147	.013	.718	.015	.276	.018	.836	.018	.167	.015

[a] In Experiments 2A and 2B, the YN test items associated with recognition confidence responses greater than 3 were treated as hits and FAs. [b] In Experiment 3, remember and familiar responses were combined to find hits and FAs.

measures are 1.00, 0.74, and 0.58 for the left, middle, and right points, respectively. Figure 4 includes the same three points and demonstrates how they can be fit by a single ROC generated by the dual-process model. Memory accuracy is the same at the three points; the only difference is the response criterion. The right-most point reflects a bias to say "yes," the left-most point reflects a bias to say "no," and the middle point is not biased in either direction. This difference, between the models represented in Figures 3 and 4, in evaluating memory accuracy arises because when calculating a simple d' value one is essentially finding the symmetrical ROC that intersects with that point. Both the dual-process and unequal-variance models can produce the type of ROC illustrated in Figure 4, and both provide more accurate accounts of the recognition ROCs than the traditional signal-detection model underlying the d' measure (for discussion of these models, see Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). Both of these models assume that there are two independent factors: The dual-process model has familiarity and recollection; the unequal-variance model has sensitivity and old-item variance, V_o. Thus, we will refer to them both as *two-factor models*.

On the basis of results from ROC studies and on the basis of the dual-process and unequal-variance models, we expected that as the participants' response criterion became more strict in a YN test, the traditional measure of d' would increase. This would explain the findings of the current experiment. That is, participants who adopted a strict criterion exhibited inflated d' values on the YN test and therefore performed better on that test then on the FC test, whereas participants who adopted a more lenient criterion either performed roughly similarly on the two tests or performed more poorly on the YN than the FC test.

However, this criterion-bias explanation of the current results was in need of further scrutiny. Although it was consistent with the results, it was not something we predicted a priori. Moreover, it was possible to construct plausible alternatives. For example, one might argue that participants who adopt a lenient response criterion during the YN test tend to use both recollection and familiarity to an equal extent on the YN and FC tests and thus exhibit roughly equivalent recognition accuracy scores. In contrast, other participants who adopt a more strict response criterion on the YN test may rely more heavily on recollection during the YN test than during the FC test. If one assumes that recollection is more accurate than familiarity, then this latter group might be expected to perform better on the YN than the FC test.

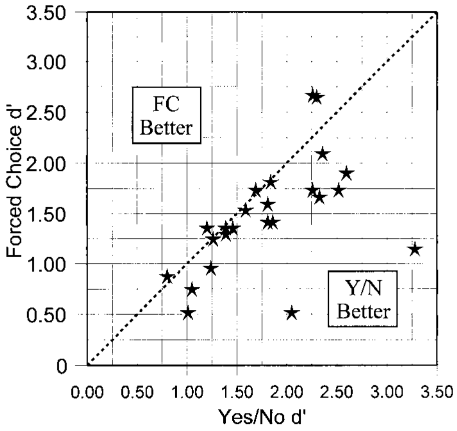

Figure 1. Each participant's d'_{FC} score plotted as a function of that participant's d'_{YN} score. If a point is located above the diagonal, the d'_{FC} score was better than the d'_{YN} score for that participant; if below, the d'_{YN} score was better. FC = forced choice; YN = yes–no.

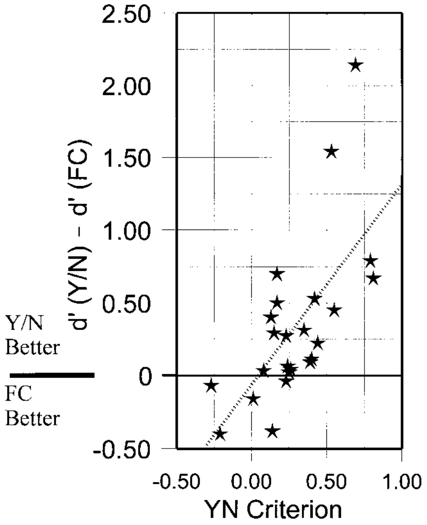

Figure 2. The YN test advantage ($d'_{\mathrm{YN}} - d'_{\mathrm{FC}}$) in Experiment 1 is plotted as a function of each participant's YN criterion. The dotted line represents the linear best fit (r = .658). FC = forced choice; YN = yes–no.

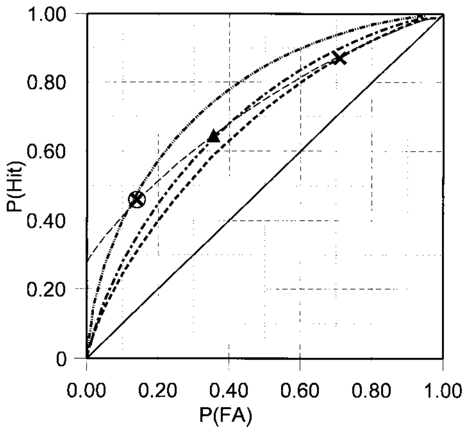

Figure 3. Three points, with a bias toward "new" (⊗), no bias (▲), and a bias toward "old" (✕), being evaluated by the signal-detection theory model. If each point is assumed to lie on a symmetric curved line, then d' = 1.00, 0.74, and 0.58, respectively, as one moves from the left-most to the right-most point. P(Hit) = proportion of hits; P(FA) = proportion of false alarms.

To test the criterion-bias explanation further, we conducted another experiment. This experiment was designed to more directly test the claims that performance is really the same on the YN and FC tests and that the apparent differences in d' are due to the fact that d' provides a distorted estimate of recognition accuracy.

Experiments 2A and 2B

Experiments 2A and 2B were similar to Experiment 1 except that participants were required to rate the confidence of their recognition judgments. The confidence responses allowed us to examine the relationship between the response criterion adopted on the YN test and the difference between FC and YN recognition accuracy. If the response bias explanation is correct and the signal-detection model underlying the d' index is incorrect, then when we use d' to measure accuracy, d' should be greater for the YN than the FC test when a strict response criterion is adopted (e.g., high-confidence YN responses), and the difference should decrease (and potentially reverse) as the YN response criterion is relaxed. If this turns out to be the case, it would demonstrate that d' does not provide a reliable index of recognition accuracy and that using this measure can lead to artifactual differences in apparent accuracy between the two tests.

The confidence judgments were also used to plot YN ROCs so that we could directly assess whether the ROCs in the study were consistent with the dual-process and unequal-variance models and inconsistent with the standard signal-detection model. Most important, this allowed us to test whether performance on the YN and FC tests was comparable or whether it was necessary to propose that performance was better on one type of test. That is, we tested whether the modified signal-detection models could accurately predict FC scores on the basis of the YN scores to determine whether the same parameters used to describe accuracy on the YN test were able to describe accuracy on the FC test.

Experiments 2A and 2B were similar except that 2B was designed to elicit lower levels of performance than 2A to verify that the results generalized across different levels of overall performance. To decrease performance, the lures in Experiment 2B were selected to be very similar to the target items. The results of the two experiments were very similar and, consequently, are described together.

Method

Participants and materials. Sixty participants (30 per experiment) participated in Experiments 2A and 2B. The participants were from the same participant pool as Experiment 1, and they participated in exchange for partial credit in an introductory psychology course.

The materials for Experiment 2A were the same as those used in Experiment 1. The materials for Experiment 2B were similar except for the changes designed to make the tests in Experiment 2B much more difficult than those in the earlier experiments. Eight lists of 90 pictures each were constructed. Each of these 720 pictures was paired with a picture that was similar to itself. The relationship between the pairs varied among the following:

1. Thematic: The pictures were of similar things (e.g., a picture of a German cockroach and a picture of a Brown-band cockroach, or two pictures of a young girl in costume doing a Scottish dance, but the two pictures showed her in different poses).
2. Zoom: One picture was an enlargement of the other.
3. Reversed: One picture was a reversal of the other. Most of these were left–right reversals (e.g., a duck swam to the left in one picture but to the right in the paired picture), but a few were up–down reversals (e.g., a salamander climbed upward in one picture and downward in the paired picture).
4. Overlap: The main item of interest was in both pictures, but one picture continued in one direction whereas the other continued in the other direction (e.g., one picture showed the Brandenburg gate on the left side of the picture and the picture continued to the right; the paired picture showed the Brandenburg gate on the right side of the picture and the picture continued to the left).

This resulted in eight lists of 90 pairs of pictures each (A–A′, B–B′, . . . H–H′). Approximately the same number of each type of relationship was represented in each of the paired lists. When pictures were modified to form the similar pair (e.g., reversed or enlarged), half of the originals were assigned to one list and half to its paired list. Any given participant saw only four of the lists and two of the paired lists. For example, 2 participants

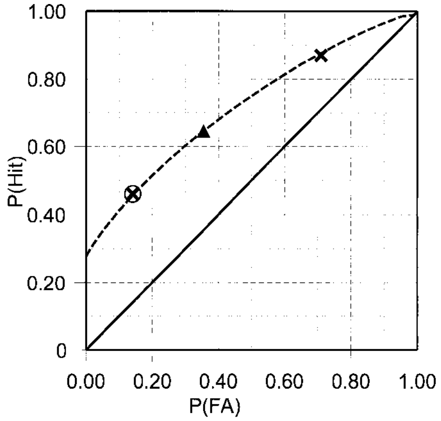

Figure 4. A single dual-process receiver operating characteristics function (d' = 0.40, R_o = 0.275) fitting the same three points shown in Figure 3. P(Hit) = proportion of hits; P(FA) = proportion of false alarms.

received lists A, C′, E′, and G during study and lists A, C, E′, and G′ during test. For 1 of these participants, the pictures in list A served as the targets and those in list C as the lures in the YN tests; the pictures in list E′ served as the targets and those in list G′ as the lures in the FC tests. Thus, a lure was always very similar to a picture that had been presented during the study phase. The assignment of lists to YN and FC conditions was counterbalanced across participants.

Design and procedure. The design and procedure for Experiment 2 were the same as for Experiment 1 except that participants were now required to make recognition confidence judgments. On each test trial, participants used the response keys *1* to *6*. On the YN test trials, a response of 1 indicated *confident new* and a response of 6 indicated *confident old*, with intermediate numbers indicating intermediate degrees of confidence. On FC test trials, 1 indicated a *very confident*, 2 a *less confident*, and 3 a *guessing* “left–old” response, whereas 4 indicated a *guessing*, 5 a *more confident*, and 6 a *very confident* “right–old” response. The relevant confidence rating scale was provided for participants during each test trial, and participants were instructed to use the entire range of confidence responses. Although the confidence responses obtained on the FC test were not used in the data analysis, confidence responses were collected to roughly equate the response and decision requirements of the YN and the FC tests.[4]

The design and procedure for Experiment 2B was the same as for Experiment 2A except for the following changes. The study phase included 360 pictures. The first 20 and last 20 study items were used as buffer items that later appeared as practice items in the test phase. The practice test phase included 60 practice test trials. After the practice test trials, participants were given 240 test trials (i.e., 80 FC, 80 YN-old, and 80 YN-new trials interleaved in a quasi-random sequence). Most important, the lures were similar to the studied items. Thus, during the test phase, the lures differed only subtly from their paired studied pictures.

Results and Discussion

The confidence responses in the YN tests were used to plot ROCs (Figure 5), such that the left-most point of the function reflected only the items that received a confidence response of 6, the next point reflected all items receiving a confidence response of 5 or higher, and so forth. The ROC in Experiment 2A was considerably higher than that in Experiment 2B, thus indicating that performance was poorer in the latter experiment. In agreement with previous studies, the ROCs were curved and asymmetrical. Moreover, when the signal-detection model and the two modified models were fit to the ROCs, the modified models both provided better fits than did the signal-detection model, and the fit of the two modified models did not differ appreciably.

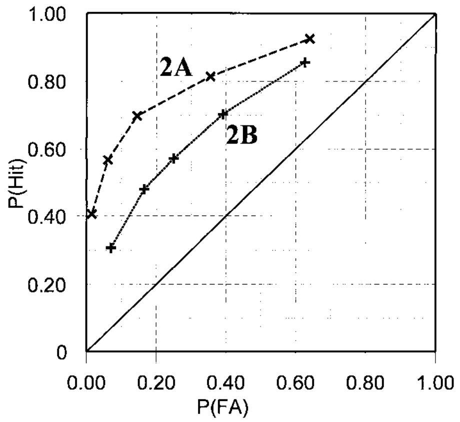

Figure 5. The average hit and false-alarm rates at each confidence level in Experiment 2A and Experiment 2B. P(Hit) = proportion of hits; P(FA) = proportion of false alarms.

A least squares method was used to find the best fit of the three models to the average ROC in each experiment (from Yonelinas et al., 1998). In Experiment 2A the sum of squared errors (*SSE*) associated with the signal-detection model (when $d' = 1.60$, $SSE = 0.01$) was significantly worse than that observed for the dual-process model (when R_o [the probability that studied items were recollected] = 0.38 and $d' = 1.03$, $SSE = 0.001$), $F(1, 3) = 15.00$, $p < .05$, and the unequal-variance model (when $V_o = 3.34$ and $d' = 2.15$, $SSE = 0.001$), $F(1, 3) = 15.00$, $p < .05$.[5] This shows that the two modified models provided a better fit of the recognition data than the standard signal-detection model. Given that the error terms of the two modified models were identical to three decimal places, it is clear that the two models provided comparable fits of the data. In Experiment 2B, the *SSE* for the signal-detection model (when $d' = 0.88$, $SSE = 0.006$) was worse, but not significantly so, than that of the dual-process model (when $R_o = 0.14$ and $d' = 0.89$, $SSE = 0.003$), $F(1, 3) = 3.00$, $p > .05$, and the unequal-variance model (when $V_o = 1.72$ and $d' = 0.97$, $SSE = 0.002$), $F(1, 3) = 6.00$, $p > .05$, and again the fits of the modified models were comparable.

Thus, the ROC analyses indicated that the standard signal-detection theory provided a poorer account of recognition memory performance on the YN test than did the two modified models. Although the differences did not reach the traditional level of significance in Experiment 2B, they did in Experiment 2A, and these results converge with the results of several previous studies showing the advantage of using the modified models (e.g., Yonelinas et al., 1996).

To examine the effects of using d' to measure accuracy, d' values were calculated from each of the points on the ROCs in Figure 5, and they were compared to the d' estimates derived from the FC tests. The d' values from the YN tests are shown in Figure 6 as solid bars, the *C* values are shown as cross-hatched bars, and the d' values from the FC tests are shown as dashed lines. An examination of Figure 6 indicates that when a strict YN response criterion was adopted, the YN d' value was greater than the FC d'

[4] It is possible to plot the ROC functions for FC performance. However, the FC ROC is based on the difference distribution and is therefore symmetric, as all three of the models under consideration would predict. Moreover, it turns out that the d's from each of the five points, within both of the experiments, are virtually identical. This latter fact reinforces the conclusion that there were essentially no FC position biases.

[5] Calculating the best fit of these models to the data of individual participants occasionally led to values not acceptable to the model in question. For example, the estimated R_o of 1 participant in Experiment 2A was −0.007, this “unacceptable” value was replaced by 0.0. Similarly, the estimated V_o values of 12 participants in Experiment 2B and 3 participants in Experiment 3 were less than 1.0 (one as low as 0.228). To obtain the best fit of the data, we allowed V_o to take on any value that resulted in the best fit. However, if one assumes that the variances are unequal because study increases the variance of the studied items, then one should replace all of the V_o values that were less than 1.0 with 1.0. Such constraints on the value of V_o change the average estimates of V_o and d' but do not greatly influence relationship between YN and FC recognition.

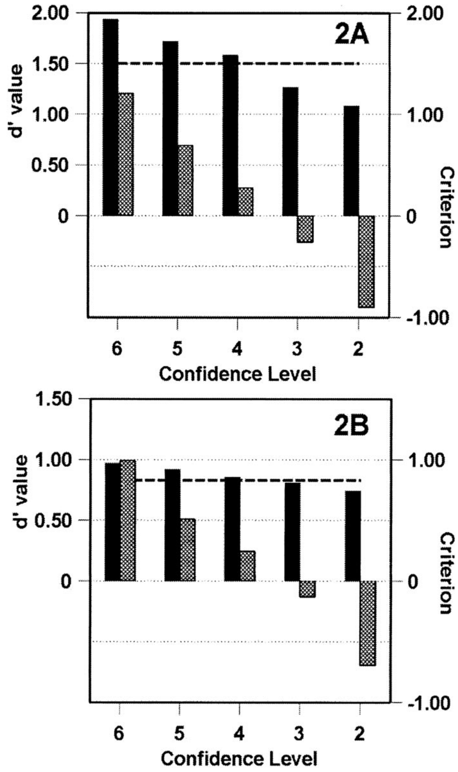

Figure 6. The signal-detection theory d' value for each of the confidence levels of the yes–no test in Experiments 2A and 2B, with the forced-choice d' value shown as a dashed line. Values were calculated on the basis of average hit and false-alarm rates. The average criterion value at each confidence level is given in cross-hatched bars.

value, but as the response criterion was relaxed this pattern reversed, so that at the most lenient response criterion the YN d' was worse than the FC d'. The same pattern was observed for both Experiments 2A and 2B. To ensure that this pattern was not an averaging artifact, we calculated d' values for every participant at each confidence level, and these participant d' values were averaged and plotted in Figure 7. Confidence Level 6 is not presented for Experiment 2A because a large number of participants in this experiment made no false alarms at this confidence level, thereby making it impossible to calculate d' for these participants. In addition, in Experiment 2A only 24 of the 30 participants were used in this set of calculations because the remaining 5 participants had either no false alarms at Confidence Level 5 or no hits at Confidence Level 1. Similarly, in Experiment 2B only 26 of the 30 participants were used in the analysis. As with the initial analyses, Figure 7 shows that when a strict response criterion was adopted in the YN test, the d' estimates in that test were greater than those seen in the FC test, and the pattern reversed as response criterion became more lenient. The exact value of a participant's response criterion in the YN test at which one should expect the d' value to be the same in the two tests depends on the degree of asymmetry in the YN ROC function (which is determined in the dual-process model by the proportion of trials relying on recollection and, in the unequal-variance model, by the difference in the variances produced by old and new items.)

If d' cannot be relied on to measure accuracy in the recognition tests because the underlying signal-detection model does not accurately describe the relationship between accuracy and response criterion, then how can one compare performance on the two types of recognition memory tests? To do so it is necessary to use recognition models that provide an accurate account of recognition accuracy and response criteria, such as the dual-process or unequal-variance models. For example, when these two modified models were fit to the YN ROC data, they produced parameter estimates that characterized recognition accuracy. In the dual-process model, one parameter reflected the increase in familiarity due to the study phase (d') and another reflected the probability that studied items were recollected (R_o), whereas in the unequal-variance model, one parameter reflected the increase in familiarity (d') and the other represented the variance associated with the old-item familiarity relative to the new-item variance (V_o). If memory accuracy was identical on the YN and FC tests, then the parameters from the YN test should predict performance on the FC test.

To predict each participant's FC performance, we entered the parameter estimates obtained from the least squares fit of the YN

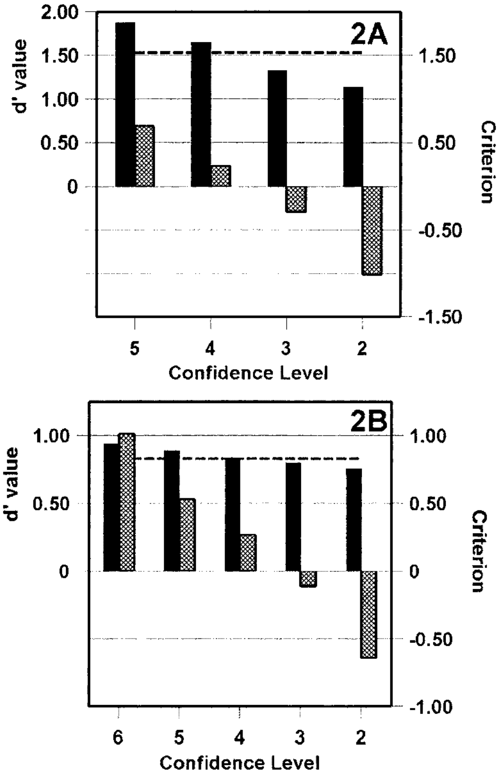

Figure 7. The d' values were calculated for each individual participant, at each confidence level of the YN test, and for each participant's forced-choice score. Then these individual-participant d' values were averaged to find the values presented in the figure. Only 24 of the 30 Experiment 2A participants and only 26 of the 30 Experiment 2B participants were included in these calculations because the remaining participants had either no false alarms at the high-confidence level or no hits at the low-confidence level. The average criterion value at each confidence level is given in cross-hatched bars.

ROC functions into the following formula (an extension of Formula 5.13 in Macmillan & Creelman, 1991, p. 130):

$$E(FC_{PC}) = R_o + (1 - R_o)\left[\Phi\left(\frac{d_2'}{\sqrt{1+\sigma^2_{old}}}\right)\right], \qquad (1)$$

where $E(FC_{PC})$ is the expected proportion correct on the FC task, R_o is the estimated recollection rate, d_2' is the distance between old- and new-item familiarity distributions in units of the old-item standard deviation, and σ^2_{old} is the variance of the old-item distribution in comparison with that of the new, which is set at 1. The bracketed portion of the equation represents the expected proportion of the distribution of familiarity differences that will fall above zero, and hence, lead to a correct response. This distribution will have a mean of d_2' and a standard deviation of the square root of the sum of old- and new-item distribution variances. The Φ represents the cumulative standard normal distribution function. In the case of the unequal-variance signal-detection model, the R_o parameter is set to 0 because recollection is not assumed to contribute to performance, but the old-item variance is free to vary. In contrast, under the dual-process model, both the recollection parameter (R_o) and distance (d') are free to vary, but because this model assumes equal variance, the old-item distribution variance (σ^2_{old}) is fixed at 1. Thus, overall, both models have two free memory parameters that must be estimated to predict forced-choice performance.[6]

In Experiment 2A the average observed FC score ($M = 0.85$, $SE = 0.01$) was almost identical to the average of the FC scores predicted by the dual-process model ($M = 0.85$, $SE = 0.01$) and the unequal-variance model ($M = 0.84$, $SE = 0.01$). The correlation between predicted and obtained FC scores across participants was also very high for both dual-process and unequal-variance models—$R^2(28) = .79$ and .75, respectively—supporting the hypothesis that the relationship between predicted and observed values was the same across the different performance levels. In Experiment 2B the average observed FC score ($M = 0.72$, $SE = 0.02$) was again almost identical to the average FC score predicted by the dual-process model ($M = 0.73$, $SE = 0.01$) and the unequal-variance model ($M = 0.72$, $SE = 0.01$). And, again, the correlation between predicted and obtained FC scores was very high for both models, $R^2(28) = .78$ and .79.

To further assess the models' predictions, we contrasted the average differences between the observed and predicted FC scores. In Experiment 2A, the mean difference between the observed FC score and that predicted by the dual-process model was approximately 0.5%, and this difference was not significant ($M = 0.006$, $SE = 0.007$, $p = .5$). The mean difference between the observed and predicted FC scores for the unequal-variance model was just under 2%, and this difference was significant ($M = 0.018$, $SE = 0.007$, $p = .03$). In Experiment 2B, the average difference between the observed FC scores and that predicted by the dual-process model was again about 0.5% and was not significant ($M = 0.005$, $SE = 0.013$, $p > .10$). Similarly, the difference for the unequal-variance model was less than 0.5% and was not significant ($M = -0.002$, $SE = 0.013$, $p = .5$). Although the unequal-variance model deviated from the observed score in Experiment 2A, it differed by less than 2%, and it predicted FC performance almost perfectly in Experiment 2B. The dual-process model predicted FC performance to within less than 0.5% in both experiments. Thus, in general, both the dual-process and the unequal-variance models were able to accurately predict FC performance on the basis of YN performance by assuming that memory accuracy was identical in the two test conditions. These results provide very little support for the claim that YN and FC tests differ appreciably in terms of overall accuracy.

In summary, Experiments 2A and 2B showed that using d' as a measure of accuracy led to a complex pattern of results such that when a strict response criterion was adopted, the YN test produced higher d' values than the FC test, whereas, when the response criterion was relaxed, the pattern reversed and the YN test produced lower d' values than the FC test. These results indicated that d' did not provide a reliable measure of recognition memory accuracy. Moreover, the examination of ROCs showed that the signal-detection theory underlying the d' measure provided a poor account of recognition performance, whereas the dual-process and unequal-variance signal-detection models were found to provide more accurate accounts of the data. Finally, both of the two-factor models were able to accurately predict FC scores on the basis of YN performance, indicating that the same memory parameters that described the YN data also described the FC data. Thus, the results provide no evidence that the two types of recognition tests differed greatly with respect to overall accuracy. On the other hand, the results do suggest that two-factor models are needed for an accurate assessment of recognition memory and, thus, for an accurate comparison of recognition memory across test formats.

Experiment 3

Although the previous experiments suggested that YN and FC recognition tests were similar in terms of overall accuracy, we wanted to directly test the claim that the two tests might differ in terms of the contribution of recollection and familiarity. That is, it has been suggested that FC performance relies more on familiarity than recollection, whereas YN performance relies more on recollection than familiarity (e.g., Aggleton & Shaw, 1996; Nolde et al., 1998). The preceding experiments do not rule out this possibility. Although the dual-process model interprets the asymmetry of the YN ROC curves as indicating a significant role for recollection, the accuracy measures in FC performance may be the result of a reduction in recollection being accompanied by an increase in performance due to the ease of comparative familiarity. Thus, it is

[6] It should be noted that the use of this formula is generated from Macmillan and Creelman's (1991) equation for the theoretical *maximum* percentage correct. However, from the perspective of the model, there can only be one reason why participants perform under this maximum value: position bias. Thus, estimating FC performance from YN performance assumes that the participant does not have a position bias on the FC test. Any such bias will systematically reduce FC performance below the estimated value. Given that our participants did not seem to have any such bias (see Footnote 6), we did not attempt to include this factor in the equation. Furthermore, the parameters we used were derived from YN performance, and because these estimates are likely to be noisy themselves (i.e., they are not perfect indicators of the underlying distributions and recollection process), it is possible in some cases that this formula will actually underestimate a participant's performance on the FC task. Consequently, we feel that it is reasonable, in the absence of any notable position bias, to refer to the result of this equation as the expected value of FC performance.

possible that the equivalence in overall accuracy of the two tests was due to a trade-off between two different underlying processes. To test this possibility, in Experiment 3 we directly measured recollection using the remember–know procedure (Tulving, 1985). This procedure has been used extensively for the purpose of separating (a) study items recognized on the basis of specifically recollected information from (b) study items recognized on the basis of feelings of familiarity (e.g., Gardiner, 1988; Gardiner & Java, 1990, 1991). Experiment 3 was exactly the same as Experiment 2A except that participants were required to make remember–know responses rather than confidence judgments. If the contribution of recollection was the same in the two recognition tests, then the proportion of remember responses should be similar in the two tests.

Moreover, as in Experiments 2A and 2B, we tested the claim that YN and FC accuracy would be the same in the two tests by testing our ability to predict the FC scores given the YN test performance. That is, if we use the remember responses as a measure of recollection (R), then, given the overall recognition performance on the YN test, we can derive estimates of familiarity (d'). On the basis of those two parameters, we should be able to predict FC performance provided the FC test relies on those two processes to a similar extent as the YN test. Similarly, if we use the unequal-variance model to fit the remember and know responses, by treating the remember responses as high-confidence responses, then we can derive estimates of d' and V_o and use those parameters to predict FC performance.

Method

Participants and materials. Thirty participants, from the same participant pool as the previous experiments, participated in exchange for partial credit in an introductory psychology course. The materials were the same as those used in Experiment 2A.

Design and procedure. The design and procedure were the same as for Experiment 2A, except for the changes resulting from the addition of the remember–know judgment. On each test trial, participants first made a yes–no judgment on YN test trials or a left–right judgment on FC test trials. Immediately following each old and each left–right judgment, the participant was asked to judge the basis for the decision. If the participant remembered some specific feature of the stimulus judged to be old, or remembered thinking something specific while seeing the stimulus during the study period, the participant pressed the *R* key. On the other hand, if the stimulus just seemed very familiar, in the YN test, or more familiar than the alternative, in the FC test, the participant pressed the *F* key.[7] On the first six practice trials and on six additional trials spread out over the remaining practice trials, participants were asked why they had chosen their particular response (*R* or *F*). If their explanation indicated that they had not understood the instructions, the instructions were repeated and questions solicited. At the end of the practice trials, participants were shown their performance as a function of their response choices, and the instructions were repeated.

Results and Discussion

The first question that needs to be addressed concerns the change in the task—that is, does the participant's remember–know judgment change recognition performance in some way? For example, does the act of making the judgment somehow change a participant's reliance on familiarity in one of the tests (e.g., the FC test) more than in the other (e.g., YN test)? To help answer this question, we first analyzed the recognition data in Experiment 3, ignoring the remember–know distinction so that the overall pattern of results could be compared with the data from the earlier, similar experiments (i.e., Experiments 1 and 2A). The average scores are again presented in Table 1. A comparison of performance on the two tests indicated that, as in Experiment 1, participants had significantly higher d' scores on the YN test ($M = 1.63$, $SE = 0.10$) than on the FC test ($M = 1.45$, $SE = 0.09$), $t(29) = 3.00$, $SE = 0.06$, and that the positive correlation, $R^2(28) = .25$, $p < .01$, between response criterion and the YN–FC accuracy advantage was again significant. Thus, it appears that making the remember–know judgment does not change the previously observed relationships between the participant's recognition performance on YN and FC tests.

The average proportions of remember and familiar responses on the YN and FC tests are presented in Table 2. The proportion of remember responses on the two tests did not differ significantly for either the old items, $t(29) = 1.69$, $p = .10$, or the new items, $t(29) = 0.54$, $p = .59$, suggesting that the contribution of recollection was similar on the two tests. Performance was examined further by plotting the proportion of remember responses to old items on the YN test as a function of the proportion of remember responses to old items on the FC test. Figure 8 shows that the proportions of remember responses on the two tests were highly correlated, $R^2(28) = .90$, $p < .01$, and that the correspondence between the two tests was observed across a wide range of remember values. This supports the hypothesis that the contribution of recollection was similar on the YN and FC tests.

As in Experiments 2A and 2B, to determine whether overall accuracy in the YN and FC tests was comparable, we used the dual-process and unequal-variance models to derive parameter estimates for the YN test. We then used these estimates to make predictions about overall FC performance using Equation 1, presented earlier. In the case of the dual-process model, we used the proportion of correct remember responses as an index of recollection [$R_o = P(R \mid old) - P(R \mid new)$][8] and then calculated d' for the items that were recognized on the basis of familiarity (e.g., Yonelinas et al., 1996). That is, we calculated the proportion of old items

[7] Although we continue to refer to the *remember–know* procedure, we have found that it is much easier to communicate to the participant concerning the required judgment if we use the term *familiar* in place of *know*.

[8] Some readers may be concerned about our correction for "remember" false alarms, assuming that the threshold nature of recollection in the dual-process model precludes the possibility of false recollections. However, threshold theory does not preclude the possibility of false alarms. A discussion of the causes of these false alarms would take us away from the point of this article; consequently, we will just mention, as one possibility, that some recollective false alarms may result from cases in which a new test item leads participants to recollect a study item that is similar enough to be confused with the test item. In any case, in most of our experiments the probability of false recollection has been less than 5%, and thus the method that one uses to correct for false remember responses does not make a major difference. However, in experiments in which the new items are highly related to the studied items (e.g., Dobbins, Kroll, & Liu, 1998; Roediger & McDermott, 1995), levels of false recollection can be much higher. Whether the threshold correction is appropriate in these latter cases is not yet clear.

Table 2
Mean Proportion and Standard Error of Remember and Familiar Responses for Targets and Lures in Experiment 3

	Hits				False alarms			
	Remember		Familiar		Remember		Familiar	
Test	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
YN	.456	.032	.185	.018	.022	.007	.102	.011
FC	.475	.035	.360	.026	.019	.004	.147	.016

Note. YN = yes–no; FC = forced choice.

that were accepted on the basis of familiarity given they were not recollected [F_o = P(F | old) / (1 – P(R | old)]. Then, we did the same for the new items [F_n = P(F | new) / (1 – P(R | new)]. Treating the F_o and F_n values as hits and false alarms, we calculated *d′* for each participant. The average observed FC score ($M = 0.83$, $SE = 0.02$) was almost identical to the average score predicted by the dual-process model ($M = 0.841$, $SE = 0.0149$, $p = .5$). The mean difference between the observed FC score and that predicted by the dual-process model was only −0.01, and this difference was not significant, $t(29) = -0.69$, $SE = 0.009$; moreover, the correlation between predicted and observed scores was highly significant, $R^2(28) = .66$, $p < .01$. Thus, combined with the finding of comparative proportions of remember responses given on the two tests, the accurate prediction of the overall proportion correct on FC from the YN responses by means of the dual-process model supports the hypothesis that FC and YN are, in fact, tapping the same memory processes, and with similar degrees of sensitivity.

The above calculations, of course, make the assumption that the recollection and familiarity estimates from the ROC calculations are tapping the same underlying processes as are the estimates from the remember–know procedure. For an extended discussion of evidence for this assumption, see Yonelinas (2001a, 2001b).

To our knowledge, proponents of the unequal-variance model have never attempted to actually fit the results of a remember–know experiment. However, at least one such proponent has suggested that the unequal-variance model assumes that remember responses simply reflect a strict response criterion and that overall recognition (R + F responses) reflects a more lenient response criterion (Donaldson, 1996, p. 524). (Notice, however, that this neither predicts nor explains that the number of remember responses should be the same across the two memory tests.) On the basis of this two-criteria rationale, we connected the resulting two *z* coordinates for each participant and used the resulting line to estimate V_o and *d′* (Macmillan & Creelman, 1991, p. 66). On the basis of these parameters, the overall FC performance was predicted for the unequal-variance model as in Experiments 2A and 2B. However, we ran into difficulties at the first step. Ten of our 30 participants had no false R responses, even though they had between 0.15 and 0.59 correct R responses (6 of these participants had over 0.40 correct R responses). Thus, the *d′* scores for one third of our participants were undefined. To try our best to fit the unequal-variance model to this data, we replaced all 0.0 false R values with 0.0091 (the equivalent of one false R response) before estimating the V_o and *d′* as above. The mean difference between the observed FC score and that predicted by the unequal-variance model was only 0.026, and this difference was not significant, $t(29) = 1.29$, $SE = 0.02$. However, unlike the values predicted from the dual-process model, the values predicted by the unequal-variance model did not correlate significantly with the observed scores, $R^2(28) = .01$.

According to the evidence, the unequal-variance model (or at least our interpretation of it) does not seem appropriate for the evaluation of the remember–know paradigm. However, if the fit is forced, the unequal-variance model also finds that FC performance is approximately equivalent to that of YN performance when tested by the remember–know paradigm. The dual-process model finds both a correspondence between FC and YN remember responses and overall performance at both the level of the overall averages and that of the individual participant.

Conclusion

The results of the current experiments suggest that YN and FC recognition memory tests do not differ greatly in terms of overall accuracy or in terms of the contribution of recollection and familiarity. The results of Experiment 1 showed that when *d′*, found by means of traditional signal-detection theory, was used to measure accuracy, *d′* appeared to be greater in the YN than FC test conditions, but this was only observed with the participants who adopted a strict criterion during the YN test. The sensitivity of the *d′* measure to the criterion suggested that *d′* did not provide a reliable measure of accuracy for the recognition tests. In Experiments 2A and 2B, performance was examined as a function of response confidence. It was demonstrated that when a strict response criterion was adopted, *d′* was greater in the YN than in the FC tests, but, as response criterion was relaxed, the pattern reversed such that *d′* was worse in the YN than in the FC test. The results verified that *d′* did not provide a reliable measure of recognition accuracy and that it should not be used to compare YN and FC accuracy. Furthermore, the signal-detection model underlying *d′* was found to provide a poor fit for the observed ROCs. In contrast, both the dual-process and unequal-variance modifications of the signal-detection model were found to provide accurate accounts of the ROC data, and they were able to accurately predict

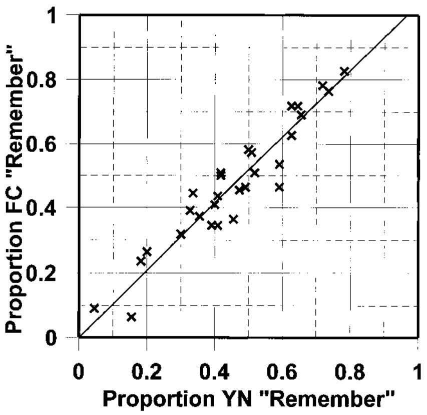

Figure 8. Proportion of remember responses on the forced-choice (FC) test as a function of the proportion of remember responses on the yes–no (YN) test.

FC performance on the basis of the YN ROC data, indicating that the same parameters that describe YN performance can be used to describe FC performance. Finally, in Experiment 3, the proportions of remember responses were almost identical in the YN and FC tests, suggesting that the two tests relied to a similar extent on recollection. Moreover, when evaluated with the dual-process model, YN performance was again found to accurately predict FC performance, indicating that overall performance on the two tests was comparable.

The results are consistent with previous studies that have reported no difference in accuracy between YN and FC recognition memory tests (e.g., Green & Moses, 1966; Khoe et al., 2000; Yonelinas et al., 1992). However, given that these earlier studies used d' as a measure of accuracy, their results must be carefully reconsidered. That is, if d' provides a distorted index of accuracy in recognition memory experiments, then those findings are inherently ambiguous with respect to overall accuracy. Without additional measures of memory performance, such as confidence judgments or remember–know judgments, it is impossible to determine whether YN and FC performance was indeed comparable. Only one previous study provided additional measures of performance. That is, in the study by Khoe et al. (2000), remember–know responses were collected, and their results indicated that the contribution of recollection was comparable across YN and FC recognition memory tests for previously studied words. Moreover, Khoe et al. examined the recognition performance of amnesics, who presumably are relying more on familiarity than on recollection, and found that performance did not differ in the two tests. Taken together with the current results, it appears that, at least in recognition memory tests for words and pictures, there is little evidence that YN and FC performance differ in overall accuracy or in the contribution of recollection and familiarity.

There may of course be conditions in which accuracy on the YN and FC test procedures are not identical. For example, there may be other encoding conditions, other types of materials, or other participant populations in which differences in YN and FC performance might arise. Future studies that examine the generalizability of the current results will be necessary. It will, however, be critical in those future studies to collect more than simple proportion correct scores if one wishes to compare performance on the YN and FC tests.

In the current set of experiments, both the dual-process and unequal-variance models led to the same conclusions regarding the relationship between YN and FC test performance, and both provided good fits for the ROC data. Thus, the conclusion that overall YN and FC accuracy levels are similar holds whether one prefers the dual-process or unequal-variance approach. Nevertheless, our preference is for the dual-process model for several reasons.

First, the unequal-variance model ran into several difficulties in accounting for the results of Experiment 3:

1. It could not explain the large number of participants who did not make any false alarms when indicating a "remember" experience.
2. It does not include a measure of recollection, and consequently, it makes no obvious predictions about remember responses. One attempt to remedy this deficiency is to argue that remember responses reflect a very strict response criterion in the YN test (e.g., Donaldson, 1996; but see Dobbins, Khoe, Yonelinas, & Kroll, 2000), but there is no principled way, of which we are aware, for predicting that the number of remember responses in the FC test should be the same as that observed in the YN test. That is, a strict response criterion on the YN test would be based on the underlying raw familiarity distributions for the studied and new items (which results in an asymmetric ROC function, supposedly because of the increased variance of the studied items). On the other hand, responses in the FC test would be based on the underlying difference distribution, which results in a symmetric ROC function. Thus, there is to our knowledge no reason to predict that the placement of a strict criterion should result in the same number of high-confidence responses. There was, for example, no such correspondence between numbers of items endorsed with a 6 in the YN tests and those items endorsed by a 1 and a 6 during the FC tests of Experiments 2A and 2B.
3. Unlike the predictions from the dual-process model, the predictions from the unequal-variance model did not fare well at the level of the individual participant.

Second, in experiments designed to discriminate between the two models, it is clear that the dual-process model provides a better account of the data than does the unequal-variance model. For example, the dual-process model predicts that under conditions in which performance is expected to rely primarily on recollection, the ROCs should become more linear because recollection is assumed to reflect a threshold process. Linear ROCs have been observed in tests of associative recognition and source recognition (e.g., Rotello, Macmillan, & Van Tassel, 2000; Yonelinas, 1997, 1999; Yonelinas, Kroll, Dobbins, & Soltani, 1999; but see Qin, Raye, Johnson, & Mitchell, 2001), indicating that the unequal-variance model is inconsistent with the recognition data. Moreover, an examination of individual differences in a remember–know experiment demonstrated that false-alarm rates were independent of remember rates across individuals and were instead related to hit rates by means of the familiarity response criterion. This indicated that performance did not reflect the operation of a single underlying familiarity process but rather was more consistent with dual-process models of recognition (Dobbins et al., 2000).

Third, a growing number of neurobiological studies indicate that recognition memory reflects the contribution of two separate retrieval processes. For example, results from lesion and activation studies in rats and nonhuman primates indicate that recognition memory judgments are supported by two functionally and anatomically distinct temporal lobe networks (Aggleton & Brown, 1999; Eichenbaum, Otto, & Cohen, 1994). Moreover, in humans, recollection- and familiarity-based recognition memory judgments are associated with event-related potentials that exhibit distinct temporal and spatial scalp distributions, indicating that that the two types of recognition judgments involve distinct neural generators (e.g., Curran, 2000; Düzel, Yonelinas, Mangun, Heinze, & Tulving, 1997). Similarly, lesion and neuroimaging studies have indicated that the hippocampus is necessary for recollection, whereas other temporal lobe regions support familiarity-based recognition (e.g., Aggleton & Shaw, 1996; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Yonelinas et al., 1996; Yonelinas, Hopfinger, Buonocore, Kroll, & Baynes, 2001).

Whether one prefers the dual-process or unequal-variance model, the current results join a growing body of recognition literature showing that single-factor models of recognition, such as the traditional signal-detection model, are incapable of accounting for recognition performance and that the continued use of these models can lead one to draw erroneous conclusions. Using the single-factor estimate of accuracy (d') in Experiment 1 would have led us to erroneously conclude that YN performance was better then FC performance. It was only when the results were more carefully examined that it became apparent there was no evidence for an appreciable difference between YN and FC accuracy levels in the current study.

What, then, is the broader message from these experiments? In our opinion, our results suggest that when researchers rely on signal-detection theory for a statistical decision model to remove the contribution of strategic response biases or guessing strategies from estimates of discriminative abilities, it is necessary to determine whether it is really appropriate for the particular application. Our suspicions were aroused in Experiment 1 when we found that the relationship between FC and YN was a function of the participant's YN criterion. We then examined the YN ROC functions. Curved, symmetric YN ROC functions are good indications that a single-factor signal-detection model is appropriate. Asymmetric ROC functions are an indication that it is not appropriate. If the YN ROC functions are asymmetric, the unequal variance approach is often a good "fix" if all that is needed is a data reduction process. If, however, one wants to have a signal-detection model that matches the model of the underlying cognitive processes, then one should consider whether the unequal-variance model has a reasonable correspondence to the cognitive processes being measured. The most straightforward way of knowing if the YN ROC functions are asymmetric is, of course, to actually find these functions by obtaining different response criteria (i.e., by manipulating the payoff function or by obtaining confidence judgments). But the results of these experiments suggest that another method of testing for asymmetry is to correlate performance differences between FC and YN tests as a function of the response criterion that a particular participant sets on the YN test.

If one believes there may be two cognitive processes involved in a participant's performance, as we believe is the case in recognition memory, then a model similar to our dual-process model may be appropriate. Abstract theoretical concepts (such as hypothesized cognitive processes) need to be reconciled with converging evidence from different kinds of experiments. As detailed above, we believe that we have provided such converging evidence in favor of the dual-process model of recognition memory.

References

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences, 22,* 425–489.

Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia, 34,* 51–62.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28,* 923–938.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing method, encoding strategy, and confidence. *American Journal of Psychology, 94,* 13–26.

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.

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

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

Donaldson, W., & Murdock, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology, 76,* 325–330.

Düzel, E., Yonelinas, A. P., Mangun, G. R., Heinze, H., & Tulving, E. (1997). Event-related brain potential correlates of two states of conscious awareness in memory. *Proceedings of the National Academy of Science, USA, 94,* 5973–5978.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Rep. No. AFCRC-TN-58–51). Bloomington: Indiana University, United States Air Force Operational Applications Laboratory.

Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral & Brain Sciences, 17,* 449–517.

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.

Freed, D. M., Corkin, S., & Cohen, N. J. (1987). Forgetting in H.M.: A second look. *Neuropsychologia, 25,* 461–471.

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

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

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 19,* 579–583.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66,* 228–234.

Griggs, R. A., & Keen, D. M. (1977). The role of test procedure in linguistic integration studies. *Memory & Cognition, 5,* 685–689.

Hirst, W., Johnson, M. K., Kim, J. K., Phelps, E. A., Risse, G., & Volpe, B. T. (1986). Recognition and recall in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 445–451.

Holmgren, J. E. (1974). Visual search in a forced-choice paradigm. *Perception & Psychophysics, 16,* 253–258.

Huppert, F. A., & Piercy, M. (1976). Recognition memory in amnesic patients: Effects of temporal context and familiarity of material. *Cortex, 12,* 3–20.

Huppert, F. A., & Piercy, M. (1978). The role of trace strength in recency and frequency judgements by amnesic and control subjects. *Quarterly Journal of Experimental Psychology, 30,* 347–354.

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia, 38,* 1333–1341.

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.

Nolde, S. F., Johnson, M. K., & Raye, C. L. (1998). The role of prefrontal cortex during tests of episodic memory. *Trends in Cognitive Sciences, 2,* 399–406.

Qin, J., Raye, C. L., Johnson, M. K., & Mitchell, K. J. (2001). Source

ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1110–1115.

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

Reed, J. M., Hamann, S. B., Stefanacci, L., & Squire, L. R. (1997). When amnesic patients perform well on recognition memory tests. *Behavioral Neuroscience, 111,* 1163–1170.

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

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

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers.* New York: Wiley.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

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

Warrington, E. R., & James, M. (1967). An experimental investigation of face recognition in patients with unilateral cerebral lesions. *Cortex, 3,* 317–326.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: Evidence for a dual-process signal-detection model. *Memory & Cognition, 25,* 747–763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434.

Yonelinas, A. P. (2001a). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London, 356,* 1–12.

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

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

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345–355.

Yonelinas, A. P., Hopfinger, J. B., Buonocore, M. H., Kroll, N. E. A., & Baynes, K. (2001). Hippocampal, parahippocampal and occipital-temporal contributions to associative and item recognition memory: An fMRI study. *Neuroreport, 12,* 359–363.

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

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I. G., & Soltani, M. (1999). Recognition memory for faces: When familiarity supports associative recognition judgments. *Psychonomic Bulletin and Review, 6,* 654–661.

Zola-Morgan, S., & Squire, L. R. (1992). The components of the medial temporal lobe memory system. In L. Squire & N. Butters (Eds.), *Neuropsychology of Memory* (2nd ed., pp. 325–335). New York: Guilford Press.

Received May 8, 2001
Revision received January 17, 2002
Accepted January 17, 2002 ■