Cognitive Psychology 64 (2012) 1–34

Contents lists available at SciVerse ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model

Jeffrey J. Starns [a,*], Roger Ratcliff [b], Gail McKoon [b]

[a] University of Massachusetts, Amherst, United States
[b] Ohio State University, United States

ARTICLE INFO

Article history:
Accepted 12 October 2011
Available online 11 November 2011

Keywords:
Recognition memory
Receiver operating characteristic (ROC)
Sequential sampling models
Signal detection models
Dual process theory

ABSTRACT

We tested two explanations for why the slope of the z-transformed receiver operating characteristic (zROC) is less than 1 in recognition memory: the unequal-variance account (target evidence is more variable than lure evidence) and the dual-process account (responding reflects both a continuous familiarity process and a threshold recollection process). These accounts are typically implemented in signal detection models that do not make predictions for response time (RT) data. We tested them using RT data and the diffusion model. Participants completed multiple study/test blocks of an "old"/"new" recognition task with the proportion of targets and the test varying from block to block (.21, .32, .50, .68, or .79 targets). The same participants completed sessions with both speed-emphasis and accuracy-emphasis instructions. zROC slopes were below one for both speed and accuracy sessions, and they were slightly lower for speed. The extremely fast pace of the speed sessions (mean RT = 526) should have severely limited the role of the slower recollection process relative to the fast familiarity process. Thus, the slope results are not consistent with the idea that recollection is responsible for slopes below 1. The diffusion model was able to match the empirical zROC slopes and RT distributions when between-trial variability in memory evidence was greater for targets than for lures, but missed the zROC slopes when target and lure variability were constrained to be equal. Therefore, unequal variability in continuous evidence is supported by RT modeling in addition to signal detection modeling. Finally, we found that a

* Corresponding author. Address: Department of Psychology, 441 Tobin Hall, University of Massachusetts – Amherst, Amherst, MA 01003, United States
E-mail address: jstarns@psych.umass.edu (J.J. Starns).

0010-0285/$ - see front matter
doi:10.1016/j.cogpsych.2011.10.002

two-choice version of the RTCON model could not accommodate the RT distributions as successfully as the diffusion model.

1. Introduction

Even the simplest decisions take time to make, and a complete account of decision making cannot ignore this temporal dimension. In recognition memory experiments, for example, participants are asked to decide whether words were previously studied ("old") or not ("new"). The resulting response time (RT) distributions show systematic changes in location and spread across experimental conditions and are invariably positively skewed in shape (Ratcliff & Murdock, 1976; Ratcliff & Smith, 2004; Ratcliff, Thapar, & McKoon, 2004). Unfortunately, recognition memory researchers have paid little attention to the rich information available in RT data; instead, theories of recognition are predominantly tested only in terms of the accuracy of memory decisions. The current work addresses a popular topic in recognition memory with the goal of showing what can be gained by considering RT in addition to accuracy.

1.1. Accuracy models and ROCs

In the early 1990s, Egan's (1958) pioneering work on recognition memory receiver operating characteristics (ROCs) was revived as a method for testing memory theories (Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, 1994). ROCs are plots of the hit rate ("old" responses to old items) against the false alarm rate ("old" responses to new items) across conditions in which response bias varies but memory evidence is constant. In many cases, the hit and false alarm rates are converted to *z*-scores, and the resulting function is called a zROC. This conversion often makes it easier to assess model predictions; for example, zROCs should be linear under the assumption that memory evidence is normally distributed.

zROC functions are usually based on confidence ratings, but they can also be formed from an "old"/"new" task in which bias is manipulated experimentally. In the current experiment, for example, we varied the proportion of targets on the test to produce different levels of bias. Specifically, participants studied multiple lists that were each followed by a 56-item "old"/"new" recognition test. Tests had either 12 (.21), 18 (.32), 28 (.50), 38 (.68), or 44 (.79) targets, and participants were informed of the target proportion after each study list just before they began the test list. To manipulate memory performance, we used high and low frequency words, and each study list included words studied once, twice, or four times.

Fig. 1 shows stereotypical zROC functions from a paradigm like our own, with the circles representing words studied once and the triangles representing words studied four times. Words studied four times should be more easily recognized than words studied once, leading to a higher hit rate in all of the conditions. Test lists with a low proportion of targets promote a bias to say "new," leading to a low hit rate and a low false alarm rate (the leftmost points). As the proportion of targets increases, participants become more willing to say "old," and the hit and false alarm rates increase for all item types. The displayed zROCs follow linear functions with slopes less than one, both of which are benchmark characteristics of zROCs from recognition experiments (Egan, 1958; Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff et al., 1992, 1994; Wixted, 2007; Yonelinas & Parks, 2007).

zROC modeling has sustained a heated debate about the nature of memory evidence, with controversy focused on two models offering contrasting explanations for why zROC slopes are less than one (Wixted, 2007; Yonelinas & Parks, 2007). The unequal-variance signal detection (UVSD) model assumes that decisions are based on a single evidence variable, frequently conceptualized as the degree of match between a probe and memory traces (Clark & Gronlund, 1996; Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997). Match values are normally distributed for targets and lures, with a higher mean and greater variability for the target items (Cohen, Rotello, & Macmillan, 2008; Heathcote, 2003; Hirshman & Hostetter, 2000; Mickes, Wixted, & Wais, 2007). Participants establish a response crite-

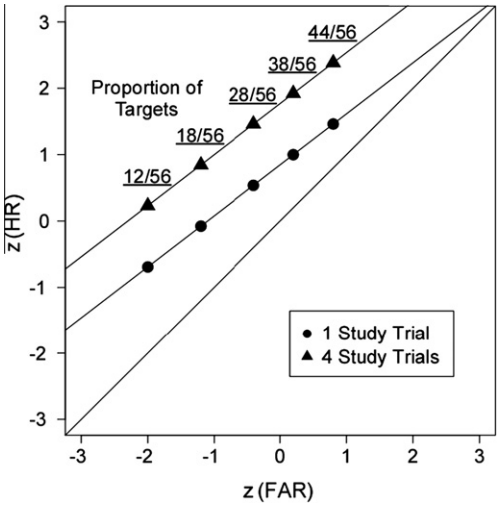

Fig. 1. Stereotypical zROC functions from a two-choice task with a target-proportion manipulation. The circles show targets studied one time and the triangles show targets studied four times. The five points on each function are the five target proportion conditions, and the proportion of targets used in the current experiment is shown above them. z(FAR) and z(HR) indicate the z-transformed false alarm rate and hit rate, respectively. The functions were generated from the UVSD model.

rion on the match dimension, and any test word with a match exceeding the criterion is called "old." The criterion accommodates response biases; for example, participants should use a more liberal (lower) criterion when test words are predominantly targets and a more conservative (higher) criterion when the test words are predominantly lures. In fitting the model, the lure distribution is scaled to have a mean of 0 and a standard deviation of 1. All other parameters are measured relative to the lure distribution, including the position of the response criterion (λ), the mean of the target distribution (μ), and the standard deviation of the target distribution (σ).

The UVSD model predicts linear zROC functions with a slope equal to the ratio of the standard deviations of the lure and target evidence distributions ($1/\sigma$). By assuming that memory match values are more variable for targets than for lures, the model can accommodate zROC slopes below one. For example, the zROC functions in Fig. 1 were generated from the UVSD model with $\sigma = 1.3$. The predicted zROC intercept is equal to the target distribution's mean divided by its standard deviation, so higher intercepts indicate better memory performance (i.e., stronger evidence for targets).

The primary competitor to the UVSD model is the dual-process signal detection (DPSD) model (Yonelinas, 1994; Yonelinas & Parks, 2007), which assumes that some recognition decisions are based on a vague sense of familiarity while others are based on recollecting a specific detail of the learning event (e.g., "this word came right before 'nurse' on the study list"). For targets, recollection always leads to an "old" response when it succeeds (with probability R) and has no influence on responding when it fails (with probability $1 - R$). Decisions for non-recollected targets and for all lure items are based on familiarity. Familiarity follows a signal detection process like the one described for the UVSD model, except that evidence must be equally variable across targets and lures ($\sigma = 1$). Thus, when responding is based solely on familiarity, the model predicts linear zROCs with a slope equal to one. When recollection succeeds for a proportion of the targets, the model predicts zROC functions that have slopes less than 1 and show slight non-linearity (although in practice the model's predictions are often very close to a linear function).

As is clear from the previous discussion, a principle difference between the UVSD and DPSD models is the mechanism for producing zROC slopes less than one. The UVSD model assumes that decisions are based on a single underlying evidence variable, and slopes are below one because targets have a higher variance than lures. The DPSD model assumes that slopes are below one because some decisions are based on recollection while others are based on familiarity. Comparing the UVSD and DPSD

models has been a primary focus of the recognition literature, along with occasional consideration of various mixture models (e.g., Decarlo, 2002; Onyper, Zhang, & Howard, 2010). However, no consensus has been reached. The various models all provide a very close fit to zROC data and show almost complete mimicry in their predictions in the range of parameter values actually observed in experiments (for reviews, see Wixted, 2007; Yonelinas & Parks, 2007). Therefore, zROC functions by themselves are not sufficiently diagnostic. Our goal is to put both the unequal-variance and the dual-process accounts to a stronger test using RT data. In the ensuing sections, we describe our strategy for extending both accounts to RTs.

1.2. RT data and unequal variance

We tested the unequal-variance account by implementing it in the diffusion model in an attempt to accommodate both zROC functions and RT distributions. The diffusion model is a sequential sampling model for accuracy and RT in simple, two-choice decisions (Ratcliff, 1978). The model has been shown to fit both response proportions and RT distributions across a wide variety of tasks (for reviews, see Ratcliff & McKoon, 2008; Wagenmakers, 2009), and it has been successfully applied in diverse fields such as aging (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004; Starns & Ratcliff, 2010), child development (e.g., Ratcliff, Love, Thompson, & Opfer, in press), individual differences in IQ (e.g., Ratcliff, Thapar, & McKoon, 2010, 2011), perceptual learning (e.g., Petrov, Van Horn, & Ratcliff, 2011), depression and anxiety (e.g., White, Ratcliff, Vasey, & McKoon, 2010), single-cell recording (e.g., Gold & Shadlen, 2000; Ratcliff, Cherian, & Segraves, 2003), and fMRI (e.g., Forstmann et al., 2010). Despite its wide application, the diffusion model has never been evaluated with zROC data. Perhaps because of this, the model has always been implemented under an equal-variance assumption, even when applied to recognition memory (Ratcliff & Smith, 2004; Ratcliff, Thapar, & McKoon, 2004). If the unequal-variance explanation of zROC slope is correct, then extending the diffusion model to zROC data should require abandoning the equal-variance assumption. That is, the diffusion model should match the zROCs and RT distributions in an unequal-variance version, but fail to do so in an equal-variance version. In the following section, we describe the diffusion model and how we used it to test the unequal-variance account.

1.3. Diffusion model

In the diffusion model, evidence accumulates over time until it reaches one of two boundaries associated with the two response alternatives, such as a_{OLD} and a_{NEW} in Fig. 2. The starting point of the accumulation process varies from trial to trial over a uniform distribution with a mean of zero and a range s_z.[1] In a given trial, the process approaches one of the boundaries with a drift rate (v) represented by the arrows in Fig. 2, but the process is subject to moment-to-moment variability resulting in actual paths represented by the wandering lines. The within-trial standard deviation is a scaling parameter, and we follow convention by setting it to .1 (Ratcliff, Van Zandt, & McKoon, 1999). As a result of the within-trial variability, the process may terminate on the boundary opposite the average direction of drift, leading to errors. The within-trial variability also results in different finishing times across trials, creating the distributions of decision times that are shown outside of each boundary. RT predictions are derived by combining the decision times and a uniformly distributed non-decision component with mean T_{er} and range s_t. The non-decision component represents the time for processes such as reading the test word before accessing memory and executing a motor response once a decision has been made.

Each parameter of the model has a direct psychological interpretation. The drift rate represents the quality of the evidence driving the decision; for example, a word studied four times should have a higher drift rate than a word studied once. The distance between the two response boundaries represents the speed–accuracy compromise: a narrow boundary separation leads to fast decisions and a

[1] In parameterizing the model, one can either set the bottom boundary to zero and estimate parameters for the starting point (z) and top boundary (a) or set the starting point to zero and estimate parameters for both boundaries (a_{OLD} and a_{NEW}). These alternative parameterizations produce equivalent models, and parameters assuming a starting point at zero can be directly translated to parameters assuming a bottom boundary at zero ($z = -a_{NEW}$ and $a = a_{OLD} - a_{NEW}$).

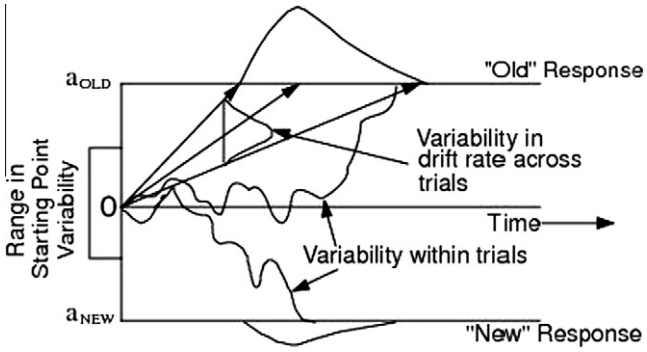

Fig. 2. The diffusion model of two-choice decision making. The horizontal lines at a_{OLD} and a_{NEW} are the response boundaries. In this example, the boundaries are for "old" versus "new" responses in a recognition task. The line at zero is the starting point, which varies from trial to trial across a uniform range. The straight arrows show average drift rates, and the wavy lines represent the actual accumulation paths that are subject to moment-to-moment variation. Three average drift rates are shown to represent the across-trial variability in drift. Predicted decision time distributions are shown at each boundary. Predicted RTs are found by adding the decision times and a uniform distribution of non-decision times with mean T_{er} and range s_t.

high probability of reaching the wrong boundary due to noise whereas a wide boundary separation leads to slower decisions and a smaller chance of reaching the wrong boundary. The relative position of the boundaries represents response biases: if one boundary is closer to the starting point than the other, then the accumulation process will be more likely to terminate at the close boundary. Decision times will also tend to be shorter for the close boundary than for the far boundary. Our target proportion variable should influence the response boundaries, with a_{OLD} approaching the starting point and a_{NEW} moving farther from the starting point as target proportion increases. Therefore, "old" responses should get faster and more frequent from the .21 to the .79 target-proportion conditions, and "new" responses should get slower and less frequent.

The diffusion model assumes that evidence from the stimulus varies between trials, creating normal distributions of drift rates (Ratcliff, 1978). Fig. 3 shows drift rate distributions for targets and lures in a recognition task, each with its own mean (μ) and standard deviation (η). The drift distributions represent across-trial variation in evidence; thus, they are analogous to distributions of evidence in signal detection theory. To evaluate the unequal-variance explanation of zROC slopes, we tested models in which the drift distribution standard deviation (η) could differ for targets and lures. Fig. 3 shows a larger η for targets. Like signal detection models, decreasing the lure/target ratio in the η parameters produces a lower zROC slope. With standard values for the other parameters, the model predicts a slope close to 1.0 with equal η values down to a slope around 0.76 when the target η is double the lure η.[2] The value of η also affects RT distributions, primarily in terms of the relative speed of correct and error responses. Specifically, higher values of η produce slower error RTs relative to correct RTs (see Ratcliff & McKoon, 2008, for a detailed discussion). The distributional effects are relatively subtle, resulting in high estimation variability in parameter recovery simulations (Ratcliff & Tuerlinckx, 2002). Because unequal η's are of particular theoretical importance in the current investigation, we collected more data than in most previous studies (i.e., 20 sessions for each participant) to obtain more reliable η estimates. We tested whether an unequal-variance model could fit the RT and zROC data from our experiments, and we compared this model to an equal-variance diffusion model in both group and individual-participant fits. In this way, we devised a novel test of the unequal-variance account of zROC slope.

[2] Notice that the lure/target ratios for the standard deviations of the drift distributions do not match the zROC slopes. In the basic UVSD model, the slope is equal to the lure/target ratio, but this is not true for RT models. RT models have sources of variability other than variability in memory evidence, most notably variability in the evidence accumulation process. These sources of decision noise affect both targets and lures, increasing the total variability for each and producing zROC slopes that are closer to 1 than the standard deviation ratio (Ratcliff & Starns, 2009).

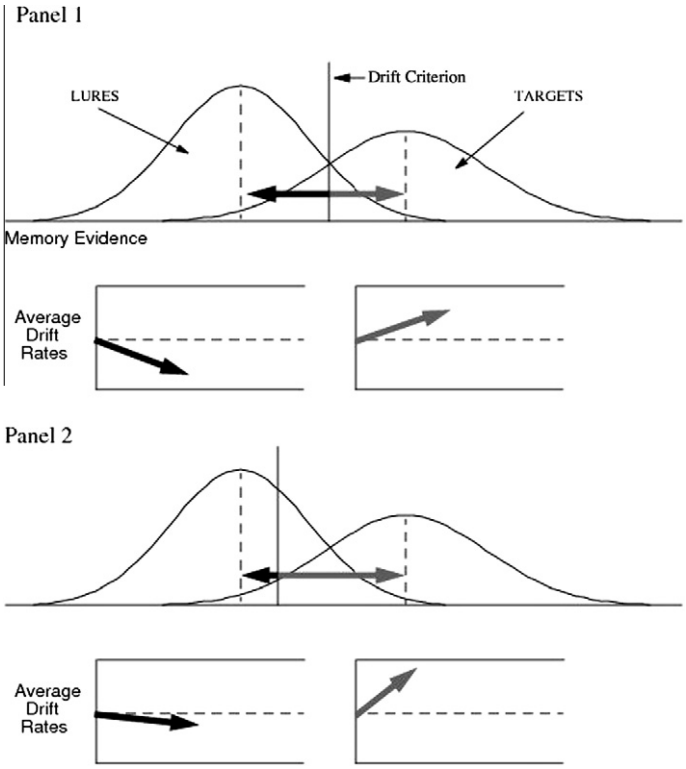

Fig. 3. Demonstration of how drift rates are determined based on the position of the drift distributions and the drift criterion. The average drift rates are equal to the deviation between the mean of the drift distribution and the drift criterion (shown as black arrows for lures and grey arrows for targets). Panel 1 shows a relatively unbiased position for the drift criterion, and Panel 2 shows a more liberal setting. The figure demonstrates a situation in which the target drift distribution is more variable than the lure distribution.

The vertical line in Fig. 3 is the drift criterion (*dc*), which is a subject-controlled parameter defining the zero point in drift rate. Specifically, drift rates (v) are determined by the distance of the evidence value from the drift criterion, with positive drifts above the drift criterion and negative drifts below (Ratcliff, 1978, 1985). The drift criterion provides an additional method for introducing response biases (besides changes in the boundary positions), and it acts the same way as the response criterion in signal-detection theory. Therefore, our target proportion manipulation might influence the drift criterion in addition to the response boundaries (although previous results with this manipulation are mixed; Criss, 2010; Ratcliff, 1985; Ratcliff & Smith, 2004; Ratcliff et al., 1999; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). For example, Panel 1 of Fig. 3 represents a test with an equal proportion of targets and lures, so the drift criterion is placed near the mid-point of the distributions to ensure that most targets have positive drift rates and most lures have negative drift rates. Panel 2 represents a test that is predominantly target items, making it advantageous to set the drift criterion to ensure that almost all targets have positive drift rates even if a substantial proportion of lures will also have positive drift rates (see Ratcliff et al., 1999, Fig. 32).

Although changing either the boundaries or the drift criterion can produce identical biases in terms of response proportion, these alternative mechanisms are identifiable because they have different effects on RT distributions. Changing the positions of the boundaries relative to the starting point has a larger effect on the leading edge of the RT distribution compared to changing the drift criterion (see Ratcliff et al., 1999, p. 289; Ratcliff & McKoon, 2008). Moreover, the two decision parameters have distinct psychological interpretations. To understand the difference, it is helpful to think of the diffusion

model as a dynamic version of the signal detection process (Ratcliff, 1978). Instead of the decision being made based on one value of match to memory, a new match to memory could be made every 10 ms with the results of these matches accumulated over time. At each 10 ms time step, the accumulation process takes a step toward the "old" boundary if the match on that time step falls above the drift criterion or takes a step toward the "new" boundary if the match falls below the drift criterion. Therefore, the drift criterion is the cutoff between the amount of memory match that supports an "old" response and the amount of memory match that supports a "new" response. In contrast, the response boundaries determine how far the accumulation process must go in the "old" or "new" direction before the corresponding response will be made. The diffusion model implements the process just described, except that it uses infinitely small time steps to model the continuous accumulation of evidence (Ratcliff, 1978).

1.4. Unequal variance in RT confidence models

The unequal-variance account has already been extended to RT in a handful of studies investigating zROCs formed from confidence ratings (Ratcliff & Starns, 2009; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004). Van Zandt (2000) developed a Poisson counter model for recognition ROCs

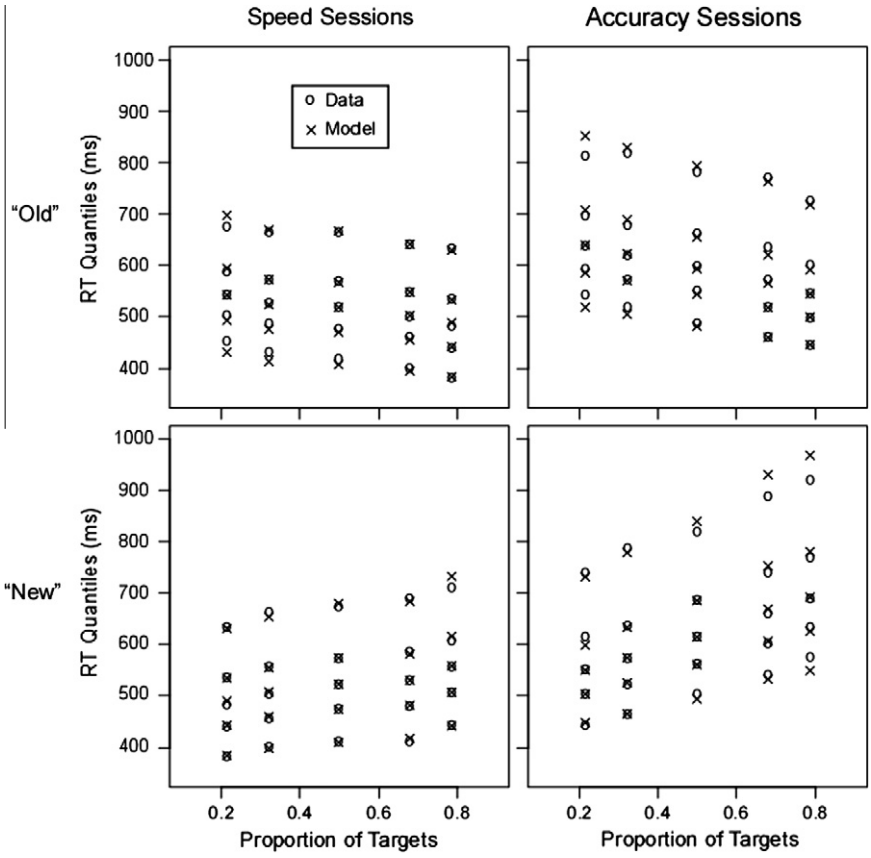

Fig. 4. Data and diffusion model predictions for the RT distributions in the speed-emphasis (left column) and accuracy-emphasis (right column) sessions based on the proportion of targets tested (collapsed across all other variables). The top row shows "old" responses and the bottom row shows "new" responses. Each set of plotting points shows the .1, .3, .5, .7, and .9 quantiles of the RT distribution.

formed from tasks in which "old"/"new" decisions are followed by confidence ratings (also see Pleskac & Busemeyer, 2010). She assumed an unequal-variance model of memory evidence, but did not compare it to an equal-variance model to determine the importance of this assumption. Ratcliff and Starns developed the RTCON model for tasks in which participants made a single response on a 6-point scale from "definitely new" to "definitely old." In RTCON, zROC slopes are affected by the position of decision criteria, so the model can predict slopes below one even with equal variance in memory evidence (Ratcliff & Starns, Fig. 4). Nevertheless, fits to data showed that the unequal variance assumption was needed to match empirical zROC slopes.

The RT models for confidence described in the last paragraph represent a small segment of the RT modeling literature. In general, the RT modeling has focused on two-choice tasks, with models for confidence and multiple-choice responding in a more nascent stage of development (Leite & Ratcliff, 2010; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2001). Therefore, the two-choice paradigm in the current study is an important extension to the existing studies. Our design also permitted the first test of the unequal-variance account using the diffusion model, which has been much more thoroughly investigated than either the Poisson-counter model or RTCON.

1.5. Dual-process explanation and RTs

Although the dual-process approach has not been implemented in a model for RT distributions, a central tenet of dual-process theory is that familiarity becomes available before recollection (Yonelinas, 2002). This tenet is supported by experiments in which a signal occurs at varying lags after the presentation of the test stimulus and a response must be made within a brief time frame after the signal (Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree, Dolan, & Jacoby, 1999). A number of studies have used the response signal paradigm to compare discriminations that can be made based on a vague sense of familiarity – such as whether or not a word was previously presented – and discriminations that require specific recollection – such as discriminating words presented together on the same study trial (intact pairs) from words presented on different study trials (rearranged pairs). Participants make familiarity-based discriminations early in processing, with performance rising above chance around 450 ms after stimulus presentation (Gronlund & Ratcliff, 1989; McElree et al., 1999; Rotello & Heit, 2000). However, discriminations that require recollection cannot be made until later in processing. For example, discrimination rises above chance around 550 ms or later for words studied on different lists (McElree et al., 1999) or intact versus rearranged pairs (Gronlund & Ratcliff, 1989). Response-signal studies also show a non-monotonic function for highly familiar lures that can be rejected by recollecting an incompatible studied item (for example, the lure word "dog" when "dogs" was one of the studied items). The false alarm rate for such lures rises across early signals followed by a decrease at later signals, with the reversal occurring around 700 ms (Dosher, 1984; Hintzman & Curran, 1994) or even as late as 900–1000 ms (Rotello & Heit, 2000). Such results show that early processing is dominated by familiarity with a delayed influence of recollection.[3]

All of the studies discussed in the last paragraph challenged participants to discriminate classes of items with the same or very similar levels of familiarity (e.g., words studied on different lists); therefore, they specifically promoted the use of recollection. Some have questioned the role of recollection in item recognition tasks, given that familiarity is sufficient to discriminate targets from lures (e.g., Gillund & Shiffrin, 1984; Gronlund & Ratcliff, 1989; Malmberg, 2008). However, the DPSD model as-

[3] Investigations using the Remember-Know (RK) procedure have found that R responses (which presumably reflect recollection) are made more quickly than K responses (which presumably reflect familiarity; Dewhurst & Conway, 1994; Dewhurst, Holmes, Brandt, & Dean, 2006), which may lead some to conclude that recollection is a faster process than familiarity. This conclusion is inappropriate for several reasons. The speed advantage for R responses is eliminated when the level of confidence in controlled, showing that the RT differences do not reflect distinct underlying processes (Rotello & Zeng, 2008). Moreover, in the RK studies, RTs for both responses are well past the point at which both familiarity and recollection have become available based on response signal studies (Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree et al., 1999), with RT means for R responses typically around 1000 ms. Therefore, these studies cannot provide information about which form of information has the earliest influence on responding.

sumes that recollection does play a role in item recognition, as evidenced by zROC slopes less than one.

We tested the dual-process account by evaluating the effect of time pressure on zROC slope. We had participants complete multiple sessions of data collection. In half of the sessions, participants were instructed to give themselves time to make an accurate decision. With these instructions, both recollection and familiarity should influence responding according to the dual-process account. For the other sessions, we pushed participants to respond very quickly. Participants were able to respond with a mean RT of 526 ms, and the studies discussed above suggest that responding should be based almost exclusively on familiarity at this pace. Therefore, if recollection is truly the factor that produces zROC slopes less than 1, then we should see slopes that are closer to 1 with speed instructions than with accuracy instructions.

1.6. Comparing RT models

A secondary goal of the current work was to compare the diffusion model with a two-choice version of RTCON, a model that has previously been applied only to confidence rating data (Ratcliff & Starns, 2009). The diffusion model has been shown to outperform several alternative sequential sampling approaches for two-choice data (Ratcliff & Smith, 2004), so the diffusion fits should set a high standard from which to judge RTCON. We will discuss RTCON in more detail when we begin evaluating the model.

2. Method

2.1. Participants

Four Northwestern University undergraduates participated, each of whom was currently serving as a research assistant. Each participant completed 20 hour-long sessions, 10 with speed-emphasis instructions and 10 with accuracy-emphasis instructions. Speed and accuracy sessions alternated, with three participants beginning with speed and one beginning with accuracy. The first session in each instruction condition was considered practice and excluded from data analyses, so each participant had experienced both the speed and accuracy condition before they contributed any data.

2.2. Design

Target proportion (.21, .32, .50, .68, or .79) varied across study-test blocks within each session. There were three levels of target strength (1, 2, or 4 presentations at study), which together with the lure items comprised four item types. The four item types were crossed with two word frequencies (high and low) and all eight factorial combinations appeared in each study-test block. Speed-emphasis versus accuracy-emphasis instructions differed across sessions, resulting in 80 conditions overall (5 probability conditions $\times$ 4 item types $\times$ 2 word frequencies $\times$ 2 instructions conditions).

2.3. Materials

For each session, a set of high frequency (60–10,000 occurrences/million) and low frequency (4–12 occurrences/million) words were randomly selected from pools of 729 and 806 words, respectively (Kucera & Francis, 1967). All study/test blocks within a session had a unique set of words. Words were studied in pairs to encourage elaborative encoding, and each pair had words from the same frequency class (high or low). For the .21, .32, and .5 target proportion conditions, each study list was composed of 26 pairs of words. The first and last pairs served as buffer items. For the critical items, there were 12 high frequency pairs and 12 low frequency pairs, and within each class there were 4 pairs presented once, 4 presented twice, and 4 presented four times. For the .68 and .79 target proportion conditions, we added filler pairs to the study list to increase the number of targets on the test. These fillers always came at the beginning of the study list, so the retention interval for the critical items was constant

across all proportion conditions. For the .68 condition, the study list began with five additional filler pairs – three pairs studied once, one pair studied twice, and one pair studied four times. The .79 study lists began with eight additional filler pairs – five studied once, two studied twice, and one studied four times.

We used different numbers of studied words across the bias conditions so we could fit more study/test cycles into the sessions; that is, so participants would not have to study additional target items even on lists where they would not be tested. We now briefly discuss whether this choice might have introduced interpretation problems. ROCs are analyzed under the assumption that all points along the function represent the same memory evidence with only response bias varying. The fact that we added items to the study list for the .68 and .79 target conditions represents a potential violation of this assumption, given that memory may be worse for longer lists (e.g., Bowles & Glanzer, 1983; Gronlund & Elam, 1994, but see Dennis, Lee, & Kinnell, 2008). However, a close inspection of list length studies suggests that any effect arising from this change should be negligible. When retention interval is controlled (as it is for all of our critical items), list length effects are quite small (Dennis & Humphreys, 2001), and these small effects are produced by adding many more items than we have added. For example, Bowles and Glanzer found that adding 120 items to the beginning of a study list decreased accuracy by .04–.10 on a forced-choice recognition test. In light of this, we expected that adding a maximum of 16 items (eight pairs) would not produce a noticeable effect. Moreover, Dennis and Humphreys report evidence that the small effect of adding a large number of items to the beginning of a study list is based on waning attention (also see Underwood, 1978). Our design limited this factor with its relatively quick pace of study/test cycles (20 cycles within an hour-long session). Finally, the zROC functions from our experiments were consistent with the existing literature; that is, they closely followed linear functions with slopes less than 1. For these reasons, we are not concerned that adding items to the high probability lists distorted the results.

Test lists were constructed of individual words, and test composition varied across the target proportion conditions. For the .21 condition, the targets were taken from one pair within each of the six strength conditions formed by crossing word frequency (high or low) with number of learning trials (1, 2, or 4). Each study pair contributed two separate target trials on the test, for a total of 12 targets. The test contained 42 critical new items (21 high and 21 low frequency) and began with two new item buffers, for a total of 44 new items. The .32 condition had the same composition, except that six additional filler targets (drawn from the strength conditions at random) replaced six of the critical new items (three high and three low frequency), for a total of 38 new and 18 old. Tests in the .5 condition had 24 critical old and 24 critical new items. There were four targets from each of the six strength conditions, and the critical new items were split evenly between high and low frequency. There were also four old and four new fillers – two fillers started the test and the rest were distributed randomly throughout the test, yielding a total of 28 new and 28 old. The .68 condition had 24 critical old items as in the .5 condition, but the critical new items were reduced to 18 and there were 14 old filler items (two serving as the beginning test buffers), for a total of 18 new and 38 old items. For the .79 condition, the critical new items were reduced to 12 and 6 old fillers were added, resulting in 12 new and 44 old items.

2.4. Experimental procedure

Initial instructions informed participants that they would study lists of word pairs with a recognition test immediately following each list. They were told that a message would appear after each study list to inform them of the proportion of targets on the upcoming test, and that they should use the proportion information to help them decide if each word was old or new. For speed sessions, they were asked to respond as quickly as they could without resorting to guessing, and their RT was displayed on the screen following each response. For the accuracy sessions, they were asked to be careful to avoid mistakes, and the word ERROR appeared on the screen after each incorrect response. In both conditions, participants were cautioned not to make responses before they had read the test word, and a TOO FAST message appeared on the screen for all RTs faster than 250 ms.

Participants completed 20 study/test blocks in each session, with four blocks randomly assigned to each of the target-proportion conditions. On the study lists, each pair remained on the screen for 1 s

followed by 50 ms of blank screen. Immediately after the last studied pair, participants were prompted to begin the test. The test message informed participants of the (approximate) target to lure ratio by signaling one of the following: "1 OLD: 4 NEW", "1 OLD: 2 NEW", "1 OLD: 1 NEW", "2 OLD: 1 NEW", or "4 OLD: 1 NEW."

2.5. Modeling procedures

All model fits were performed using the SIMPLEX fitting algorithm (Nelder & Meade, 1965) to minimize either χ^2 or G^2. We compared models with different numbers of parameters using *BIC* (Schwarz, 1978): $BIC = -2L + P\ln(N)$. *BIC* combines a model's optimized log likelihood (L) with a penalty term based on the number of free parameters (P). The penalty for free parameters becomes more severe with increasing sample size (N). Lower values of *BIC* indicate the preferred model. We computed the log likelihood for each model based on the χ^2 or G^2 statistic resulting from the fits. G^2 is a direct transformation of the multinomial likelihood of the observed counts in each frequency bin given the proportions in each bin predicted by the model: $G^2 = -2 * (L_{SAT} - L_{FIT})$, where L_{FIT} is the log likelihood of the model being fit and L_{SAT} is the log likelihood for a saturated model that has as many degrees of freedom as the data (so the predicted proportions are equal to the observed proportions). Thus, minimizing G^2 is equivalent to maximizing the multinomial likelihood, and the latter can be directly calculated from the former: $L_{FIT} = L_{SAT} - G^2/2$. When χ^2 was used in the initial fits, we simply used this value as an estimate of G^2, as the former is a very close approximation to the latter with large sample sizes like our own.

3. Results

We first briefly discuss the empirical results, and then we assess the unequal-variance and dual-process accounts of zROC slope. Table 1 shows the RT medians for "old" and "new" responses across all 80 conditions. Word frequency and study repetition had relatively small effects on RTs, although participants were slightly faster to accept targets that received additional study trials (the difference in "old" RT medians between targets studied once and four times was about 5 ms in the speed sessions and 15 ms in the accuracy sessions). In contrast, both target proportion and instructions (speed versus accuracy) produced large RT differences.

Fig. 4 shows full RT distributions for "old" and "new" responses across the target proportion manipulation (collapsed across frequency and item type). The left column shows results from the speed sessions and the right column shows results from the accuracy sessions. Each column of points shows the .1, .3, .5, .7, and .9 quantiles of the RT distribution (the .1 quantile is the point at which 10% of responses have already been made, etc.). The distributions were positively skewed, as can be seen in the increased spread between the .7 and .9 quantiles compared to the other adjacent quantiles. Participants responded more slowly overall in the accuracy-emphasis sessions than in the speed-emphasis sessions. RTs for "old" responses decreased as the proportion of targets increased, whereas "new" RTs increased. The effect of proportion was more pronounced with accuracy than speed instructions. The magnitude of the target proportion effect was similar for the leading edges (.1 quantiles), medians (.5 quantiles), and tails (.9 quantiles) of the RT distributions. This indicates that target proportion produced a shift in the location of the distributions with little effect on the shape or the spread of the distributions.

Table 2 shows the response proportion results. The proportions were strongly influenced by all of the independent variables. As intended, participants became more willing to make "old" responses as the proportion of targets on the test increased. The proportion manipulation had a larger effect for accuracy sessions than for speed sessions. Also, targets were more likely to be called "old" if they were presented more times on the study list. Compared to high frequency words, low frequency words had a higher proportion of "old" responses for targets and a lower proportion for lures (demonstrating a word frequency mirror effect, Glanzer & Adams, 1985). Accuracy sessions also led to higher memory performance than speed sessions.

Table 1
RT medians for "old" and "new" responses across the 80 conditions.

Instructions and target proportion	Word frequency and number of study presentations							
	High frequency				Low frequency			
	4	2	1	New	4	2	1	New
"Old" responses								
Speed								
.21	535	515	527	549	541	556	557	548
.32	520	527	525	517	520	531	542	524
.50	507	524	519	516	513	518	520	511
.68	497	489	490	488	502	510	516	498
.79	476	478	465	472	486	487	491	479
Accuracy								
.21	621	637	633	671	606	604	630	683
.32	612	625	611	648	589	602	613	648
.50	590	576	596	630	573	586	597	620
.68	559	559	578	582	552	569	574	584
.79	526	536	534	541	536	544	551	570
"New" responses								
Speed								
.21	490	474	474	487	443	479	495	501
.32	495	495	488	506	478	501	519	520
.50	504	500	529	524	506	529	524	537
.68	524	516	528	539	513	515	538	549
.79	540	541	547	559	536	546	578	583
Accuracy								
.21	538	537	536	553	546	559	566	550
.32	561	565	574	571	547	594	581	576
.50	596	607	618	619	612	629	620	611
.68	673	666	651	670	616	665	679	652
.79	683	691	695	704	673	694	685	670

Table 2
Proportion of "Old" responses across the 80 conditions.

Instructions and target proportion	Word frequency and number of study presentations							
	High frequency				Low frequency			
	4	2	1	New	4	2	1	New
Speed								
.21	.473	.407	.340	.185	.676	.568	.489	.146
.32	.547	.545	.410	.284	.763	.681	.500	.229
.50	.666	.616	.494	.348	.757	.728	.646	.286
.68	.690	.648	.548	.454	.837	.742	.643	.354
.79	.787	.753	.694	.563	.871	.814	.763	.448
Accuracy								
.21	.476	.367	.290	.099	.700	.575	.437	.074
.32	.600	.481	.427	.182	.798	.686	.589	.124
.50	.707	.656	.574	.338	.835	.802	.673	.223
.68	.867	.809	.753	.533	.921	.858	.772	.343
.79	.940	.907	.867	.721	.967	.915	.868	.465

Table 3 reports the zROC slopes and intercepts across all conditions from the group data. We used the response frequency data to construct a zROC plot for each condition, and we fit the UVSD model to the data in each plot. The free parameters in the fits were the mean (μ) and standard deviation (σ) of

Table 3
zROC slopes and intercepts.

Instructions and word frequency	zROC measure and number of presentations					
	Intercept			Slope		
	4	2	1	4	2	1
Speed						
High	.65 (.04)	.55 (.04)	.30 (.04)	.82 (.09)	.81 (.09)	.87 (.08)
Low	1.21 (.07)	.98 (.06)	.77 (.06)	.74 (.11)	.73 (.10)	.86 (.10)
Accuracy						
High	1.00 (.04)	.79 (.04)	.59 (.04)	.86 (.05)	.90 (.05)	.89 (.05)
Low	1.78 (.08)	1.45 (.07)	1.14 (.06)	.87 (.08)	.85 (.07)	.87 (.07)

Note: Values in parentheses are the standard errors of the intercept and slope estimates from a bootstrap procedure (see the text for more details).

the target distribution as well a response criterion (λ) for each of the five target proportion conditions, with the lure distribution fixed at a mean of 0 and a standard deviation of 1 (Appendix A gives the equations for the model predictions). The best-fitting parameters were used to define the intercept and slope for each zROC function (intercept = μ/σ, slope = $1/\sigma$). We also performed a bootstrap procedure to estimate the degree of variability in the slopes and intercepts (Efron & Tibshirani, 1985). To create each bootstrapped dataset, we randomly sampled trials with replacement. Specifically, we sampled N trials from each condition where N is the original number of observations for the condition. We generated 1000 bootstrapped datasets and fit each dataset with the UVSD model to produce estimates of the zROC slope and intercept. The standard deviation of these estimates across the bootstrap runs gave the standard errors for the parameters.

There are several things to note from the empirical zROC data. First, word frequency and number of learning trials had their intended effects on memory performance. Intercepts were higher for low- versus high-frequency words, and intercepts increased with additional study trials. Intercepts were also higher with accuracy than speed instructions. zROC slopes did not change much based on number of presentations. Slopes were generally lower for low- versus high-frequency words, although the effect was small in some comparisons and even reversed in one (words presented four times with accuracy instructions). Slopes were also slightly lower in general with speed than accuracy instructions. All of the slope differences were quite small in relation to the variability in the estimates.

3.1. Unequal variance and the diffusion model

3.1.1. Fit for the unequal-variance diffusion model

We will begin by evaluating the fit of the unequal-variance diffusion model to the response proportion data as well as the RT distributions, and then we will directly compare unequal- and equal-variance versions of the model. For each condition, the .1, .3, .5, .7, and .9 quantiles of the RT distributions were used to segment the data into RT bins, and the model was fit to the frequencies in each bin. The model had 66 free parameters to fit 880 freely varying response frequencies. Details on the diffusion model fitting and a full list of parameter values can be found in Appendix A. Here, we briefly summarize how the parameters were constrained across conditions. Starting point variability was held constant across all conditions. The mean and range of the non-decision times could change between speed and accuracy sessions, but were fixed across all other variables. The decision parameters (response boundaries and drift criteria) varied across the instruction and target proportion variables but were fixed across word frequency and item type (lures and targets studied once, twice, or four times). The means and standard deviations of the drift distributions could change across word frequency and item type but could not change across the target proportion conditions.

In initial fits, we tried versions in which memory evidence was constant between speed-emphasis and accuracy-emphasis sessions and versions in which evidence was free to vary across the instruction variable. Past reports have been able to fit speed/accuracy manipulations in recognition memory

with the same evidence parameters (Ratcliff & Starns, 2009; Ratcliff, Thapar, & McKoon, 2004), but the current data were better accommodated by a model with free evidence parameters (*BIC* = 7886) than by a model with constrained evidence parameters (*BIC* = 7916). Without free evidence parameters, the model predicted zROC intercepts that were too low for all of the low frequency functions in the accuracy conditions; that is, the model could not fully accommodate the change in memory performance from speed to accuracy sessions. In our speed sessions, participants maintained a quick pace even compared to the speed conditions in previous experiments; for example, young subjects in Ratcliff, Thapar, & McKoon (2004) had RT means close to 580 ms with speed instructions, compared to an overall RT mean of 526 ms in the current experiment (Ratcliff et al. had two sessions as opposed to 20 for the current experiment, so our participants may have benefitted from more practice making fast responses). The extremely fast pace in the current speed condition may have affected memory evidence by impairing participants' ability to form effective retrieval cues, and we discuss this possibility in more depth when we discuss the results for the parameter values. For now, we simply note that we used a model with free evidence across the instruction conditions and that this choice was data-driven.

The χ^2 value from the fit to group data was 2418 for the unequal-variance version of the diffusion model (note that a χ^2 distribution cannot be assumed for group fits, so the group χ^2 value cannot be used for a significance test). The individual subject χ^2 values had a median of 1493 with a range of 1445–2097. With 814 degrees of freedom (880 free response frequencies – 66 free parameters), the χ^2 critical value is 881.5 ($\alpha = .05$). Thus, the χ^2 value for all of the subjects exceeded the critical value, but this is typical for datasets with a high number of observations and many conditions (Ratcliff, Thapar, Gomez, et al., 2004). Appendix B lists all of the best-fitting parameter values, and results for the parameters of greatest interest are summarized below. The parameters were quite consistent between the individual-subject and group fits. The group parameters were within 10% of the

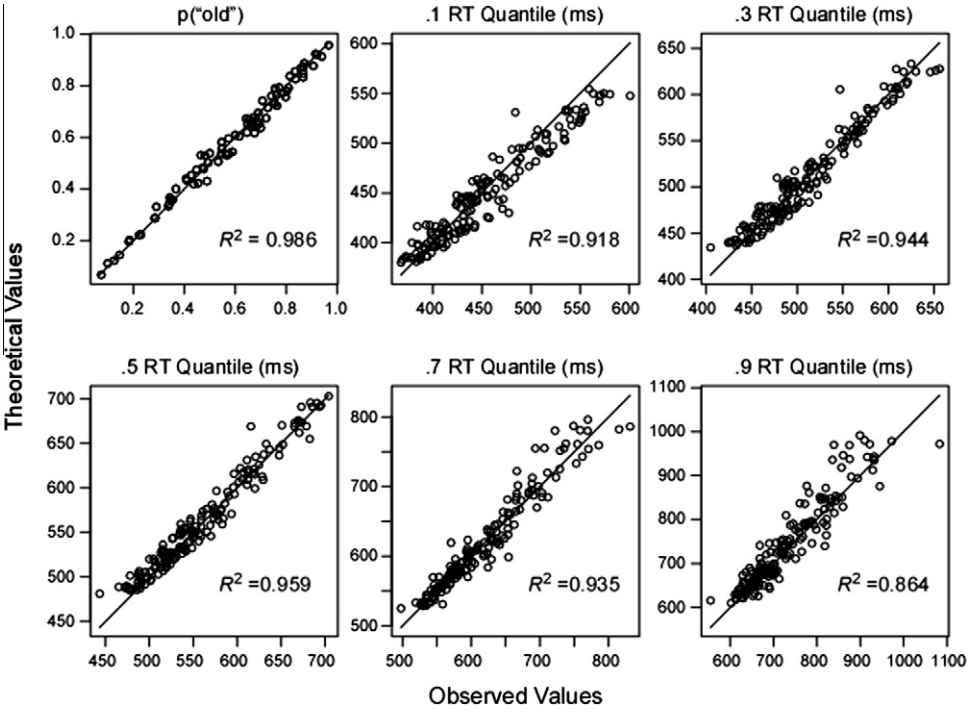

Fig. 5. Observed versus theoretical values for the diffusion model. The first panel shows the proportion of "old" responses for each of the 80 conditions. The next 5 panels show the .1–.9 quantiles of the response time distributions for both "old" and "new" responses, so each plot has 160 points (80 conditions × 2 responses). The numbers in each plot are the proportion of variance in the data accounted for by the model predictions.

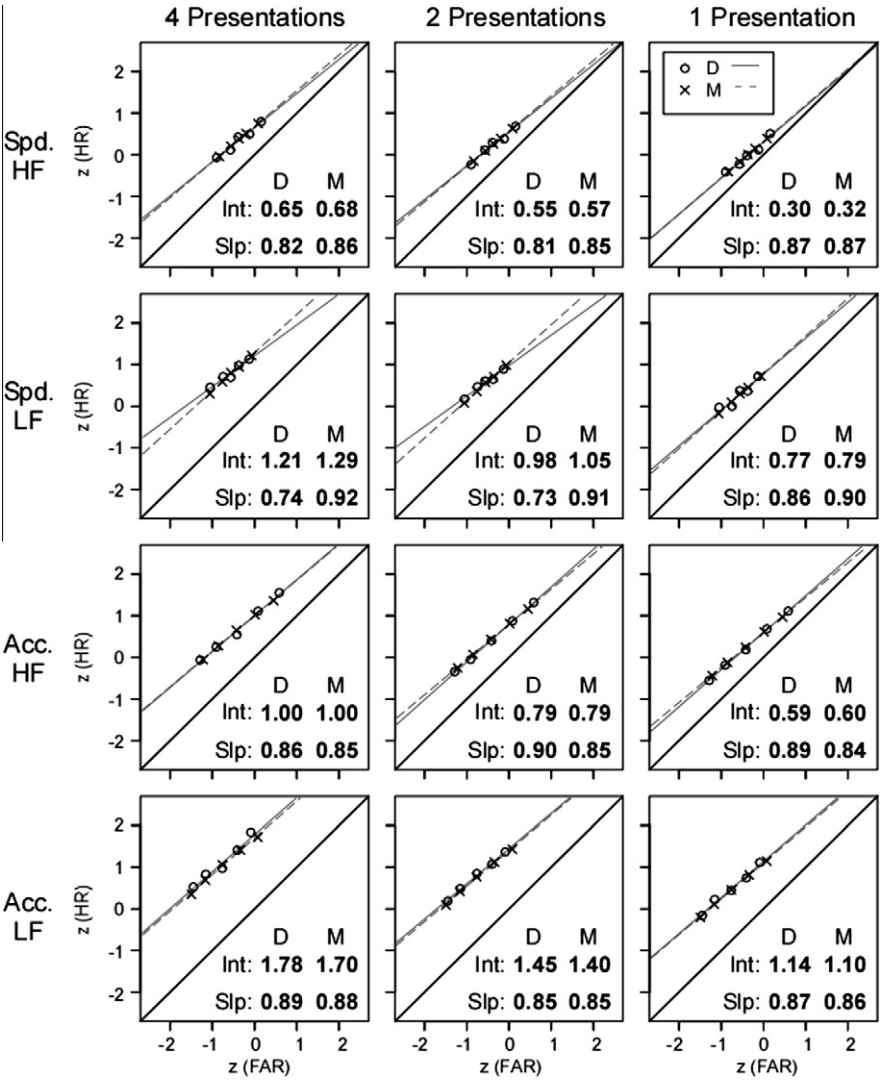

Fig. 6. Fit of the unequal-variance diffusion model to the zROC data. The rows show the results for high-frequency (HF) and low-frequency (LF) words from the speed-emphasis (Spd) and accuracy-emphasis (Acc) sessions. The columns show results for targets studied four times, twice, or once. On all plots, the circles are the observed values and the *x*'s are the model fit. The lines were produced by fitting a UVSD model to both the data (solid lines) and diffusion model predictions (dashed lines). The numbers within each plot are the zROC intercept (''Int.'') and slope (''Slp.'') from the UVSD fits to both the observed data (''D'') and the model predictions (''M''). z(FAR) and z(HR) indicate the z-transformed false alarm rate and hit rate, respectively.

individual averages for 53 of the 66 parameters. For the 13 remaining parameters, none had a deviation larger than 25%.

Fig. 5 displays the group fits for response proportion and RT. Each scatterplot shows the theoretical values plotted against the observed values across the 80 conditions of the experiment. The RT quantile plots show results for both ''old'' and ''new'' responses, so they each have a total of 160 points. The diagonal lines show where the points would fall if the model predictions perfectly matched the data, and each panel shows the proportion of variance in the data accounted for by the predictions. The un-

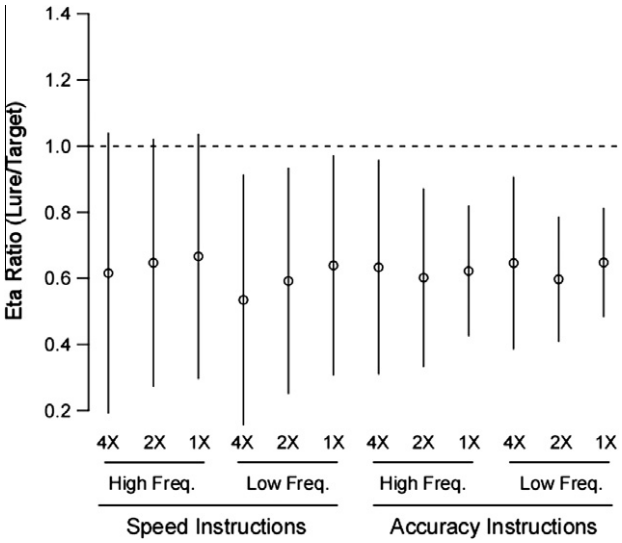

Fig. 7. Average η ratios (lure η/target η) from the individual subject fits of the diffusion model with 95% confidence intervals. "4X" indicates targets with four study trials, etc.

equal-variance diffusion model provided a good fit to the response proportions and all of the RT quantiles. Predictions accounted for over 90% of variance for all aspects of the data except the .9 quantiles.

The scatterplots in Fig. 5 show that the model generally followed the data across all 80 conditions, but a more important test is whether the model correctly accommodated the effects produced by the experimental variables. As noted, the variables that produced the biggest RT effects were target proportion and speed versus accuracy instructions. Fig. 4 shows the observed and predicted RT distributions for these variables averaged over the strength conditions. The model closely matched the shapes of the distributions and correctly accommodated the effects of both variables. Specifically, the model matched the slower RTs for speed than for accuracy instructions as well as the faster "old" and slower "new" responses produced by increasing the proportion of targets on the test.

We explored model fits for response proportion by evaluating the zROC data. Fig. 6 shows the fit to the zROC data for the unequal-variance diffusion model, along with the intercepts and slopes of the zROC functions for the data (D) and the model (M). Again, the model impressively matched the observed effects. For both the data and model results, target proportion produced larger response biases for accuracy sessions than for the speed sessions. As a result, the points are more spread out along the zROC function for accuracy sessions. For both model and data, the intercepts show that memory performance improved for low versus high frequency words, for accuracy versus speed sessions, and for more versus fewer learning trials. Finally, for both model and data, the zROC slopes were all below 1 and did not vary much across conditions relative to the standard error in the empirical slope estimates (see Table 3). Only two functions showed deviations between the observed and predicted slopes that were larger than the estimation error: Low frequency words studied 2 and 4 times in the speed sessions both had misses of .18, whereas the standard error was about .10 for both. Although these slopes were missed, we note that the individual data points do not show large misses.

3.1.2. Tests for unequal variance

The η parameters showed the pattern predicted by the unequal-variance account of zROC slope. For the group data, all of the η parameters for targets were higher than the η parameters for the corresponding lure items. Parameters from the individual subject fits also supported an unequal-variance model. Fig. 7 shows the average η ratios (lure η/target η) in each condition for the individual fits, and the lines show 95% confidence intervals around the means. Target evidence was more variable than lure evidence in every condition, leading to ratios close to .6. Moreover, the hypothesis of equal var-

iance was rejected at the 5% level for 9 of the 12 conditions (i.e., only 3 of the confidence intervals include a ratio of 1).

To further explore the role of unequal-variance in matching zROC slopes, we tested a model in which η was constrained to be equal across targets and lures. This model produced a χ^2 value of 2646 compared to 2418 for the unconstrained model, and *BIC* preferred the unequal-variance model (7886) over the equal-variance model (7981). More importantly, the equal-variance model clearly failed to match the empirical zROC slopes. Fig. 8 shows the zROC fit for the equal-variance model,

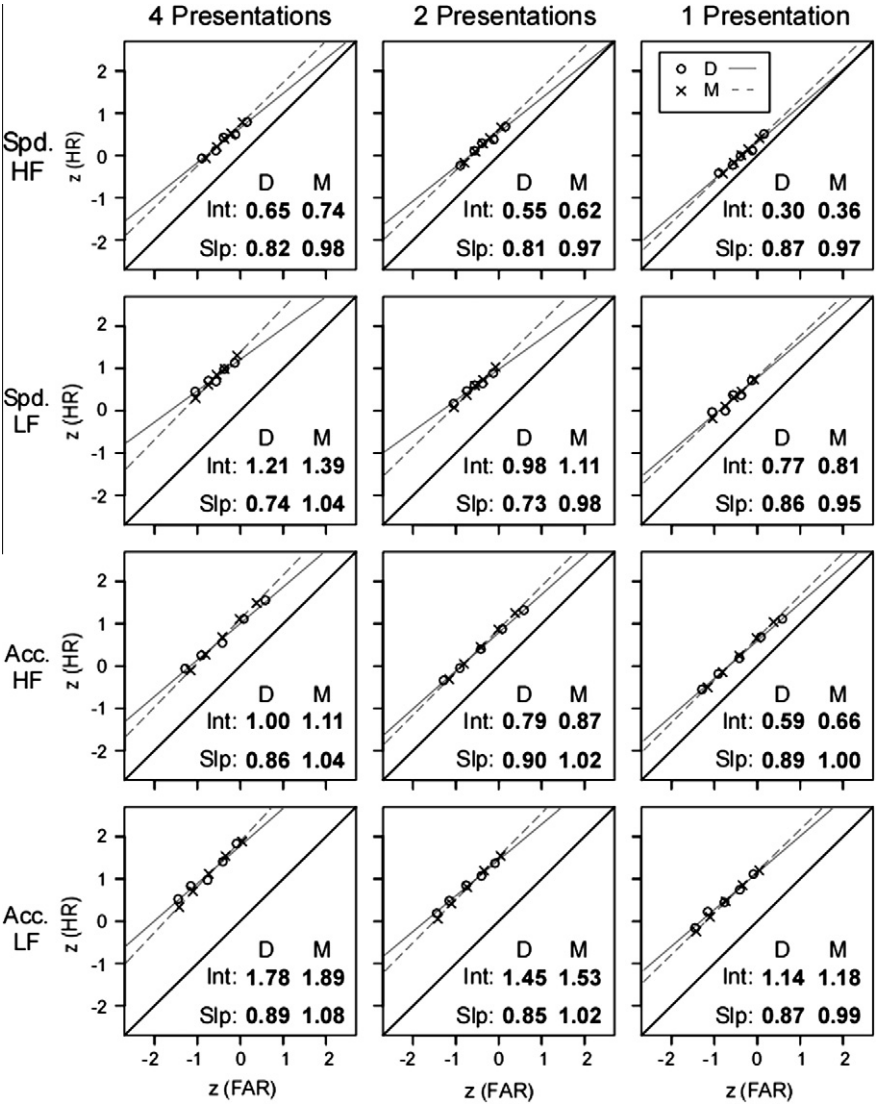

Fig. 8. Fit of the equal-variance diffusion model to the zROC data. The rows show the results for high-frequency (HF) and low-frequency (LF) words from the speed-emphasis (Spd) and accuracy-emphasis (Acc) sessions. The columns show results for targets studied four times, twice, or once. On all plots, the circles are the observed values and the *x*'s are the model fit. The lines were produced by fitting a UVSD model to both the data (solid lines) and diffusion model predictions (dashed lines). The numbers within each plot are the zROC intercept ("Int.") and slope ("Slp.") from the UVSD fits to both the observed data ("D") and the model predictions ("M"). z(FAR) and z(HR) indicate the z-transformed false alarm rate and hit rate, respectively.

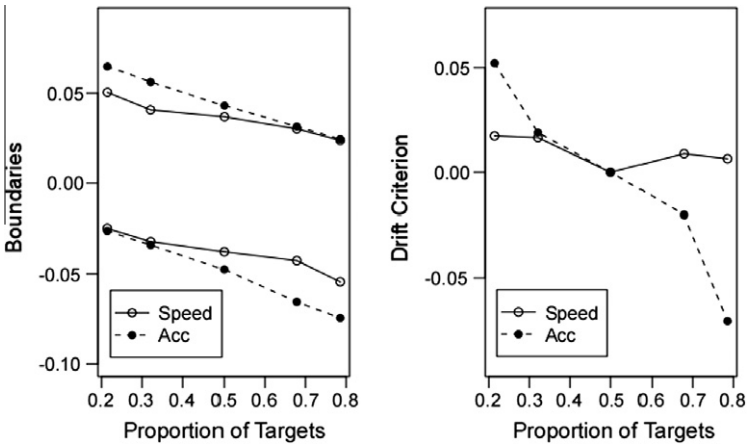

Fig. 9. Boundary and drift criterion results for the unequal-variance diffusion model across the target proportion conditions. "Acc" indicates the accuracy-emphasis instructions.

and large slope misses are apparent for nearly all of the conditions. The predicted zROC functions all had slopes close to 1, in contrast to both the current data and an extensive literature on recognition zROCs (Egan, 1958; Glanzer et al., 1999; Ratcliff et al., 1992, 1994; Wixted, 2007; Yonelinas & Parks, 2007). Therefore, our diffusion model results support the unequal-variance account: the model provided a good fit to the data when target evidence was more variable than lure evidence, but could not fit the data with equal target and lure variability.

3.1.3. Results for other parameters

The left panel of Fig. 9 shows the response boundary results from the unequal-variance diffusion model fit to the group data. As expected, the "old" boundary approached the starting point (0) as the proportion of targets on the test increased, and the "new" boundary moved away. In this way, the model explains why "old" responses were made more quickly and more frequently as the proportion of targets on the test increased, whereas the opposite pattern held for "new" responses. Boundary separation was wider with accuracy than with speed instructions, allowing the model to accommodate the slower responding in accuracy sessions and contributing to the lower error rate for accuracy sessions. The average non-decision time was about 50 ms slower in accuracy (482) than speed (431) sessions, which also contributed to the RT difference between the two (for direct evidence that speed/accuracy instructions affect non-decision processing, see Rinkenauer, Osman, Ulrich, Müller-Gethmann, & Mattes, 2004).

The right panel of Fig. 9 shows the drift criterion parameters. With accuracy instructions, the drift criterion became more liberal as target proportion increased; that is, with many targets on the test participants were willing to accept a lower memory match value as evidence for an "old" response. With speed instructions, the drift criterion showed little change based on target proportion. *BIC* values preferred a model with free drift criteria (7886) over a model with drift criteria fixed across the target proportion conditions (7963). Previous fits to target-proportion manipulations sometimes suggest that this variable only affects response boundaries (Ratcliff & Smith, 2004; Wagenmakers et al., 2008) and sometimes suggest that it affects both boundaries and the drift criterion (Ratcliff, 1985; Ratcliff et al., 1999). These alternative outcomes can even vary from one participant to the next within a single experiment (Criss, 2010). Thus, the current results are consistent with the picture offered by previous literature: proportion manipulations always affect boundaries and sometimes affect the drift criterion as well.

Fig. 10 shows the average drift rates from the unequal-variance fit to the group data. As expected, lure drift rates were below zero and target drift rates were above (except for high frequency words

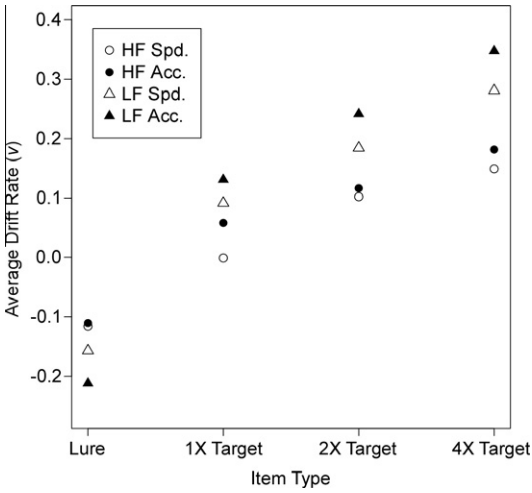

Fig. 10. Average drift rates for the unequal variance diffusion model across the instruction, word frequency, and study presentation variables. The displayed drift rates are from the 50% target condition. The drift rates in the other target proportion conditions can be derived by subtracting the relevant drift criterion parameter from the drift rate in the 50% condition. "Spd." = speed emphasis sessions; "Acc." = accuracy emphasis sessions; "1X Target" = targets studied once, etc.

studied once in the speed sessions). Low frequency words had higher target drift rates and lower lure drift rates than high frequency words. Target drift rates also increased with extra presentations on the study list. For low frequency words, the drift rates were consistently higher in absolute value in accuracy sessions than in speed sessions. The high frequency lures had consistent drift rates across the speed and accuracy sessions, but the high frequency targets did show some evidence of an increase with accuracy emphasis.

The results suggest that pushing participants to respond very quickly in the speed sessions may have impaired their ability to construct effective memory cues. Indeed, some recent models assume that memory probes have few active features early in a test trial, with additional features filling in over time (Diller, Nobel, & Shiffrin, 2001; Malmberg, 2008). Such a mechanism would produce more complete memory probes in the accuracy sessions than in the speed sessions, and the more complete probes would yield better evidence from memory. Exploring the possibility that time pressure affects memory probes in addition to speed/accuracy criteria is an interesting avenue for future research.

3.2. Dual-process account

The zROC data provide no support for the dual-process prediction that slopes should be closer to 1 with speed versus accuracy sessions. Indeed, slopes were numerically lower in the speed sessions (see Table 3). We explored this issue further by directly fitting the DPSD model to our response proportion data to see how the familiarity and recollection parameters changed across instructions. To be consistent with the timing of recollection versus familiarity, the model results should show that speed stress impairs the former but has a smaller effect on the latter.

We fit the model across all of the conditions to both the averaged data and to data from each individual participant. The model had 38 parameters to fit 80 freely varying response frequencies (removing the RT data results in a much smaller dataset than the one fit by the diffusion model). Appendix A lists the model parameters and gives the prediction equations. Here we will simply note how the parameters were constrained across conditions. The probability of recollecting a target (R) was allowed to vary based on word frequency, number of learning trials, and instructions, but did not change based on the proportion of targets on the test. Similarly, the means of the familiarity distributions (μ) were allowed to vary across word frequency, item type, and instructions, but not across target proportion. The response criterion for familiarity-based responding (λ) changed based on the proportion of

Table 4
Parameter values from the DPSD model.

Frequency and number of presentations	Parameter and instructions			
	R		μ	
	Speed	Accuracy	Speed	Accuracy
High				
4	.22	.22	0.45	0.82
2	.15	.14	0.43	0.66
1	.12	.12	0.18	0.48
New	–	–	0*	0*
Low				
4	.36	.36	0.85	1.46
2	.31	.31	0.60	1.06
1	.18	.18	0.43	0.76
New	–	–	−0.30	−0.65

Note: Parameters marked with an asterisk were fixed; *R* – probability of recollection; μ – mean of the familiarity distribution. The standard deviation of the familiarity distributions for low frequency words was 1.16 in speed sessions and 1.36 in accuracy sessions (high frequency fixed at 1 for both). Speed session response criteria: −.16, .16, .35, .57, .90. Accuracy session response criteria: −.57, −.09, .42, .90, 1.30.

targets on the test, and we also allowed different criteria for the speed and accuracy sessions. The criteria were fixed across word frequency and item type.

The DPSD model generally has no free parameters for the standard deviations of the familiarity distributions, but we found it necessary to have different variability parameters (σ) for high and low frequency words to adequately fit the data. The target proportion manipulation had a smaller effect on false alarms for low-frequency lures than high-frequency lures, and this pattern could not be accommodated by a model with equal variance across word frequency. Indeed, *BIC* statistics showed that the model with variability free to change across word frequency (996) was preferred to the constrained model (1070). Critically, the variability parameters were still constrained to be equal across targets and lures, so recollection was still the only process that could produce slopes below one within a frequency class (e.g., when high-frequency targets were contrasted with high-frequency lures). The recollection parameter cannot be estimated without this constraint, because unequal variance provides a redundant mechanism for matching zROC slopes. Although recollection and unequal variance technically predict different zROC shapes (with the former predicting slightly u-shaped as opposed to linear functions), this difference is often too subtle to tease apart the processes in fits to data.

The DPSD model produced a G^2 of 45.34, and Table 4 shows the best fitting parameter values. Compared to high-frequency targets, low-frequency targets had higher recollection and familiarity parameters. Similarly, increasing the number of learning trials increased both the recollection and familiarity parameters for targets. Word frequency also affected lure familiarity estimates, with low-frequency lures less familiar than high-frequency lures. The result of primary interest was the effect of speed versus accuracy instructions on familiarity and recollection parameters. Target familiarity estimates increased going from speed to accuracy sessions, and familiarity estimates for low-frequency lures decreased. Thus, the results suggest that familiarity better discriminated targets from lures when participants allowed themselves extra decision time. When the familiarity parameters were constrained to be equal across the speed and accuracy sessions, the G^2 value nearly tripled to 116.71, demonstrating that the model could not fit the data without positing changes in familiarity. In contrast, recollection estimates changed little across speed and accuracy sessions. Indeed, constraining the recollection parameters to be equal across the instruction variable produced a G^2 of 45.40, nearly identical to the fit with recollection free to vary (45.34). Moreover, BIC preferred the equal-recollection model (930) over the unconstrained model (996). Thus, the results contradict the dual-process prediction that time pressure should impair recollection with relatively little impact on familiarity.

To ensure that the conclusion of no change in recollection across the speed and accuracy sessions was not based on distortions due to averaging, we also performed ANOVAs on the familiarity and *R*

parameters from the individual-participant fits. Familiarity parameters were converted to d' scores to measure how well this process discriminated targets from lures [d' = (target μ − lure μ)/σ for each condition]. The individual fits confirmed the conclusions from the group analysis. Recollection showed practically no change between accuracy (.27) and speed (.26) sessions, $F(1,3) = .05$, *ns*, $MSE = .013$. Recollection changed significantly based on word frequency, $F(1,3) = 68.52$, $p < .05$, $MSE = .004$, and number of learning trials, $F(2,6) = 17.72$, $p < .05$, $MSE = .007$. In contrast to recollection, familiarity d' did change significantly from accuracy (1.02) to speed (.62) sessions, $F(1,3) = 25.86$, $p < .05$, $MSE = .074$. Familiarity also varied based on word frequency, $F(1,3) = 13.92$, $p < .05$, $MSE = .135$, and degree of learning, $F(2,6) = 50.96$, $p < .05$, $MSE = .008$.[4]

The problems for the dual-process account can also be seen by evaluating the speed-instruction data in isolation. The recollection estimates from the speed sessions were surprisingly high given the pace of responding. For example, based on the exponential growth function for recollection in McElree et al.'s (1999) Experiment 1, recollection became available around 608 ms, reached a third of its asymptotic value around 654 ms, and reached two thirds of the asymptotic value by around 746 ms. For comparison, consider the low-frequency targets studied four times in our speed sessions. The median response time for "old" responses to these items was 509 ms, with 90% of the responses made within 625 ms. That is, well over half of the responses were made before the onset of recollection as estimated by McElree et al., and over 90% of responses were made before the point that recollection reached 33% of its asymptotic level. The DPSD model produced an *R* estimate of .36 for this condition, suggesting that over a third of the responses were based on recollection. This value is difficult to reconcile with the RT data. Even if only the slowest responses had time to be influenced by recollection, *every* response made after 537 ms would have to have been recollection-based to equal the model's *R* estimate.

3.3. RTCON

The RTCON model is similar to other sequential sampling approaches, such as the dual-diffusion model (Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007; Ratcliff & Smith, 2004; Smith, 2000) and the Leaky-Competing Accumulator (LCA) model (Usher & McClelland, 2001). RTCON was developed to accommodate RT distributions from a 6-choice confidence judgment task, and applying the model highlighted the need for significant changes in the interpretation of confidence-rating ROCs (Ratcliff & Starns, 2009). Ratcliff and Starns were not able to test RTCON against alternative RT models, because RTCON is currently the only RT model to be applied to one-shot "definitely new" to "definitely old" confidence ratings (although other RT approaches have been extended to confidence responses following an initial two-choice decision; Pleskac & Busemeyer, 2010; Van Zandt, 2000). We adapted RTCON to a two-choice procedure in the current work, introducing the possibility of comparative fitting with a well-established model of two-choice decision making. Here we present the two-choice version of RTCON and note the changes from the model reported by Ratcliff and Starns.[5]

The model assumes that the evidence driving a decision (in this case memory evidence) is normally distributed across trials, as in signal detection theory. The bottom panel of Fig. 11 displays these between-trial distributions, one for targets and one for lures, each with its own mean ($\mu_{BETWEEN}$) and standard deviation ($\sigma_{BETWEEN}$). On each trial, an evidence value is sampled from the appropriate between-trial distribution, and a within-trial distribution with a standard deviation of 1 is centered on the sampled value. In the original model, 5 confidence criteria segmented the within-trial distribution into regions associated with each confidence response. For the two-choice model, a single confidence criterion establishes regions for "new" and "old" responses. The position of the confidence criterion on each trial is a random draw from a Gaussian distribution with mean *c* and standard deviation σ_C.

[4] We also ran these analyses using parameters from model fits in which the standard deviation in familiarity was fixed across all item types (high and low frequency targets and lures). Parameters from these fits also showed that the instruction variable influenced familiarity [Speed d' = .48, Accuracy d' = .98, $F(1,3) = 102.53$, $p < .05$, $MSE = .029$] but not recollection [Speed R = .31, Accuracy R = .26, $F(1,3) = .94$, $p = .41$, $MSE = .035$]. Parameters from the equal-SD fits actually showed nominally more recollection in speed sessions than accuracy sessions.

[5] Although no confidence judgments were made in this experiment, we continued to use the name "RTCON" for continuity with our past work and to highlight that nothing has changed in the model except for the number of response alternatives.

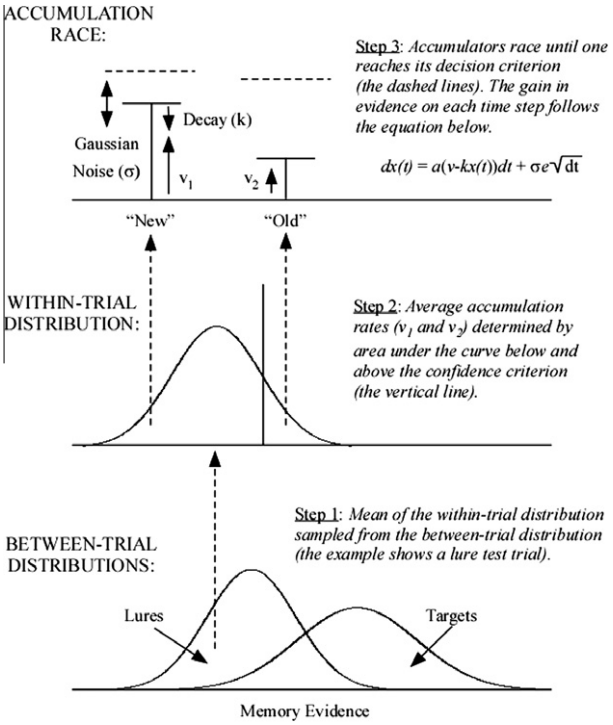

Fig. 11. Procedure for simulating a trial of the two-choice RTCON model. The bottom panel shows between-trial evidence distributions for both lures and targets, with higher variability in target evidence. The middle panel shows the within-trial evidence distribution on a single lure test trial, and the top panel depicts the accumulation race. In the equation governing the position of the counters across time, $x(t)$ is the position of the process at time step t, $dx(t)$ is the change in evidence at time step t, a is the scaling factor (fixed at .1), v is the average accumulation rate, k is the decay term (the proportion of the accumulator's current activation that is lost on each time step), dt is the length of the time step, σ is the standard deviation in accumulation noise, and e is a random normal variable.

The original model had six accumulators for the six confidence levels. The two-choice model has just two accumulators for "new" and "old" responses (top panel of Fig. 11). The proportion of the within-trial distribution below and above the confidence criterion determines the average drift rate (v) for the "new" and "old" accumulators, respectively. The accumulators race with moment-to-moment Gaussian variation around the average drift rates (with a standard deviation of σ). The activation of the accumulators is subject to decay; that is, each accumulator loses a proportion k of its activation on every time step. Each accumulator has a decision criterion (d_{OLD} and d_{NEW}) that varies across trials over a uniform distribution with range s_D. When one of the accumulators reaches its decision criterion, the corresponding response is made.

Analytical solutions of the model are not available because the activation of the accumulators is truncated at zero, creating a non-linear process (see Usher & McClelland, 2001). Predictions from the model are derived by Monte Carlo simulation, and we ran 20,000 simulated trials of the accumulation race to define the predictions for each condition. For more details on the fitting procedure, see Appendix A.

For the full, 80-condition dataset, we could not find fits for RTCON that were anywhere near the quality of the diffusion model fits. The lowest χ^2 we were able to find for RTCON was 5533 compared to 2418 for the diffusion model. However, we were concerned that the difference in fit might reflect difficulties in finding the optimal parameter values for RTCON. This model must be simulated, which introduces error in the predicted values from one model run to the next. Specifically, each time the fitting program evaluates the predictions of RTCON, the model must go through 80 simulation runs of 20,000 simulated trials each. Although having 20,000 trials ensures a low degree of variability between runs, this variability builds up across the 80 conditions, making it difficult for the fitting algorithm to

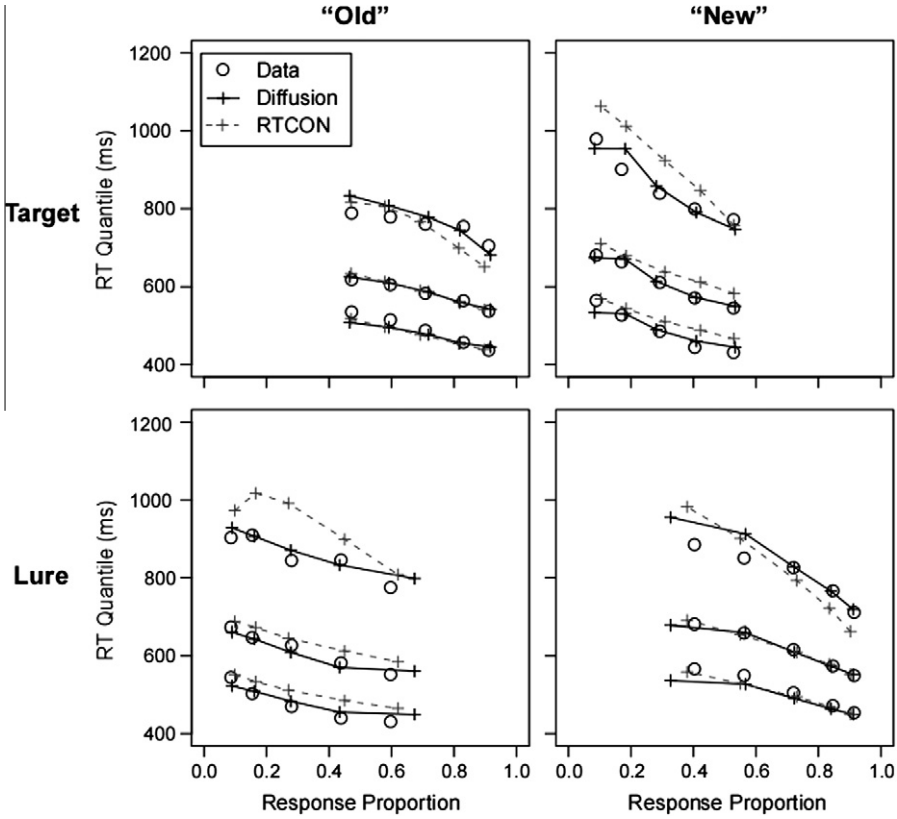

Fig. 12. Fit of RTCON and the diffusion model to the 10-condition dataset. Each panel shows the .1, .5, and .9 quantiles plotted on the proportion of responses. The .3 and .7 quantiles were included in the model fits, but they are not displayed. The five columns of data points in each panel are from the five target proportion conditions. For the "old" responses (left column), the .21-targets condition is the set of scores furthest to the left and the .79-targets condition is furthest to the right. This ordering is reversed for the "new" responses (right column). The lines show each model's quantile predictions, and the "+" symbols mark the location of the model's probability predictions.

determine which parameter changes are truly improving the fit. Increasing the number of runs to decrease the variability quickly becomes computationally infeasible with datasets of this size. Moreover, the large dataset forced us to use 10 ms time steps instead of the 1 ms time steps used by Ratcliff and Starns (2009), meaning that the simulations do not as closely approximate a continuous model.

To make sure that the model selection results were not unduly influenced by simulation error, we compared RTCON to the diffusion model on a much smaller dataset. We used only the data from the accuracy sessions, and we collapsed over the word frequency and number of learning trials variables. This resulted in a dataset with 10 conditions – targets and lures across the five target proportion conditions – and 110 degrees of freedom. The diffusion model applied to this dataset had 22 parameters, and the RTCON model had 24 (see Appendix A for more details). Previous work with RTCON demonstrates that the model is able to recover parameters for datasets of this size (Ratcliff & Starns, 2009). The smaller dataset also allowed us to simulate RTCON with 1 ms as opposed to 10 ms time steps. Therefore, any differences in model fit for this smaller dataset should reflect true differences in the models themselves and not the effectiveness of the fitting procedure.

Fig. 12 shows the fits of RTCON and the diffusion model to the 10 condition dataset (only the .1, .5, and .9 quantiles are shown to avoid clutter, but the .3 and .7 quantiles were also fit). Both models provided a good fit to the response proportions; that is, the "+" symbols on the model functions closely line up with the data points. RTCON provided a slightly better fit to the accuracy data, in that the diffusion

model showed a fairly large miss for lure items in the condition with the highest proportion of targets on the test (the model predicted more "old" and fewer "new" responses than in the data). However, the diffusion model much more closely matched the RT quantiles. A general problem for RTCON was that the model predicted too much change in the spread of the RT distributions going from the fast to the slow conditions. For the fastest sets of quantiles, RTCON tended to predict .9 quantiles that were too low or .1 quantiles that were too high; that is, the predicted distributions were more compact than the data. In the slower conditions, RTCON consistently predicted .9 quantiles that were much too high; that is, the predicted distributions were more spread than the empirical distributions. The diffusion model also tended to predict too much spread in the distributions for slow conditions, but not nearly to the same extent as RTCON. Another big miss for RTCON was that the model consistently predicted slower error RTs than observed. These differences in the ability to account for the RT quantiles led to a much better fit for the diffusion model, which had a χ^2 of 823 compared to 1586 for RTCON.

Clearly, RTCON did not perform up to the standards of the diffusion model, even for a limited dataset. Given that the models are relatively similar in structure, it is useful to think about differences between the models that might explain their differential success. One difference is how the within-trial variation in drift rate is implemented. In the diffusion model, there is one accumulation process tracking the difference between evidence for one response and evidence for the other. As a result, the noise in accumulated evidence for the two responses is perfectly correlated: a step toward the "old" boundary is an equal-sized step away from the "new" boundary. In RTCON, separate accumulators have their own independent noise in accumulation rates; for example, in a particular cycle of the race, both the "old" and the "new" accumulator could have a particularly large gain in activation.

This difference in structure leads to an important difference in predicted RTs. By accumulating differences, the diffusion model naturally produces the appropriate positive skew in RT distributions without a decay term; in fact, when a decay term is added it hovers near zero in fits (Ratcliff & Smith, 2004). In contrast, RTCON produces distributions that are far too symmetrical unless the decay term is added to produce the appropriate skew (Ratcliff & Starns, 2009; Usher & McClelland, 2001). In the Ratcliff and Starns fits, adding decay was an acceptable solution for modeling confidence ratings made under time pressure. However, the RT distributions from their experiments showed little change in location or spread across ratings. The current dataset suggests that simply adding decay is not an acceptable solution when the observed distributions have a range of locations, as the decay produces inappropriately large differences in spread from the fast to the slow conditions. Either the positive skew in RT distributions reflects a process other than decay, or decay must be implemented in an alternative model architecture.

4. General discussion

We tested the unequal-variance and dual-process accounts of zROC slopes with a two-choice recognition memory task. The two accounts have proven difficult to distinguish when implemented in signal-detection models to fit only zROC data (Wixted, 2007; Yonelinas & Parks, 2007). We tested the unequal-variance account by fitting zROCs and RT distributions with the diffusion model. The model produced a good match to the data with unequal variability in target and lure evidence, but produced large misses to the zROC slopes in an equal-variance version. We tested the dual-process account by evaluating zROC slopes from decisions made under time pressure. Violating the predictions of this account, zROC slopes were not closer to one in the speed-emphasis sessions than in the accuracy-emphasis sessions. Moreover, the DPSD model could only fit the data by proposing that speed pressure affected familiarity with no effect on recollection, which is inconsistent with the time course of the two processes (Dosher, 1984; Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree et al., 1999; Rotello & Heit, 2000).

4.1. RT data and model constraint

Our results show that RT data dramatically increase model constraint. For example, the diffusion model had less than twice as many parameters as the DPSD model (66 versus 38) to fit a dataset with

more than 10 times as many freely varying response frequencies (880 versus 80). Much more important than the numbers, though, is the fact that almost all the diffusion model parameters are controlled by multiple aspects of the data. Changing the response boundaries, for example, affects response proportions as well as the location and spread of each RT distribution. With these extra constraints, RT modeling alleviates the current problem of model identifiability that characterizes zROC research (Wixted, 2007; Yonelinas & Parks, 2007).

4.2. Slopes in the diffusion model

Previous fits of the diffusion model to recognition data used the same η values for targets and lures (e.g., Ratcliff & Smith, 2004; Ratcliff, Thapar, & McKoon, 2004). Most previous applications to recognition are from two-choice tasks with only a single hit and false alarm rate. Without a bias manipulation, the value of η only influences relatively subtle aspects of the RT distributions; thus, the assumption of equal versus unequal η's is not critical for fitting data or for the estimation of the other model parameters. By fitting the model to zROC data across our target proportion manipulation, we demonstrated the need for unequal variance in drift distributions for recognition memory. Curiously, Ratcliff and Smith (2004) report a recognition memory experiment with a target proportion manipulation similar to our own, and they were able fit the data with an equal-variance version of the model. Ratcliff and Smith did not plot zROC functions, but a re-analysis of their data showed that the zROC slopes were close to one in all of the conditions. This is an unusual finding in recognition memory, but one that illuminates why unequal η's were not needed. In the current project, we observed the more standard finding of zROC slopes less than 1, and the model accommodated this finding by proposing greater variability in target versus lure drift rates.

The recognition experiment fit by Ratcliff and Smith (2004) used different instructions than the current studies. Specifically, in the current experiments we asked participants to use the target proportion to help guide their decisions, whereas the Ratcliff and Smith participants were asked simply to be accurate without following target probability. The Ratcliff and Smith data showed less influence of target proportion on responding compared to the current accuracy sessions, although responding did change enough to define zROC functions. We cannot be sure if these instructional differences played a role in the slope results, but slopes should not be seen as reflecting basic properties of memory if they are indeed sensitive to instructional manipulations.

4.3. What can zROC's tell us?

Both signal detection and sequential sampling models often assume that evidence distributions are Gaussian in form, but it is important to realize that it would be difficult – perhaps impossible – to justify this assumption empirically (Rouder, Pratte, & Morey, 2010). For example, deviating from normal variation in drift rates in the diffusion model would have minimal effects on the predicted RT distributions, given that the between-trial variability would be swamped by considerable within-trial variability. The assumed distributions for variation in starting point and non-decision time also have little influence on predictions. ROC data place some constraints on distributions of memory evidence, but these constraints are also relatively lax. ROCs show how responding changes for different types of items (say, targets versus lures) across multiple levels of bias. If the evidence distributions for the two item types differ in some way, then this will lead to differences in how responding changes across the bias levels. For example, if target evidence is more variable than lure evidence, then the hit rate will tend to change more gradually across bias compared to the false alarm rate. However, ROCs provide no distributional information outside of the range of bias achieved in an experiment, and they provide only crude distributional information within this range.

Given that specific distributional forms cannot be justified with data, one should avoid conclusions that rely heavily on distributional assumptions. For example, our estimates of the relative variability of target and lure drift distributions could change substantially if other distributional forms were assumed, so confidence in the precise values is unwarranted. However, the data do have sufficient constraint to rule out a model in which target and lure evidence values are distributed identically, given that responding for targets is less affected by the bias shifts than responding for lures. Using Gaussian

distributions with unequal variability is a convenient way to accommodate this, but the success of such a model does not imply that the distribution shape has been correctly specified. In our view, the fact that zROC slopes are less than 1.0 properly supports a rather mundane conclusion: target evidence is more spread out than lure evidence over the range of bias.

An alternative view makes a bolder claim: zROC slopes reflect different processes subsumed by separate neurophysiological systems (Yonelinas & Parks, 2007). Estimates of these processes from ROC data are just as dependent on distributional assumptions as estimates of specific standard deviation ratios. Indeed, recollection estimates from the dual process approach rely not only on the assumption of Gaussian distributions, but also on the assumption that the target and lure variability are equal. Thus, the specific values of these estimates can never be interpreted with confidence without some independent confirmation of distributional form (and it is difficult to imagine how that confirmation would be achieved).

Of course, zROC data might provide evidence for separate processes even if it cannot support the exact estimation of these processes, just as we have claimed that it provides evidence for more variable target evidence even though it cannot support exact estimates of this variability. However, the current results reveal no link between zROC slopes below 1.0 and the recollection process. Recollection is disrupted by time pressure (Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree et al., 1999; Rotello & Heit, 2000), but our slopes showed no evidence of this. With a confidence-rating task, Ratcliff and Starns (2009) also found no change in zROC slope between accuracy- and speed-emphasis sessions. These results build on research demonstrating that recollection estimates produced in zROC fits do not reflect the same recollection that forms the basis of associative judgments such as pair recognition or list discrimination. Starns and Ratcliff (2008) had the same participants complete item recognition with confidence ratings and a pair recognition test (i.e., "were these words studied *together*?"). They applied the DPSD model to the item recognition zROCs to produce estimates of recollection (R) and familiarity (d'). The R and d' parameters were equally predictive of a participant's performance on the pair recognition test. If the recollection estimate from the zROC data were measuring the same type of information as required for pair discrimination, then this parameter should have been more predictive than the familiarity parameter. Results like these suggest that the various phenomena that have been explained by appealing to recollection are not actually produced by the same process (see Malmberg, 2008, for similar arguments).

4.4. RT modeling of confidence rating data

A critical – and elusive – goal of zROC modeling is a model that can accommodate accuracy data and RT distributions from both two-choice and confidence rating tasks. Although RTCON has been successfully fit to confidence rating data (Ratcliff & Starns, 2009), it did not perform well for our two-choice dataset. RTCON missed key aspects of the RT distributions and produced a χ^2 that was nearly twice as large as the diffusion model. As mentioned, a critical difference between RTCON and the diffusion model could be uncorrelated (RTCON) versus correlated (diffusion) noise in evidence accumulation, together with the need to add a decay term for the uncorrelated noise model. The two-choice version of RTCON could be revised to incorporate correlated noise for the two counters; that is, evidence producing an increase in the activation of one counter could trigger a corresponding decrease in activation for the other. Indeed, under this assumption, a two-accumulator model can be developed that is mathematically identical to the diffusion model (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ditterich, Mazurek, & Shadlen, 2003). The question would be whether such a model would also work for confidence rating data. We are currently exploring this strategy for updating the RTCON approach.

To date, RTCON is the only model to be applied to a confidence rating task with the goal of modeling response proportion and RT distributions for each level of the rating scale. However, other models have been developed for confidence judgments that are made after a two-choice decision (Pleskac & Busemeyer, 2010; Van Zandt, 2000). The Van Zandt model is an extension of the two-choice Poisson race model (Townsend & Ashby, 1983; Vickers, 1979). Unfortunately, even the two-choice version of this model produces RT distributions that are too symmetrical to match empirical distributions (Ratcliff & Smith, 2004). The Pleskac and Busemeyer model extends the diffusion model by assuming that the two-choice decision is followed by a fixed time period of additional evidence accumulation, with

the confidence judgment based on the position of the diffusion process after this period. Pleskac and Busemeyer suggest that this model could be adapted to a single rating-scale response by assuming that an implicit two-choice decision precedes the selection of a confidence level. Thus, it is possible that this model could be extended across all paradigms, but work will be needed to determine if the proposed two-stage process will be fast enough to match confidence ratings made under time pressure. The Pleskac and Busemeyer model is particularly promising, because it is an extension of an already well-validated two-choice model.

4.5. RT shifts and non-decision time

One interesting aspect of our RT results is that the target proportion variable produced shifts in the RT distributions; that is, the size of the proportion effect was similar for the leading edges (.1 quantiles) and tails (.9 quantiles) of the distributions. The only diffusion model parameter that produces pure shifts in the RT distributions is the duration of non-decision processes, T_{er}. Unfortunately, T_{er} does not affect the response proportion data, so this parameter cannot accommodate the effect of target proportion on bias to use the "old" response. The model parameters that can accommodate this bias are the response boundaries and the drift criterion, and changing either of these parameters has a larger effect on the .9 quantiles than on the .1 quantiles. Ratcliff and McKoon (2008) offered the general rule that the .9 quantiles should change about twice as much as the .1 quantiles across different boundary settings and about four times as much across different drift criterion settings. For the current data, the model predicted a larger effect than observed for the .9 quantiles and a smaller effect than observed for the .1 quantiles (see Fig. 4). The model might match the data more closely if T_{er} were allowed to vary across conditions, perhaps representing differences in responses preparedness (i.e., frequent responses can be made more quickly). However, the misses were fairly small, and they were limited to conditions with few observations and thus relatively poor estimation of the RT quantiles (the .21 condition for "old" responses and the .79 condition for "new" responses). For these reasons, we do not think that introducing additional complexity in the model is warranted at this point. If pure shifts prove to be a consistent empirical finding, then accommodating them will be an important goal of future modeling work.

4.6. Relationship between decision and memory models

Like signal detection models, sequential sampling models are models of the decision process and not models for the evidence underlying the decision; for example, the same model is applied whether the relevant evidence is perceptual (Ratcliff & Smith, 2010), lexical (Wagenmakers et al., 2008), or mnemonic (Ratcliff, Thapar, & McKoon, 2004). The diffusion model that we applied makes basic assumptions about the nature of memory evidence – e.g., that it can be expressed as a single continuous value (drift rate) – but does not address how this evidence is generated by the memory system. A variety of memory models attempt to make explanations at this level and can address questions such as why target evidence is more variable than lure evidence (Clark & Gronlund, 1996; Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) or how recollection and familiarity are produced (Elfman, Parks, & Yonelinas, 2008; Norman & O'Reilly, 2003; Reder et al., 2000).

Developing memory models is an important goal, but this goal cannot be achieved without developing appropriate models of decision making. That is, understanding decision processes is not only interesting in its own right, but also necessary before theorists can take the further step of testing process models. In a recognition memory task, we cannot directly observe the evidence driving decisions; we observe only the outcome of those decisions. Thus, the evidence generated by memory models must pass through a decision process before the models' predictions can be compared to data, and the details of this decision process can dramatically change how a model is assessed. Memory models have traditionally assumed a signal detection model for decision making (Clark & Gronlund, 1996), but a handful of models are beginning to adopt a sequential sampling approach (Malmberg, 2008; Nosofsky & Stanton, 2006). Our results suggest that switching to a sequential sampling decision process will be an important advance in the process model literature. More generally, our results exemplify how understanding decision making and understanding memory are complimentary goals.

5. Conclusion

The ROC literature has needed more powerful ways to discriminate alternative models, in particular the UVSD and DPSD models (Wixted, 2007; Yonelinas & Parks, 2007). Accommodating RT distributions places much more constraint on models, increasing the effectiveness of model selection. Our results showed that the dual-process account of zROC slope was not consistent with the RT data. In contrast, the unequal-variance account matched zROCs and RT distributions when implemented in the diffusion model. Extending the account in this way not only puts it to a stronger test than can be made with zROC data alone, but also has the potential to change key zROC interpretations (Ratcliff & Starns, 2009; Van Zandt, 2000). Therefore, future developments in RT modeling will be integral to theoretical progress in recognition memory.

Acknowledgements

Preparation of this article was supported by NIMH Grant R37-MH44640 and NIA Grant RO1-AG17083.

Appendix A

A.1. Formal definition and parameter list for each model

A.1.1. Unequal variance signal detection (UVSD) model

All model parameters are measured relative to the mean of the lure distribution in units of the standard deviation of the lure distribution, so memory evidence for lures has a mean of 0 and a standard deviation of 1. The model parameters include the position of the response criterion (λ), the mean of the target distribution (μ), and the standard deviation of the target distribution (σ). Model predictions are derived by the equations

$$p(\text{“old”}|\textit{lure}) = 1 - \Phi(\lambda)$$

$$p(\text{“old”}|\textit{target}) = 1 - \Phi\left(\frac{\lambda - \mu}{\sigma}\right)$$

where Φ is the cumulative distribution function of a standard normal. Put simply, the predicted hit rate is the proportion of the target distribution above the response criterion and the predicted false alarm rate is the proportion of the lure distribution above the criterion. This model was used to derive slope and intercept estimates for each zROC function (slope = $1/\sigma$; intercept = μ/σ). Each fit had one μ parameter, one σ parameter, and five λ parameters for the five target proportion conditions.

A.1.2. Dual-process signal detection (DPSD) model

For our fits, all familiarity parameters (the response criterion, the familiarity distribution means, and the familiarity distribution standard deviations) were measured relative to the mean of the high-frequency lure distribution in units of the standard deviation for high frequency words. Thus, the high frequency lure distribution was fixed at a mean of 0 and standard deviation of 1. The model assumes that responding for lures is always based on familiarity. Therefore, predictions for lure items are derived by the equation

$$p(\text{“old”}|\textit{lure}) = 1 - \Phi\left(\frac{\lambda - \mu}{\sigma}\right)$$

where Φ is the cumulative distribution function of a standard normal, λ is the position of the response criterion, μ is the mean of the familiarity distribution, and σ is the standard deviation of the familiarity distribution. We have used the more general equation (compared to the UVSD lure predictions) because the evidence distributions for low and high frequency lures could have different means and different standard deviations in our fits.

The model assumes that responses for targets can be based on recollection in addition to familiarity, and the two processes are independent. Specifically, with probability R the participant recollects studying the word and produces an "old" response. When recollection fails $(1 - R)$, decisions are based on the familiarity of the target. Therefore, the prediction equation for targets is

$$p(\text{"old"}|target) = R + (1 - R) \times \left[1 - \Phi\left(\frac{\lambda - \mu}{\sigma}\right)\right]$$

where R is the probability of recollection and the other symbols have the same meaning as in the lure equation (except that μ and σ now denote target distributions instead of lure distributions).

The DPSD model was fit across all conditions for both the group data and for each individual participant. Thus, there were 80 freely varying response frequencies; that is, two response frequencies ("old" and "new") for each of the 80 conditions, one of which was fixed because the responses had to add up to the total number of observations. Overall, the model had 38 parameters, including 14 means (μ) of the familiarity distributions [2 word frequencies $\times$ 4 item types $\times$ 2 instruction conditions with the high frequency lures fixed at zero for both speed and accuracy]; 2 standard deviations (σ) for the low-frequency evidence distributions [2 instruction conditions with the high-frequency σ fixed at 1 for both]; 12 recollection (R) parameters [2 word frequencies $\times$ 3 target strengths $\times$ 2 instruction conditions]; and 10 response criteria [5 probability conditions $\times$ 2 instruction conditions].

In fits to the group data, we tested for differences between the speed and accuracy sessions by constraining either the familiarity or recollection parameters to be equal across this variable (the results are reported in the main text). Constraining familiarity to be equal eliminated 8 free parameters to produce a 30 parameter model (the full model had 14 free μ parameters and 2 free σ parameters across speed and accuracy, whereas the constrained model had 7 μ parameters and 1 σ parameter). Constraining recollection to be equal eliminated 6 free parameters to produce a 32 parameter model (the full model had 12 R parameters across speed and accuracy versus 6 for the constrained model).

A.1.3. Diffusion model

The diffusion model was fit to both the group and individual-participant data using the χ^2 method (Ratcliff & Tuerlinckx, 2002). For each participant, we computed the .1, .3, .5, .7, and .9 RT quantiles for "old" and "new" responses within each condition. We averaged the quantile values across participants to derive the group quantiles (Ratcliff, 1979). For each response, we divided the frequency into six bins based on the five quantiles (i.e., .10 of the responses below the .1 quantile, .20 between the .1 and .3 quantiles, etc.). This resulted in 12 frequencies for each condition, six each for "old" and "new" responses. χ^2 values were computed based on the observed and predicted frequencies in the RT bins. A degree of freedom is lost for each condition because the frequencies have to sum to the total number of observations for the condition. Therefore, the dataset as a whole has 880 degrees of freedom (80 conditions $\times$ 11 degrees of freedom per condition). For further details and for the equations defining the predicted proportion in each RT bin, see Ratcliff and Tuerlinckx (2002, Appendix B).

Drift rates depend only on the relative positions of the drift distributions and the drift criterion, so one of them can be fixed at zero without loss of generality. In our analyses, we set the drift criterion from the equal-probability condition equal to zero and estimated parameters for the drift criteria in the other probability conditions. Because drift rates (v) are defined relative to the drift criterion, the average drift rates for the even-probability condition are equal to the drift distribution means, and the average drift rates for the other proportion conditions are equal to the drift distribution means minus the drift criterion for that condition.

The unequal-variance diffusion model had 66 free parameters, including 10 "old" (a_{OLD}) and 10 "new" (a_{NEW}) boundary parameters [5 probability conditions $\times$ 2 instruction conditions]; one range of stating point variability (s_Z); 16 means (μ) and 16 standard deviations (η) for the drift distributions [2 word frequencies $\times$ 4 item types $\times$ 2 instruction conditions]; 8 drift criteria (dc) parameters [5 probability conditions $\times$ 2 instruction conditions with the even probability condition fixed at zero for both speed and accuracy]; 2 means (T_{er}) and 2 ranges (s_t) of non-decision times for the two instruction conditions; and one parameter for the proportion of trials with RT contaminants (p_O). The last parameter accommodates trials with RT delays resulting from lapses in attention (Ratcliff & Tuer-

linckx, 2002). As in most previous applications, the proportion of contaminated trials was estimated to be very low (.0001), which is expected given that the participants were well practiced in making quick responses.

To test the unequal variance explanation, we fit a version of the model in which target and lure items within each condition were constrained to be equally variable. This constraint eliminated 12 free parameters for a total of 54 parameters (what was 4 η parameters for lures and targets studied once, twice, and four times became 1 η parameter across all item types, and this collapsing was applied to the low-frequency speed, low-frequency accuracy, high-frequency speed, and high-frequency accuracy conditions). As mentioned in the text, we also tried model versions with evidence parameters constrained to be equal across speed and accuracy sessions. This constraint eliminated 15 free parameters resulting in a model with 51 parameters. One might expect that the evidence constraint would eliminate 16 parameters: i.e., 8 means and 8 standard deviations of the drift distributions (4 item types $\times$ 2 word frequencies) would be fixed across speed and accuracy. However, fixing the distributions in this way necessitates a free drift criterion parameter for the .5 target probability condition with speed sessions, given that the position of the drift criterion can vary between speed and accuracy. (A free parameter was not needed in the original fits, because with all the distributions free to move it would be redundant to let the criterion move as well). Thus, the net loss is 15 parameters.

For the fit to the smaller dataset used to compare the diffusion model and RTCON, the diffusion model had 22 free parameters: 5 "old" and 5 "new" boundary parameters across the 5 target proportion conditions; 1 range of stating point variability; 2 means and 2 standard deviations for the drift distributions (1 for targets and 1 for lures); 4 drift criteria for the 5 target proportion conditions (with the .5 targets condition fixed at zero); 1 mean and 1 range for the distribution of non-decision times; and 1 parameter for the proportion of trials with RT contaminants.

A.1.4. RTCON

RTCON was fit with the same χ^2 method used to fit the diffusion model. Predictions from RTCON were derived with Monte Carlo simulations. For each condition, the prediction program simulated 20,000 runs of the accumulation race. Each run had a different random sample from the between-trial memory evidence distribution with a mean μ_{BETWEEN} and standard deviation σ_{BETWEEN}. Also, a different random position of the confidence criterion was sampled from a normal distribution with a mean c and a standard deviation σ_{C}. The between-trial evidence sample and the criterion position were used to determine the average accumulation rates (v) for each counter as shown in Fig. 11. Each run also had different positions of the decision criteria sampled from a uniform distribution with means d_{NEW} and d_{OLD} and a range s_{D}. Within each run, the position of each counter was incremented at each time step, with the increment determined by a random draw from a normal distribution with a mean equal to the average accumulation rate and a standard deviation σ. More specifically, the change in evidence for each accumulator was governed by the equation

$$dx(t) = a(v - kx(t))dt + \sigma e\sqrt{dt}$$

where $dx(t)$ is the change in evidence at time step t, a is the scaling factor (fixed at .1), v is the average accumulation rate, k is the decay term, dt is the length of the time step, σ is the standard deviation in accumulation noise, and e is a random normal variable. Across the simulated trials, response proportion predictions were determined by the proportion of trials won by each accumulator and RT predictions were determined by the number of cycles needed to complete the races. A uniformly distributed non-decision component with mean T_{er} and range s_{t} was added to the RTs from the decision process to represent the latency of non-decision processing.

For the smaller dataset used to compare RTCON and the diffusion model, RTCON had 24 parameters. The parameters included 5 "old" and 5 "new" positions of the decision criteria across the 5 target proportion conditions; 1 range of across-trial variability in the decision criteria; 1 mean for the target between-trial evidence distribution (with the lure mean fixed at 0); 2 standard deviations for the between-trial evidence distributions for targets and lures; 5 confidence criteria for the 5 target proportion conditions; 1 standard deviation for across-trial variability in the decision criterion, 1 decay rate;

1 standard deviation for the variability in the accumulation rate; and 1 mean and 1 range for the distribution of non-decision times.

Appendix B

B.1. Full parameter results for the diffusion model and RTCON

Note: Throughout the appendix, parameter values marked with an asterisk were fixed in fits.

Unequal-variance diffusion model full dataset fit

Range in starting point variation (s_Z): .045
Mean of non-decision time distribution (T_{er}): speed = 431, accuracy = 482
Range of non-decision time distribution (s_t): speed = 197, accuracy = 180
Proportion of trials with RT contaminants (p_O): .0001

Parameters varying across target proportion and instructions:

Instructions	Proportion of targets				
	.21	.32	.50	.68	.79
"Old" boundary (a_{OLD})					
Speed	.051	.041	.037	.030	.024
Accuracy	.065	.056	.043	.032	.024
"New" boundary (a_{NEW})					
Speed	−.025	−.032	−.038	−.043	−.054
Accuracy	−.026	−.034	−.047	−.066	−.074
Drift criterion (dc)					
Speed	.017	.017	0*	.009	.006
Accuracy	.052	.019	0*	−.020	−.071

Parameters varying across word frequency, number of presentations, and instructions:

Instructions and frequency	Number of study presentations (0 for lures)			
	4	2	1	0
Drift distribution means (μ)				
Speed				
High	.149	.102	−.001	−.116
Low	.281	.184	.092	−.157
Accuracy				
High	.182	.116	.058	−.111
Low	.347	.241	.131	−.212
Drift distribution standard deviations (η)				
Speed				
High	.288	.288	.259	.159
Low	.222	.193	.170	.100
Accuracy				
High	.211	.205	.202	.105
Low	.259	.250	.228	.150

Diffusion model smaller dataset fit

Drift distribution means (μ): target = .203, lure = −.177
Drift distribution standard deviations (η): target = .300, lure = .185
Range in starting point variation (s_Z): .057
Mean of non-decision time distribution (T_{er}): 482
Range of non-decision time distribution (s_t): 183
Proportion of trials with RT contaminants (p_O): .0001

Parameters varying across target proportion:

Parameter	Proportion of targets				
	.21	.32	.50	.68	.79
"Old" boundary (a_{OLD})	.066	.059	.046	.034	.029
"New" boundary (a_{NEW})	−.029	−.037	−.050	−.069	−.071
Drift criterion (dc)	.090	.036	0*	−.027	−.176

RTCON model

Between-trial distribution means ($\mu_{BETWEEN}$): target = .851, Lure = 0*
Between-trial standard deviations ($\sigma_{BETWEEN}$): target = .839, lure = .689
Range of decision criteria variation (s_D): .898
Standard deviation for confidence criteria variation (σ_C): .100
Standard deviation of variation in the accumulation process (σ): .063
Decay (k): .142
Drift rate scaling factor (a): .1*
Mean of non-decision time distribution (T_{er}): 474
Range of non-decision time distribution (s_t): 219

Parameters varying across target proportion:

Parameter	Proportion of targets				
	.21	.32	.50	.68	.79
"Old" decision criterion (d_{OLD})	5.222	4.573	3.885	2.942	2.323
"New" decision criterion (d_{NEW})	2.903	3.753	4.496	5.411	5.943
Confidence criterion (c)	.474	.533	.539	.554	.501

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*, 700–765.
Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition, 11*, 307–315.
Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37–60.
Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15*, 906–926.
Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 484–499.
DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109*, 710–721.
Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*, 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*, 361–376.
Dewhurst, S. A., & Conway, M. A. (1994). Pictures, images, and recollective experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1088–1098.
Dewhurst, S. A., Holmes, S. J., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness and Cognition: An International Journal, 15*, 147–162.
Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 414–435.
Ditterich, J., Mazurek, M. E., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience, 6*, 891–898.
Dosher, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed–accuracy study. *Cognitive Psychology, 16*, 519–555.
Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika, 17*, 1–35.
Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51). Hearing and Communication Laboratory, Indiana University.
Elfman, K. W., Parks, C. M., & Yonelinas, A. P. (2008). Testing a neurocomputational model of recollection, familiarity, and source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 752–768.
Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., et al (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 107*, 15916–15920.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.
Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8–20.
Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 500–513.
Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature, 404*, 390–394.
Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1355–1369.
Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 846–858.
Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1210–1230.
Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*, 1–18.
Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relation between presentation duration and slope. *Memory & Cognition, 28*, 161–166.
Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception and Psychophysics, 72*, 246–273.
Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*, 335–384.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760.
McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 563–582.
Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858–865.
Nelder, J. A., & Meade, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313.
Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110*, 611–646.
Nosofsky, R. M., & Stanton, R. D. (2006). Speeded old-new recognition of multidimensional perceptual stimuli: Modeling performance at the individual-participant and individual-item levels. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 314–334.
Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General, 139*, 341–364.
Petrov, A. A., Van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual learning mechanisms revealed by diffusion-model analysis of the patterns of specificity. *Psychonomic Bulletin and Review, 18*, 490–497.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two stage dynamic signal detection theory: A dynamic and stochastic theory of confidence, choice, and response time. *Psychological Review, 117*, 864–901.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86*, 446–461.
Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review, 92*, 212–225.
Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. *Journal of Neurophysiology, 90*, 1392–1407.
Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. (in press). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child Development*.
Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97*, 1756–1774.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
Ratcliff, R., McKoon, G., & Tindall, M. H. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 763–785.
Ratcliff, R., & Murdock, B. B. Jr., (1976). Retrieval processes in recognition memory. *Psychological Review, 83*, 190–214.
Ratcliff, R., Sheu, C-F., & Gronlund, S. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.
Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–367.
Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*, 70–94.
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83.
Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*, 278–289.
Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*, 408–424.
Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*, 127–157.
Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*, 464–487.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review, 9*, 438–481.
Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*, 261–300.
Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320.
Rinkenauer, G., Osman, A., Ulrich, R., Müller-Gethmann, H., & Mattes, S. (2004). On the locus of speed–accuracy trade-off in reaction time: Inferences from the lateralized readiness potential. *Journal of Experimental Psychology: General, 133*, 261–282.
Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*, 370–392.
Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28*, 907–922.
Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember-know paradigm. *Psychonomic Bulletin & Review, 15*, 825–832.
Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review, 17*, 427–435.
Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.
Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44*, 408–463.
Starns, J. J., & Ratcliff, R. (2008). Two dimensions are not better than one: STREAK and the univariate signal detection model of RK performance. *Journal of Memory and Language, 59*, 169–182.
Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25*, 377–390.
Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.
Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society, 12*, 89–91.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.
Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 582–600.
Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1147–1166.
Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.
Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*, 641–671.
Wagenmakers, E-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.
White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*, 39–52.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.
Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.
Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*, 800–832.