Receiver Operating Characteristics (ROCs) in Recognition Memory: A Review

Andrew P. Yonelinas and Colleen M. Parks
University of California, Davis

Receiver operating characteristic (ROC) analysis is being used increasingly to examine the memory processes underlying recognition memory. The authors discuss the methodological issues involved in conducting and analyzing ROC results, describe the various models that have been developed to account for these results, review the behavioral empirical literature, and assess the models in light of those results. The empirical literature includes studies of item recognition, relational recognition (e.g., source and associative tests), as well as exclusion and remember–know tasks. Nine empirical regularities are described, and a number of unresolved empirical issues are identified. The results indicate that several common classes of recognition models, such as pure threshold and pure signal detection models, are inadequate to account for recognition memory, whereas several hybrid models that incorporate a signal detection-based process and a threshold recollection or attention process are in better agreement with the results. The results indicate that there are at least 2 functionally distinct component/processes underlying recognition memory. In addition, the ROC results have various implications for how recognition memory performance should be measured.

Keywords: recognition, receiver operating characteristics, signal detection theory, remember–know, dual process

Editor’s Note. Janet Metcalf served as a guest editor for this article.—HC

In tests of recognition memory, individuals must discriminate between target and lure items, such as items that were previously studied and items that are new to the experiment. The simplicity of these tests has made them one of the cornerstones of memory research. However, despite the enormous amount of work aimed at uncovering the processes underlying recognition memory, much remains to be learned, and debates regarding these processes still abound. One tool that is being increasingly used to examine the processes underlying recognition memory is the analysis of receiver operating characteristics (ROCs). A sufficient body of ROC research has now accumulated that it is useful to review those results and consider the theoretical implications of this body of work. As the following review shows, the examination of recognition ROCs is useful in testing a wide range of different memory models, and it speaks directly to various current debates regarding the processes underlying recognition memory.

An ROC is the function that relates the proportion of correctly recognized target items (i.e., the hit rate) to the proportion of incorrectly recognized lure items (i.e., the false alarm rate) across variations in response criterion (i.e., the propensity to make a positive recognition response). In a test of item recognition memory, the hit rate is equal to the probability of correctly accepting an old item as old, and the false alarm rate is equal to the probability of incorrectly accepting a new item as old. Example ROCs are presented in Figure 1. The leftmost point on each function reflects the hit rate plotted against the false alarm rate when adopting a strict response criterion. Each subsequent point reflects performance at a more and more relaxed response criterion.

ROCs have been examined for several reasons. First, given that a subject must adopt some response criterion to make an old–new recognition memory judgment, theories of recognition performance must be able to characterize the relationship between accuracy and response criterion if they are to provide a comprehensive account of recognition performance. Second, ROC results prove to be much more constraining than standard old–new recognition results. That is, memory studies in which subjects are required only to make simple old–new discriminations produce an ROC with only one point (a hit rate and a false alarm rate). To appreciate the usefulness of ROC studies, one need contemplate only how many different theories might account for such a single point—it is difficult to imagine any theory having any difficulty. However, examining various points along the ROC provides much greater constraint, and thus many fewer theories are able to account for these results. Third, if one can find a theory that adequately accounts for the relationship between accuracy and response bias, then it could be used to derive estimates of accuracy that are not distorted by the particular response criterion that the subject selects at a given time. Without such a theory, it may be impossible to determine if an experimental manipulation has influenced accuracy or response bias, or both. So determining which theory can accurately account for ROCs has implications not only...
for theory development but for measuring recognition memory performance itself.

ROC results can also be used to address important theoretical questions about the nature of the processes underlying memory. For example, ROC studies have already had a profound influence on memory research. In fact, it is fair to say that the dominance of memory theories based on signal detection theory that we currently see in the memory literature is a direct consequence of ROC studies conducted in the late 1950s and early 1960s. That is, it became apparent that recognition ROCs were curvilinear, leading to the rejection of various threshold theories that predicted linear ROCs and providing support for signal detection-based theories that naturally predicted curvilinear ROCs, points we address in detail below (for earlier reviews of this issue, see Murdock, 1974; Kinchla, 1994). In addition, ROC results speak directly to several other questions, such as does recognition memory reflect the contribution of one or more underlying components or processes? What is the nature of the memory signal or signals underlying recognition performance? Do different recognition tasks such as tests of item, source, and remember–know (RK) recognition rely on similar or different processes?

The current review is divided into three main sections. First, we begin by discussing the methodological and analytical issues related to conducting ROC studies of recognition memory. Although the ROC methods are relatively simple, there are several aspects of experimental design and analysis that require careful attention because of potential methodological problems that can seriously bias or limit the conclusions that are drawn from the ROC results.

In the second section, we describe the various quantitative theories that have been proposed to account for ROC results, discuss their underlying assumptions, and delineate their predictions (those models are illustrated in Figures 2 and 3). These theories fall into three general classes—threshold, signal detection, and hybrid models. Although each of these various models can be treated either as a purely descriptive/measurement model or as a psychological/explanatory model, we will focus on the latter. That is, because each of these models can be formalized as a mathematical equation that relates the hit rate to the false alarm rate, they can be treated as purely descriptive in the sense that one can ask if they provide an accurate mathematical description of a given ROC. If they do provide an accurate description of an ROC, then the model equations can be used to fit the data and provide measures of the model parameters, such as $d'$ in signal detection models or recollection and familiarity in dual process models. A descriptive model may work well for one type of recognition experiment and thus can be treated as quite useful, even if it fails completely in another type of recognition experiment. Although it is important to have good descriptive models of performance, most psychologists wish to know if a model serves as a good explanatory (or process) model. That is, our goal is often to determine whether a model provides a meaningful explanation of behavior and if it accurately captures the underlying psychological components or processes. As such, one can ask questions such as do the model’s parameters act in psychologically meaningful ways? Is a model able to account for results from a wide range of recognition tasks and conditions? Does a model make any novel predictions that have been verified? Our primary goal will be to assess these models as psychological explanatory theories, and thus we will attempt to answer these latter questions. Nonetheless, it is also useful to understand the conditions in which each model provides a reasonable description of the data, even if they do not provide particularly compelling explanations of the results.

One class of models that we will not attempt to evaluate here are computational models such as the global matching models (e.g., theory of distributed associative memory [TODAM], Murdock, 1993) and neuroanatomical models (e.g., Complementary Learning Systems, McClelland, McNaughten, & O’Reilly, 1995). Although ROC results do present some important challenges to such models (e.g., Ratcliff, Shu, & Gronlund, 1992), these models have not yet been adequately evaluated with respect to the types of ROCs that they produce, making it difficult to draw firm conclusions about how well these models account for ROC results. Nonetheless, because many of these computational models adopt the signal detection and threshold assumptions of the simpler quantitative models, the current review has implications for the computational models as well, and we will briefly comment on these models in the discussion.

In the third section, we review the results from empirical ROC studies of recognition memory and discuss the theoretical implications of those results. The most general empirical regularities are listed in Table 1. Several types of recognition tasks will be discussed. First, in item recognition tests, subjects study a list of items such as words or pictures and are then presented with a mixture of studied and nonstudied items and are required to indicate if each item was old or new. In contrast, in relational recognition tests, subjects must indicate whether the test stimulus was associated with a specific aspect of the study event rather than just indicate whether the item was old or new. For example, in source memory tests subjects are required to discriminate between items from one source and those from another (e.g., male/female speaker, right/left side of screen, presented in green/red, etc.). In associative memory tests, subjects are required to discriminate between pairs of items that were studied together from pairs in which each item was studied with some other item (i.e., Study: A–B, C–D, E–F, etc.; Test: C–D?, A–F?, etc.). Another example of a relational recognition test is the plurality-reversed test in which subjects must discriminate between studied words and words that were studied in the reversed plurality (e.g., Study: apple; Test: apples or apple?). A related test is the word conjunction test, in which subjects study compound words and are then tested on repeated and rearranged compounds (e.g., tested on blackbird when blackmail and jailbird had been studied). Another general type of recognition task in which ROCs can be examined is the exclusion task (e.g., Jacoby, 1991) in which subjects are instructed to respond old to one class of items (e.g., items from list 1) and to respond no to (or exclude) items from another class (e.g., items from list 2), as well as to new items not previously studied. Finally, in the RK recognition memory task (Tulving, 1985), subjects are asked to indicate whether they recognize test items on the basis of conscious recollection of information about the study experience (i.e., a remember response) or on the basis that it is familiar in the absence of recollection (i.e., a know response).

In the current empirical review, we focus on studies that have directly examined behavioral recognition ROCs to characterize how the ROCs behave in various conditions and recognition tasks. Although some of the models we consider make additional predictions about the neural correlates of the processes underlying ROC shape, we limit our review to purely behavioral studies. For
a consideration of ROC theories in light of neuroimaging and electrophysiological ROC data, several recent articles are available (e.g., Azimian-Faridani & Wilding, 2006; Daselaar, Fleck, Dobbins, Madden, & Cabeza, 2006; Montaldi, Spencer, Roberts, & Mayes, 2006; C. M. Parks & Yonelinas, 2007; Ranganath et al., 2004; Rugg & Yonelinas, 2003; Wixted, 2007; Woodruff, Hayama, & Rugg, 2006; Yonelinas, Otten, Shaw, & Rugg, 2005).

A final point on terminology is in order. The terms that have been used to describe ROC results vary considerably across studies. In the current review, we will use the term ROC to refer to ROCs that are plotted in probability space (e.g., hit rates vs. false alarm rates). In contrast, we will use the term zROC to refer to the function when it is plotted in z-space and will use the terms z-slope and z-intercept to refer to the slope and intercept of the y-axis of the zROC. We use the terms linear, U-shaped, and inverted-U-shape to describe the shapes of the ROCs.

**ROC Methods**

**Collecting ROC Data**

ROCs can be obtained in various ways, but by far the most common is the confidence rating method. For example, after studying a list of items, subjects are presented with a mixture of studied and new items and are required to indicate how confidently they recognize each item on a continuous scale, such as a six-point scale ranging from 1 (sure new) to 6 (sure old). A slightly different procedure is to have subjects first make an old–new response then rate the confidence of each response on a scale from 1 to 3. The ROC is then constructed by plotting hit and false alarm pairs beginning with the most confidently recognized items (e.g., hits = P[6|old]; false alarms = P[6|new]) then repeatedly recalculating the values by including the next most confidently recognized items (e.g., hits = P[6|old] + P[5|old]; false alarms = P[6|new] + P[5|new], etc.). It is important to note that the function is cumulative and thus both the hits and false alarms are constrained to increase or remain constant as the scoring criterion is relaxed. Figure 1 presents confidence-based ROCs from one of the first ROC studies of recognition memory (Experiment 1, Egan, 1958). Note that chance performance would be reflected by a function lying on the diagonal (i.e., hits = false alarms), and increasing accuracy is associated with a function moving toward the upper left, such that the greater the area under the curve, the greater the memory sensitivity or discriminability.

![Figure 1](image_url)

**Figure 1.** A: Item recognition receiver operating characteristics (ROCs) for weak (one study presentation) and strong (two study presentation) items from Experiment 1 in Egan (1958; from Figure 20) plotted in probability space and z-space. The figure illustrates that item ROCs are concave in probability space and linear in z-space, they are asymmetrical along the chance diagonal, appearing to be pushed upward on the left side and they have a slope in z-space of less than 1.0. In addition, increasing item strength leads to an increase in sensitivity but does not influence the degree of asymmetry (i.e., the z-intercept increases but the z-slope remains constant). B: Item recognition ROCs and zROCs for items encoded deeply or shallowly (Experiment 1, Yonelinas et al., 1996). Increasing depth of processing increases sensitivity and leads the ROC to become more asymmetrical (i.e., the z-intercept increases and the z-slope decreases).
ROCs can also be obtained by varying the test demands such that subjects adopt different response criteria in different conditions. For example, subjects can be instructed to respond yes if they are “absolutely sure,” “very sure,” or “reasonably sure” in different conditions, or they could be incorrectly told the test list consists of 30%, 50%, or 70% old items. Alternatively, the proportion of old and new items in the test list can be varied (i.e., if most test items are old, subjects tend to adopt a lax criterion and respond old to a greater proportion of the test items), or the “payoff” or cost of hits and false alarms can be varied (e.g., by providing various monetary rewards for avoiding false alarms). These methods are less common than the confidence method, probably because it is more time consuming to collect responses separately for each level of response criterion than it is to use a confidence scale. One additional concern with these latter methods is that the experimental manipulations may have effects on memory sensitivity as well as response bias. For example, under strict response instructions subjects may adopt different retrieval strategies, such as waiting longer for retrieved information to accrue. Although this possibility probably deserves further investigation, comparisons of the ROCs obtained from the confidence method with the proportion old method are quite similar, suggesting that these methods are comparable (e.g., Ratcliff et al., 1992; but see Van Zandt, 2000, for evidence that manipulating stimulus probabilities and payoffs may lead the ROCs to become slightly more asymmetrical as the response bias becomes strict).

Analyzing ROC Data

ROC results can be analyzed in various ways. To quantify the shape of the ROCs, it is common to plot them in z-space by simply taking the z-score (i.e., the inverse of the standard cumulative normal distribution assuming a mean of 0 and a standard deviation of 1) of each hit and false alarm rate (e.g., see Figure 1). The theoretical motivation for doing so is grounded in signal detection theory and is discussed below. If the zROC is linear, then the intercept can be used as a rough index of recognition accuracy, and the slope can be used to index the asymmetry of the ROC. The ROC in Figure 1 is asymmetrical along the diagonal such that it appears to be pushed up along the left side. If the ROC were perfectly symmetrical, the z-slope would be 1.0, but when it is asymmetrical as in Figure 1 the z-slope will be less than 1.0, and if it is asymmetric in the other direction it will have a z-slope greater than 1.0.

Z-slope and z-intercept are the most common measures derived in ROC studies. Note, however, that there are various alternative measures of ROC accuracy that may be slightly more appropriate in specific cases. For example, the sensitivity measure da (or the associated area measure Az) should be used when the zROC is linear and asymmetrical, and other measures are more appropriate when the zROCs are not linear (see Macmillan & Creelman, 2005).

One common analytic approach is to conduct standard regression analyses on either the ROCs or the zROCs to assess the slope and intercept of the functions. Note that if the functions are not linear then the slope and intercept are not particularly useful because they will depend on which portion of the function is assessed. The regression method can also be used to assess the linearity of the ROCs or zROCs, by adding a quadratic component to the regression equation and testing to see if it provides a significant improvement over the fit of the linear equation (see Glanzer, Hilford, & Kim, 2004; Heathcote, 2003; Hilford, Glanzer, Kim, & DeCarlo, 2002; Qin, Raye, Johnson, & Mitchell, 2001; Yonelinas, 1997). A potential limitation of this approach, however, is that the standard regression method assumes that the ROC points vary only in the y-dimension (they can, in fact, vary in both x- and y-dimensions) and that each point on the function is independent (which is not true in a cumulative confidence-based ROC). A potentially useful method would be based on maximum likelihood, but to date, this method has not been adapted to fitting functions in z-space.

Rather than simply assessing the linearity or curvilinearity of the ROCs, an alternative (or complementary) analytic approach is to directly fit theoretically based models to the observed ROC data. For example, various algorithms are available for fitting signal detection models to the ROCs (e.g., Metz, Herman, & Roe, 1998, http://krlindex.htm; Dorfman & Alf, 1969, http://perception.radiology.uiowa.edu; Macho, 2002, 2004, http://www.unifr.ch/psycho/general/english/people/macho/research.php; for a discussion of various methods, see Stanislaw & Todorov, 1999), as well as Excel spreadsheets that can be used to fit various signal detection and dual process models to ROCs (http://psychology.ucdavis.edu/labs/Yonelinas; see also http://www.unifr.ch/psycho/general/english/people/macho/SDT.php). There are currently two main ways to fix these models to the data. One way is to use a regression method that assumes variability in x- and y-dimensions (e.g., Yonelinas, 1999). Another is to use maximum likelihood estimate methods to fit linear and nonlinear equations to the observed confidence responses rather than the derived ROCs (e.g., Healy, Light, & Chung, 2005; Macho, 2002, 2004). Although these various methods are somewhat different, in the cases in which they have been compared, they have led to comparable results (e.g., Hilford et al., 2002; Van Zandt, 2000; Yonelinas, 1999; although we have found that the maximum likelihood method can provide rather peculiar fits to the ROCs when performance becomes very high).

The advantage of these model-based approaches is that they provide estimates of theoretically motivated parameters. For example, the signal detection methods provide measures of sensitivity and old–new variance ratio parameters, and dual process methods provide estimates of recollection and familiarity. These methods can be used to fit individual subject ROCs or to fit average ROCs.

One benefit of using the model-based approach to fitting ROCs is that the parameter estimates derived from the models can be assessed using standard statistical methods like analysis of variance to determine how the parameters are affected by experimental variables. So, for example, one can determine whether an experimental variable has a significant effect on recollection and/or familiarity in dual process models, or if it affects sensitivity and/or variance ratio in the signal detection models, simply by deriving estimates of the parameters for each subject in the various experimental conditions and directly comparing those estimates.

Another advantage of the model-based approach is that it is possible to directly contrast the fits of alternative models. If the models are nested (e.g., the equal variance signal detection [EVSD] model is a simplified version of the unequal variance signal detection [UVSD] model), then one can contrast the
goodness-of-fit measures of each model (e.g., $R^2$ or $\chi^2$ values). If
the models are not nested, then it is possible to use goodness-of-fit
indices such as Akaike’s information criterion (Akaike, 1973) or
the Bayesian information criterion (Schwarz, 1978). These
goodness-of-fit measures take into account the number of param-
eters estimated by the model, penalizing models with more param-
eters. Although these latter measures do not indicate if one model
provides a significant improvement over another, they can show if
one provides a numerically better fit than does another. Note,
however, that although these latter two measures of fit are often in
agreement with respect to which model provides the best fit, in
some cases they can be in disagreement (e.g., Healy, Light, &
Chung, 2005). Moreover, the fits of competing models can often
all be exceptionally good, such that all models differ from the
observed data by less than one or two percent. In such cases, the
fit measures are not found to be particularly useful in assessing the
adequacy of the competing models (for discussion of this issue, see
Healy et al., 2005; Macho, 2002, 2004; Yonelinas, 1999).

**ROC Measurement Issues**

In conducting an ROC experiment, several practical issues need
to be kept in mind. First, it is necessary to obtain a sufficient
number of response categories to accurately characterize the shape
of the ROC. The majority of the existing recognition ROC studies
have examined six levels of recognition confidence, which leads to
five ROC points because the most extreme point is constrained to
be 1,1. Using fewer categories will make it more difficult to
characterize the shape of the function. Additional response cate-
gories may be added, but the statistical advantage of increasing
beyond about six categories is diminishingly small (Nishisato &
Torii, 1970).

Second, it is necessary to obtain a sufficient number of re-
sponses in each condition. Because the aim is to determine the
shape of the ROC, there should be a reasonable number of re-
sponses per point to estimate that function. More systematic em-
pirical and theoretical studies are needed to address what the
effects of having too few responses might be. However, we have
found that when ROCs are based on less than 60 items per
subject/condition (i.e., 60 new items plus 60 old items from each
study condition) the functions can become poorly behaved in the
sense that subject ROCs become noisy and irregularly shaped.
Macmillan, Rotello, and Miller (2004) conducted various simula-
tions based on a signal detection model and found that accuracy
could be estimated well with as few as 50 responses per condition,
but slope estimates were more variable and require more responses
per condition. They also found that as performance levels became
very high the variability in slope estimates increased, and so they
recommended that $d$ scores be kept below 2.0 (see Macmillan &
Creelman, 2005, for $d$ and other alternative measures of sensitiv-
ity).

Third, if groups of subjects are tested (rather than traditional
psychophysical studies of one or two subjects), it may be useful to
examine the “average” ROC by aggregating data across subjects,
but it is important to determine if the average ROC provides an
accurate representation of the individual subject ROCs. This is
critical because of the potential for averaging artifacts (Brown &
Heathcote, 2003; Wickens, 2002). For example, the average of two
or more nonlinear functions may be a function that does not look
like either of the individual functions. As an illustration, imagine
averaging an $S$-shaped function and a $C$-shaped function together.
So, in principle, a model may provide an accurate account of the
average ROC but fail to provide an acceptable account of the
individual subject ROCs. This can be addressed by collecting a
sufficient number of responses from each subject so that the results
of each subject can be analyzed and modeled. To determine
whether the average ROC is representative, one can compare the
parameter estimates from a selected model based on the average of
the ROCs to that obtained by fitting each subject individually and
averaging those parameter values. If there is a discrepancy be-
tween the two sets of parameter estimates, this is a good indication
that there may be averaging artifacts. Regardless, a careful visual
examination of the individual subject ROCs is critical, and an
analysis at the subject level is preferable to an analysis at the
aggregate level.

A related issue is that some subjects may not use the entire
response scale or may spread their responses across the confidence
scale in some idiosyncratic manner, such as using one response
category for a majority of their responses. In this way, individual
subject ROCs may not be useful in inferring the ROC shape and
could artifactually distort the average ROC if they are included.
One option is to remove these subjects from the analysis, but we
have found that such problems can generally be avoided if there
are a sufficient number of trials and if subjects are instructed to
spread responses across the scale, to the extent that they can, given
their experience so that they use the whole scale.

A fourth important issue is related to floor and ceiling effects.
Although floor and ceiling effects can cause problems in any
psychological study, they can be particularly problematic in ROC
studies. The first problem is that as overall recognition perfor-
mance becomes very poor the ROC approaches the chance diag-
onal, at which point it will necessarily have a slope of 1.0. Thus,
assessments of slope when performance is very low are not par-
ticularly meaningful. Ratcliff et al. (1992) suggested that when
overall $d'$ scores fall lower than .5, slopes are clearly constrained
to approach 1.0, so such low levels should be avoided. Very high
levels of performance can also cause problems. For example, as
mentioned earlier it may be difficult to derive stable parameter
estimates when performance becomes too high (Macmillan et al.,
2004).

A fifth important issue, and one of the most common problems
in ROC studies, is what might be referred to as the truncation
problem that can occur when the hits or false alarm rates on the
extreme ends of the ROC approach 0.0 or 1.0. Because ROC
studies examine performance across changes in response bias, the
left- and rightmost points on the ROC will tend to approach the x-
or y-axes, thus hit rates can approach 1.0 and false alarms can
approach 0.0. It is important to realize that these ceiling effects can
occur even when overall recognition accuracy is not very high.
These floor and ceiling effects can lead the mean values of the
extreme ROC points to under- or overestimate their true values. It
may be possible to estimate the true values by using statistical
methods for dealing with truncated or censored data (e.g., Long,
1997), however we are not aware of any ROC studies in which this
has been attempted. Alternatively, one can simply be aware of this
potential problem and be cautious about drawing conclusions that
rest solely on the location of these most extreme points, particu-
larly when they are within 5% of floor or ceiling.
Besides these basic measurement issues, there are several potential measurement artifacts that have been discussed in the ROC literature. For example, Ratcliff, McKoon, and Tindall (1994) conducted a number of model simulations and showed that if the response criterion varied across the experimental session, or if some portion of the responses were random, then this could lead the observed ROCs to exhibit an exaggerated inverted U-shape. Similar observations were made by Yonelinas (2001b), with slightly different models. We would also point out that the floor and ceiling effects produced when hits approach 1.0 or false alarms approach 0.0 could also produce such an artifact. Moreover, Malmberg and Xu (2006) showed that the manner in which the ROC will be distorted by random responses can be influenced by the way these responses are distributed across response confidence, as well as by the overall level of performance and the related response distributions. Further studies designed to determine if these potential distortions do show up in empirical ROCs will be important. In any case, it would seem prudent to instruct subjects to respond carefully on the basis of memory strength and to avoid inducing any dramatic changes in response criterion across the test.

Memory Theory

In this section, we provide a brief outline of the various models that have been proposed to account for recognition ROCs, point out the motivations underlying their core assumptions, and discuss their predictions. The hierarchical relationship between these various models is illustrated in Figures 2 (models) and 3 (predictions), which present the memory strength distributions for each of these models along with the predicted ROCs and zROCs. The models fall into three general classes (i.e., threshold, signal detection, and hybrid models), each with several major variations ranging from simple to more complex (moving from left to right in Figure 2). Note that a brief summary of these models is presented at the end of this section.

Threshold Models

The origins of threshold theory are obscure, but they can be traced back to the psychophysical work of Fechner (see Boring, 1929), who proposed that there is some minimum sensory signal strength (i.e., the threshold or limina) that must be attained before a subject is able to perceive a stimulus. Thus, in the context of recognition memory tasks, threshold theory treats memory as probabilistic in the sense that the memory strength of some of the test items will fall below the threshold and so memory will fail, whereas the memory strength of other items will fall above the strength threshold and these items will be recognized.

The high threshold (HT) model. One of the simplest threshold models is the HT model. The model assumes that studied items are stronger than nonstudied items, but there is a memory threshold that is exceeded only by the old items. The proportion of old items (or target items) above the threshold is the probability \( R_T \). Memory decisions are made by selecting some level of memory strength as a response criterion and accepting items that exceed that level of strength as having been studied. If the response criterion is set exactly at the threshold, then the hit rate will be equal to \( R_T \), and the false alarm rate will be zero. For a particular set of experimental conditions the threshold is fixed, but subjects are free to vary their response criterion. Note that threshold distributions are often represented as rectangular, but this is done only for the sake of simplicity, and the distributions can actually take on various different shapes (e.g., see Macmillan & Creelman, 2005).

Critically, as the response criterion is relaxed (i.e., shifted to the left of the threshold), the hit rate and false alarm rate will increase proportionally, producing a linear ROC like that seen in Figure 3, and if it is plotted on z-coordinates then this zROC exhibits a U-shape. Thus, assessing the linearity of the ROC and zROC provides a direct test of this threshold model. Moreover, the model predicts that as performance increases, the ROC should become more asymmetrical. That is, as memory increases, the left y-intercept will increase and thus the slope of the ROC (and the zROC) will decrease.

The high–low threshold (HLT) model. Another common threshold model is the HLT model (sometimes referred to as the “two-HT” or the “double HT” model), in which a second memory component is added to represent the probability that new items, or lure items, can be recognized as new (\( R_L \), e.g., “I would have remembered my name if it had been in the study list”). As illustrated in Figure 2, the high–low threshold model is identical to the HT model except that the new item distribution extends further to the left than does the old item distribution. The low threshold falls at the leftmost point of the old item distribution, such that any new item falling below this threshold can be recognized as new.

The HLT model generates ROCs like the HT model, except that the right end of the ROC moves up and intersects the upper x-axis at a point that is \( R_L \) from the 1,1 intercept (see Figure 3). Thus, like the HT model, the HLT model predicts linear ROCs that are U-shaped in z-space, but it is not constrained to generate an asymmetrical ROC. In fact, the degree of ROC asymmetry can vary independently of overall performance. If \( R_T \) is greater than \( R_L \), then the ROC will have a slope less than 1; if \( R_T \) is less than \( R_L \), then it will have a slope greater than 1; and if \( R_T \) is equal to \( R_L \), then the ROC slope will be equal to 1 (i.e., a symmetrical ROC).

Although the HT and the HLT models predict linear ROCs, it is possible to produce various nonlinear ROCs by introducing additional parameters or thresholds. For example, Luce (1963) proposed a “two-state high-threshold” model that produces an ROC made up of two joined linear segments (also see Green, 1960; D. A. Norman & Wickelgren, 1965). Thus, threshold models can produce nonlinear ROCs. In fact, by adding multiple steplike thresholds or decision rules, one can produce an ROC that is effectively curvilinear (e.g., Buchner, Erdfelder, & Vaterrodts-Plunnecke, 1995; Krantz, 1969; Malmberg, 2002). However, the latter approach is rarely adopted because it requires an additional free parameter for each new threshold (Hilford et al., 2002). Nonetheless, a consideration of these models is important because it shows that although ROC experiments may prove useful in testing specific threshold models, results that disconfirm one threshold model may not be problematic for another. In the current article, we focus on the HT and HLT models because they are the two most commonly adopted threshold models.
Figure 2. Strength distributions for threshold, signal detection, and hybrid recognition memory models. The models include the high threshold (HT), high–low threshold (HLT), equal variance signal detection (EVSD), unequal variance signal detection (UVSD), two-dimensional signal detection (2DSD), sum–difference theory of remembering and knowing (STREAK), dual process signal detection (DPSD), variable recollection dual process (VRDP), some-or-none (SON), and the mixture model. \(R_T\) = proportion of old items above threshold; \(R_L\) = proportion of new items recognized as new; Signal Detection Models: \(d'\) = strength of old items; \(V_T\) = variance of target (old) items; \(d'_{L,T}\) = lure source strength relative to new items; \(d'_{T,L}\) = target source strength relative to new items; \(V_L\) = lure distribution variance; \(S\) = source; \(RV_f\) = recollection strength; \(F_f\) = familiarity strength; \(FV_f\) = variance of familiarity; \(RV_{T,f}\) = variance of recollection; Hybrid Models: \(R_{T,F}\) = recollection of targets; \(R_{L,F}\) = strength of recollection of targets; \(R_{L,F}\) = recollection of lures; \(R_A\) = associative strength; \(AV_{T,f}\) = variance of associative memory; \(R_{L,F}\) = strength of recollection of lures.
There is one important caveat about using tests of linearity to assess the threshold models. Technically, the ROCs produced by the HT and HLT models (and models in which these theories are nested such as several dual process models) predict ROCs that are kinked at their extremes. For example, when the response criterion moves to the right of the HT in these models, the otherwise linear ROC intersects the y-axis and is forced to drop and approach the 0,0 intercept. For the HLT model, the same will also occur as

Figure 3. Predicted receiver operating characteristics (ROCs; left panel of each pair) and zROCs (the ROC function when it is plotted in z-space; right panel of each pair) for threshold, signal detection, and hybrid recognition memory models. The models include the high threshold (HT), high–low threshold (HLT), equal variance signal detection, unequal variance signal detection (UVSD), dual process signal detection (DPSD), variable recollection dual process model (VRDP), some-or-none (SON), and the mixture model. Note that the two-dimensional signal detection and sum–difference theory of remembering and knowing models make the same ROC and zROC predictions as the UVSD model for item recognition. \(R_T\) = proportion of old items above threshold; \(R_L\) = proportion of new items recognized as new; Signal Detection (SD): \(d'\) = strength of old items; \(V_O\) = variance of target (old) items; DPSD: \(R_L\) = recollection of lure; \(R_T\) = recollection of targets; \(d'\) = familiarity strength; VRDP: \(d'\) = familiarity strength; \(R_T\) = recollection of targets; \(R_{d'T}\) = strength of recollected of targets; \(RV_{d'}\) = variance of recollection; SON: \(R_L\) = recollection of lure; \(R_T\) = recollection of targets; \(d'\) = item strength; \(d'_{L}\) = associative strength; \(V_{T}\) = associative variance; Mixture: \(R_L\) = recollection of lure; \(R_T\) = recollection of targets; \(d'\) = strength of target source; \(d'_{L}\) = strength of lure source; \(d'\) = item strength.
response criterion becomes very lax and the function moves toward the right—both hits and false alarms are forced to approach 1.0. This means that if the extreme points on an ROC approach floor or ceiling levels, the ROC might appear curvilinear, even if it is perfectly predicted by a threshold model. Thus, when assessing threshold-based models, it is important to determine if the extreme points in the ROC are approaching floor or ceiling levels. In these cases, an evaluation of ROC linearity may not provide a valid assessment of these models. It should also be pointed out that this problem is not limited to tests of linearity but also holds for assessments of models that incorporate a threshold process. We are not aware of any studies that have accounted for such kinked ROCs when assessing the fits of threshold-based models.

**Signal Detection Models**

Signal detection theory (e.g., Swets, Tanner, & Birdsall, 1961; Tanner & Swets, 1954) is a statistical decision model that has been applied to studies of item recognition (e.g., Banks, 1970; Egan, 1958; Murdock, 1965; T. E. Parks, 1966) and source recognition (e.g., Hoffman, 1997; Marsh & Bower, 1993). In item recognition tasks, it is typically assumed that studied items have greater memory strength than do nonstudied items but that there is variability in memory strength such that the old and new items form overlapping Gaussian (or normal) distributions as in Figure 2. The distance between the old and new distributions measured in z-scores is $d'$, which represents how much stronger the studied items are than the new items. Recognition decisions are made by setting a response criterion and responding old to all items exceeding that criterion. In contrast to the threshold models, these models are deterministic in the sense that they assume that there is a relevant memory signal for every test item, whether the item is old or new. Moreover, the signal detection models almost always assume that the old and new item distributions are normal (or Gaussian) in shape. The overlapping Gaussian distributions underlying the model lead to curved ROCs that are perfectly linear when plotted in z-space (see Figure 3). Thus, plotting ROC data on z-coordinates and assessing the linearity of the zROC provides a direct test of the Gaussian assumption. Note that finding linear ROCs would provide support for the model’s Gaussian assumptions, but it would not rule out all alternative models, because very similar ROCs can also be produced by models that assume non-Gaussian strength distributions (e.g., see Lockhart & Murdock, 1970, and the hybrid models below).

*The EVSD model.* The simplest signal detection model is the EVSD model. The model assumes that the variance associated with the target items is equal to that associated with the new items. Because the old and new distributions are assumed to have the same shape, the model generates a symmetrical ROC that has a slope of 1.0 in z-space. Thus, assessing the slope of the zROC provides a direct test of the equal variance assumption.

Although the core assumptions underlying the EVSD model are the Gaussian and equal variance assumptions, a number of auxiliary assumptions have sometimes been adopted. For example, many theories have assumed that the memory signal reflects a continuous scalar index of memory strength or familiarity (e.g., such as in global memory models like TODAM, Murdock, 1974; and SAM, Gillund & Shiffrin, 1984). Alternatively, the memory signal could also reflect how many different aspects or features of the test item are remembered (e.g., Johnson, Hashtroudi, & Lindsay, 1993), or it may reflect the products of two or more separate memory processes such as recollection and familiarity (e.g., Wixted & Stretch, 2004; Wixted, 2007). Another assumption, which is often adopted to account for subjective reports of Remembering and knowing from the RK procedure, is that remember responses simply reflect stronger or more confident memories than do know responses (e.g., Donaldson, 1996; Dunn, 2004; Hirshman & Master, 1997). Thus, when RK scores are plotted in ROC space, they should fall along the same symmetrical ROC that is expected to be observed in item recognition, but the remember point (i.e., the proportion of correct remember responses vs. the proportion of incorrect remember responses) would simply fall to the left of recognition point (i.e., remember plus know responses) along that function.

*The UVSD model.* A common modification of the EVSD model is the UVSD model, in which a second memory component is added—the variance of the old (or target) item distribution relative to the new item distribution ($V_{VT}$). If the old item variance is greater than that of the new distribution, then the ROC will appear to be pushed up on the left side, as in Figure 3. If the old item variance is less than that of the new item distribution, the ROC will be pushed up on the right side (this is not illustrated, and it is rarely observed). Like the EVSD model, the UVSD model predicts curved ROCs that are linear when plotted in z-space. However, because the old and new item variances can differ, the model can produce asymmetrical ROCs (i.e., slopes in z-space greater or less than 1.0). Moreover, because the UVSD model has one parameter indexing sensitivity, and another indexing symmetry, the model suggests that the two aspects of the ROC might be experimentally separable. That is, there may be variables that influence $d'$ while leaving $V_{VT}$ unaffected, whereas other variables might influence $V_{VT}$ while leaving $d'$ unaffected. As far as we know, however, no clear predictions have been made about which experimental variables might produce such dissociations.

One property of the UVSD model that is often overlooked (although see Green & Swets, 1966; DeCarlo, 2002) is that if the variance of the old item distribution is greater than that of the new item distribution, the model predicts a curved ROC that drops below the chance diagonal (see Figures 2 and 3). The reason for this is that as the old item variance becomes large, some portion of old item distribution will be pushed below that of the new item distribution, meaning that the encoding phase must have decreased, rather than increased, the memory strength of some of the studied items.

The UVSD model does not specify why the old item variance differs from that of the new items, but it is sometimes argued that the old item distribution is more variable than the new item distribution because of encoding variability (e.g., Hilford et al., 2002; Wixted, 2007). That is, because not all studied items are expected to increase in strength by the same amount, the old item distribution will be more variable than the new item distribution. Such an account leads to the expectation that the ROCs should be asymmetrical such that the z-slopes are less than 1, rather than being equal to or greater than 1.

As with the EVSD model, one can potentially explain RK reports by assuming that remember responses simply reflect high confidence recognition responses (Wixted & Stretch, 2004; Wixted, 2007). The UVSD model accordingly predicts that the RK
data should fall on the same function that is observed in recognition ROC studies, but in contrast to the EVSD model, the recognition ROC can be asymmetrical.

The two-dimensional signal detection (2DSD) model. To account for performance in item and source recognition tests, the UVSD model has been extended to include one memory strength dimension for each of two sources (Glanzer et al., 2004, Hilford et al., 2002; for earlier development of these types of models, see Ashby, 1992; Banks, 2000; Macmillan & Creelman, 1991; Tanner, 1956). Thus, studying an item in one source will increase its strength along that dimension, whereas studying an item in the other source will increase its strength in the second dimension. Because the strength and variance of each source can vary, the model requires four memory parameters (target strength, \( d'\_T \); target variance, \( V\_T \); lure strength, \( d'\_L \); and lure variance, \( V\_L \)). In addition, another parameter is needed to index the distance between the two source distributions (\( d'\_\text{ov} \)) or the angle between the two strength dimensions. Note that the two source dimensions are presented as orthogonal in Figure 2, but they are generally expected to be offset by much less than 90 degrees to account for the observation that accuracy tends to be lower for source than item recognition.

Item recognition discriminations are made by placing a linear response criterion between the new item distribution and the two source distributions, whereas source memory discriminations are made by placing a linear response criterion between the two source distributions. A core assumption of the model is that item and source judgments are based on the same underlying strength distributions, thus the model predicts that performance on these tasks should be directly related. That is, manipulations that increase source recognition will necessarily also increase item recognition. Note, however, that the model does not predict exactly how closely item and source recognition will be related because this will depend on the angle between the two source dimensions and the types of source information that the subject brings to bear when making the item discrimination.

The 2DSD model makes ROC predictions that parallel those made by the UVSD model. That is, because the model is based on Gaussian strength distributions, it predicts that item and source ROCs should be curved in probability space and linear in z-space. Because the model includes free parameters for the variance of the old item distributions, it can produce ROCs with slopes less than 1, and it can produce dissociations between ROC accuracy and asymmetry.

The sum–difference theory of remembering and knowing (STREAK). STREAK is a two-dimensional signal detection model that has been proposed to account for RK and item recognition ROC results (Rotello, Macmillan, & Reeder, 2004). In this model, one dimension represents global familiarity, and another orthogonal dimension represents recollection of specific details associated with the item. Studied items are assumed to have greater recollection and familiarity strengths than the new items (indexed as \( R\_d' \) and \( F\_d' \), respectively). The recollection and familiarity strength distributions for studied items are assumed to have equal variance, but the old item variance is expected to be greater than that of the new items. For the model to be identifiable in standard RK experiments, the old item variance is set to 1.25, and the new item variance is set to 1 (to approximate item recognition z-slopes of .8), but in studies in which RK and confidence responses are collected, the old item variance is treated as a free parameter. Note that in standard item ROC studies, the model is not identifiable because there are an infinite number of recollection/familiarity combinations that could lead to a given level of overall recognition performance, but if RK and confidence judgments are collected concurrently, the model is identifiable. According to the model, item recognition decisions are made by setting a linear response criterion between the new and old item distributions that is parallel to the line intersecting the \( R\_d' \) and \( F\_d' \) values on their respective axes, and RK decisions are made by selecting a perpendicular linear response criterion that is used to determine if the item is more remembered or more familiar. Thus, the sum of the recollection and familiarity strength values is used to make the old–new response, whereas the difference between the two strength values is used to make the RK response.

Because the model is based on Gaussian strength distributions, like the UVSD model it predicts curved item ROCs that are linear in z-space. If the \( V\_T \) parameter is fixed at 1.25, it predicts asymmetrical ROCs with z-slopes of .8. However, if the old item variance is treated as a free parameter, then the item recognition process becomes equivalent to UVSD model. The unique aspect of this model, however, is that because RK judgments are based on different decision rules than are old–new judgments, the model can produce RK z-slopes that differ from the ROC z-slopes. That is, the model can produce a remember ROC point that can fall below, above, or along the confidence ROC, whereas the remember plus know point has to fall exactly on the confidence ROC because it corresponds to all the items exceeding the old–new response criterion. In this way, the RK slope (i.e., the line joining the \( R \) point to some point to the right along the ROC) can be greater, less than, or equal to that of the ROC slope.

Hybrid Models

Various models have been proposed that combine the assumptions of signal detection theory and threshold theory. These models assume that a signal detection process is supplemented by a threshold process. Several of these models assume that these two processes reflect familiarity and recollection, whereas one model assumes that they reflect familiarity and attention.

The dual process signal detection (DPSD) model. The DPSD model (e.g., Yonelinas, 1994, 2001a, 2001b) was developed to account for item recognition ROCs, and it assumes that recognition memory judgments are based on a recollection process whereby qualitative information about the study event is retrieved (e.g., where or when an item was studied), or if recollection fails, recognition is based on a familiarity assessment process like that underlying the EVSD model. Recollection is indexed as the probability that subjects correctly recollect some aspect of the study event (\( R\_k \)), whereas familiarity is indexed as the increase in familiarity related to the study phase (\( d' \)). Subjects are assumed to recollect various different types or amounts of information about a study event, but because recollection will fail for some items it is described as a threshold process. In relational recognition tests where subjects are assumed to be able to recollect the occurrence of the lure items (i.e., the item was in list 2 rather than list 1), an additional recollection parameter is introduced (\( R\_k \)).

Because familiarity is assumed to reflect a signal detection process, the model can produce curved symmetrical ROCs that are
linear in z-space. However, recollection is expected to increase high confidence hits in tests of item recognition, and this will lead the ROCs to become asymmetrical so that the zROCs have a slope less than 1.0. Because increasing recollection will lead the ROCs to become more asymmetrical, whereas increasing familiarity will not, it should be possible to find dissociations between the degree of asymmetry and overall performance. In relational recognition tests, the degree of ROC asymmetry will be determined by the relative probabilities of $R_T$ and $R_{F}$. Because recollection is a threshold process, the model predicts slightly flatter ROCs than do pure signal detection models, and thus the ROCs can exhibit a slight U-shape in z-space, particularly under conditions expected to rely heavily on recollection.

The DPSD model makes a number of additional predictions about how different experimental variables and test conditions should influence the shape of recognition ROC shape. First, when subjects rely heavily on recollection, such as in relational tests, the ROCs should become more linear and will begin to exhibit a more pronounced U-shape in z-space. This is because recollection is a threshold process and because, in tests like source memory, recollection is expected to occur for targets and lures (e.g., items from source 1 and source 2). Note that although tests of relational recognition are expected to rely heavily on recollection, they are not process pure measures of recollection, and therefore the resulting ROCs should not be expected to be perfectly linear. Moreover, it has been assumed that although familiarity is less useful at making relational than item recognition judgments, there are various conditions in which it is expected to contribute to a much greater extent, such as when there are differences in familiarity between items from different sources or when pairs of items are unitized in associative recognition (Yonelinas, 1997, 1999; Yonelinas, Kroll, Dobbins, & Soltani, 1999). Thus, under these conditions the relational ROCs should become more curved and, thus, less U-shaped in z-space.

Second, under conditions in which recollection and familiarity are put in opposition, the model predicts that the ROCs should be curvilinear and negative going (see Figure 3). For example, in Jacoby’s exclusion paradigm (e.g., Jacoby, 1991) subjects respond yes to studied items but respond no to an item if they recollect that it was from a specified source (e.g., list 1). Familiarity will lead the ROCs to be curvilinear like those predicted by signal detection theory, but recollection will be used to reject the excluded-source items, which will effectively push the ROC downward. The resulting ROC crosses the chance diagonal as the familiarity response criterion is relaxed (Yonelinas, 1994; Yonelinas, Regehr, & Jacoby, 1995). In z-space, the exclusion ROC is generally linear and will have a slope of less than 1.0, but as the relative contribution of recollection increases, the zROC will begin to exhibit an inverted U-shape.

Third, if subjects are able to report when they recollect details about a study event, then RK reports should be useful in providing an index of recollection and familiarity (Yonelinas & Jacoby, 1995; Yonelinas, Dobbins, Szymanski, Dhalwal, & King, 1996). Thus, remember reports should be associated with high confidence recognition responses, whereas know reports should reflect familiarity when recollection fails and should be associated with a range of recognition confidence. The RK responses should therefore fall along the same function as the item recognition confidence ROC. That is, the remember ROC point (i.e., the proportion of remember responses to old items plotted against the proportion of remember responses to new items) should fall to the left of the recognition point (i.e., remember plus know responses), and both points should fall along the confidence ROC. The remember point should be close to the highest confidence ROC point but will be shifted slightly to the right if there are high confidence responses based on familiarity. Because the RK and ROC points should fall along the same function, the z-slope of the RK and ROC results should be comparable. Note that Lampinen and colleagues (Lampinen, Odegard, & Neuschatz, 2004; Lampinen, Watkins, & Odegard, 2006) have modified the model to include a false recollection component. Although the modification leads to changes in the predicted shape of the ROCs, it has not yet been applied to RK reports.

Fourth, because manipulations like semantic versus perceptual levels of processing and full versus divided attention increase recollection much more than does familiarity (Yonelinas, 1994, 2002), the model predicts that these manipulations should lead the z-slope of the ROCs to decrease as performance goes up. In contrast, manipulations that have comparable effects on both processes, such as study duration, should lead the z-slopes to be roughly constant as performance goes up.

Fifth, because recollection and familiarity reflect distinct memory processes, it has been assumed that they exhibit partially distinct neural correlates. Consistent with previous dual process models, it has been assumed recollection relies more heavily than does familiarity on the medial temporal lobes (necessary for linking arbitrary aspects of study episodes together) and the prefrontal cortex (necessary for elaboration and strategic search). Thus, because recollection is expected to be particularly disrupted in these patient groups, these patients’ ROCs should be much more symmetrical than are those of the control subjects, and the ROC parameter values should indicate that the patients have disproportionately large recollection deficits relative to familiarity. In fact, if the hippocampus is critical for recollection, but not for familiarity (e.g., Aggleton & Brown, 1999; Eichenbaum, Otto, & Cohen, 1994; Yonelinas, 2002), then an ROC analysis should indicate that individuals with selective hippocampal damage should exhibit a deficit in recollection, but not familiarity.

The variable-recollection dual process (VRDP) model. A modification of the DPSD model was proposed in which recollection is assumed to be a thresholded signal detection process (see Figures 2 and 3; Sherman, Atri, Hasselmo, Stern, & Howard, 2003; for related modifications of the DPSD model, also see Healy et al., 2005; Macho, 2002). As in the DPSD model, familiarity is treated as an EVSD process ($d'$), and recollection is a threshold process in the sense that only some of the studied items will be recollected ($R_T$). The critical modification is that the items that are recollected produce a strength distribution that is Gaussian in shape with a mean level of recollection strength ($Rd'$) and some variability around that mean ($RV_{Rd'}$). Thus, items will be recognized if they are recollected and their recollection strength exceeds the response criterion or if familiarity strength exceeds the response criterion. An important aspect of this model is that it allows for the possibility that an experimental manipulation might result in a change in the probability that an item is recollected and/or a change in the strength of the recollected information, something that is not explicitly captured by the DPSD model.

The model can produce item recognition ROCs that are identical to the DPSD model because if the recollection strength distribution
falls above the high confidence response criterion, then all the recollection responses lead to high confidence responses, and the model collapses into the original DPSD model. Thus, the model can be used to generate all the same predictions as that of the DPSD model. However, if some portion of the recollection distribution falls below the high confidence response criterion, then some of the recollected items can receive lower confidence responses than can the high confidence familiarity items. Thus, the model can produce ROCs that can exhibit a slight \( \sim \)-shape in \( z \)-space, such that the ROCs are slightly \( U \)-shaped across most of the range but then bend downwards as response criterion becomes strict (see Figure 3). We note that the VRDP model has not been applied to relational recognition tasks, but one possibility is to apply the model in a manner similar to that of the DPSD model; that is, add an additional set of recollection parameters to allow for the possibility that lure items are recollected.

The some-or-none (SON) model. Kelley and Wixted (2001) proposed another variation of the DPSD model to account for associative recognition. The model assumes that memory judgments are based on assessments of associative memory strength and item memory strength, both of which are well described by signal detection theory. However, recollection is a threshold process in the sense that the retrieval of associative information is probabilistic and only some proportion of pairs will be recollected. Item strength is assumed to reflect an EVSD process, thus it requires one parameter (\( d'_{\lambda} \)), whereas associative strength is assumed to reflect a UVSD process, thus it requires two parameters (\( d'_{\lambda} \) and \( \lambda \)). Because the model assumes that there is a threshold on associative information such that the retrieval of associative information can sometimes fail, recollection is referred to as some or none. Moreover, as in the DPSD model the probability of retrieving associative information about a studied pair (\( R_T \)) can differ from the probability of retrieving associative information about a rearranged pair (\( R_R \)). It is important to note that the SON model assumes that item and associative strength are combined and equally weighted when making associative recognition judgments. In this way, an intact pair will be recognized if associative retrieval is successful (\( R_T \)) and the sum of the associative (\( A \)) and individual (I) item strengths (the I + A distribution) exceeds the response criterion or if associative retrieval fails but the item strength still exceeds the response criterion. In contrast, a rearranged pair will be correctly rejected if recollection that the pair is rearranged occurs (\( R_R \)) and the item strength minus the associative strength is lower than the response criterion or if recollection fails and the item strength is lower than the response criterion.

The SON model can produce ROCs that are similar to that of the VRDP model, in the sense that the zROCs can be linear, \( U \)-shaped, or \( \sim \)-shaped. The degree of asymmetry is determined by the two recollection parameters, and the degree of curvilinearity is determined by the relative contribution of the signal detection and threshold processes. Generally, the model predicts that associative ROCs should be \( U \)-shaped in \( z \)-space, and asymmetrical. However, when \( R_T \) and \( R_R \) approach 1.0, the predicted ROC will become symmetrical and the zROC will become linear. That is, as recollection approaches unity, the strength distributions become equal variance Gaussian distributions. Although it has not been explicitly applied to exclusion ROCs, it has been somewhat described in the UVSD model, the SON model produces a negative-going curved ROC in this case.

The SON model has not been applied to tasks other than associative recognition, so it is not clear how well it would perform, but it is one of the most complex models and it shares many assumptions with the other hybrid models, so it likely can be applied in similar ways. However, with the larger number of parameters it is not clear whether it would be identifiable or how the parameters would behave.

The mixture model. Another hybrid signal detection model that has been proposed to account for item and source recognition ROCs is the mixture model (DeCarlo, 2002, 2003; Hilford et al., 2002). In tests of item recognition it is assumed that memory judgments are based on the assessment of item strength in a manner consistent with the EVSD model (e.g., it assumes equal variance Gaussian memory strength distributions). However, an attention process is also included such that only some proportion of the studied items will increase in memory strength (i.e., the attended items). DeCarlo uses the term \( \lambda \) to designate the probability that a target item is attended, but to be consistent with the other models we will use the term \( R_T \). In this way, the new items form a normal strength distribution, but the old items form a mixture of two equal variance normal strength distributions; one at the same location as the new item distribution and the other shifted to the right by some constant (\( d' \)).

One of the motivations for this model was to allow for encoding variability while avoiding the theoretical problems of negative memory that the UVSD model encountered (DeCarlo, 2002). That is, according to the UVSD model, if one simply increases the old item variance relative to the new item variance, this leads to negative memory in the sense that it requires that the study phase must have decreased the memory strength of some of the test items. The mixture model avoids this by only allowing the memory strength to be increased or to remain unaffected by the study event.

The same model is used in tests of source memory except that memory strength is assumed to reflect how strongly each item matches one of two sources. That is, studying items in one source (i.e., the target source) increases source strength and shifts the items to the right (\( R_S \)), whereas studying items in the other source (i.e., the lure source) decreases source strength and shifts the items to the left (\( R_L \)). As in the item recognition model, only items that are attended at study will be associated with a change in memory strength, thus the source model requires two strength parameters and two attention parameters. Both the item and source models can be extended by adding additional parameters to allow for different levels of attention, but the effects of such modifications have not been fully explored.

The model produces ROCs that are similar to those produced by the VRDP and SON models. In tests of item recognition, the model predicts asymmetrical curved ROCs that are approximately linear in \( z \)-space, with slopes of less than 1.0. The equal variance Gaussian distributions lead the model to generate curved symmetrical ROCs, but because only some of the old items increase in strength this effectively increases the variance of the old item distribution relative to the new item distribution, leading the ROC to be asymmetrical and to have a \( z \)-slope of less than 1. Although the ROCs are approximately linear in \( z \)-space, they tend to exhibit a slight \( U \)-shape and can even exhibit a subtle downward trend at the extreme criterion values, resulting in a \( \sim \)-shape zROC. The non-
linearities arise because of the probabilistic attention process that effectively divides the old items into two distributions (i.e., the attended and nonattended items). When memory strength is low, the mixed distribution is effectively normal, but as strength increases, the two portions of the mixture distribution move farther apart, leading the overall old item distribution to be non-Gaussian.

In tests of source memory, the predicted ROCs are similar to those predicted in item recognition, but because there are two, rather than one, probabilistic attention parameters influencing performance, the ROCs tend to be flatter in probability space and more noticeably U-shaped in z-space (see Figure 3). The model predicts symmetrical ROCs if the strength and attention parameters are comparable for the two sources but can produce asymmetrical ROCs either by making the two strength parameters unequal or by making the two attention parameters unequal.

Because the model assumes that one of the memory components underlying the ROCs is an attention process operating during encoding, manipulations that occur after the time of encoding should affect the strength parameters and not the attention parameters. Because increasing the strength parameter leads the old item distribution to become less normal in shape, any manipulation that increases strength should lead the ROCs to become more nonlinear. In tests of item recognition, such manipulations should make the ROCs more asymmetrical (z-slope should decrease). Determining the expected effects of encoding manipulations on ROC shape is more difficult because attention has nonmonotonic effects on the expected nonlinearities and because the same overall effects can often be produced by varying either strength or attention. Nonetheless, DeCarlo (2002, 2003) has argued that manipulations like increasing study duration, levels of processing, attention, and word frequency should primarily increase attention and thus should lead item ROCs to become more asymmetrical.

The model has not been directly applied to exclusion data, but its application seems fairly straightforward. That is, as source memory performance is currently modeled, some proportion of the target items increase in strength from some baseline (e.g., items from source-1 increase), whereas some proportion of the lure items decrease in strength relative to baseline (items from source-2 decrease). If the new item distribution (nonstudied items) were assumed to have a source memory strength that is at baseline, then the new item distribution would be situated between the two source-item distributions. In this way, the probability of incorrectly accepting an item from the lure source will be less than that of accepting an item from the new distribution. Thus the exclusion ROC will fall below the chance diagonal and would look like a mirror reflection of the ROC predicted in item recognition.

Model Summary

The existing ROC models vary considerably in their assumptions, their complexity, their predictions, and the paradigms in which they have been applied. The threshold models assume that there is a threshold, below which subjects are unable to make accurate recognition discriminations. The HT model assumes a single HT for old items, whereas the HLT model assumes an additional threshold for new items.

In contrast, the signal detection models assume that memory can never truly fail or fall below a threshold but that old and new strength distributions are overlapping and Gaussian in shape. The EVSD model is the simplest signal detection model and assumes that the old and new item strength distributions have the same variance, whereas the UVSD model allows the old item variance to be greater than the new item variance, possibly because of encoding variability. The 2DSD model is an extension of the UVSD model that incorporates a separate strength dimension for each of two memory sources. Similarly, the STREAK model incorporates two unequal variance strength dimensions, but in this model one dimension reflects unfamiliarity and the other reflects recollection.

The hybrid models generally assume that recognition reflects the contribution of a signal detection-based familiarity process and a threshold recollection process. The simplest hybrid model is the DPSD model, which assumes that item recognition reflects the contribution of an EVSD-based familiarity process and a threshold recollection process. In relational recognition, the model requires an additional parameter to allow for the recollection of two different types of studied items. The VRDP model makes the additional assumption that the recollected information strength distribution is Gaussian, and thus it incorporates additional parameters for the strength and variance of recollection. The SON model assumes that performance reflects threshold recollection processes, as well as an equal variance item familiarity process, and an unequal variance associative familiarity process. The mixture model assumes that familiarity reflects an EVSD process, but that there is an additional attentional process at encoding, rather than recollection, that either succeeds or fails. In relational tests, additional familiarity and attention components are required to allow for the recognition of two types of studied items.

With respect to the ROC predictions, probably the most fundamental difference in the models is that the threshold models predict linear ROCs that are U-shaped in z-space, whereas the signal detection models predict inverted U-shaped ROCs that are linear in z-space. In contrast, the hybrid models can accommodate both of these patterns. Of the hybrid models, the DPSD model can produce either linear or U-shaped zROCs, whereas the VRDP, SON, and mixture models can produce linear, U-shaped, and ⊖-shaped zROCs.

Another significant difference in the models’ predictions is in how they deal with ROC asymmetry. The EVSD model predicts symmetrical ROCs because the shape of the old and new item distributions is assumed to be the same. In contrast, all the other models can produce asymmetrical ROCs, but they do so in somewhat different ways. According to signal detection models (e.g., UVSD, 2DSD, and STREAK), the degree of ROC asymmetry is due to differences in target and lure variance. For example, it has been suggested that because of encoding variability, the old item variance will be greater than the new item variance. In contrast, the threshold and most hybrid models assume that the degree of ROC asymmetry will be related to the relative likelihood of recollecting (or attending to) the targets and lures. For example, in item recognition, the dual process models (e.g., DPSD, VRDP, SON) predict greater old than new item variance because they assume that new item responses rely on familiarity, whereas old item responses rely on familiarity and recollection. Similarly, in the mixture model, new item responses are influenced by a familiarity process, whereas old item responses are influenced by familiarity as well as an attention process.

Most models predict that RK responses should fall on the confidence ROC, but the signal detection models tend to treat
remember responses as synonymous with the strongest memories, whereas the dual process models tend to treat remember responses as relying more on recollection than familiarity. The one exception is the STREAK model, which assumes that different decision rules are used to treat RK and confidence responses and thus the remember ROC point can fall either above, on, or below the confidence ROC.

In general, the dual process models (e.g., DPSD, VRDP, SON) predict that exclusion ROCs should be curved and negative going because recollection is assumed to be able to oppose the effects of familiarity and thus push the ROC down. None of the other models have been directly applied to the exclusion paradigms.

Several models make additional assumptions about the underlying components of recognition, and thus they make predictions about the effects of various experimental variables on ROC shape. For example, the dual process models make several ROC predictions about the effects of various manipulations and brain damage based on expected effects of these variables on recollection and familiarity. Similarly, the mixture model makes a number of ROC predictions based on the expected effects of manipulations on attentional mechanisms during encoding.

The DPSD model has been applied across the broadest range of ROC paradigms, including item, relational, exclusion, and RK recognition tasks. The modifications of that model have been applied to only a restricted range of paradigms; the VRDP has been directly applied only to item recognition and the SON model only to relational recognition. The mixture model has been applied to item and relational recognition, whereas the threshold and signal detection models have been applied to item and relational recognition as well as RK paradigms. Finally, the STREAK model has been applied only to RK and item confidence results.

The simplest models are the HT and EVSD models, which have only one memory parameter each. Dual component models include the HLT and the UVSD models. The DPSD model also requires two free parameters for item and RK recognition but requires an additional recollection parameter for relational recognition. The mixture model is also a two parameter model for item recognition but requires four to account for relational recognition. STREAK, which is applicable only when RK responses are collected, requires two parameters but assumes that the old–new variance ratio is fixed at 1.25 (i.e., a slope of .80). This parameter can be freed though if confidence responses are also collected. The VRDP model requires four parameters to account for item recognition, and the 2DS and SON models use five free parameters to account for item and relational recognition. These three models are not uniquely identifiable in standard item or relational confidence studies, but they do become identifiable in more complex data sets or when parameter restrictions are made. Thus, while all the models make testable predictions as outlined in detail above, some are clearly easier to fit than others. In the next section, we concentrate on the empirical regularities that have emerged from ROC literature and evaluate which models are capable of accounting for these general patterns, regardless of model complexity. However, we also review studies that have formally contrasted the different models by using statistical methods to compensate for the number of free parameters, and we find that they generally support the same conclusions as those based on the examination of the overall patterns of results.

ROC Results and Theoretical Implications

The ROC studies will be reviewed in the following order: studies of item recognition, relational recognition (e.g., source and associative recognition), exclusion recognition, RK recognition, and ROC studies of amnesia. Our aim was to identify the empirical regularities (i.e., patterns of results that have been observed across various different studies), and highlight the theoretical implications for each of these. Nine such regularities are identified (see Table 1). Results that are less well established (i.e., a pattern of results that has not yet been replicated across labs) will also be briefly described. We focus on the more well-established findings because of potential concerns about individual studies that may have suffered from measurement problems, such as insufficient numbers of trials, random responding, and so on. We will point out cases in which the findings may have been influenced by such measurement issues. In addition to identifying the reliable patterns of results, we also aim to identify the related empirical issues that have not yet been adequately addressed. In the empirical review, our goal was to be as inclusive as possible and to include all the recognition ROC studies of each of these general types that we could find. Several somewhat related ROC studies that were not included because of space limitations, however, include studies of frequency estimates (Hintzman, 2004), lag-recency effects (Schwartz, Howard, Jing, & Kahana, 2005), forced-choice recognition (Smith & Duncan, 2004), false recollection (Lampinen et al., 2006; Westerberg & Marsolek, 2003), and neural monitoring studies (e.g., Woodruff et al., 2006; Yunelinas et al., 2005).

The measures of ROC shape that we report below, like z-slope and z-intercept, were taken from the published articles in the following manner: If the relevant measures were provided in the text, those values were used. If average values were reported then we used those values; otherwise if the values from the aggregate data were reported we used those; and finally, if neither was reported, but ROCs were included in a figure, we extracted parameter estimates with DataThief (http://www.datathief.org).

Item Recognition

Three general behavioral findings have been established in the item recognition ROC literature.

1. Item ROCs exhibit an inverted U-shape and are approximately linear in z-space. In studies of item recognition, the observed ROCs are almost always curvilinear such that they exhibit an inverted U-shape. This pattern was first reported by Egan (1958; see Figure 1A) and has now been observed in countless experiments (for earlier discussions of these results, see Glanzer, Kim, Hilford, & Adams, 1999; Murdock, 1974; Ratcliff et al., 1992, 1994). Note that when performance is at chance the ROC is constrained to follow the diagonal, and thus it is forced to be linear, but with cases of chance performance excluded, virtually every item recognition study that has been published has reported a curved ROC.

In general, when the item recognition ROCs are plotted in z-space they are found to be linear. Although this linearity has simply been noted in most recognition studies, several studies have directly assessed the linearity of the zROC's and have found that they are fit well by linear functions, and the fits are not significantly improved by introducing a quadratic component (e.g., Arndt
Slight deviations from linearity have sometimes been reported in studies of item recognition, however. For example, significantly U-shaped zROCs have been reported for elaboratively encoded complex visual photographs (Howard, Bessette-Symons, Zhang, & Rotello, 2006) found that both young and elderly subjects exhibited comparably U-shaped zROCs in a recognition test for photographs.

The nonlinearities sometimes observed in item recognition zROCs could be important, as only some models can account for these results. However, these effects are quite subtle and not very common. Moreover, the conditions that lead to these nonlinearities are not yet understood and might simply reflect methodological artifacts like random responding or variability in response criterion (e.g., Malmberg & Xu, 2006; Ratcliff et al., 1994; Yonelinas et al., 1996). Nonetheless, future studies designed to determine when and why the zROCs in item recognition sometimes deviate from linearity will be of considerable empirical and theoretical importance.

The shape of the ROCs observed in item recognition indicates that threshold models are unable to account for recognition memory performance, and they provide support for signal detection models and hybrid models. The finding that item ROCs are invariably curved in probability space is problematic for threshold models such as the HT and HLT models, which predict that the ROCs should be linear in probability space. The aspect of these models that leads them to incorrectly predict linear ROCs is that they assume that there is a sensory limit (i.e., a threshold) on the true memory strength of test items. Thus, the models assume that there should be some test items that fall below the memory threshold, and for those items subjects should be completely at chance at discriminating between old and new items. However, the curved ROCs make it clear that in item recognition tests there is no memory threshold between old and new items. However, these effects are quite subtle and not very common. Moreover, the conditions that lead to these nonlinearities are not yet understood and might simply reflect methodological artifacts like random responding or variability in response criterion (e.g., Malmberg & Xu, 2006; Ratcliff et al., 1994; Yonelinas et al., 1996). Nonetheless, future studies designed to determine when and why the zROCs in item recognition sometimes deviate from linearity will be of considerable empirical and theoretical importance.

The shape of the ROCs observed in item recognition indicates that threshold models are unable to account for recognition memory performance, and they provide support for signal detection models and hybrid models. The finding that item ROCs are invariably curved in probability space is problematic for threshold models such as the HT and HLT models, which predict that the ROCs should be linear in probability space. The aspect of these models that leads them to incorrectly predict linear ROCs is that they assume that there is a sensory limit (i.e., a threshold) on the true memory strength of test items. Thus, the models assume that there should be some test items that fall below the memory threshold, and for those items subjects should be completely at chance at discriminating between old and new items. However, the curved ROCs make it clear that in item recognition tests there is no memory threshold below which performance is completely at chance. Rather, subjects exhibit some ability to discriminate between old and new items.

Table 1

<table>
<thead>
<tr>
<th>Empirical Regularities</th>
<th>Threshold</th>
<th>Signal Detection</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HT</td>
<td>HLT</td>
<td>SD</td>
</tr>
<tr>
<td>Item recognition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Inverted U-shaped item ROCs (approx. linear zROCs)</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>2. Asymmetrical item ROCs (zROC slope &lt;1)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3. Dissociation of item ROC sensitivity and asymmetry</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>4. Linear and inverted U-shaped relational ROCs (U-shaped zROCs)</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>5. Symmetrical and asymmetrical relational ROCs</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>RK recognition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Curved, negative-going exclusion ROCs</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>7. Overlapping RK and confidence ROCs</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>8. Exceptionally high RK slopes (z-slope &gt;1)</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Memory impairments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Symmetrical ROCs in amnesia (z-slope approaching 1)</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Note. The ✓ symbol indicates that the model can account for the effect, the × indicates that it cannot, and the ?(✓) symbol indicates that the model does not account for the effect, but that a natural extension of the model could account for the effect. ROC = receiver operating characteristic; HT = high threshold; HLT = high–low threshold; SD = signal detection; UVSD = unequal variance signal detection; 2DSD = two-dimensional signal detection; STREAK = sum–difference theory of remembering and knowing; DPSD = dual process signal detection; VRDP = variable recollection dual process; SON = some-or-none; RK = remember–know.

a To account for relational recognition several additional parameters are required.
b As long as $V_o$ is treated as a free parameter.
c But leads to problematic parameter values.
d But treats the overlap as a coincidence.

Table 1: Empirical Regularities Observed in ROC Studies of Recognition Memory and the Ability of Various Recognition Memory Models to Account for These Effects
new items even when their response criterion becomes very liberal.

The results suggest either that the threshold notion is inap-
propriate for item recognition or that it must be modified in
such a way that performance is no longer predicted to be at
chance when the memory strength is below the threshold. One
such modification is to introduce additional thresholds such that
a continuous memory strength distribution could be mimicked
by a set of small steplike functions. However, as discussed in
the model section earlier, such a post hoc approach is not par-
nicularly satisfying on theoretical grounds, and it is not very
useful because of the large number of parameters that it would
require to produce curved ROCs. Another approach is to sup-
plement the threshold process with some other process that does
generate curved ROCs, such as a Gaussian familiarity process,
as in several of the hybrid models.

The findings provide support for signal detection models (i.e.,
SD, UVSD, 2DSD, and STREAK). All these models assume that
the old and new strength distributions reflect overlapping Gaussian
distributions, and it is this Gaussian assumption that leads the
models to predict curved ROCs that are linear in z-space. Gaussian
distributions produce curvilinear ROCs because changes in re-
sponse criteria result in disproportional changes in the hits and
false alarms. That is, when the response criterion is strict, changes
in criterion will have large effects on the proportion of the old item
distribution that will be recognized but will have relatively modest
effects on the proportion of the new item distribution that will be
recognized. In contrast, as the response criterion becomes more
lax, changes in criterion tend to have larger effects on the propor-
tion of new items that are recognized and decreasing effects on the
proportion of old items that are recognized. Thus, the slope of the
predicted ROCs will decrease as the false alarm rate increases,
leading to a curvilinear ROC.

The hybrid models that supplement signal detection theory
with a probabilistic threshold process are also generally consis-
tent with the curved item ROCs (e.g., DPSD, VRDP, mix-
ture, and SON). Although these models can predict a slightly
nonlinear zROC, the predicted deviation from linearity in tests
of item recognition can be extremely subtle (e.g., DeCarlo,
2002; Yonelinas, 1999), and so the linear zROCs are in good
agreement with these models. Note that the slight nonlinearity
sometimes observed in item zROCs can, in principle, be ac-
counted for by these hybrid models. For example, the slight
U-shape sometimes seen in zROCs could reflect the contribu-
tion of recollection according to the DPSD model (Yonelinas,
1994; Yonelinas et al., 1996) or could reflect the contribution of
a threshold attention process occurring during encoding accord-
ing to the mixture model (DeCarlo, 2002). However, as dis-
cussed above, the conditions under which nonlinear zROCs are
observed in item recognition are not yet well understood. Thus,
at this point, the item recognition ROCs appear to provide
support for the pure signal detection theories and the hybrid
models alike, and an examination of the overall shape of the
item recognition ROCs does not clearly support one class of
these models over the other.

These general conclusions are also supported by studies that
have directly contrasted the statistical fits of various models to
results from item recognition experiments. For example, in tests of
item recognition, the UVSD and DPSD models provide excellent
fits to the data, with the best fit going to one model in some studies
and the other model in other studies (Heathcote, 2003; Howard,
Fotedar, Datey, & Hasselmo, 2005; Kelley & Wixted, 2001; Prull
et al., 2006; Yonelinas et al., 1996; 1999). Moreover, these two
models consistently fit item recognition data better than do the SD,
HT, and HLT models (e.g., Kelley & Wixted, 2001; Slotnick,
Klein, Dodson, & Shimamura, 2000; Yonelinas et al., 1996).
Finally, Sherman et al. (2003) contrasted the DPSD and VRDP
models and found that they fit control subjects’ data equally well
and led to virtually identical parameter estimates. However, the
VRDP model provided a better fit for subjects given the amnestic
drug scopolamine prior to encoding.

2. Item ROCs are asymmetrical with a z-slope of less than 1.0.
In a vast majority of item recognition tests, the ROCs are asym-
metrical along the diagonal such that they are pushed up along
the left side (e.g., see Figure 1). When plotted in z-space, the
slope of the function is less than 1.0. The left panel of Figure 4
plots z-slopes against z-intercepts for 149 published conditions
from standard item recognition experiments.1

An examination of Figure 4 shows that the z-slopes are well
below 1.0 in almost every condition (mean z-slope <.76), indi-
cating that the item ROCs were almost always asymmetrical. In
fact, in the few cases in which the z-slope was found to be 1.0 or
greater (e.g., Gronlund & Elam, 1994; Heathcote, 2003; Hirshman
& Hostetter, 2000), the same conditions were examined in other
experiments within those studies, and in all cases the high z-slope
value did not replicate. Whether the cases of high z-slope were due
to noise or methodological properties of those specific experiments
is not known.

Figure 4 also indicates that the z-slopes tend to decrease as the
intercept increases. This is due in part, however, to the fact that as
the intercept decreases, the slope is mathematically constrained to
decrease toward 1.0. The relationship between z-intercept and
z-slope is examined in more detail below.

The asymmetrical ROCs observed in item recognition studies
indicate that the old items vary more in overall memory strength
than do the new items. This means that it is inaccurate to describe
memory encoding as resulting in a simple increase in memory
strength. Such findings reveal a basic failure of models like the
EVSD model that assume that old and new strength distributions
take on the same shape. Moreover, the results indicate that sensi-
tivity indexes provided by measurement models assuming similar
old and new strength distributions, including the d’ index of the
EVSD model and the A’ statistic sometimes referred to as a
nonparametric measure of sensitivity (see Macmillan & Creelman,
1996; 2005), cannot be relied on to provide unbiased measures of
recognition performance (we elaborate on this point in the Dis-
cussion).

So why do the old and new item distributions differ in shape?
The pure signal detection models account for these results by
including a parameter specifically designed to allow for the vari-
ability difference between old and new items. That is, the UVSD,
2DSD, and STREAK models treat the old item variance as a free
parameter, so these models can produce zROCs with slopes less
than or greater than 1.0. These models generally do not provide

---
1 Item recognition ROCs that were collected concurrently with RK
reports or source memory reports were not included, but they did not
appear to differ in any appreciable way from the data plotted in Figure 4.
insight into why the old item variance is greater than that of the new items, but one suggestion has been that the greater old item variability is due to encoding variability (e.g., Hilford et al., 2002; Wixted, 2007).

The DPSD and VRDP models can also account for the asymmetrical item ROCs, but they explain it by assuming that the greater old item variance seen in item recognition arises because the new item distribution reflects only one process (i.e., familiarity) whereas the old item distribution reflects the combination of two processes (i.e., recollection and familiarity). Because recollection is expected to lead to high confidence responses, it will generally pull the old item distribution to the right and thus increase the overall old item variance. Similarly, the mixture model assumes that the new item distribution reflects only one familiarity process, whereas the old item distribution reflects the contribution of two processes (familiarity and attention), and thus the old item distribution reflects a mixture of two different distributions (i.e., the attended and the unattended items).

Although other explanations of the ROC asymmetry might be possible, currently the most viable explanations appear to be the encoding variability, recollection, and attention accounts. Although these accounts seem quite different, they are difficult to differentiate empirically on the basis of the item ROC data alone because they can all produce ROCs that are in general agreement with those seen in item recognition.

3. ROC sensitivity and asymmetry are dissociable in item recognition. Two aspects of the ROC shape can be varied separately in tests of item recognition: sensitivity, which is the distance from the negative diagonal (roughly captured by the z-intercept), and asymmetry, which is the degree to which the ROC is pushed up along the left side (indexed by the z-slope). For example, as illustrated in Figure 1A, strength manipulations, such as increasing the number of times an item was studied, increase the z-intercept but do not significantly alter the z-slope. This pattern was examined extensively by Ratcliff et al. (1992) who argued that as long as floor effects were avoided (i.e., when performance decreases, the z-slope of the ROCs is forced to be equal to 1.0), increasing memory strength by increasing study repetitions or by increasing study duration did not affect z-slope. Strength manipulations have been examined in 31 published comparisons (e.g., Egan, 1958; Glanzer et al., 1999; Heathcote, 2003; Hirshman & Hostetter, 2000; Ratcliff et al., 1992, 1994; Yonelinas, 1994). On average, increasing memory strength leads to an increase in the z-intercept from 1.23 to 1.68, accompanied by only a slight decrease in z-slope from .77 to .74. Moreover, the slight drop in slope can be attributed to the few cases in which performance was at floor. That is, when the experiments with the lowest levels of performance are excluded (i.e., Hirshman & Hostetter, 2000; Experiments 1 and 2, Ratcliff et al., 1994), the average slope across 26 published comparisons was .74 for both weak and strong items. Note that it has been suggested that increasing study duration might have a slightly greater effect on z-slope than has increasing the number of study repetitions (Glanzer et al., 1999), but the studies reviewed above do not appear to support this claim.

In contrast to the strength manipulations, most other experimental manipulations that have been examined increase ROC sensitivity and lead the ROCs to become more asymmetrical (e.g., as the z-intercept increases the z-slope decreases). For example, Figure 1B presents ROCs from a study examining the effect of varying levels of processing on item recognition ROCs (Yonelinas et al., 1996; for similar results, also see Glanzer et al., 1999). The figure indicates that meaningful compared with perceptual processing at encoding leads to an increase in sensitivity and at the same time leads the ROC to become more asymmetrical. Similar effects have been reported for various other manipulations such as word frequency (Arndt & Reder, 2002; Glanzer & Adams, 1990; Glanzer et al., 1999; Ratcliff et al., 1994), word concreteness (Glanzer & Adams, 1990), list length (Egan, 1958; Glanlund & Elam, 1994; Ratcliff et al., 1994; Yonelinas, 1994); although Strong, 1912, as reanalyzed by Gronlund & Elam, 1994, and Gronlund & Elam, 1994, report studies in which slope remained largely unchanged across list length), dividing attention (Yonelinas, 2001b), orthographic similarity (Heathcote, 2003), and aging
Although the degree of ROC asymmetry can be functionally dissociated from sensitivity in item recognition (i.e., increases in intercept leave slope unaffected in some cases but lead the slope to decrease in others), the complementary dissociation has not been reported. That is, we are not aware of any manipulation that consistently increases item recognition performance while making the ROC more symmetrical (i.e., increase intercept and increase slope). Whether such an effect can be observed is unclear.

The dissociation of ROC sensitivity and asymmetry in item recognition indicates that two or more functionally independent memory components are required to account for recognition memory performance. To account for the experimental separability of z-slope and z-intercept, no fewer than two separate memory components are necessary. This presents a serious challenge for all single component models of recognition including the HT and EVSD models. That is, there is no way to account for the experimental separation of z-slope and z-intercept without postulating at least two separate memory components.

However, these results do not specify exactly what those two components are or what psychological processes might give rise to them. Accordingly, there are various two-component models that can produce the observed dissociation. The dissociation can be accounted for most transparently by the pure signal detection models (i.e., UVSD, 2DSD, and STREAK) because those models have one parameter for sensitivity (i.e., memory strength) and another parameter for the degree of asymmetry (i.e., old–new variance ratio). However, these models do not provide insight into which experimental manipulations should influence only one parameter and which should influence both. Nonetheless, the SDT models are successful at accounting for these results descriptively and suggest that studying an item once increases the strength and variance of the studied item. However, the models indicate that studying the items a second time, or for a longer duration, will further increase the old item strength but not the variance. If the increase in old item variance is interpreted as arising because of encoding variability, it would appear that there is encoding variability only during the first presentation of an item. In contrast to manipulations of strength, however, most other manipulations that increase performance appear to increase the strength component and the variance component together.

In contrast, the DPSD and VRDP models account for the strength/asymmetry dissociation by assuming that different manipulations have different effects on recollection and familiarity. For example, manipulations that primarily increase recollection, such as elaboration, should make the ROC more asymmetrical (i.e., the z-slope should decrease), whereas manipulations that increase both recollection and familiarity, such as increasing study time, should increase sensitivity, and the degree of asymmetry should be less affected (i.e., the slope should remain relatively constant). Thus, the general patterns of strength/asymmetry that are observed in item recognition are in good agreement with these models. The mixture model can also produce the observed dissociations. That is, it predicts that manipulations that increase familiarity will increase sensitivity and the degree of asymmetry because increasing familiarity moves the two portions of the old item distribution farther apart, effectively increasing the old item variance. In contrast, because increasing attention has nonmonotonic effects on asymmetry, it is possible that selectively increasing attention could increase sensitivity while leaving the degree of asymmetry unaffected. Thus, strength manipulations may have increased the attention parameter, whereas other manipulations may have increased the strength parameter.

As with the asymmetrical ROC findings, it may be possible to develop alternative accounts for the dissociation of ROC sensitivity and asymmetry, but currently the most viable accounts are the encoding variability, recollection, and attention explanations.

Relational Recognition: Tests of Associative and Source Recognition

Previous studies using relational recognition tests, such as source recognition, associative recognition, plurality recognition, and conjunctive recognition, have shown that performance on these tests is functionally and neuroanatomically dissociable from that seen on item recognition tests (e.g., Johnson et al., 1993; Yonelinas, 2002). In line with these earlier findings, the examination of the relational ROCs shows that they can be very different from item recognition ROCs. We note that although the different types of relational tests listed above differ from one another in several ways, for the most part, previous behavioral studies have suggested that they are functionally quite similar (for a review, see Yonelinas, 2002). With a few notable exceptions, the ROC literature supports this observation as well.

4. Relational ROCs can be linear or curved in probability space, but they are typically U-shaped in z-space. Unlike the ROCs observed in tests of item recognition, ROCs in relational recognition tests can be either linear or curved downward when plotted in probability space, and they are almost always U-shaped when plotted in z-space. For example, Figure 5A presents associative ROCs for word pairs along with item ROCs for single words (Experiment 1 from Yonelinas, 1997). The item ROC is significantly curved downward in probability space and does not differ significantly from linearity in z-space. Conversely, the associative ROC is close to linear in probability space and U-shaped in z-space. Similar findings have been reported in other associative recognition studies (e.g., Healy et al., 2005; Kelley & Wixted, 2001; Yonelinas et al., 1999), in tests of source recognition (see Figure 5B, which is from Yonelinas, 1999; also see DeCarlo, 2003; Glanzer et al., 2004; Hilford et al., 2002; Qin et al., 2001; Slotnick et al., 2000; Slotnick & Dodson, 2005), and in plurality-reversed recognition tests (see Figure 5D, which is from Rotello et al., 2000; see also Arndt & Reder, 2002).

A close examination of Figures 5A, 5B, and 5D indicates that although the relational ROCs are fit well by linear functions, they do tend to exhibit a slightly inverted U-shape. In fact, several studies have reported relational ROCs with a very pronounced downward curve. For example, Figure 5C presents a source ROC from an experiment where the target items were studied just prior to the test, whereas the lure items were studied 5 days previously (Experiment 4, Yonelinas, 1999). Similarly curved relational ROCs have been reported in other source memory studies (e.g., Glenzer et al., 2004; Qin et al., 2001; Slotnick & Dodson, 2005; Slotnick et al., 2000), as well as associative recognition studies (e.g., Healy et al., 2005; Kelley & Wixted, 2001; Yonelinas et al., 1999).
Unlike item recognition, however, in almost all of the relational studies, the ROCs exhibit a U-shape when plotted in z-space. For example, Figure 6 presents the quadratic estimates for 59 conditions from published studies of source, associative, as well as plurality-reversed recognition. For each type of relational test, the quadratic term was positive in a vast majority of the cases, indicating that the relational ROCs were U-shaped in z-space. An examination of Figure 6 indicates that a positive quadratic component was observed across a wide range of z-intercepts, but in the few cases in which performance was very high (i.e., the rightmost points in Figure 6), the quadratic component appears to decrease. Although the pattern is based on only a few points, the figure suggests that as performance increases above an intercept of about 1.5 the U-shape observed in the zROCs begins to decrease, and an inverted U-shape starts to appear. However, whether the effect of high performance on the ROC shape is reliable, and whether it is due to ceiling or truncation effects as described in the methods section above, is not yet clear.

Several studies have begun to explore the factors that influence the extent to which the relational ROCs are curved or linear, but the critical factors are not yet fully understood. First, Yonelinas (1999) proposed that source ROCs could become more curved if the familiarity of the items from the two different sources were made markedly different, such that differences in item familiarity became diagnostic of source membership. Support for this hypothesis was provided by experiments showing that the source ROCs became more curvilinear when items from one source were presented more recently, or more frequently, than were those from the other source.

A second hypothesis is that relational ROCs will become more curvilinear when the components to be encoded (e.g., two words or a word and a source) are treated in a unitized or holistic manner (Yonelinas et al., 1999; Quamme, Yonelinas, & Norman, 2007). In support of this hypothesis is that associative ROCs for internal–external face features were linear when the faces were studied and

**Figure 6.** Quadratic components plotted as a function of sensitivity in experiments examining relational recognition receiver operating characteristics (ROCs) from C. M. Parks and Yonelinas (2007). Positive values indicate cases in which the zROC (the ROC function when it is plotted in z-space) exhibited a U-shape. In almost all cases, the relational ROCs were U-shaped in z-space. The figure is reproduced from “Moving Beyond Pure Signal-Detection Models: Comment on Wixted (2007)” by C. M. Parks and A. P. Yonelinas, 2007, *Psychological Review, 114*, p. 188. Copyright 2007 by the American Psychological Association. Reprinted with permission.
tested upside down, but they became curved when the faces were studied and tested in an upright orientation. The results were interpreted as indicating that the upright faces were treated holistically as a single item, whereas the upside-down faces were treated as a collection of separate features. Similarly, random word pairs that were encoded as new compound words led to more curvilinear associative ROCs than did word pairs treated as two separate items (Quamme et al., 2007).

A third hypothesis is that source ROCs may become more curved as the complexity of the study event increases. That is, Qin et al. (2001) found highly curved source memory ROCs for complex verbal statements spoken by different individuals. To account for these findings and the previous results showing linear source ROCs, they suggested that in standard source memory tests in which the stimuli are relatively simple (e.g., a word spoken by a male voice) the ROC can be linear and threshold-like because it is likely that subjects will be able only to recollect one aspect of the study event linking the item to the source. However, as the complexity of the event increases, various different aspects of the event may be retrieved, leading the ROC to become more curvilinear.

A fourth hypothesis is that the associative ROCs may become more curved as overall performance increases. Kelley and Wixted (2001) found that as pairs of words were repeated during study, performance increased and the associative ROCs became more curvilinear (and changed from U-shaped to more linear in z-space). On the basis of this result they developed the SON model that produces U-shaped zROCs when performance is low (because judgments are based on distributions that are non-Gaussian when recollection is low), whereas it produces more linear zROCs when performance is high (because the strength distributions become increasingly more Gaussian as recollection approaches 1.0). Note that this account can explain why the quadratic seems to decrease as performance increases (see Figure 6).

Finally, Slotnick et al. (2000; Slotnick & Dodson, 2005) suggested that source ROCs could become more linear if a guessing process contributed to source memory performance. In support of this hypothesis, they found that when conditional source ROCs were plotted such that only items that received a high confidence recognition response were included, the source ROCs became more curved.

It is now well established that ROCs in source, associative, and plurality-reversed tests can vary from linear to concave and that they are almost always U-shaped in z-space. Thus, models of recognition must be able to produce these types of ROCs. The factors that influence the degree of linearity in these ROCs, however, are only beginning to be understood. There appear to be several different factors that influence the curvature of these zROCs; however, none of these have yet been extensively tested, and it is not known how these different accounts are related. Future studies that test these various accounts further will be essential in understanding the processes involved in relational recognition memory tests.

The ROCs observed in relational recognition tests indicate that pure threshold and pure signal detection models are not adequate to account for recognition, and they support hybrid models that incorporate signal detection and threshold processes. The U-shaped zROCs that are consistently observed in relational recognition studies cannot be accounted for by a pure signal detection model (i.e., the SD, UVSD, 2DSD, and STREAK models) because these models are based on Gaussian strength distributions and are constrained to predict ROCs that are curvilinear in probability space and linear in z-space. Conversely, although the linear ROCs that have been observed are consistent with threshold models (e.g., HT and HLT), the curved ROCs that are often observed in these studies indicate that even these models are not adequate to account for performance because these models always predict perfectly linear ROCs in probability space.

These findings can, however, be accounted for by hybrid models that combine the Gaussian distributions of signal detection theory with a threshold process. For example, the DPSD, VRDP, and SON models include a probabilistic recollection process that leads the ROC to exhibit a somewhat linear shape in probability space and a U-shape in z-space, as well as an equal variance Gaussian signal detection process that leads the ROC to be moreconcave downward in probability space and linear in z-space. The mixture model also includes an equal variance signal detection process but supplements it with a probabilistic attention process during encoding rather than a recollection process. The Gaussian familiarity distribution leads the model to predict curved ROCs that are approximately linear in z-space, but because attention operates as a threshold during encoding (i.e., some items are attended whereas some are not), it leads the ROCs to become more linear in probability space and U-shaped in z-space.

Although it is tempting to conclude that a hybrid approach is necessary to account for the relational recognition results, it is possible that alternative models might be developed that do not build on either of these theoretical frameworks. For example, if the Gaussian strength distributions central to signal detection theory were replaced by skewed distributions like those generated by the hybrid models, such a model should be able to account for the observed ROCs. Whether such a parameterization would produce results that differ from the existing hybrid models, and whether the parameters would be psychologically meaningful, would of course be important questions to address. The current results, however, do strongly support the hybrid notion that both signal detection and threshold processes are operating in recognition.

These conclusions are further supported by studies directly contrasting the fits of the various models to relational ROC results. That is, in tests of relational recognition, the DPSD, SON, and mixture models are found to provide good accounts of the data, and they are generally better than those provided by the UVSD and 2DSD models (Healy et al., 2005; Hilford et al., 2002; Yonelinas, 1999), and the UVSD model tends to fit relational ROCs better than does the HLT model (e.g., Slotnick et al., 2000; Slotnick & Dodson, 2005), though under some conditions the HLT model can provide an adequate account of the data (e.g., Kelley & Wixted, 2001). A comparison of the DPSD and SON models for associative ROC data from previously published studies showed that the two models provided excellent fits of the data “with one type of model providing slightly better fits of the data and another being slightly superior with respect to another data set” (Macho, 2004, p. 95). However, in some cases the parameter estimates pointed to potential problems for each model. For example, in one data set the DPSD model parameters suggested that increasing the number of times pairs were presented led to a decrease in recollection, which makes little sense. In another data set, the SON model parameters suggested that the variance of the item distribution was greater than that of the item plus associative distribution, which should not
occur when distributions are added, and contradicts the assumptions of the model. Thus, the hybrid models are not problem free, but they do consistently provide appreciably better accounts of relational recognition than do either pure threshold or pure signal detection models. A critical question for future studies will be to determine whether the threshold nature of the ROCs seen in relational recognition tasks arises because of a recollection process as assumed by several dual process models or because of an attention process at encoding as assumed by the mixture model.

5. Relational ROCs are symmetrical in some cases ($z$-slope = 1) but asymmetrical in others ($z$-slope < 1). Symmetrical ROCs have been reported in several studies of source memory (see Figure 5B; Glanzer et al., 2004; Hilford et al., 2002; Yonelinas, 1999), but asymmetrical ROCs (i.e., ROCs that are pushed up on the left side with $z$-slope of less than 1.0) have been reported in other source memory tasks (Figure 5C; Yonelinas, 1999), word-pair recognition (Healy et al., 2005; Kelley & Wixted, 2001; Yonelinas, 1997), face-feature recognition (Yonelinas et al., 1999), and plurality-reversed tests (Rotello et al., 2000).

The degree of asymmetry in relational ROCs is directly related to the relative memory strengths of the targets and lures. That is, in tests of source memory, when the encoding conditions are comparable for items from two different sources, the resulting source ROCs are symmetrical, whereas when the strength of one source is greater than that of the other, the ROC becomes asymmetrical. For example, when items from each source are studied for an equal study duration the source ROCs are symmetrical, whereas when items in one source are stronger (e.g., twice- vs. once-presented or after a short vs. long delay) the source ROC becomes asymmetrical, such that the intercept of the strengthened source is greater than that of the weaker source (e.g., Yonelinas, 1999). Similarly, in plurality-reversed tests, the ROCs were found to be linear and approximately symmetrical when subjects were instructed to use a recall-to-reject strategy (e.g., if you studied frog then you can be sure frogs was not studied; Experiment 1, Rotello et al., 2000). In contrast, when subjects were not explicitly instructed to reject similar lures, the ROC remained linear but was more asymmetrical (slope less than 1), as expected if the likelihood of using memory for the targets was greater than that for the related lures (Experiment 2, Rotello et al., 2000). In tests of associative recognition, the ROC slopes are less than one, which has been interpreted as indicating that the likelihood of recognizing a studied pair is greater than the likelihood of recognizing that a lure pair was not studied (Kelley & Wixted, 2001; Yonelinas, 1999). Moreover, when the memorability of the lures was decreased further by pairing each study item with two study items rather than just one other study item (so that subjects must retrieve two associations to reject the lures as new), the ROCs became even more asymmetrical (Yonelinas, 1997). Finally, Kelley and Wixted (2001) found that as associative recognition performance became very high with multiple study repetitions, the associative ROCs became more symmetrical. This was interpreted as indicating that when the pairs were strengthened subjects were able to recollect all of the targets and lures.

The symmetrical and asymmetrical relational ROCs indicate that the shape of the memory strength distributions for targets and related lures can be varied independently, and they present problems for models that assume a fixed relationship between the target and lure strength distributions. Although there are cases in which it is appropriate to describe the difference between the targets and related lures as a simple difference in memory strength (e.g., two equally encoded sources might lead to two identical strength distributions differing only in mean strength), when the relative strengths of the two sources differ, the shapes of the two distributions must also change. These findings are problematic for models that predict symmetrical ROCs like the EVSD model. Most of the other models are able to account for these results because they assume that the shape of the target and lure strength distributions can vary independently. For example, the DPSD, VRDP, and SON models assume that the probability of recollecting items from the sources can differ. Thus, only when recollection is equal for the two sources will the ROCs be symmetrical. The mixture model assumes that the ROCs can be made asymmetrical either by varying the amount of attention given to the two different sources or by changing the familiarity strength of the two sources.

The UVSD and 2DSD models can produce symmetrical and asymmetrical relational ROCs by changing the relative variance of the two source-item distributions, but these results do raise some thorny theoretical issues for these models. For example, according to these models, changes in ROC symmetry cannot be caused simply by increasing the memory strength of one source over the other because this would simply move the two strength distributions farther apart, leaving the ROC symmetry unaffected. The asymmetry can be influenced only if the strengthening manipulation also affects the variances of the source items. The problem, however, is that strength manipulations like increasing study duration do not generally lead to an increase in memory variance in tests of item recognition, but they do increase variance in source memory tests. The UVSD model might account for this discrepancy by assuming that the strengthening manipulations simply have different effects on item recognition (i.e., it increases target strength) and relational recognition (i.e., it increases one source’s variance). This, however, raises the question of why the same manipulation has such different effects on these two tests. But of more noted importance, it indicates that the model needs to include at least two different types of memory strength to account for recognition ROCs. The 2DSD assumes two different types of memory strength (e.g., source-1 strength and source 2), but it also has difficulty dealing with these effects because it aims to account for item and source memory by using the same set of underlying memory strength distributions. In this way, experimental manipulations are expected to have parallel effects on tests of item and source recognition, and thus it should not be the case that a manipulation like study duration would have one effect on item strength (i.e., increasing strength) and a different effect on source strength (i.e., increasing variance).

One set of relational ROC results that seems worthy of additional study is from Rotello et al. (2000), who showed that the degree of asymmetry in the relational ROC was dependent on whether subjects were instructed to use a recall-to-reject strategy. If those results can be replicated, it suggests that the degree of asymmetry observed in the relational ROCs is to some extent under the subjects’ control. Although the existing models do not rule out the possibility that a subject’s retrieval orientation can affect the underlying memory processes, such a finding would be useful in further characterizing the functional nature of these processes. It also raises the question of whether retrieval orienta-
6. Exclusion ROCs are curved and negative going. In relational recognition tests, subjects must reject or “exclude” lures that are related in some way to the studied items or pairs. If completely new items or pairs are included in the test list, then it is possible to plot the proportion of incorrectly accepted related lures against the proportion of incorrectly accepted nonstudied items or pairs, and this produces what is referred to as an exclusion ROC. Although these exclusion ROCs are often overlooked in ROC studies of relational recognition, they do differ from standard relational ROCs in several ways and thus can be quite informative.

Like standard item recognition ROCs, the exclusion ROCs are curved downward; however, they quickly approach the diagonal (i.e., the point at which the likelihood of accepting an exclude item is equal to that of accepting a new item) and in some cases cross it. Figure 7 presents representative exclusion ROCs from studies of source memory (Yonelinas, 1999), word-pair recognition (Healy et al., 2005; Kelley & Wixted, 2001), and word conjunction recognition (Lampinen et al., 2004). In the source test, the exclusion ROC represents the probability of incorrectly accepting an item from the lure list (e.g., list 1) when told to accept items only from a target list (e.g., list 2) plotted against the probability of accepting a nonstudied item. In the word-pair experiment, the exclusion ROC represents the probability of accepting a recombined word pair plotted against the probability of accepting a pair of nonstudied items. In the conjunction recognition experiment, the ROC represents the probability of accepting a recombined word (e.g., blackbird when blackmail and jailbird had been studied) against the probability of accepting a completely new compound word. Each of the exclusion ROCs indicates that when the subjects adopt a strict criterion they accept more related lure items than new items (i.e., the ROC is above the chance diagonal), but as their criterion becomes more lax they are equally or more likely to accept a new item than a related lure item (i.e., the ROC approaches or goes below the chance diagonal).

Curved, negative-going exclusion ROCs have been observed in three source memory experiments (Yonelinas, 1994), five associative recognition experiments (Healy et al., 2005; Kelley & Wixted, 2001), and two conjunction experiments (Lampinen et al., 2004). In addition, Heathcote, Ramond, and Dunn (2006) recently examined exclusion ROCs in source and plurality recognition and found downward-going ROCs in both tasks. Thus, the pattern appears to be fairly robust. Note that when the exclusion ROCs are plotted in z-space, they resemble item ROCs in the sense that they are relatively linear. However, the z-slopes range from .4 to .6 and thus appear to be considerably lower than those typically seen in item recognition tests.

The curved negative-going exclusion ROCs indicate that encoding does not lead to a simple increase or decrease in memory strength and that recognition involves at least two functionally independent memory components—at least one of which appears to be a signal detection process. The curvilinearity of the exclusion ROCs is problematic for threshold models and supports signal detection and hybrid models. In fact, as predicted by the signal detection-based models, the ROCs are approximately linear in z-space.

The observation that exclusion ROCs tend to cross the chance diagonal, however, is problematic for models that treat memory encoding as leading to either a simple increase or decrease in memory strength. That is, if encoding simply increases memory strength, then the ROC must remain above the chance diagonal as response criterion is varied, whereas if encoding simply decreases memory strength then the ROC must remain below the chance diagonal. For this reason, the results are problematic for the EVSD and mixture models, which predict that if performance is above chance at one point on the ROC it must also be above chance at all other points. That is, the probability of incorrectly accepting a lure item as coming from the target source should always be less than the probability of accepting a completely new item. This means that the exclusion ROC will always be below the chance diagonal, which conflicts with the observed ROCs.

So how can the exclusion ROCs be explained? Two general accounts have been provided. First by the DPSD and VRPD models, because studied items are expected to be more familiar than nonstudied items, the familiarity process leads the ROCs to be curvilinear and above the chance diagonal, but recollection that the lure item is not a target (e.g., if it was in list 1 then it could not have

Figure 7. Exclusion receiver operating characteristics (ROCs), which plot the probability of incorrectly accepting a related lure against the probability of accepting a new item, for (a) source recognition (short list condition from Experiment 1 in Yonelinas, 1994), (b) word-pair recognition (strong item condition from Experiment 1 in Kelley & Wixted, 2001), and (c) conjunction recognition (the multiple presentation condition from Experiment 2 in Lampinen et al., 2004). The exclusion ROCs are concave and negative going.
been in list 2) pushes the ROC downward, such that as response
criterion is relaxed the ROC crosses the chance diagonal. Figure 7
presents the fit of the DPSD model to the observed exclusion
ROCs and indicates that the model can produce the observed
pattern of results. The critical assumption in this model is that
recollection can operate in opposition to familiarity. It is not clear
whether the SON and STREAK models can account for these
results, but because they both include recollection and familiarity
processes, it seems they might be able to do so in a manner similar
to that of the DPSD model.

Alternatively, by the UVSD model (as well as extensions of
this model like the 2DSD model), the negative-going ROCs can
be produced if the study phase has very little (or even a
negative) effect on the overall strength of the lure items relative
to the new items, but it leads to a pronounced increase in the
variance of the lure items, such that some studied items increase
in memory strength whereas others decrease in strength. For
example, the model was fit to the conjunction ROC in Figure 7
(plotted as a dashed line). An examination of that figure indicates
that the model provided a good descriptive fit for the
observed data. However, to do so the model parameters indicated
that the study phase led to an average decrease in memory
strength for the related lures ($d' = -1.20$) relative to the new
items. Moreover, it suggests that the study phase led to a two-
to threefold increase in the variance of the related lures relative
to the new items ($V_0 = 2.71$). The fits of the UVSD model to
the other exclusion ROCs are even more surprising. For example,
in the associative test, the model suggests that the study phase had virtually no effect on the average memory strength of
the re-paired items (i.e., $d' = .096$), yet it almost doubled the
variance of those items relative to the new pair distribution (i.e.,
$V_0 = 1.83$). A similar pattern was seen in the source exclusion
ROC (e.g., $d' = 0.16$, $V_0 = 1.95$). To accomplish this, it is
necessary to assume that the study phase led to an increase in
strength for about half of the exclusion items and a decrease in
strength for the other half of those items, yet it did not alter the
Gaussian shape of the strength distributions. These results
suggest that although the model provides a good mathematical
description of the exclusion ROCs, it does not provide a par-
ticularly compelling psychological explanation of those results.
We do note, however, that one might take the model’s param-
eters at face value and conclude that they provide evidence for
a form of memory repression or suppression (e.g., Anderson,
Bjork, & Bjork, 1994). However, we are not aware of any
attempts to explain the results of relational recognition tests
using such an approach. An alternative interpretation is that the
results indicate that there may be two different processes oper-
ating on item strength or two different types of strength signal
that can act in opposition to one another (Heathcote et al.,
2006). This possibility is consistent with the dual process
models discussed above, which assume that recollection can act
in opposition to familiarity.

In sum, there appear to be two major accounts for the exclusion
results, the recollection/familiarity account and high variance ac-
count. The psychological plausibility of the recollection/familiarity
account argues in favor of that interpretation, but further theo-
retical and empirical work exploring the variance account would be
informative.

**Figure 8.** Confidence receiver operating characteristics (ROCs) and
remember–know (RK) ROCs from Stretch and Wixted (1998) and Yoneli-
nas et al. (1996). The figure illustrates that when subjects make RK and
confidence responses to each test item the RK ROC points fall along the
same function as the confidence ROC. The figure presents the fits of the
dual process signal detection (DPSD) and unequal variance signal detec-
tion (UVSD) models and indicates that both models fit the empirical data
quite well.

**7. RK and confidence ROCS overlap.** In a number of studies,
confidence judgments and RK reports have been directly com-
pared when the encoding conditions and materials have been held
constant. Results from these studies have shown that RK scores
typically fall along the same ROC that fits the recognition confi-
dence judgments. For example, Figure 8 shows “remember,” and
“remember plus know” responses plotted along with confidence-
based ROCs from the same subjects. As can be seen in Figure 8,
the RK points fall along the same functions that fit the confidence
ROCs. Thus, overall sensitivity ($z$-intercept) and degree of asym-
metry ($z$-slope) are comparable for recognition confidence and RK
judgments. Wixted and Stretch (2004) reexamined several previ-
sous studies that had collected RK and confidence judgments and
found that the $z$-slopes from the two procedures were similar in 15
out of 16 different experiments from six different published studies
(Gardiner & Java, 1990; Rajaram, 1993; Rotello et al., 2004;
Stretch & Wixted, 1998; Yonelinas, 2001b; but see Rajaram,
Hamilton, & Bolton, 2002). Malmberg and Xu (2006) also found
that RK and confidence ROCs were indistinguishable. However,
Rotello, Macmillan, Hicks, and Hautus (2006) reported that the
RK and ROC slopes were similar for a shallow encoding condi-
tion, but the RK slopes were slightly higher than the ROC slopes
for a deep encoding condition. In any case, direct comparisons of
RK and ROC results indicate that in a vast majority of cases, the
two methods produce ROCs that overlap.

In some of these studies, the majority of remember responses
(e.g., 95%) were associated with the highest level of recognition
confidence (Yonelinas, 2001b; Yonelinas et al., 1996), whereas in other studies, remember responses were associated with a wider range of recognition confidence judgments (Rotello et al., 2004; Stretch & Wixted, 1998). These differences are relevant for several models discussed below. Although there are several methodological factors that will need to be further examined to determine why these differences arose, one critical factor is the nature of the RK test instructions. In the studies in which remember responses were limited primarily to the highest level of recognition confidence, subjects were told that they should only make a remember response if they could, if asked, report what they recollected about the item. This was done to ensure that subjects were not making a remember response on the basis of high levels of familiarity. In contrast, the studies that found remember responses associated with lower levels of recognition confidence did not include this latter instruction. In fact, Rotello, Macmillan, Reeder, and Wong (2005) directly compared these test instructions and replicated the differences just described, within a single study. Thus, the results suggest that when remember responses are made only when the subjects can report some qualitative information about the study event, then remembering is associated with the highest recognition confidence. However, under less strict instructions, subjects sometimes make remember responses even when they cannot remember any aspect of the event that they can report, and these items can be associated with lower recognition confidence. In addition, Rotello et al. (2005) found that under the less strict instructions there were only very small differences in the models’ ability to fit RK and confidence data but that in general the models were ordered EVSD, UVSD, STREAK, and DPSD from best to worst. However, some subjects were fit better by single process models, whereas others were fit better by dual process models, which was interpreted as indicating that the extent to which two or one process was used might vary across subjects.

The overlapping RK and confidence ROCs suggest that the two component/processes that support RK responses are similar to those supporting standard recognition confidence judgments. Although it is conceivable that the RK instructions might bias subjects to engage in different retrieval strategies than those used in standard recognition tests, the current results suggest that the RK instructions do not fundamentally alter the processes that the subjects bring to bear when making recognition judgments.

But what are the processes underlying RK and confidence responses? Single-component models have been argued to account for the RK results (e.g., Donaldson, 1996), but the results from the studies examining both RK and confidence responses are clear in ruling out single-component models such as the EVSD and HT models. That is, to account for the curved asymmetrical confidence ROCs, one requires no less than two functionally independent memory components. Even accounting for just the RK results from these studies requires two components, as the slope and intercept of the RK zROCs behave in a manner similar to that of the confidence ROC data, which demands two components.

The convergence between the RK and confidence data is consistent with several different memory models, each leading to slightly different interpretations of the data. For example, the results are consistent with models that assume that remember responses simply reflect high levels of recognition strength (e.g., UVSD, 2DSD). Because both RK and confidence judgments are expected to be based on memory strength, the RK points should fall on the same function as that generated from recognition confidence responses. Note that the UVSD model is a two-component model, but it assumes that there is no direct relationship between the two memory components underlying the model and RK reports. That is, it assumes that RK and confidence ROCs will be determined by the combined effects of familiarity strength and the old–new variance ratio. Because the remember response criterion is thought to be arbitrary, there should be no simple relationship between either of these memory components and RK reports.

The results are also consistent with dual process models that assume that recollection leads to remember responses and high confidence old judgments (e.g., DPSD). In these models, remember reports are not thought to reflect simply strong recognition responses, but rather to provide an indirect index of the recollection process. Nonetheless, because recollection and familiarity are assumed to contribute to RK judgments and recognition confidence judgments, the RK and confidence ROCs should take on the same shape, thus the results of the two methods should converge. The VRDP and SON models have not been applied to RK results, but because they include recollection and familiarity components they could accommodate these results in a manner similar to the DPSD model.

The convergent RK and ROC results can also be accounted for by the STREAK model, but this convergence is treated as coincidental from the standpoint of this model. The model can produce a remember ROC point that falls on, above, or below the confidence ROC. Therefore the results present somewhat of a puzzle for STREAK and suggest that the model does not capture this regularity in recognition.

8. The z-slopes observed in RK studies can be extremely high (z-slope > 1). Although in experiments that have directly contrasted RK and ROC results, the two methods are found to be in close agreement; in many pure RK experiments, the z-slopes are much higher than would be expected from the item recognition ROC literature. For example, Rotello et al. (2004) reviewed 373 published RK conditions and found that the average RK slope was close to 1.0 (also see Dunn, 2004, for a similar analysis), and in many cases the slopes were well above 1.0. The right panel in Figure 4 plots the z-slope against z-intercept for 403 RK conditions described by Dunn (2004), in the same way as was done in the left panel for item recognition confidence ROCs. An examination of the figure highlights just how different the ROC and RK results are. As previously described, item recognition z-slopes are almost always below 1. In contrast, the RK data contain studies with z-slopes less than 1, but they also contain a comparable number of studies with z-slopes well above 1.0, a pattern never seen in item recognition. Thus, it is clear that RK results are not always consistent with those from recognition ROC studies. Although the materials and test procedures may differ across the RK and ROC studies, thus complicating the comparison of the z-slopes seen on those studies, it is difficult to attribute such a large difference to this factor alone (see Kelley & Wixted, 2001).

The high RK z-slopes indicate that RK judgments can be quite different from recognition confidence judgments, and they either suggest that RK and confidence judgments can sometimes involve different processes or decision rules, or that they can be differentially affected by false recollection or measurement artifacts. These high RK z-slopes are problematic for almost all of the current models. For example, the results are problematic for signal
detection models such as the EVSD and UVSD models, which have been used to account for RK results by assuming that remember responses simply reflect strong or high confidence recognition responses. Similarly, hybrid models such as the SON, mixture, and the DPDS models can account for cases in which RK and ROC results converge, but these models would then fail to account for the cases in which the test results diverge.

Exactly why the RK z-slopes are often so high is not yet clear, but several different accounts have been suggested. The first possibility is that the discrepancy is related to the different decision rules involved in RK and confidence judgments (i.e., the STREAK model of Rotello et al., 2004). If RK responses are based on the difference between recollection and familiarity strength, and the old–new responses are based on the sum of recollection and familiarity strength, then the RK z-slope can be either greater than, equal to, or less than that of the recognition confidence ROC. Second, Rotello et al. (2004) also suggested that the DPDS model (Yonelinas, 1999) might be able to produce high RK slopes if it incorporates a nonmnemonic guessing process or false recollection. That model predicts that remember responses should lie along the confidence ROC, but if some of the new items are falsely recollected then the model can produce high RK slopes. That is, false recollection should primarily increase the remember false alarm rate, leading the RK z-slope to be higher than the confidence z-slope.

Several measurement artifact accounts of this effect have also been put forward. For example, if subjects were to adopt a variable response criterion between remember and know responses this could produce an artificial increase in RK z-slope (Wixted & Stretch, 2004). Alternatively, examining the z-slopes of the aggregate data, as was done in the previous reanalyses of the RK literature, rather than examining the subject z-slopes could produce increases in the apparent RK z-slope (Wixted & Xu, 2006). Finally, C. M. Parks (2007) suggested that the remember responses might fall below the ROC if subjects adopt a strict definition of remembering such that only some aspects of the study event are treated as adequate to support a remember response. In addition, we point out that because false alarms associated with remember responses often approach zero that truncation problems may also have influenced the analysis of RK slopes. However, as of yet, none of these theory-based or methodological hypotheses have been extensively tested, and thus the reason for the high z-slopes remains somewhat of a mystery.

Obviously, the alternative accounts of the discrepancy between RK and confidence results are quite different, and further work that aims to differentiate between these explanations will be critical in testing various theories of recognition. At the very least, these results indicate that the relationship between RK and confidence judgments is not a simple one. Moreover, the alternative accounts force one to think about some potentially interesting questions regarding the involvement of alternative decision rules, measurement artifacts, and false recollection—questions that probably have not been adequately considered in the ROC literature.

Memory Impairments

ROC methods are being used increasingly to examine the neural substrates of recognition memory and to test the predictions of the various ROC models. For example, lesion studies have been used to determine if various brain regions are necessary for different recognition memory processes/components. Although this area of research is just beginning, at least one general pattern has been established and several related findings are beginning to emerge.

9. Item ROCs in amnesics are often symmetrical such that the z-slope approaches 1.

Damage to the medial temporal lobe decreases overall recognition performance but produces recognition ROCs that are relatively symmetrical (see Figure 9). For example, Yonelinas, Kroll, Dobbins, Lazzara, and Knight (1998) examined item recognition ROCs in amnesic stroke patients with medial temporal lobe damage that included the hippocampus and the surrounding parahippocampal gyrus. Unlike control subjects who exhibited curved and asymmetrical ROCs, the amnesics exhibited symmetrical ROCs (i.e., a z-slope of .97, see Figure 9A). It is important to note that the high ROC slope could not be attributed simply to low levels of performance, as a group of control subjects that were matched in performance using shorter study durations still exhibited asymmetrical ROCs (z-slope = .75). Item recognition ROCs were also examined in a group of hypoxic cardiac arrest patients (see Figure 9B; Yonelinas et al., 2002), and again the amnesics’ ROCs were close to symmetrical (z-slope = .93). In a recognition memory study of rats, a sham lesion group exhibited curved and asymmetrical ROCs like those seen in human subjects (z-slope = .74), whereas rats with lesions restricted to the hippocampus exhibited symmetrical ROCs (z-slope = 1.08, see Figure 9C). Again, these results could not be attributed simply to lower overall levels of performance because a control group was matched to the hippocampal group in overall performance by testing after a longer delay and were found to still exhibit asymmetrical ROCs (z-slope = .86).

A similar pattern was reported in a patient with relatively selective hippocampal atrophy related to meningitis (Figure 9D; Aggleton et al., 2005). The patient exhibited a z-slope that was closer to 1.0 than the control subjects (z-slope = .87 and .62, respectively). The patient’s results, however, are noteworthy for two reasons. First, despite the difference in z-slope compared with the control subjects, the patient’s overall recognition performance did not differ substantially from that of the control subjects (the ROCs clearly overlap). Thus, the difference in ROC shape seen in this patient indicates that his or her recognition deficit cannot be attributed to a simple decrease in memory strength or overall memory performance. As a practical aside, the results show that using a simple yes–no recognition test would have failed to show that the patient was impaired on recognition. Rather it was only through the use of the ROC analysis that the patient’s recognition deficit was revealed. Second, it is noteworthy that the patient’s ROC was not perfectly symmetrical (i.e., z-slope of .87). Although the z-slope of this patient’s recognition ROC was much greater than the control subjects in this study (i.e., z-slope of .62), z-slopes comparable with that of this patient have sometimes been reported in healthy control subjects under other conditions. This difference may be related to the rather subtle recognition deficit exhibited by this patient. In any case, the result indicates that amnesics’ item ROCs are not always perfectly symmetrical.

Another case study led to largely consistent conclusions (Cipollo et al., 2006) in the sense that the patient exhibited ROCs with a much higher slope (z-slope = 1.06) than that seen in control subjects (z-slope = .80). Note that the ROC plotted in Figure 9E includes all types of materials the patient was tested on (i.e.,
words, buildings, landscapes, and faces). However, the patient was
least impaired on the face memory test and produced an ROC that
looked closest to that of the control subjects in this one condition.
An additional study examined item recognition ROCs in a mixed
etiology group (e.g., heroin overdose, carbon monoxide poisoning,
cardiac arrest, and unknown etiology) of amnesic patients (Wais,
Wixted, Hopkins, & Squire, 2006). Because very few responses
were collected from each patient, subject-level ROC analyses
could not be conducted, but the aggregate data were in general
agreement with the previous studies, with the z-slope higher for the
patient ROC (z-slope = 1.16) than the control ROC (z-slope = .80).
However, when the amnesics were tested with short 10-item
lists, their performance increased and the slope of the aggregate
ROC decreased (z-slope = .83). The latter result is interesting in
suggesting that under short list conditions amnesics’ ROCs may
become more asymmetrical.

In addition to studies of medial temporal lobe amnesia, item rec-
ognition ROCs have also been examined in patients with thalamic
lesions and in normal control subjects while under the amnesic effects
of scopolamine. Like medial temporal lobe damage, thalamic lesions
appear to decrease sensitivity while leading the ROCs to become
more symmetrical (e.g., Kishiyama et al., 2005; Zoppelt, Koch,
Schwarz, & Daum, 2003). In addition, administration of scopolamine
prior to encoding—a drug that influences cholinergic receptors and
has amnestic effects on long-term memory—was found to impair
recognition for pictures (Sherman et al., 2003). Although further
studies are needed to verify these results, the ROCs suggested that the
drug decreased sensitivity (z-intercept dropped from .96 to .72) and
led the ROCs to become slightly more symmetrical (z-slope increased
from .63 to .70). The authors also noted that the zROCs for the control
subjects were U-shaped whereas for the drug group they exhibited a
slight ~-shape.

Figure 9. Item recognition receiver operating characteristics (ROCs; left panels) and zROCs (right panels) for
medial temporal lobe amnesics and control subjects. A: ROCs for amnesics with damage to the hippocampus and
surrounding parahippocampal gyrus, control subjects, and performance-matched control subjects (Yonelinas et al.,
1998). B: ROCs for amnesics with suspected selective hippocampal damage and control subjects (Yonelinas et al.,
2002). C: ROCs for rats after selective hippocampal lesions, sham surgeries, and matched control subjects (Fortin
et al., 2004). D: ROCs for a patient with selective hippocampal damage and for control subjects (Aggleton et al.,
2005). E: ROCs for an amnesic with hippocampal and parahippocampal damage and for control subjects (Cipolotti
et al., 2006). F: ROCs for a mixed etiology group of amnesic patients, control subjects, and patients tested with short study
lists (Wais et al., 2006). The patient ROCs are generally symmetrical (z-slope approaching 1.0) relative to the control
subjects, even when patients and control subjects are matched for overall performance.
Note that an important limitation of memory impairment studies is that they often have included very few subjects. This is particularly critical in studies in which very few responses per subject condition have been collected (e.g., Wais et al., 2006). However, with the exception of the latter study, these studies have tended to include large numbers of trials to conduct adequate subject-based ROC analyses. Thus, the common finding that item ROCs in amnesia tend to be symmetrical is reasonably well established. In contrast, conclusions about the effects of thalamic lesions or the effects of scopolamine await additional study. Additional neuro-psychological studies as well as animal lesion studies making use of ROC methods will prove useful not only in understanding the neural substrates of recognition memory but also for furthering our understanding of the recognition deficits of memory impaired individuals.

The relative symmetry of the item recognition ROCs observed in amnesic patients indicates that medial temporal lobe damage does not simply reduce memory strength, rather it disproportionately disrupts recollection compared with familiarity, or it disproportionately disrupts old–new variance compared with memory strength. The amnesic results indicate that the medial temporal lobes are particularly important for one of the two components/processes underlying recognition. For example, according to the DPSD model, the symmetrical ROCs seen in these patients indicates that they suffer disproportional deficits in recollection, which is responsible for the asymmetry observed in the control subjects’ ROCs. The highly symmetrical item ROCs often observed in amnesia suggests that recollection was effectively eliminated under these conditions. Nonetheless, there is some evidence that when the amnesia is mild (e.g., Aggleton et al., 2005) or when only short lists of items are tested (Wais et al., 2006) recollection can be operative in these patients.

The amnesia results are also consistent with the SON model. Because SON contains recollection and familiarity processes, and because it assumes that item familiarity reflects an EVSD process, if amnesia is assumed to disrupt recollection, then the model predicts more symmetrical ROCs in amnesia. The STREAK model also predicts that recollection should be more disrupted than familiarity in amnesia, but it currently has no mechanism linking the old–new item variance to the level of recollection or familiarity strength. Therefore, it can describe the results, but it does not provide an explanation for differences in ROC slope seen in these patients relative to control subjects.

The VRDP model is also consistent with the amnesia data in the sense that recollection reductions are expected to lead the ROCs to become more symmetrical. Moreover, in an experiment comparing control subjects with those given the amnesic drug scopolamine (Sherman et al., 2003) the VRDP model fit the data from the drug group better than did the DPSD model. In addition, whereas the DPSD model simply indicated that the drug led to a decrease in recollection, the VRDP model indicated that the drug resulted in a decrease in the strength of the recollected information but in no change in the probability of recollection. If these results can be replicated, it would suggest that one may need to measure two properties of recollection rather than just one.

The amnesia results are also broadly consistent with some signal detection models, such as the UVSD and 2DSD models, which can describe the data quite well. However, the psychological explanation they provide for the results has yet to be adequately articulated. According to these models, the amnesic patients suffer a severe impairment in the old–new variance component and a less severe impairment in the memory strength component. That is, amnesics would appear to be able to increase the strength of studied items relative to new items, but unlike control subjects, they do not tend to increase the variance of the old items relative to new items. However, why the medial temporal lobe would be more important for the variance component than the memory strength component in recognition is not yet clear. In any case, the UVSD model supports conclusions similar to that of the dual process models in the sense that amnesia cannot be interpreted as simply reflecting a deficit in memory strength. In fact, if anything, memory strength would seem to be relatively preserved, whereas it is the variance component that is the most affected.

The dissociative effects of amnesia on the two memory components in recognition present problems for all of the single-component models of recognition and present some important challenges for several more complex models as well. For example, although the mixture model has not been applied to studies of amnesia, the amnesics’ results present the model with some difficult questions. The model can produce symmetrical ROCs that are above chance only when the attentional parameter approaches 1.0 (i.e., all studied items are perfectly attended). The model could account for the asymmetrical ROCs in the control subjects by assuming that subjects do not attend to all the studied items. However, to account for the amnesics’ symmetrical ROCs, it would be necessary to assume that although the amnesics exhibited a deficit in overall recognition performance, their attention at encoding was better than that of the control subjects, which seems to be a rather odd assertion.

The amnesia results also provide preliminary support for an additional assumption of the DPSD model, which is that within the medial temporal lobe, the hippocampus is particularly critical for recollection but not for familiarity. In each of the ROC studies of amnesia described above, the DPSD model was used to estimate the effects of amnesia on recollection and familiarity. In cases in which there was documented damage to the hippocampus as well as the surrounding medial temporal lobe cortex, the patients exhibited deficits in recollection and similar, although somewhat smaller, deficits in familiarity (Cipolotti et al., 2006; Yonelinas, 2002; Yonelinas et al., 1998). In contrast, in cases in which the damage was expected to be restricted to the hippocampus (Aggleton et al., 2005; Fortin, Wright, & Eichenbaum, 2004; Yonelinas et al., 2002), the amnesics exhibited a selective deficit in recollection. The results from the Wais et al. (2006) study suggested that both recollection and familiarity were reduced in patients with hippocampal damage, but they are somewhat difficult to interpret because a subject analysis was not conducted and because it is not known what role extrahippocampal damage played in those patients (Yonelinas et al., 2004). Fortin et al.’s (2004) study is perhaps the least ambiguous with respect to lesion location because the experimenters had control over lesion location, and in this case it is clear that hippocampal damage selectively affected recollection. Thus, in general, the ROC results suggest that damage to the hippocampus disrupts recollection but has little if any effect on familiarity, whereas when damage extends into the surrounding parahippocampal gyrus, both recollection and familiarity are severely reduced.
Discussion

ROCs methods have been successfully applied to various different recognition memory paradigms and have proven useful in furthering our understanding of the memory processes involved in recognition. The empirical review has revealed at least nine well-established regularities, each of which was found to have important theoretical implications, along with numerous other findings that if further confirmed will also have important implications. Without reiterating all of those points here, at the most general level there are four important conclusions that can be drawn from the ROC results.

Single-Component Models Do Not Provide an Adequate Account of Recognition Memory

Perhaps the most general implication of the ROC results is that they present problems for all single-component models of recognition memory. One illustrative example is the finding that in item recognition tasks, sensitivity and ROC asymmetry are functionally dissociable. There is no way to account for this dissociation without postulating at least two separable memory components. Thus, models such as the HT and EVSD models that contain only a single memory component (or a single index of memory performance) cannot account for the existing data. The same conclusions were supported by the results from various other recognition paradigms such as tests of relational recognition, exclusion tasks, and RK tasks, as well as results from studies of amnesics. These results provide strong support for various two-component models of recognition such as the unequal variance, dual process, and mixture models.

The importance of this finding should not be underestimated. At the most basic level, it tells us that recognition memory does not reflect a monolithic or unified type of memory. Rather there are at very least two separate aspects of recognition memory that can vary independently of one another. There is of course a legitimate debate regarding exactly what those two components reflect (e.g., the processes of recollection and familiarity, strength and variance ratio components, or the processes of familiarity and attention). However, the ROC results leave no question about whether recognition reflects one or more components.

Given that the ROC literature indicates that there are at least two components underlying recognition memory judgments, it makes little sense to ask questions like “What is the effect of a given manipulation on recognition memory?” Rather, we should be asking “How does a manipulation influence the two components underlying recognition memory?”

An important practical implication of these results is that a single point in ROC space in a given condition (i.e., a hit and false alarm rate) is necessarily ambiguous, or overinterpretable. For example, two experimental conditions may result in identical hit and false alarm rates but can be associated with very different ROCs (e.g., different levels of sensitivity and asymmetry). This is true regardless of one’s preferred theory of recognition memory (i.e., how one interprets sensitivity and asymmetry). Thus, we see the recent trend to move away from single-point methods like yes–no recognition and their single-component measures of accuracy such as $d’$ and $A$ to multiple-point methods like inclusion/exclusion, RK, and ROC methods, as a positive one. Such measures permit much stronger conclusions regarding similarities or differences between experimental conditions than do single-point measures which are insufficient to determine the level of memory performance.

Note that it is sometimes suggested that forced-choice recognition tests are useful in providing a single index of recognition accuracy (i.e., proportion correct). However, this approach simply ignores the problem because, as with yes–no recognition, a given level of performance can be arrived at via various different combinations of the underlying memory components.

When measuring performance in relational recognition tests like source recognition and associative recognition, it is even more critical to collect more than simple yes–no responses. In these paradigms, no less than three functionally independent components (and by some models up to five free components) were required to account for relational recognition. At the very least, one needs an index of the strengths or variances of the two types of items in the relational tests, plus some index of the degree of curvilinearity in the ROC. By the dual process models, this reflects the difference in familiarity between the two sources and the probability that items from each source are recollected; by the mixture model, this would include measures of familiarity strength and attention for each type of item. Note that the issue would seem to become even more complicated if the relational tests required a sourceA/sourceB/new response. In this case, it would seem that two components are needed to account for item recognition and at least three are needed to discriminate between the sources.

Pure Threshold Models Do Not Provide Adequate Accounts of Recognition

The second general conclusion one can draw from the ROC results is that pure threshold models are inconsistent with the existing recognition data (i.e., HT and HLT models). For example, the curved ROCs almost always observed in item recognition rule against the threshold models, which predict linear ROCs. Note that the threshold assumption is also shared by multinomial models (Batchelder & Riefer, 1990, 1999), and thus the curved item ROCs present problems for these models as well (for earlier discussion of this point, see Kinchla, 1994). The threshold assumption also underlies the common practice of subtracting false alarm rates from hits to correct for guessing, and thus the current results undermine these estimates. The aspect of the threshold models that leads them to incorrectly predict linear ROCs is that they assume that there is a sensory limit on the true memory strength of test items. The curved ROCs make it clear that, at least in item recognition tests, there is no memory threshold below which performance is completely at chance; rather subjects exhibit some ability to discriminate between old and new items even when their response criterion becomes very lax. In contrast, the ROCs observed in relational recognition tests are sometimes fit quite well by threshold models, indicating that these models might serve as good descriptive models of relational recognition and suggesting that in some sense it may be useful to think of relational recognition tasks as involving some type of a threshold component (i.e., some items fall below the memory threshold and the subject is unable to retrieve any relevant memory information about the source or associations made with that item). However, even if restricted to relational recognition tasks, the threshold models still
run into problems because curvilinear ROCs are quite often observed in these tests, indicating that a pure threshold model is not entirely adequate even for these tasks.

**Pure Signal Detection Models Do Not Provide Adequate Accounts of Recognition**

The third general conclusion to follow from the observed results is that pure signal detection models are inconsistent with the existing results (e.g., EVSD, UVSD, 2DSD, and STREAK). The core assumption of these models is that the underlying strength distributions are Gaussian in shape—an assumption that is assessed by testing for linear zROCs. Results from most, although not all, item recognition studies indicate that the zROCs are approximately linear, suggesting that these models do provide a good mathematical description of these data. However, the finding that the ROCs observed in relational recognition tests are almost always U-shaped in z-space indicates that the underlying memory strength distributions cannot be truly Gaussian. Although the basic signal detection model has been modified in various ways, such as increasing the number of strength dimensions (e.g., 2DSD, Glanzer et al., 2004) and incorporating dual process assumptions (e.g., Wixted & Stretch, 2004; Wixted, 2007), these models maintain the Gaussian assumption of signal detection theory and thus remain in conflict with the recognition data.

The problems that arise for signal detection theory have far reaching implications because so many current theoretical accounts of recognition memory have been built on this framework. For example, the d’ statistic used to measure memory sensitivity depends critically on the validity of the Gaussian assumption. Although this assumption appears to be approximately right in tests of item recognition, it is certainly not appropriate in relational recognition tests. Even if one opts to adopt such a statistic in item recognition studies, it alone is not sufficient there either, as one necessarily needs to include measures of a second memory component (i.e., variance ratio, recollection, or attention).

The observed ROCs also have implications for various computational models. For example, global memory models such as TODAM and SAM generally base recognition judgments on an assessment of a Gaussian memory strength signal, and thus they are not consistent with the U-shaped zROCs observed in relational recognition tasks (Malmberg & Xu, 2006). However, as Malmberg and Xu (2006) have shown, these models may be able to account for nonlinear zROCs by introducing noise to the models. In addition, some of these models include recall mechanisms that might be used to supplement the standard recognition familiarity mechanism in such a way that they may lead to non-Gaussian memory strength distributions. One computational model that does incorporate familiarity and recollection processes is the Complementary Learning Systems model (e.g., McClelland et al., 1995; K. A. Norman & O’Reilly, 2003; O’Reilly & Rudy, 2001). The model is based on the assumption that the hippocampus supports recollection by developing minimum overlapping representations of prior episodes, whereas the surrounding cortex gradually tunes populations of cortical units to respond strongly to different stimuli in such a way that it can discriminate between familiar and new items. A review of the model goes beyond the scope of the current article (for a detailed discussion of the model, see K. A. Norman & O’Reilly, 2003), but the results from preliminary simulations with that model are promising because they indicate that it can account for the differential importance of the hippocampus versus the surrounding neocortex in recollection and familiarity, and it successfully predicts behavioral properties of recollection that other computational models do not. For example, the hippocampus produces a threshold output such that it can produce linear ROCs, whereas the cortex produces curvilinear ROCs (Elman, Parks, & Yonelinas, 2007). It is not yet clear whether the model is able to account for the full body of results discussed above; however, models such as this one are particularly promising because they aim to incorporate behavioral and neuroanatomical knowledge within the same theoretical framework.

**Hybrid Models Provide a Good Account of the Recognition ROC Data**

The fourth general conclusion is that the ROC results support hybrid memory models that include both signal detection and threshold processes. For example, dual process models (e.g., DPSD, VRDP, and SON) assume that familiarity is a signal detection process and so can produce curved ROCs, whereas a threshold recollection process can lead to the U-shape often observed in z-space. Similarly, the mixture model assumes that there is a signal detection process that is supplemented by a threshold attention process at encoding. The signal detection process leads to curved ROCs, whereas the attentional process can lead the ROCs to exhibit a U-shape in z-space.

Which of the hybrid models provides the best account of the existing data? In our view, all of the hybrid models provide good, but not perfect, accounts of the existing data, and at this point the existing results do not provide conclusive evidence in favor of any one hybrid model over the others. For example, the mixture model is a promising model that accounts very well for the ROCs observed in tests of item and relational recognition. The model has not been applied to exclusion, RK, or amnesia data, so it is not yet known if it can account for those results. A consideration of that model suggests that it likely can account for some of these ROC results, but, there do appear to be important challenges for the model as well. For example, it likely could account for the overlapping RK and confidence ROCs. However, it is not clear how it would account for the RK z-slopes much greater than 1 or the negative-going exclusion ROCs. The amnesia data may also be a challenge for the model in the sense that it is not clear how it would account for the symmetrical ROCs in amnesia. It may be possible, however, to account for some of these latter findings by incorporating both source strength and item strength mechanisms into the same model.

The other hybrid models that did quite well were dual process models that included separate recollection and familiarity processes. For example, the DPSD model provided a good account for the major empirical findings from the item and relational recognition studies as well as the exclusion and amnesia data. The model also provided an account for the convergence of RK and confidence results. However, it did not provide a good account for the high z-slopes observed in RK studies. Although it was suggested that it might be able to account for these results by assuming that subjects falsely recollect some proportion of the test items, additional empirical and theoretical work will be necessary to assess this possibility. Another potential problem was that repetition was
found to lead to a decrease in recollection estimates (Macho, 2004). Although this pattern has not been seen with other manipulations that increase associative recognition performance (Healy et al., 2005), or with manipulations that increase source strength (e.g., Yonelinas, 1999), further studies exploring this result will be important.

The VRDP model is a direct extension of the DPSD model that maintains the threshold assumption about recollection but incorporates additional parameters to describe the strength and variance of the recollection strength distribution. Modeling recollection strength with a Gaussian distribution seems quite promising, and it allows the model to predict not only linear and U-shaped zROCs but also ~-shaped zROCs. It would also provide a natural way of accounting for the cases in which remembered items are associated with various different levels of recognition confidence. Although there is little evidence for ~-shaped zROCs as of yet, in at least one study such an ROC was observed in subjects under the influence of scopolamine (Sherman et al., 2003). Further work exploring the generalizability of these results will be extremely useful in verifying the usefulness of this model. One potential challenge for the VRDP model is related to its number of free parameters. In tests of item recognition, it requires four free parameters compared with the DPSD model, which accounts for item recognition results with only two free parameters. This increase in complexity may be demanded by the data in some cases (e.g., Sherman et al., 2003). However, further studies showing that the model provides a significant improvement over the simpler models will be necessary before this more complex model can be adopted. Similar to other hybrid models, the VRDP model must also become more complex to deal with relational recognition data because at least one more recollection parameter is required to account for the probability of recollecting lure items. In fact, to fit the relational data it may also be necessary to include strength and variance parameters for the lure recollection distributions. Therefore, the number of free parameters may become even more problematic for the model. Even for item recognition data, we found several different sets of parameters that all provided excellent fits to the data when we fit the model to observed item ROCs. For the model to be useful in examining ROC data, additional work will be necessary either to put constraints on some of the model parameters or to find ways of weeding through these various different model solutions.

The SON model is another modification of the DPSD model that is similar to the VRDP model. Like the VRDP model it can produce linear, U-shaped, and ~-shaped zROCs. However, it has only been applied to associative recognition tasks, so it is not yet clear if it can account for other recognition results. Given the number of free memory parameters (i.e., five), it may be able to account for many of the results. However, the challenges faced by this model are similar to those faced by the VRDP model. Namely, future studies will be necessary to determine if its number of free memory parameters are justified compared with simpler models and to determine if the model parameters behave reasonably. As discussed earlier, one such test of the model suggested that it leads to conclusions that are inconsistent with the underlying assumptions of the model (Macho, 2004), but only additional tests of the model will tell whether it provides an acceptable account of recognition memory. Moreover, as with the VRDP model, given the number of free parameters it is not clear if the model will provide unique solutions when used to account for recognition results.

Conclusions

The examination of ROC data from a wide range of recognition memory paradigms indicates that single-component models of recognition memory are inadequate and that there are at least two functionally distinct component/processes involved in recognition. To account for ROC results, current models have incorporated several different theoretical divisions such as distinctions between recollection/familiarity, item/associative information, attention/familiarity, and strength/variance. Although there is support for all of these distinctions, in general, only the hybrid models assuming the contribution of signal detection and threshold processes were successful at accounting for the existing literature (e.g., recollection/familiarity, attention/familiarity).

Disagreements still exist, however, about the nature of these two processes that require further empirical work to resolve. The existing results argue strongly against pure threshold and pure signal detection models, but we do not consider the current review to have provided definitive evidence for the superiority of one hybrid model over the others. Our hope, however, is that by directly comparing these alternative theories we have come to more clearly see the important empirical questions that need to be answered. In so doing, we hope that the next phase of ROC research will focus on testing competing predictions of these various hybrid models and, hopefully, will lead to an even deeper understanding of recognition memory.

References


Received February 15, 2006
Revision received January 25, 2007
Accepted February 22, 2007