Jeffrey Starns and Andrew Heathcote

Evidence Accumulation and Decision Processes

Event memory is a process of interpretation. The rememberer acts less like an archivist retrieving records of past events and more like a detective piecing together clues to build a case about what must have happened (Johnson, Hashtroudi, & Lindsay, 1993). The information stored in memory provides the evidence for the case, not the verdict. As such, researchers cannot develop a complete understanding of memory without considering the decision processes that map evidence states onto explicit answers to memory questions ("Have I ever met this person before?"; "Did I see that movie with Paula?"; "Have I told this story at a previous work party?"). To complete the courtroom analogy, decision processes play the role of the jury by integrating the available evidence and making a judgement about past events. Decision processes are a worthy subject of investigation in themselves, and researchers need to understand decision making to correctly interpret results from memory tasks that have a decision component. Such decision tasks include recognition, in which one decides whether or not an event was previously experienced (often the event of seeing a word in a list); source memory, in which one decides which context or attribute was paired with a stimulus in an earlier event (such as whether a word was heard in a male or female voice); and many other tasks used by memory researchers (even recall tasks involve deciding whether or not a generated item should be reported; Anderson & Bower, 1972).

The last few decades have marked substantial theoretical progress for mathematical models of decision making, and *evidence-accumulation models* have provided the main

impetus for this progress (e.g., Brown & Heathcote, 2008; Ratcliff, 1978; Smith & Vickers, 1988; Usher & McClelland, 2001). Broadly, evidence-accumulation models assume that decisions are made by adding or integrating samples of noisy evidence over time until the total evidence reaches a threshold, so they are also commonly called sequential-sampling models. When applied to memory tasks, these models are consistent with the metaphor in the preceding paragraph: they assume that information retrieved from memory provides evidence about past events, and they implement the decision process that translates this evidence into a response.

Currently, a number of models are available that consistently match all aspects of behavioral data from decision tasks: choice proportions; the location, shape, and spread of response time (RT) distributions for both correct and error responses; and the manner in which all of these characteristics (co)vary across experimental conditions or individual differences (e.g., Ratcliff & Smith, 2004; Brown & Heathcote, 2008). These models have proven to be valuable research tools across a wide variety of tasks by mapping complex patterns of data onto psychologically meaningful processes (Donkin & Brown, 2018; Ratcliff & McKoon, 2008). They are also becoming increasingly influential in memory research, where they are slowly replacing decision models that address accuracy but do not consider RT data, such as models based on signal-detection theory (Green & Swets, 1966; Tanner & Swets, 1954).

To illustrate the role of dynamic decision processes in memory we will focus on two evidence-accumulation models that have been widely applied, the diffusion-decision model (DDM; Ratcliff & McKoon, 2008) and the Linear Ballistic Accumulator (Brown & Heathcote, 2008). We will first provide details of the DDM and then describe the LBA.

A simple diffusion model was first proposed by Stone (1960) as a general model of binary perceptual choice. It assumes that the evidence for a two-alternative choice is sequentially sampled and integrated until it exceeds either an upper or lower threshold, where each threshold corresponds to a choice.

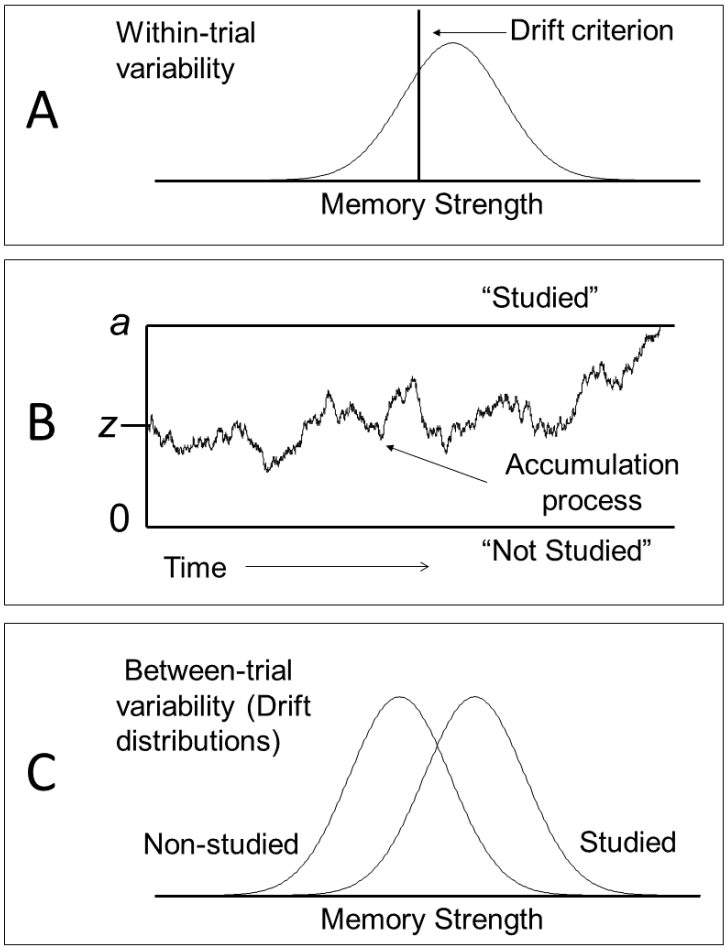

Figure 1. The diffusion model as applied to a recognition memory task. Panel A shows a Gaussian distribution of within trial variability and the drift criterion defining the zero-point in drift rates. Panel B shows an example evidence accumulation process, which begins at a starting point (*z*), accumulates (sums together) samples from the within trial distribution (Panel A), and terminates when one of the two response boundaries is

reached. Panel C shows Gaussian distributions of between-trial variability in drift rates. The average drift rate (i.e., mean of the within-trail distribution) on each trial is a random sample from this distribution.

Let's consider an example of this sequential sampling process relevant to a common recognition memory task; that is, deciding if a word was or was not studied in an earlier list. To make this decision, one needs an information source; that is, some signal that tends to be different for words that were and were not encountered on the study list. Some candidates for such a signal could be the "sharpness" of neural representations in the perirhinal cortex or the fidelity of pattern completion in the hippocampus (Norman & O'Reilly, 2003), but we will remain agnostic as to the source of the signal and simply refer to it as "memory strength." The diffusion model assumes that evidence is subject to moment-to-moment variability (neural noise, perhaps; Gold & Shadlen, 2007) that can be represented by a Gaussian distribution such as the one shown in Figure 1A. The values in the distribution represent the memory strength experienced at different points in time within a single trial of a recognition task, with higher values on the continuum representing stronger evidence that a word was previously studied. The vertical line is a model parameter called the drift criterion, and it defines how strength values are mapped to evidence that the word was studied or not studied. Strength values above the drift criterion provide evidence that the item was studied, and strength values below the drift criterion provide evidence that it was not. The memory strength experienced in a given moment is a random sample from the within-trial distribution seen in Figure 1A, and due to the

considerable within-trial variability, the strength value will sometimes fall above and sometimes below the drift criterion, creating ambiguity as to the correct response.

The model assumes that the decision maker deals with this within-trial variability by taking multiple evidence samples over time and accumulating evidence until a criterial level of support for one decision alternative is achieved. Figure 1B shows a continuum of accumulated evidence over time, with higher (lower) values indicating more evidence supporting a studied (not studied) response. The wavy line in Figure 1B shows an example evidence accumulation process, with the "waves" representing the moment-to-moment variability in strength as each evidence sample affects the total accumulated evidence. The offset produced by each new sample is determined by its distance from the drift criterion; a momentary strength sample far above the drift criterion moves the accumulation process up by a lot, a sample just below the drift criterion moves it down by a little, and so forth. The decision maker continues to consider new momentary strength samples until the accumulation process reaches one of the two boundaries shown in Figure 1B, triggering a "studied" response for the top boundary or a "not studied" response for the bottom boundary.

The speed and accuracy of responses are influenced by several factors. First, moving the within-trial strength distribution relative to the drift criterion affects the rate of approach to the top or bottom boundary, commonly referred to as drift rate, *v* (Ratcliff, 1978). A within-trial distribution centered near the drift criterion tends to produce an accumulation path with many vacillations in direction (drift rate near zero); a within-trial distribution far above the drift criterion produces an accumulation path that consistently takes big steps towards the top boundary (strong positive drift rate); and a within-trial

distribution far below the drift criterion produces an accumulation path that consistently takes big steps towards the bottom boundary (strong negative drift rate). So, the drift rate for an individual trial represents the overall memory strength of the item being considered; for example, a word studied 4 times should tend to have a higher positive drift rate than a word studied once. The distance between the response boundaries, often denoted by *a*, represents the level of response caution, or the speed-accuracy tradeoff, adopted by the decision maker. Narrow boundaries produce fast responding but high susceptibility to errors produced by within-trial variability; wide boundaries produce slow responding but reduce within-trial errors by allowing more time for within-trial variability to "average out." The position of the accumulation process at the beginning of the trial is called the starting point, *z*, and it represents response biases that are unrelated to an item's memory strength. For example, a decision maker who knows that the majority of tested items are studied might start accumulation near the top boundary to represent the fact that they are leaning towards a "studied" response even before they attempt to remember the test item.

Building on earlier work by Laming (1968) in perceptual choice, Ratcliff (1978) applied an elaborated version of the simple diffusion model to recognition memory data. The elaboration adds between-trial variability to the within-trial variability of the simple diffusion model. The most important added assumption for memory theorists is between-trial variability in drift rate. That is, the Ratcliff diffusion model assumes that drift rate varies from trial to trial according to a Gaussian distribution for both studied and non-studied items. Figure 1C shows an example of these drift distributions, one for studied and one for non-studied items. For a given trial, the position of the within-trial distribution in Figure 1A is a random draw from the appropriate between-trial distribution in Figure 1C;

e.g., a draw from the distribution on the right for a studied item. So, two different trials with a studied item would have two different positions of the within-trial distribution (i.e., two different drift rates). This added layer of variability is very plausible for memory tasks like recognition. For example, even if every word on a study list is studied once for one second, say, some words will be more memorable than others because they have distinctive orthography, have a meaning that is personally relevant to the participant, are presented at a time when the participant happens to be paying close attention to the study task, etc. Moreover, even words not studied on the list can vary in strength for a number of reasons; for example, some of these words might be a close associate of a word that was studied. This between-trial variability in drift can produce errors even with a very cautious speed-accuracy tradeoff (i.e., wide boundaries). For example, a studied word presented during a lapse in attention could fall in the lower tail of the drift distribution and end up with a negative drift rate; thus, the accumulation process would tend to move toward the bottom boundary and the participant would be likely to incorrectly decide that the item was not studied even if they adopted a slow pace of responding.

The diffusion model also has some ancillary parameters that are less likely to factor into research questions addressed with the model. One of these in the average duration of non-decision processes, including stimulus encoding and response production times. For example, non-decision time could represent the time needed to read a word before trying to remember if it was on the study list or the time needed to press a computer key once a decision has been made. RT is the sum of decision time (i.e., the time to accumulate evidence to a response boundary) and non-decision time. The full version of the DDM also includes uniformly distributed variability from trial to trial in non-decision time and in the

starting point of evidence accumulation (Ratcliff & Tuerlinckx, 2002). The latter type of variability affords the full DDM the ability to accommodate data in which error responses are faster than correct responses, which happens when people are pressed to respond very quickly but the decision task is otherwise very easy (Ratcliff & Rouder, 1998). The error-faster-than-correct pattern is rare for memory tasks, which tend to be difficult. Trial-to-trial variability in drift rates also allows the DDM to accommodate slower error than correct RT, which happens when accuracy is emphasized over speed. These two types of trial-to-trial variability together make the DDM able to accommodate most observed patterns of correct vs. error speed (Ratcliff & Rouder, 1998).

Even with multiple sources of across-trial parameter variability, the model is still quite testable in that there are many patterns of RT data that it *cannot* produce. Artificially changing RT data in even fairly subtle ways severely impairs model fits (Ratcliff, Thapar, Gomez, & McKoon, 2004), and one could list many properties of RT distributions that the model **must** predict for any parameter set, such as the fact that predicted distributions must be positively skewed and, further, must maintain the same shape across any manipulation that affects decision difficulty by changing the information available for the decision (Ratcliff & McKoon, 2008). Observed RT distributions honor these tight constraints with impressive consistency, providing strong support for the model (e.g., Ratcliff & McKoon, 2008; Ratcliff & Smith, 2004).

Although the diffusion model has been incredibly influential (Wagenmakers, 2009), it is by no means the only evidence accumulation model available to decision modelers. The LBA model was proposed more recently by Brown and Heathcote (2008) and is illustrated in Figure 2. It shares the DDMs assumptions about across-trial variability in the starting

point and rate of evidence accumulation but differs in three respects. First, it assumes the effect of within-trial variability is so much smaller than that of across-trial variability that accumulation within a trial can be approximated as having a constant rate, from which it derives the "Linear Ballistic" components of its name and the straight line accumulation illustrated in Figure 2. Second, there are separate accumulators for each response, with the first one that reaches its threshold determining the choice that is made and the decision time. Figure 2 illustrates a binary recognition memory trial in which a lure response is made because the lure accumulator reaches its boundary before the target accumulator. As we discuss in Section 3, racing accumulator models like the LBA are not limited to accounting for only binary choice like the DDM because it is straightforward to expand their architectures to have an accumulator for each possible response. Finally, in most applications of the LBA, non-decision time variability has been assumed to be so much smaller than variability in decision time that it can be treated as a constant.

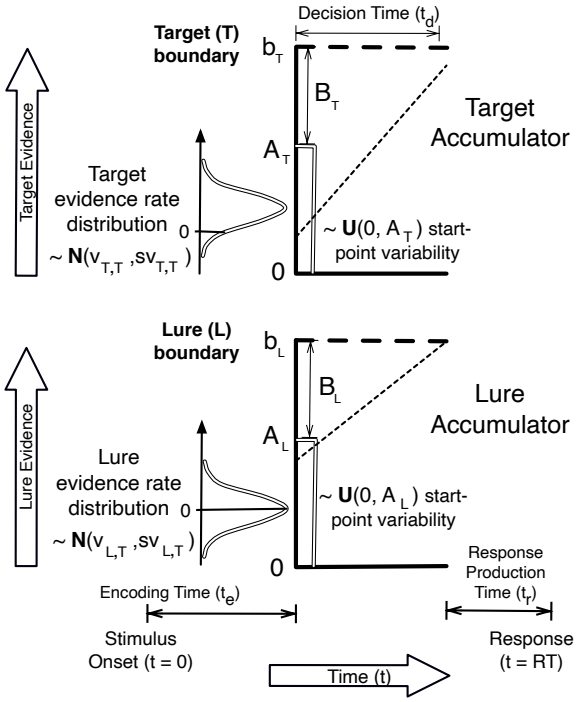

Figure 2. An LBA model of a binary recognition memory trial. Target and lure accumulators have separate parameters indicated by T and L subscripts respectively: *b*, the evidence total required to trigger a response, *v*, the mean rate of evidence accumulation, *sv*, the standard deviation of across-trial rate variability (which has a normal distribution), and *A*, the width of the uniform distribution of evidence starting points. Non-decision time is the sum of encoding and response production times. In the trial illustrated a target is presented, so the mean rate for the target accumulator given a target stimulus ($v_{T,T}$) is greater than the mean rate for the lure accumulator ($v_{L,T}$), so the slope of the dashed line representing within-trial accumulation for the target accumulator is steeper than the slope for the lure accumulator. However, random variability in the starting point of evidence accumulation gives the lure accumulator a sufficient head start to win. This illustrates how speed-accuracy tradeoff is accounted for the LBA: raising the threshold would eventually result in a correct response.

Although these different assumptions make the LBA simpler than the DDM, both mathematically and computationally, Brown and Heathcote (2008) showed that it can provide the same comprehensive account of benchmark binary choice phenomena as the DDM (e.g., Figure 2 illustrates how it accounts for speed-accuracy tradeoffs). At the same time, its greater architectural flexibility allows it not only to be applied beyond binary choice, but also to model a wider range of paradigms and cognitive processes than the DDM, as we illustrate in Section 3.

Most applications of the DDM and LBA do not account for the memory processes that give rise to drift rates. Instead, they are used as a *measurement models*, which combine the information in accuracy and RT measurements in order to separate out the effects of

memory processes, which determine drift rates, from the effects of decision processes, such as the total amount of evidence required to make a choice. However, some theorists have also advocated a closer integration of the idea of evidence accumulation with models of memory processes. In the final section of the paper we discuss three recent examples.

To set up our discussion of evidence accumulation models, the next section describes some ways that applying accuracy-based decision models has influenced memory research. Our goal for this section is to paint a broad-strokes picture of the types of research questions that have been addressed with accuracy-only decision modeling, which will serve as context for the next section describing applications that use evidence-accumulation models as measurement models in recognition memory research. As such, we will omit discussion of research questions that have not been re-assessed with evidence-accumulation models. For more thorough reviews of this literature, see Wixted (2007), Yonelinas and Parks (2007), and Pazzaglia, Dube, and Rotello (2013, with commentary by Batchelder & Alexander, 2013).

Section 1: Accuracy-based decision modeling in memory research

Decision modeling in memory research has a long history, stretching back at least to Egan's (1958) signal detection modeling of recognition memory data. In the early 1990's, seminal papers from the Ratcliff, Glanzer, and Yonelinas laboratories sparked a wave of studies that attempted to test memory theories by fitting receiver operating characteristic (ROC) functions (Glanzer, Adams, Iverson, & Kim, 1993; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, 1994). ROCs are a way of representing data and theoretical predictions that was borrowed from perceptual signal detection research (Tanner & Swets, 1954). A recognition memory ROC shows the relationship between the

hit rate and the false-alarm rate, with the former defined as the proportion of studied items ("targets") correctly identified as studied and the latter defined as the proportion of non-studied items ("lures") mistakenly identified as studied (see Figure 1). Although recognition memory ROCs are the most common, ROCs can be defined for any memory task that requires discrimination of different stimulus classes; for example, in a source memory task, researchers can form an ROC by plotting the proportion of "Source 1" responses for items actually studied in Source 1 (correct responses) on the proportion of "Source 1" responses for items studied in Source 2 (incorrect responses; e.g., Yonelinas, 1999).

ROCs can be defined by manipulations that affect response biases; most commonly, this involves changing the proportion of the different stimulus types on the test list or changing the gains and losses associated with the different response options (e.g., Starns, Ratcliff, & McKoon, 2008). For example, a participant might complete a recognition test across different test blocks that have either 20%, 35%, 50%, 65%, or 80% studied words. Figure 3A shows an example of five ROC points and labels the corresponding proportion conditions for each. Participants tend to be more willing to say "studied" when a higher proportion of studied items are on the test list, producing increases in both the hit rate and the false alarm rate. Accordingly, the ROC points move higher on both axes from 20% to 80% targets.

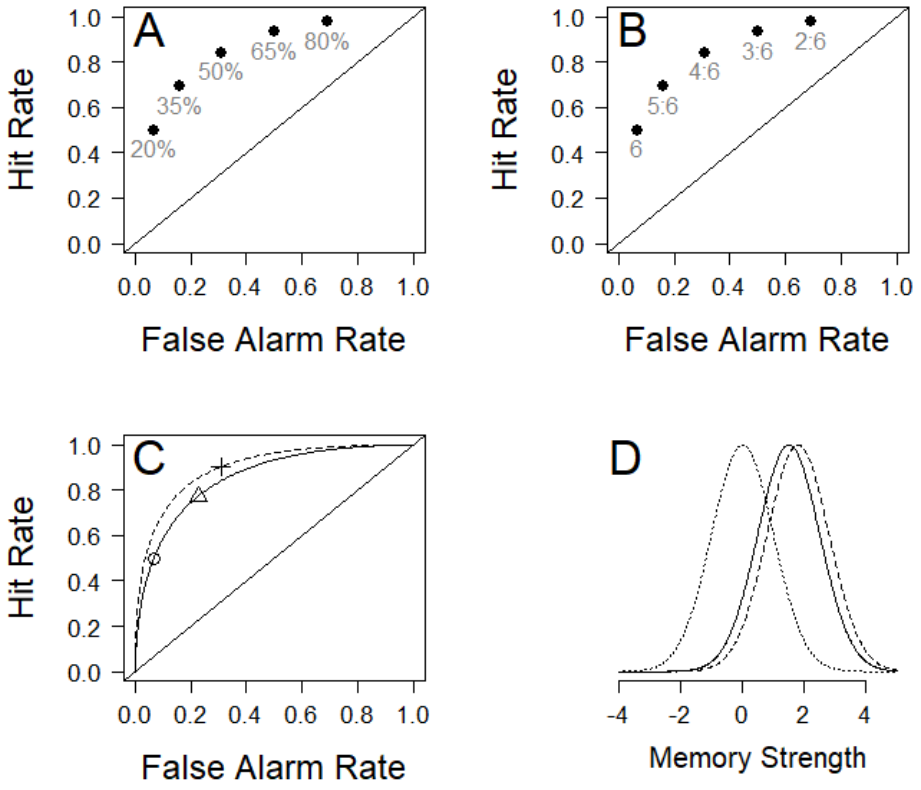

Figure 3. Recognition memory Receiver Operating Characteristics (ROCs) and Signal Detection Theory (SDT) distributions. Panels A and B show ROC points defined by 5 different bias conditions (proportion of targets on the test list) or a 6-point confidence-rating scale, where each point marks the hit rate and false alarm rate at a given level of response bias. Panel C shows three pairs of hit and false alarm rates, with two pairs indicating the same level of memory discriminability and the third indicating higher discriminability. Panel D shows the signal detection strength distributions that generated the ROCs in Figure C, with a dotted line for the lure distribution, a solid line for the weak target distribution, and a dashed line for the strong target distribution.

ROCs can also be created from a confidence scale (e.g., Egan, 1958). For instance, participants could be asked to respond to each test word with a 6-point scale ranging from 1 (Definitely Not Studied) to 6 (Definitely Studied). In this case, the first ROC point represents the most stringent policy for mapping confidence levels onto a binary "studied"/"not studied" decision, so only high-confidence "studied" responses (6's) are counted. The next point represents the next most stringent policy, so high- and moderate-

confidence "studied" responses (5's and 6's) are counted, and so forth. Figure 3B shows a confidence-rating ROC and labels the confidence-scale levels counted as "studied" for each point.

ROCs as a Measurement Tool

One strand of ROC research uses these functions as a tool to disentangle changes is memory discriminability and response bias across different levels of an independent variable or predictor variable. Discriminability refers to how much information a person has to distinguish words that were previously studied from words that were not; in everyday terms, it is a measure of how well a person remembers the words on the study list. Response bias refers to the overall predilection to say "studied" or "not studied." It is often difficult to distinguish changes in discriminability and response bias when performance is defined by a single hit and false-alarm rate. For example, imagine that the circle, triangle, and plus sign in Figure 1C show performance from Conditions 1, 2, and 3 in an experiment. Conditions 2 and 3 both have higher hit rates and higher false alarm rates than Condition 1, but determining whether or not this indicates a difference in memory discriminability requires a decision model to define the ROC.

The theoretical curves displayed in Figure 3C are based on the signal detection model (Tanner & Swets, 1954) displayed in Figure 3D. Signal detection theory posits distributions of evidence strength that are quite similar to the drift distributions discussed above as shown in Figure 1C, but the strength values are used in a one-step decision process in which a given item's strength value is simply compared to a response criterion to decide between "studied" and "not studied" responses (strength values above and below the criterion, respectively). In Figure 3D, the dotted curve represents the memory strength

distribution for non-studied words, and the solid and dashed curves represent strength distributions for studied items at two different levels of discriminability. The ROCs are produced by sweeping a criterion across the strength distributions, with the hit rate and false alarm rate for a given criterion position determined by the proportion of the studied and non-studied strength distributions above the criterion, respectively. For example, if we start the criterion at an extremely high value (all the way to the right in Figure 3D), then almost none of the strength distributions are above the criterion and we are near the (0,0) point on the ROCs. Moving the criterion lower and lower increases the proportion of the distributions above the criterion, moving to the (1,1) point on the ROCs. Any difference in performance for conditions that fall on the same ROC can be accommodated by a change in response bias without a change in memory ability.

According to the displayed signal detection model, the difference between Condition 1 and 2 (the circle and the triangle) is a function of response biases: participants in the two conditions have the same potential to discriminate studied and non-studied items (performance lies on the same ROC), but participants in Condition 2 made "studied" responses more liberally (adopted a lower criterion value). The difference between Condition 1 and 3 (the circle and the plus sign) is produced by both a bias shift and a difference in discriminability, with participants in Condition 3 displaying better ability to discriminate studied and non-studied items. In technical terms, the average strength for studied items was farther above the average strength for non-studied items in Condition 3 compared to Condition 1, so performance lies on a different ROC farther from the chance diagonal. Different decision models trace different functions through these points and might produce different conclusions about whether or not memory discriminability

changed between conditions (Kinchla, 1994). Thus, obtaining ROC data helps to distinguish discriminability and bias by providing information about whether the ROC assumed by a particular model is appropriate for the data.

A number of memory studies have employed either confidence ratings or bias manipulations to define ROCs with the goal of distinguishing memory discriminability and response bias (e.g., Gombos, Pezdek, & Haymond, 2012; Healy, Light, & Chung, 2005; Verde & Rotello, 2003). This practice has been generally successful for standard recognition tasks, and is now being explored in the eyewitness memory literature (e.g., Mickes, Flowe, & Wixted, 2012). That said, we wish to note two caveats for the use of ROCs as measurement tools.

The first caveat is empirical: a recent study sent unlabeled data to a number of published recognition memory researchers and asked them to determine if memory discriminability was manipulated across two conditions that might also vary in terms of response bias (Starns et al., 2019). The results showed no indication that the ability to make inferences about discriminability improved when confidence rating data were available to define ROCs compared to two-choice ("studied"/"not studied") data without confidence ratings. This suggests that ROC data are not always helpful, but future research will be needed to explore this pattern in more depth.

The second caveat is theoretical. Differences in memory discriminability can be masked not only by response biases, but also by differences in response caution, i.e., the speed-accuracy tradeoff adopted by the decision maker. For example, one subject population might emphasize quick responding while another takes the time to retrieve as much information as possible, which could produce lower performance in the first

population even if members of this population can remember the target events just as well as the comparison population. As we will review below, evidence-accumulation models can measure memory performance independently of differences in response caution as well as bias.

ROCs as Tests of Memory Theories

A different strand of research has focused on using ROCs to answer basic questions about memory processes. Again, we will briefly characterize this work without attempting a comprehensive review. We will consider three general research questions that have been influential in this literature and that are currently being re-assessed with evidence-accumulation models: whether there are independent memory systems or a single coherent system; whether the relative variability of memory evidence for studied and non-studied items is consistent with various mathematical models of memory; and whether retrieval processes produce continuous or discrete information states. In the following paragraphs, we will briefly describe efforts to investigate these questions and highlight a few example studies. The examples in this section demonstrate that memory researchers have a history of making theoretical conclusions based on properties of ROCs.

Dual Process Debate. Pioneering work by Yonelinas (1994) linked ROCs to earlier multiple-system theories of memory (see Yonelinas, 2002, for a review of these theories) and quickly became a powerful force promoting the popularity of ROC modeling. Yonelinas created a decision model that distinguished recollection – the controlled, contextualized recovery of specific details from an earlier experience – and familiarity – the error-prone, automatically-generated sense that a stimulus has been recently encountered (Yonelinas, 2002). This dual-process signal detection (DPSD) model assumes that familiarity is a

continuous signal-detection process; that is, familiarity values follow Gaussian distributions across trials with higher familiarity on average for studied items due to their recent presentation (like the distributions in Figure 3D). Recollection, in contrast, follows a high-threshold process that succeeds for a proportion of studied items and fails for the rest. The model further assumes that recollection and familiarity are independent (i.e., the familiarity distributions are identical for recollected and non-recollected targets) and that the familiarity distributions are equally variable for studied and non-studied words.

Yonelinas (1994) demonstrated a simple correspondence between parameters of the DPSD model and properties of the predicted ROC. In the model, familiarity-based responding produces a curved ROC that is symmetrical around the negative diagonal, whereas recollection-based responding produces a linear ROC with a slope less than 1 (thus making the function asymmetrical around the negative diagonal). As a result, the relative contribution of recollection and familiarity can be measured based on the extent to which the observed ROC is flattened and asymmetrical. This conceptualization of ROC shape triggered a flood of studies that either applied these measurement properties of the model to various research questions or tested the DPSD model against alternatives (for reviews of this literature, see Heathcote, 2003; Wixted, 2007; Yonelinas & Parks, 2007).

The dual-process approach is often contrasted with the unequal-variance signal detection (UVSD) model in which memory strength values are normally distributed for both studied and non-studied items, with a higher mean and higher variability for studied items (Egan, 1958). For some researchers, this approach was motivated by global matching models in which memory evidence is represented by a single, continuous match strength (e.g., Ratcliff et al., 1992). For others, it was motivated by dual-process models in which

recollection and familiarity are both continuous variables and are added together on every trial to represent the total evidence that an item was studied (e.g., Wixted, 2007). This approach predicts curved ROCs without the flattening produced by recollection in the dual-process model, and it links the degree of ROC asymmetry to the ratio of the standard deviation in evidence values for studied and non-studied items, with symmetrical functions for equal variability and increasingly asymmetrical functions for variability ratios farther from 1. Given that the DPSD model uniquely predicts flattened ROCs, one would expect that ROC shape provides information capable of discriminating the models. For recognition memory tasks, however, this aspect of the data has proven to be too subtle for confident conclusions. For example, Glanzer, Kim, Hilford, and Adams (1999) noted that conditions producing more asymmetrical recognition-memory ROCs do not also have more flattened ROCs as predicted by the DPSD model, but Yonelinas (1999) responded by noting that the DPSD model closely fits empirical ROCs.

Generally, evaluating ROC data has not successfully resolved the process debate. Instead of enumerating the many volleys exchanged by the competing theoretical camps, we will highlight research on source memory ROCs as an example. Early source ROCs seemed to support the DPSD model, as they were substantially flatter than recognition memory ROCs and flatter than predicted functions from the UVSD model (Yonelinas, 1999). This pattern of results is expected under the DPSD model based on the reasonable proposition that source tasks rely primarily on recollection. However, subsequent studies argued that flat ROCs are produced by uninformed guesses for the subset of items that were not recognized as being studied in any source, and this account is supported by the fact that source ROCs show the curvature predicted by the UVSD model if analyses are

limited to items that the rememberer is confident were previously studied (Slotnick & Dodson, 2005). More recent studies demonstrate that people are less willing to provide high-confidence source responses when they are less certain that an item was studied, and neither the DPSD or UVSD model can accommodate joint recognition and source confidence data without using decision criteria that implement this confidence heuristic (Starns et al., 2014; Starns, Pazzaglia, Rotello, & Hautus, 2013). Notably, the confidence heuristic produces flattened ROCs even for models that otherwise predict curved ROCs, so ROC flattening does not provide clear information about the nature of memory retrieval.

Testing Memory-Process Models. The previous section described efforts to use ROCs to test decision models that make a few basic assumptions about the nature of memory retrieval, such as assuming that strength values follow Gaussian distributions. ROCs have also been used to test "process" models that attempt to simulate the mechanisms involved in memory encoding, storage, and retrieval. This trend began with studies that used ROCs to measure the relative variability of evidence strength values for studied and non-studied items to test global matching models of memory (Ratcliff et al., 1992, 1994). These researchers measured evidence variability in terms of ROC asymmetry as defined by the UVSD model. Results showed that ROCs were asymmetrical to a degree corresponding to roughly 25% higher variability for studied than non-studied items (replicating Egan, 1958), but the degree of ROC asymmetry did not increase with additional learning. This pattern provided evidence against existing global matching models, which tended to predict either increasing unbalanced variability ratios with additional learning or equal variability for studied and non-studied items (Clark & Gronlund, 1996; Ratcliff et al., 1992, 1994).

Retrieval Format. Recently, the ROC literature has focused on the format of memory evidence, where one alternative is that a retrieval attempt produces a continuous strength value representing the total evidence available in memory and the other alternative is that retrieval results in one of a small number of discrete evidence states (for a review, see Pazzaglia et al., 2013, and commentary by Batchelder & Alexander, 2013). Discrete-state, or threshold, models have been dismissed by some researchers based on the observation that ROCs are curved, not linear (Kintsch, 1967; Wixted, 2007; Yonelinas & Parks, 2007). However, most recognition ROCs are based on confidence rating data, and discrete-state models can predict curved ROCs if responses from a given evidence state are distributed across confidence levels (Banks, 1970; Broder & Schutz, 2009; Erdfelder & Buchner, 1998; Malmberg, 2002). A number of papers have attempted to distinguish discrete and continuous models by evaluating ROCs produced by bias manipulations instead of confidence ratings (e.g., Bröder & Schutz, 2009; Dubé & Rotello, 2012; Dubé, Starns, Rotello, & Ratcliff, 2012; Kellen, Klauer, & Bröder, 2013). ROC data have not succeeded in resolving this debate, and one can find papers claiming to support continuous models and discrete models in roughly equal measure.

Summary

Perhaps the clearest take away from the ROC literature is that fitting ROCs is not a great way to test models of memory decisions, a conclusion that was presaged by Banks (1970) and Lockhart and Murdock (1970). A wide range of models have been successful in matching empirical ROCs despite offering dramatically different characterizations of memory retrieval (DeCarlo, 2002; Egan, 1958; Kellen & Klauer, 2015; Yonelinas, 1994). On a more positive note, ROCs can be a useful tool for distinguishing effects on memory

discriminability versus response bias (Banks, 1970), but this methodology is limited as a measurement technique as it has no way to identify speed-accuracy tradeoffs.

Section 2: Evidence-accumulation models in memory research

Although decision modeling has played a prominent role in many aspects of memory research, memory researchers have primarily relied on models that fit accuracy and/or confidence data without considering RTs. We now consider more recent studies that capitalize on the rich processing information provided by RT distributions while addressing the same sorts of research questions discussed in Section 1. Modeling RT distributions has been a fairly recent focus for these specific research questions, but RT data have long played an important role in various aspects of memory research. To cite some prominent early examples, Atkinson and Juola (1974) developed and tested a model for average recognition RTs based on a process in which a fast familiarity assessment was supplemented by a slower memory search for items with ambiguous familiarity values; Norman and Wickelgren (1969) explored the possibility that recognition response times are an increasing function of the distance between an item's memory strength and the response criterion; and Thomas and Myers (1972) used RT data to construct RT-ROCs with the goal of testing decision models assuming continuous or discrete evidence states (see Weidemann & Kahana, 2016, for a recent follow up). The specific studies discussed here differ from these past efforts by directly modeling all aspects of the data, including response proportions and RT distributions for both correct and error responses.

Measurement of Underlying Processes

Just as ROCs are sometimes used as a tool to distinguish memory and bias, advocates of evidence-accumulation models have emphasized their ability to separately

measure distinct psychological constructs (Donkin & Brown, 2018), and a growing number of studies have used evidence-accumulation models as measurement tools for memory tasks (e.g., Criss, 2010; Starns, Ratcliff, & White, 2012; Ratcliff, Thapar, and McKoon, 2004). We first highlight the Ratcliff, Thapar, and McKoon (2004) study as an excellent example of the value of decision modeling in memory research. These researchers used the DDM to explore aging effects in recognition memory. Previous results indicated that young and older adults have similar accuracy levels in recognition memory, but older adults take substantially more time to decide. This pattern had been interpreted in terms of a "general slowing" mechanism whereby aging slows a wide range of cognitive processes by a similar factor (Salthouse & Somberg, 1982). However, considering the decision requirements of the task reveals that there are a number of possible interpretations for the observed results. The slowing experienced by older adults could be driven by any combination of impaired evidence quality (i.e., lower drift rates), greater response caution (i.e., more widely separated thresholds), or slowed non-decision components like pressing a key once a decision has been made. The roughly equivalent accuracy values could indicate similar levels of recognition memory ability, or it could indicate impaired memory for older adults that is counteracted by increased response caution. These different theoretical possibilities produce different patterns of aging effects on accuracy data and RT distributions in the diffusion model, so applying the model can provide information about which scenario is the most credible.

In fits to data, Ratcliff et al. (2004) found that older adults had drift rate estimates that were very similar to younger adults, indicating similar levels of recognition memory ability. The age-related slowing was driven by two factors: older adults were more cautious

and took longer to complete the non-decision task requirements (e.g., hitting keys). One might expect that older adults would be more accurate than younger adults if they have similar memory acuity and are more cautious, but that incorrect expectation shows why it is critical to formalize ideas in explicit decision models instead of relying on intuition. As response caution increases, accuracy increases in a negatively accelerated function that approaches an asymptote imposed by between-trial variation in evidence (i.e., some items have drift rates that approach the wrong boundary due to the variability illustrated in Figure 1C, and these items will consistently produce errors even in conditions of high caution). Younger adults tend to set boundaries that achieve accuracy levels near asymptotic performance, and older adults tend to set boundaries well past the flat part of the curve, meaning that they could speed up substantially without a meaningful drop in accuracy (Garton, Reynolds, Hinder & Heathcote, 2019; Starns & Ratcliff, 2010). Thus, this example illustrates the multiple benefits of applying decision models: the modeling efforts not only clarified the factors driving observed effects, but also led to new discoveries that were not evident in the raw performance data, such as the fact that older participants tend to adopt a maladaptive level of caution for decision tasks (Starns & Ratcliff, 2010, 2012).

RT Models of ROC Data

Confidence ROCs. Theorists have developed a variety of models that attempt to jointly accommodate confidence and RT data (and the relationship between the two; Merkle & Van Zandt, 2006; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009; Smith & Vickers, 1988; Van Zandt, 2000). Some of these models focus on a two-step decision task in which participants first make a two-choice response and then rate their confidence in that response (Pleskac & Busemeyer, 2010; Smith &

Vickers, 1988; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979). Other models focus on a one-step paradigm in which participants make a single response across multiple confidence levels; for example, selecting one of six keys for six confidence levels from "Definitely Not Studied" to "Definitely Studied" (Moran et al., 2015; Ratcliff & Starns, 2009).

The list of memory studies that attempt to jointly model RT distributions and confidence data remains sparse (Osth, Bora, Dennis & Heathcote, 2017, Osth, Dennis & Heathcote, 2017; Osth, Jansson, Dennis & Heathcote, 2018; Moran et al., 2015; Ratcliff & Starns, 2009, 2013; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004), but this literature has already established some important conclusions. Perhaps the most important contribution of these studies is demonstrating that all properties of the ROC can be influenced by decision processes, and thus ROCs do not provide direct insight into memory processes. Section 1 highlighted the practice of interpreting certain ROC properties in terms of memory processes, with particular focus on asymmetry and shape as an indication of underlying processes such as the relative contribution of recollection and familiarity or the relative variability of memory evidence for studied versus non-studied items. The link between ROC properties and memory processes was inspired by accuracy-only models based in signal detection theory, such as the DPSD (Yonelinas, 1994) and UVSD (Egan, 1958) models. In these models, changing decision processes (i.e., response criteria) shifts performance along the same ROC without affecting properties like asymmetry and shape. Unfortunately, this clean dissociation breaks down when models are extended to RT data, and several empirical results show that decision processes do indeed influence ROC

properties in the manner predicted by RT models (Mueller & Weidemann, 2008; Ratcliff & Starns, 2009; Van Zandt, 2000).

Van Zandt (2000) provided a compelling demonstration of the uncertainties inherent in ROC interpretation. In this study, participants completed a recognition test across conditions designed to manipulate response biases (e.g., changing the proportion of studied items on the test) and were also asked to rate their confidence in each response. As expected, conditions that encouraged people to make "not studied" responses (e.g., a test list in which only 25% of the items were studied) had lower hit and false-alarm rates than conditions that encouraged people to make a "studied" response (e.g., 75% studied items). A more surprising result was that the bias manipulations also changed the asymmetry of the ROC. Van Zandt demonstrated that an evidence-accumulation model predicted the change in ROC asymmetry and also correctly predicted the effect of the bias manipulation on RT medians. Thus, this study offers both an empirical demonstration that ROC asymmetry is not a pure measure of memory processes and a theoretical demonstration that predicted ROCs can change dramatically when accuracy-only decision mechanisms are replaced with mechanisms capable of accommodating RT data.

Work by Ratcliff and Starns (2009) reinforces the message that dynamic decision processes can complicate ROC interpretation. They developed the RTCON model for the one-step confidence procedure with the goal of accommodating the proportion of responses and the full RT distribution at each level of the confidence scale. The model is similar to a signal-detection approach to confidence, as it assumes Gaussian distributions of memory strength with criteria to segment the evidence continuum into regions associated with each confidence level. However, the proportion of the strength distribution in each

region does not directly determine the distribution of confidence responses; instead, the proportions are translated into drift rates for evidence accumulators associated with each confidence response. The accumulators race and RT is the time taken by the first accumulator to reach its response boundary plus non-decision time. The accumulators differ from the DDM in that they have only one boundary. This type of accumulator has a different distribution of the time to reach the boundary than the DDM (a Wald distribution, see Heathcote, 2004) and has been used in a range of applications where there are more than two possible responses (e.g., Leite & Ratcliff, 2010; Logan, Van Zandt, Verbruggen & Wagenmakers, 2014).

Ratcliff and Starns (2009) demonstrated that the RTCON model could match data from a speeded confidence judgement experiment and noted a number of ways in which switching to an evidence accumulation mechanism changes the interpretation of ROCs. Just like the evidence accumulation model used by Van Zandt (2000), the RTCON model predicts that response-bias manipulations should affect ROC asymmetry, and empirical results again confirmed this prediction for sequential-dependency biases (i.e., a bias to repeat the previous response). Moreover, ROC shape is also affected by the relative position of the accumulation boundaries across confidence levels in RTCON. These changes in accumulation boundaries should be accompanied by changes in the RT distributions across the confidence levels, and Ratcliff and Starns (2013) showed that individual differences in ROC shape are linked to differences in the RT-confidence relationship in the manner predicted by RTCON (they actually explore an extension of this model that they creatively dubbed RTCON2).

Overall, RT models of confidence ROCs show that the ROC properties used to support conclusions about memory are also affected by decision processes. Thus, developing models with appropriate decision mechanisms is critical for making accurate conclusions about memory, and RT data provide important constraints to test whether or not a model appropriately characterizes the decision process. Many of the critical theoretical advances are still ahead of us when it comes to jointly modeling RT and confidence, and certainly none of the available models have establishing anything like the widespread success of accumulation models for two-choice tasks (see Wagenmakers, 2009). However, there is no reason to doubt that future models will share the critical property of those reviewed here, namely that decision processes can affect the asymmetry and curvature of ROCs.

Bias Manipulation ROCs. RT modeling also plays an increasingly important role in studies exploring ROCs formed from bias manipulations as opposed to confidence ratings (Dube, Starns, Rotello, & Ratcliff; Heck & Erdfelder, 2016; Klauer & Kellen, 2018; Osth, Bora, Dennis, & Heathcote, 2017; Starns, Ratcliff, and McKoon, 2012). For example, Starns et al. (2012) modeled data from a recognition task in which the proportion of studied items was manipulated across test lists in an attempt to encourage a range of response biases. In addition, participants responded under either speed or accuracy stress as a manipulation of response caution, and different levels of memory performance were created by manipulating the number of study attempts and the natural-language frequency of the stimulus words. The diffusion model was able to closely match the response proportions and RT distributions with psychologically meaningful parameter changes, and this success held across the many conditions defined by the factorial combination of the bias, caution,

encoding strength, and word frequency variables. The model matched ROC asymmetry with the same mechanism as the UVSD model; that is, the between-trial variability was higher for studied items than non-studied items.

Much like the confidence-ROC results in the last section, applying an RT model changed critical conclusions about memory processes. For one, the relative variability of memory evidence for studied and non-studied items was quite different in the diffusion results compared to a standard accuracy-only UVSD model fit to the same ROCs, with a more unbalanced ratio for the diffusion model. Evidence variability has figured prominently in several theoretical debates (e.g., Ratcliff et al., 1992; Glanzer et al., 1993; Wixted, 2007; Yonelinas, 2007), so it is important to acknowledge the possibility that conclusions about this quantity will change when a more complete decision model is applied. For another, results showed that speed pressure had the intended effect of making participants less cautious (lower boundary width in the diffusion model), but did not produce more symmetrical ROCs compared to accuracy emphasis conditions. The diffusion model was able to match the ROC asymmetry in both the speed and accuracy conditions, but this result could be challenging for dual-process models. For example, the DPSD model assumes that disrupting recollection should produce a more symmetrical ROC, and speed pressure is often assumed to affect recollection to a greater extent than familiarity (Yonelinas, 2002). As such, developing a version of the DPSD model that is capable of fitting RT distributions is an important step in testing this model against alternatives.

Corroborating ROC Conclusions

As an alternative to jointly modeling RT and ROC data, some recent studies have attempted to use RT modeling as an independent way to validate conclusions from the ROC

literature (Osth et al., 2017; Osth, Dennis, & Heathcote, 2017; Starns, 2014; Starns & Ratcliff, 2014). These studies have focused on measuring the relative variability of memory strength for studied and non-studied items in recognition and source tasks. As noted above, assessing item-to-item variability in memory strength informs a number of central theoretical questions. Notably, assessing variability plays a key role in attempts to distinguish dual process theories positing independent memory systems from "strength" models that base memory decisions on a single combined strength value produced by a coherent memory system (e.g., Wixted, 2007). The diffusion model falls into the "strength" category, because the memory strength for a given item is represented by a single drift rate. Figure 1C shows distributions of these drift rates across items. The studies reviewed in this section applied models in which the variability of drift distributions could vary between studied and non-studied items to determine if RT distributions provide evidence that variability in memory strength is higher for studied items. This finding would bolster "strength" models that match ROC asymmetry using an unequal-variance mechanism as opposed to dual-process models that match the asymmetry by combining decisions based on separate memory systems.

Starns and Ratcliff (2014) approached this task by estimating the between-trial variability in drift rate for the diffusion model (the standard deviation of the distributions shown in Figure 1C; often labeled η). They noted that RT distributions can be used to estimate between-trial variability in memory evidence, but that RT data place only subtle constraint on variability estimates (see also Ratcliff & Tuerlinckx, 2002). Basically, increasing variability in drift rates slightly increases the extent to which errors are slower than correct responses, but this change is quite small relative to the typical variability in

response times. To counteract the high uncertainty in RT-based variability estimates, Starns and Ratcliff (2014) compiled a large set of recognition memory studies that reported RT data. These studies were fit with versions of the diffusion model that had separate between-trial variability estimates for studied and non-studied items. Matching the results from ROC studies, the diffusion fits consistently indicated that between-trial variability was higher for studied than non-studied items. However, the ratio of the two variances was more extreme in the diffusion results than in accuracy-only ROC models (but similar to the ratios obtained by jointly fitting the diffusion model to RT and ROC data simultaneously, Starns, Ratcliff, & McKoon, 2012). The RT-based estimates were consistent with ROC estimates in that the variability ratio remained constant across learning variables like providing extra encoding attempts. However, the RT-based estimates did not show variability differences across natural-language word frequency, in contrast to the large effects of this variable on ROC-based estimates (e.g., Glanzer et al., 1993). In summary, fitting RT distributions supported the same general conclusions about variability in memory strength that are needed to account for ROC data, with a few differences in specific details.

Several subsequent studies replicated the finding that RT-based estimates show higher between-trial variability for studied than non-studied items (Osth, Bora et al., 2017; Osth, Dennis, & Heathcote, 2017; Starns, 2014). Starns (2014) replicated the recognition result and also explored RT-based variability estimates in a source task. The source results showed that evidence variability was higher for a strong than a weak source (defined by a higher or lower number of learning trials, respectively), which matches the results of ROC studies (Yonelinas, 1999). To make sure that the model could accurately estimate drift

variability, Starns also tested a task that allowed for strict control over the strength of evidence available for a given judgment. In the task, subjects quickly decided whether there were under or over 50 asterisks on the screen, and evidence strength was varied by changing the number of asterisks appearing on a given trial (close to 50 = low evidence strength; far from 50 = high evidence strength). Some conditions were designed to have high trial-to-trial variability in evidence strength (high variability in the number of asterisks that appeared) and other conditions had low variability. Diffusion fits accurately discriminated data sets with high and low between-trial variability, which lends credibility to the conclusions derived from applying the model to memory tasks.

Osth, Dennis, and Heathcote developed a likelihood ratio version of the diffusion model, and fits of this new model again showed higher between-trial variability for studied than non-studied items. Osth, Bora, et al. showed that the diffusion model and the linear ballistic accumulator (LBA) model produce different between-trial variability estimates when applied to the same data sets. Notably, they also fit both models to the Starns (2014) asterisk-task data and found that the diffusion model produced more accurate variability estimates than the LBA model (Brown & Heathcote, 2008, see Section 3 for further discussion), suggesting that estimates from the former model should be given higher credibility. Their diffusion results showed higher variability for studied than non-studied items, so this has proven to be a highly consistent result.

Information Format in RT Models

Modeling RT distributions is also beginning to play a role in the "retrieval format" debate contrasting models that assume continuous versus discrete evidence states. Before

discussing the modeling efforts here, we will summarize the theoretical issues at stake. As an analogy to deciding if an event was experienced or not, imagine that the decision task is to determine whether or not a young adult has a disease that does not produce symptoms until later in life. To analogize a discrete-evidence model, we could imagine that this disease is based on a single gene mutation, so people with one allele of a certain gene will get the disease and people with a different allele will not. If a genetic test identifies which allele the patient has, then the patient can be unambiguously categorized as having or not having the disease. Errors are produced only when the genetic test fails to identify the allele, thus providing no information about disease status. To analogize a continuous-evidence model, we could imagine a disease that has imperfect risk factors as opposed to a single marker. Perhaps people with high blood pressure are more likely to develop the disease later in life, but there is considerable overlap in blood pressure between people who will and will not develop the disease. In this scenario, errors are based on the inherent uncertainty in the information used for the decision; for example, sometimes a person without the disease will have very high blood pressure for other reasons.

So the basic question for this section is whether memory retrieval is like a genetic test that either succeeds and gives certain evidence or fails and gives no evidence or like a fuzzy risk factor that provides a continuous range of more-or-less ambiguous evidence states. Discrete-state theorists hold that the former analogy is more appropriate; for example, they commonly apply a two-high-threshold (2HT) model with a *Detect Old* retrieval state that gives certain evidence that an item was studied, a *Detect New* states that gives certain evidence that it was not studied, and a *Guess* state that represents the failure to retrieve diagnostic evidence (e.g., Snodgrass & Corwin, 1988). Continuous theorists hold

that the latter analogy is more appropriate, and they use models with overlapping evidence distributions. For example, the diffusion model assumes that some studied items can have very low memory strengths (i.e., drift rates) and some non-studied items can have very high memory strengths, as represented in the distributions seen in Figure 1C. Distinguishing these possibilities has a number of theoretical implications, such as how measures of memory ability should be corrected for differences in response biases. Another related question is whether memory errors can be produced by misleading retrieval; i.e., is it possible to "remember" things that didn't happen in the same way one remembers real events?. Continuous models allow for this sort of compelling false memory by assuming that non-studied items sometimes produce high memory strength values.

Although discrete-state theorists have considered RT data in the past (Hu, 2001; Kellen, Singmann, Vogt, & Klauer, 2015; Province & Rouder, 2012), discrete-state models that are capable of fitting full RT distributions have only recently appeared (Heck & Erdfelder, 2016; Klauer & Kellen, 2018; Starns, 2018). These models employ different strategies for accommodating RTs. The most direct approach is to simply model separate RT distributions for responses from each underlying evidence state (e.g., Heck & Erdfelder, 2016). For example, one could describe the RT distribution associated with each evidence state using an exGaussian distribution, which has a steep Gaussian "ramp up" on the low side of the distribution and a long exponential tail on the high side. Together, these features make the exGaussian a very good match for most empirical RT distributions (Van Zandt, 2000). The predicted RT distributions from this sort of discrete-state model are mixtures of the distributions for the different evidence states; for example, if 80% of targets called "old" are based on successful retrieval and 20% are lucky guesses, then the RT distribution

for targets called "old" is produced by mixing the *Detect Old* and *Guess* RT distributions with mixing weights of .8 and .2, respectively. Even this simple approach is a substantial theoretical advance over accuracy-only models. For example, Heck and Erdfelder showed that an RT-extended model allows researchers to address new research questions, such as characterizing whether responses based on guessing are slower than responses based on detection. Adding RT also creates a more stringent test of discrete-state models, because the models can only match the RT data if the RT distributions from all conditions can be produced by making different mixtures of the few underlying distributions for each discrete evidence state (Province & Rouder, 2012). A variant of this approach, the discrete race model, allows for the possibility that decision makers will sometimes make a guess response even for items that could have been detected if they waited longer to respond, implementing a speed-accuracy tradeoff (Starns, 2018).

Another approach to modeling RT in a discrete-state model involves specifying finishing-time distributions for each step in the decision tree that maps the process of determining a response (Klauer & Kellen, 2018). For example, one simple way to characterize a discrete-state decision process is to assume that the decision maker first determines if they will be able to detect the item (i.e., whether retrieval will succeed), responds based on detection if it succeeds (e.g., hits the "old" key if the item produces the *Detect Old* state), or selects a guess response if detection fails. One could posit that the time needed to determine if detection succeeds and the time needed to select a guess both follow an exponential distribution, say (Klauer & Kellen, 2018). The decision time for a trial is determined by adding the finishing times for all the component processes; for example, the decision time for a guess in the simple model outlined above would be the sum of two

exponentially distributed random variables representing the time needed to figure out that detection failed and the time needed to select a guess response. Klauer & Kellen (2018) developed a complete framework of discrete-state models using this approach and demonstrated that these models can be used to investigate processing questions, such as questions about the order of component processes in a decision task.

In summary, discrete-state theorists are developing sophisticated approaches to modeling full RT distributions in addition to response proportions. Most of the discoveries still remain on the horizon, but RT data are beginning to play a much larger role in this research area. Future studies will be needed to contrast the different approaches to accommodating RT distributions in a discrete-state model and to thoroughly compare these models to RT models assuming continuous evidence distributions, like the diffusion model (Ratcliff, 1978).

Section 3: Evidence-accumulation process models in memory research

In this section we review three recent developments that, to different degrees, specify the processes that give rise to the inputs to evidence-accumulation models of decision processes. Our coverage is far from exhaustive, and focuses only on episodic memory. Evidence accumulation models have been applied to learning in other areas, most notably Nosofsky and Palmeri's (1997) Exemplar-Based Random-Walk model, which has been used to explain skill acquisition and the development of automaticity (Palmeri, 1997, 1999) and incorporated into Logan's (2002) broader Instance Theory of Attention and Memory.

The first example, Osth, Dennis and Heathcote (2017), took a process-oriented approach to specifying DDM drift rates in recognition memory, assuming that memory

strength is first translated into an estimate of the relative likelihood of the observed match assuming the test item was vs. was not studied (Glanzer, Hilford, & Maloney, 2009). The second example, Cox and Shiffrin (2017), describes a comprehensively dynamic process model of recognition memory, assuming that within-trial variability in decision processing is caused by sampling different features of the test probe over time. Although on the surface resembling the DDM, their approach differs fundamentally in dropping the assumption that the decision process integrates its inputs. Like Osth et al., the inputs are based on log-likelihood estimates, but the estimates approach an asymptote after initially increasing, and so resemble a leaky-integration dynamic (e.g., Usher & McClelland, 2001). Another difference is that the within-trial fluctuations are much less pronounced than the DDM, so their smoother ballistic trajectories more closely resemble nonlinear deterministic accumulation (e.g., Brown & Heathcote, 2005). The final example, Osth and Farrell (2019), completely drops the single-decision-unit architecture and within-trial noise of the DDM, and instead uses racing LBA evidence-accumulation processes (Brown & Heathcote, 2008) to provide a dynamic characterization of free recall as multi-alternative decision making. The seminal SAM global-memory model (Raaijmakers & Shiffrin, 1981, See Chapter 1) assumed memory strength determines each memory trace's success in competing for retrieval, with recall probability described by Luce's choice rule. More recent recall theories (Polyn, Norman, & Kahana, 2009; Sederberg, Howard, & Kahana, 2008) have modeled trace competition as a race between leaky competing evidence accumulators (i.e., Usher & McClelland's, 2001, LCA model). Race processes produce a similar outcome in terms of recall probability to Luce's choice rule (Bundesen, 1983) but, because they can account for decision time based on the winning runner's time, the newer models were

applied to mean RT as well as recall-probability findings. Osth and Farrell used simpler independent race models, where each racer was either an LBA or the same single-barrier diffusion or Wald processes used in the RTCON models discussed previously. The LBA assumes a linear accumulation mechanism that drops within-trial variability in favor of purely between-trial variability; normally distributed in the case of accumulation rates and uniformly distributed for the initial evidence value on each trial, with both independently sampled for each accumulator. Although simple, these race models are capable of providing a comprehensive account of benchmark phenomena in choice RT (for the LBA race see Brown & Heathcote, 2008, and for the Wald race see Tillman, Van Zandt & Logan, in press). The computational tractability of these models makes them ideal for complex applications such as free recall, and their mathematical tractability allowed Osth and Farrell to use powerful Bayesian hierarchical methods to model the full distribution of RT.

Evidence as log-likelihoods

Almost all modern process models of memory share a common feature: they base recognition decisions on the logarithms of estimated likelihood ratios (Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997). The likelihood ratios are based on the estimated probability that a studied item (numerator) or an unstudied item (denominator) would produce a memory strength value that matches the memory strength of the test item. This approach arose from Glanzer and Adam's (1990) proposal that in order to explain pervasive mirror effects in recognition memory – the observation that many performance manipulations produce opposite effects on the hit rates and false alarm rates – signal-detection theory based on memory strength should be replaced with signal detection theory based on log-likelihood ratios.

Figure 4 illustrates how the log-likelihood approach explains one type of mirror effect, the strength-based mirror effect (Stretch & Wixted, 1998), which occurs in a comparison between study-test list conditions that differ in the memory strength of studied (old) items when false alarm rates for un-studied (new) items in the stronger condition are reduced relative to the weak condition. The strength-based mirror effect occurs with a variety of manipulations to create weak and strong conditions (e.g., study time and study repetitions); we will use the example of retention time, where strength is greater for more recently studied lists (Singer & Wixted, 2006). At first the false alarm difference seems paradoxical as new items should have the same memory strength in both conditions. This is illustrated in Figure 4A, where the memory-strength distributions for new items in the weak and strong conditions are identical standard normal distributions, as is conventional in signal-detection analysis. However, on reflection it is clear that the false alarm difference could arise if the rememberer takes account of the retention interval when setting their decision criterion. Otherwise they would too often reject test items for lists studied further in the past or too often accept test times for recently studied lists.

This problem can be solved by making recognition decisions based on the relative probabilities or likelihoods that the observed memory strength arose from a new vs. old item rather than directly based on memory strength. An ideal unbiased decision is based on a criterion where this likelihood ratio is one (or equivalently where the logarithm of the ratio is zero), which corresponds to the memory strength at which the new and old densities cross (as the height of the density is equivalent to the likelihood). Figure 4A indicates these “ideal observer” criteria as a dotted vertical lines, on left for the weak condition and the right for the strong condition. Figure 4B illustrates the densities plotted

as a function of log-likelihood (logarithms are used by convention; the same conclusions hold for raw likelihood ratios), which brings the two ideal decision criteria into registration and produces a mirror effect pattern where the new-strong distribution lies to the left of the new-weak distribution (and so the strong condition has a lower false-alarm rate) while the old-strong distribution lies to the right of the old-weak distribution (and so the strong condition still has a higher hit rate).

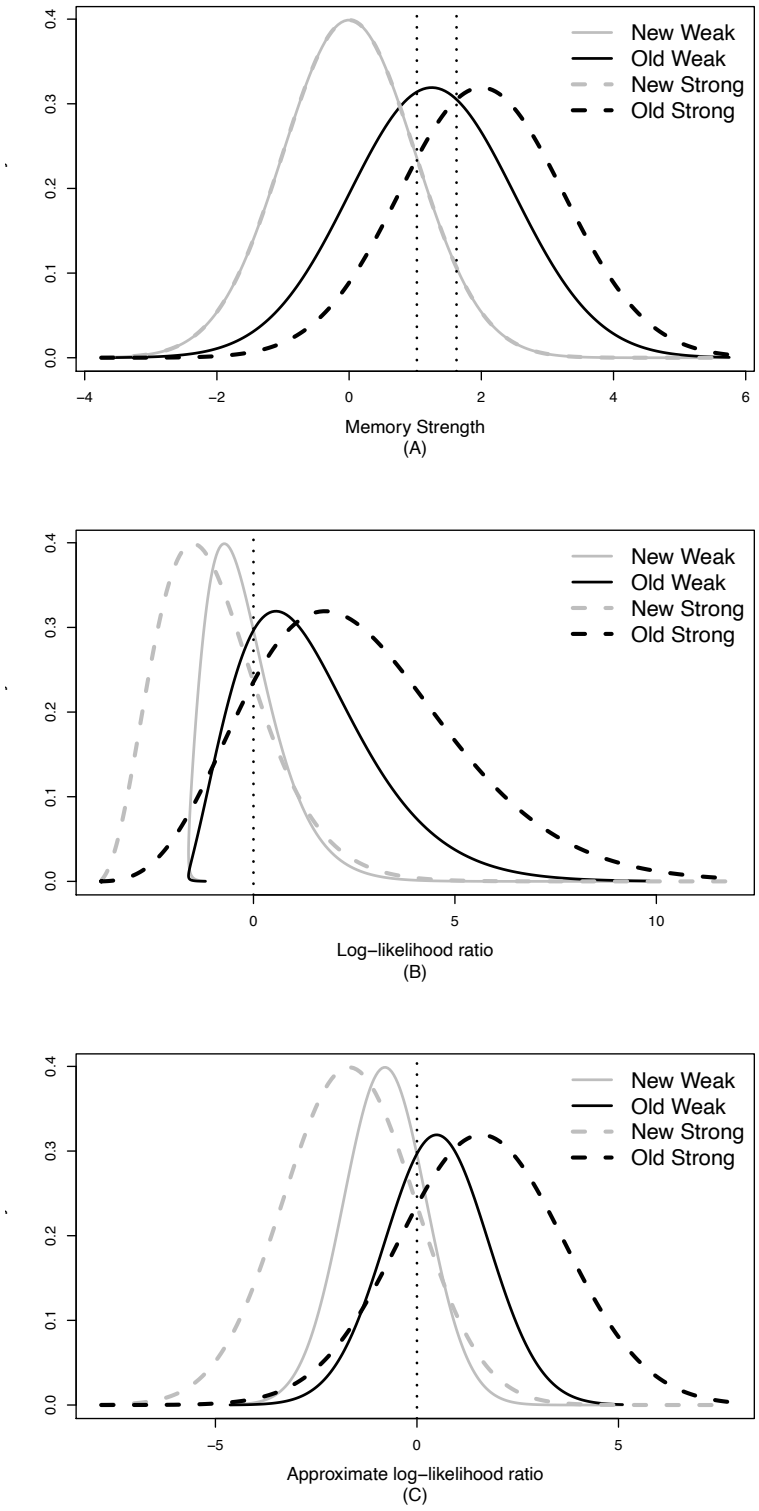

Figure 4. Memory strength and log-likelihood distributions for a list strength paradigm. Decision criteria are indicated as dotted vertical lines.

As summarized by Glanzer et al. (2009), a very large number of studies have confirmed further predictions of likelihood-ratio signal-detection theory, ranging from concentering (a version of the mirror effect applying to two-alternative forced choice), to regularities revealed by ROC analysis, the zROC length effect (i.e., plots of z transformed ROCs are shorter when performance is higher) and the variance effect (zROC slopes indicate greater underlying variance for both studied and unstudied items in conditions with higher performance). In the neurosciences it has been suggested that the evidence accumulated in perceptual-decision tasks corresponds to the log- likelihood of each option in a racing-accumulator model (Carpenter & Williams, 1995) or the log-likelihood ratio in a diffusion model (Gold & Shadlen, 2001).

Osth et al. (2017) applied this approach to modeling recognition memory with the DDM. They noted that if memory strength has a Gaussian distribution and the variance of the old and new distributions is unequal then the log-likelihood ratio is not compatible with the Gaussian trial-to-trial distribution of drift-rates assumed by the DDM because the required transformation is quadratic. This is evident in Figure 4B in the positively-skewed nature of the distributions[1]. However, they developed a linear approximation to the transformation, illustrated in Figure 4C, that both preserves a Gaussian distribution and all of the mirror-effect and ROC regularities documented by Glanzer et al. (2009). They tested the combination of this approximation and the DDM, which they dubbed the LR-DDM by fitting it to recognition memory data from over 150 participants in four recognition

[1] With unequal variance the likelihood transformation is also non-monotonic as evidenced by the left-hand edge of the weak distributions in Figure 3B. This occurs because very low values of memory strength are more likely to arise from old items because of their higher variance. This rather unintuitive property does not arise with Osth et al.'s (2016) linear approximation.

memory experiments. The LR-DDM required fewer parameters to specify drift rates than the standard DDM and provided a good fit to the data that was preferred by Bayesian model selection methods that weight model simplicity along with goodness-of-fit. Their results indicate that log-likelihood based models are an elegant explanation of the regularities of recognition memory – not only in terms of choice accuracy but also in terms of RT – and suggests that it may be useful to use the LR-DDM as a more process-plausible measurement model for recognition memory.

The approximate log-likelihood ratio transformation developed by Osth et al. (2017) has also been combined with the comprehensive process model of memory proposed by Osth and Dennis (2015). In Osth, Fox et al. (2018) it was used to replace the exact but computationally expensive transformation used in the original Osth and Dennis model and found to provide a good account of list-length effects on accuracy data in a source-memory paradigm. Osth, Jansson et al. (2018) combined the Osth and Dennis model with the LR-DDM to investigate the dynamics of recognition testing. The DDM component enabled one account that explains the test-position effect—a decline in performance over the course of testing—in terms of a speed-accuracy tradeoff whereby participants gradually require less evidence (i.e., they lower their thresholds). The model enabled them to partial out threshold effects and to adjudicate between explanations in terms of interference caused by learning of test items and a shift in the context representation used to cue memory caused by test items, and results favored the latter explanation. This application illustrates that accounting for RT using evidence-accumulation modeling can be the key to testing process-model explanations of memory performance.

A dynamic account of recognition memory

Cox and Shiffrin (2017) proposed a model that accounts for both RT and choices and provides a process account of how recognition memory log-likelihoods arise. It shares some similarities with the DDM in that choices correspond to which of two boundaries are crossed by a value that evolves over time, but the value ("familiarity") is a function of the log-likelihood that a test item was studied rather than the integral of that log-likelihood. Episodes are represented as context and content features that are probabilistically encoded in to memory traces. Context features represent times, locations and the participant's internal states. Content features represent perceptual information and knowledge associated with the percept. Recognition testing proceeds by constructing a memory probe, at first containing only the test context, with perceptual and knowledge content features derived from the test item then being probabilistically added over time. The probe is constantly matched in parallel to all memory traces and for each an estimate of the likelihood that it corresponds to the test item is computed. Logarithms of the likelihood values are summed to create a "familiarity" value that fluctuates as the features in the probe change. Recognition decisions are based on familiarity, which begins at a delay after the test item it presented, and which fluctuates over time thereafter due to the changing probe, providing a process explanation of within-trial noise. In particular, decisions are based on the difference in familiarity between its initial and current state, controlling for baseline differences in familiarity between items (Cox & Shiffrin, 2012).

Early in a test trial in a typical study-test list paradigm context dominates the probe, so the best matches are to traces from items on the study list, which for a studied test item correspond to both the target-item trace and non-target item traces. However, as content

features are added non-target matches decrease and matches to target-item traces (both from the study episode and earlier episodes) increase. As a result, familiarity for target items tends to increase over time relative to its initial level when the probe contained only context features and matches to non-target items tend to decrease. In a free-response paradigm, where participants decide when to respond, the familiarity difference is compared to positive and negative thresholds corresponding to studied and non-studied responses. As the probe can only contain a fixed maximum number of features of each type, familiarity approaches a maximum or minimum value. In order to ensure a response occurs the thresholds collapse together at a rate related to the expected rate of change in familiarity over time. In a response-signal paradigm, the probability of a choice is determined by whether the familiarity is above or below a threshold that does not change with time when the response signal occurs.

Cox and Shiffrin (2017) compared different explanations for dynamic recognition phenomena. They first investigated word frequency effects, showing that the model provided a good account of interactions with speed-accuracy tradeoffs—both choice probabilities in response-signal paradigms and choice probabilities and the distribution of RT in free-response paradigms—and supported an account in terms of low frequency words having more distinctive features rather than interference from past episodes. They next found that the model was able to account for response-signal performance in a list-discrimination paradigm on the assumption that the initial features in the probe came from the current context, but with context features representing the list-before-last accruing gradually when this list was the target context. The initial makeup of the probe also accounted for both accuracy and RT in masked-priming paradigms, due to incorporation of

features from the prime and they also found that identity primes could affect the delay before new features were sampled into the probe, which they attributed to pre-activation of the lexical trace of the test item. Their final investigations were of the effects of different types of content features entering the probe at different times. In recognition memory they accounted for non-monotonic false-alarm response-signal functions as due to delayed encoding of plurality and modality information and contrasted it with the early availability of word-form and semantic information corresponding to monotonic false-alarm response-signal functions. In associative recognition they accounted for non-monotonic false-alarm response-signal functions for rearranged pairs in terms of the later entry of association features into the probe because sampling them requires item features in the probe to be sufficiently encoded first. They also explored extensions of the model to explain recall-based phenomena in signal-respond probabilities based on trace-sampling mechanisms similar to those of the SAM global memory model (Raaijmakers & Shiffrin, 1981). In the next section we explore another recent approach to recall that focuses on RT.

Free recall as multi-alternative decision making

Osth and Farrell (2019) avoided the complexities associated with sequential dependencies in full recall series (Farrell, 2012) by modeling only the first free-recall response. They focused on two such phenomena pertaining to serial-position curves (i.e., recall probability as a function of an item's study position), the *recency effect*, an advantage for the later list items, and the *primacy effect*, a weaker advantage for the early items. Their modeling addressed three sets of hypotheses about: 1) the shape of the reduction in memory strength with increased study test lag that mediates the recency effect, either a power or exponential function; 2) primacy being caused by strength, rehearsal, or

reinstatement mechanisms; and 3) whether a recall cue initiates the recall process, or instead whether it can be started before the cue.

Their two modeling frameworks used either standard LBA or shifted Wald accumulators discussed earlier, with the exception that accumulation-rate standard deviations (i.e., for within-trial variation in the Wald and from trial-to-trial variation in the LBA) were a linear increasing function of their means (see also Ratcliff, Voskuilen & Teodorescu, 2018; van Ravenzwaaij, Brown, Marley, & Heathcote, submitted). This modification was necessary to ensure that accumulators with zero rates could not finish a race. Both Wald and LBA models supported essentially the same conclusions for the 14 data sets fit by Osth and Farrell (2019). In most of the experiments the onset of the recall cue was predictable, and there was clear evidence that at least some participants-initiated recall prior to its appearance, as evidenced by negative non-decision time estimates relative to the cue onset.

Their results also clearly supported an exponential recency function, which might appear surprising because recall probability is typically better fit by a power function (e.g., Averell & Heathcote, 2012). However, Osth and Farrell's finding applies to a latent quantity, accumulation rate, that is nonlinearly mapped to recall probability and an exponential function is consistent with the TCM model of free recall (Howard & Kahana, 2002). On the other hand, Donkin and Nosofsky (2012b) also compared recency functions for LBA rates and supported a power over an exponential function in probed item-recognition tasks. Power functions can result from a mixture of exponential functions decaying on different time scales (Brown & Heathcote, 2003) and Howard et al. (2015) suggested that they might emerge if memory is cued at a range of temporal scales. Osth and Farrell proposed this framework could provide a unified account of findings

from different tasks, with their results suggesting participants in the experiments they examined used focused single-scale cues.

The first *strength* account of primacy tested by Osth and Farrell (2019) assumes primacy items receive extra strength that exponentially reduces with increasing serial position, with both primacy and recency gradients influencing the race. The second *rehearsal* account assumes that the first item is sometimes rehearsed through to the end of the list so had the greatest memory strength of any item, because it is effectively at the head of the recency function. The third *reinstatement* model assumes that participants sometimes use a start of list retrieval cue instead of an end of list cue, in which case primacy items race without any competition from the recency gradient. They showed that these three accounts produce very similar effects of serial position on recall probability but very different effects on RT because, in a race model, the RT distributions for each serial position are mostly determined by the fastest competitor, meaning that RTs can be very similar for each serial position despite very different recall probabilities. Consequently, RT changes little with serial position for the strength model. The same is true for the rehearsal model, except for the first position, which is slightly faster in proportion to the probability of rehearsing the first item. For the reinstatement model, in contrast, there is a decrease with serial position, particularly for slower RT, in this case in proportion to the probability of using the start-of-list cue. Model selection results that took advantage of the Bayesian estimation methods to account for both goodness-of-fit and model complexity, favored the context reinstatement model with exponential gradients for both primacy and recency.

Osth and Farrell's (2019) results in long-term recall underline the utility of RT distributions, and hence of evidence-accumulation models, in discriminating among distinct theoretical positions that cannot be discriminated by recall probability. In this

regard they join the results of Donkin and Nosofsky (2012a) and Ratcliff and Murdock (1976) who showed that the shapes of RT distributions provided critical constraints allowing then to adjudicate among different models of short-term recall. These specific results with respect to primacy favor theories such as Farrell's (2012) grouping model that can selectively focus on primacy or recency items at recall initiation. However, they are also compatible with models in the temporal context family, as Osth and Farrell showed that like the independent Wald and LBA race models, RT in the LCA model is also largely determined by the fastest competitor, and models within this family have incorporated context reinstatement mechanisms (e.g., Morton & Polyn, 2016).

Conclusion

Remembering an event is essentially a process of making a series of decisions about what likely happened in the past. Understanding how these decisions are made is just as important as understanding the memory processes that inform them. A rich literature demonstrates that accounting for RT data in addition to response proportions dramatically improves researchers' ability to identify the most effective models of decision making as well as their ability to map empirical results onto psychologically meaningful processes. Evidence-accumulation models of RT data are beginning to play a prominent role in memory research because of the many advantages they offer. In terms of measurement, they provide an important advance over older accuracy-only models by revealing that some data patterns previously attributed to memory processes are likely produced by decision processes. Although it is still far from standard practice, modeling RT distributions from memory tasks has proven to be an effective way to advance theoretical understanding, and the evidence-accumulation models that underpin this enterprise are

becoming integral to process theories of memory. Like rememberers who must piece together clues to build a case about what happened in the past, memory researchers are quickly catching on to the value of dynamic models of decision processes to provide clues about how memory works.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*, 97-123.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81-99.

Batchelder, W. H., & Alexander, G. E., (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). Psychological Bulletin, 6, 1204-1212.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*(4), 548-564.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 587-606.

Brown, S.D. & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments & Computers*, *35*, 11-21.

Brown, S.D. & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117-128.

Brown, S. D., Heathcote, A. (2008). The Simplest Complete Model of Choice Response Time: Linear Ballistic Accumulation. *Cognitive Psychology*, *57*, 153-178.

Bundesen, C. (1993). The relationship between independent race models and Luce's choice axiom. Journal of Mathematical Psychology, 37 , 446–471.

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature, 377*, 59-62

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.

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science*, *4*, 135–150.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, *124*, 795-861.

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*(2), 452-478.

Donkin, C., & Brown, S. D. (2018). Response Times and Decision-Making. In E.-J. Wagenmakers (Ed.), *The Stevens Handbook of Experimental Psychology and Cognitive Neuroscience* (4 ed., Vol. 5).

Donkin, C., & Nosofsky, R. M. (2012a). The structure of short-term memory scanning: an investigation using response time distribution models. *Psychonomic Bulletin & Review*, *19*, 363–394.

Donkin, C., & Nosofsky, R. M. (2012b). A power-law model of psychological memory strength in short- and long-term recognition. Psychological Science, 23 (6), 625–634.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol 38*, 130-151.

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and Language, 67*, 389-406.

Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58-51). Hearing and Communication Laboratory, Indiana University.

Erdfelder, E., & Buchner, A. (1998). Comment: Process-dissociation measurement models: Threshold theory or detection theory? Journal of Experimental Psychology: General, 127, 83-96.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*, 223–271.

Garton, R., Reynolds, A., Hinder, M. R. & Heathcote, A. (2019). Equally flexible and optimal response bias in older compared to younger adults. *Psychology and Aging*, *34*, 821-835. http://doi.org/10.1037/pag0000339

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*, 10-16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *3*, 546-567.

Glanzer, G., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin and Review*, *16*, 431-455.

Gold J. I., & Shadlen M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*, 535–574.

Gombos, V., Pezdek, K., & Haymond, K. (2012). Forced confabulation affects memory sensitivity as well as response bias. Memory & Cognition, 40, 127-134.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Oxford, England: John Wiley.

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-Process Models of Associative Recognition in Young and Older Adults: Evidence From Receiver Operating Characteristics Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 768-788.

Heck, D. W.; Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review, 23*, 1440-1465.

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

Heathcote, A. (2004). Fitting the Wald and Ex-Wald Distributions to Response Time Data, *Behavior Research Methods, Instruments & Computers*, *36*, 678-694.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 268–299.

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. Psychological Review, 122 (1), 24–53.

Hu, X. (2001). Extending general processing tree models to analyze reaction time experiments. *Journal of Mathematical Psychology*, *45*(4), 603-634.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*, London, Academic Press.

Jacoby, L. L., & Hollingshead, A. (1990). Toward a generate/recognize model of performance on direct and indirect tests of memory. Journal of Memory & Language, 29, 433-454.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Kellen, D., Erdfelder, E., Malmberg, K. J., Dube, C., & Criss, A. H. (2016). The ignored alternative: An application of Luce's low-threshold model to recognition memory. Journal of Mathematical Psychology, 75, 86-95.

Kellen, D., & Klauer, C. K. (2015). The flexibility of models of recognition memory: The case of confidence ratings. *Journal of Mathematical Psychology, 67*, 8-25.

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review, 20,* 693-719.

Kellen, D., Singmann, H., Vogt, J., & Klauer, K. C. (2015). Further evidence for discrete-state mediation in recognition memory. *Experimental Psychology, 62,* 40-53.

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology, 82,* 111-130.

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. Psychological Review, 101, 166–171.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*, 496–504.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, *72*(1), 246–273. http://doi.org/10.3758/APP.72.1.246

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376–400. http://doi.org/10.1037//0033-295X.109.2.376

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, *121*(1), 66–95. http://doi.org/10.1037/a0035230.supp

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100-109.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*, 61–79.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724-760.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. Journal of Experimental Psychology: General, 135, 391–408.

Mickes, L., Flowe, H. D., Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. Journal of Experimental Psychology: Applied, 18, 361-376.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. Cognitive Psychology, 78, 99-147.

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language*, *86* , 119–140.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. Psychonomic Bulletin and Review, 15, 465–494

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.

Osth, A., Bora, B., Dennis, S. & Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, *96*, 36-61.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260–311.

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio evidence-accumulation models of recognition memory. *Cognitive Psychology*, *92*, 101-126.

Osth, A. and Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation, *Psychological Review*, *126*, 578-609.

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory_ Data and a global matching model. *Journal of Memory and Language*, *103*, 91–113.

Osth, A., Jansson, A., Dennis, S. & Heathcote. A. (2018). Modeling the dynamics of recognition memory testing with a combined model of retrieval and decision making. *Cognitive Psychology*, 104*,* 106-142.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 324.

Palmeri, T. J. (1999). Theories of automaticity and the power law of practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 543–551.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173-1203.

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*.*

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 1591-1607.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 109*, 14357-14362.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59-108.

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., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, *111*, 333-367.

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

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*, 59-83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*, 697-719.

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., & 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., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks_ Accounting for both magnitude and difference effects. *Cognitive Psychology*, *103*, 1–22.

Salthouse, T. A., & Somberg, B. L. (1982). Isolating the age deficit in speeded performance. Journal of Gerontology, 37, 59-63.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34,* 125–137. doi:10.3758/BF03193392

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33*, 151-170.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34–50.

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. Memory and Cognition, 42, 1357-1372.

Starns, J. J. (2018). Adding a speed–accuracy trade-off to discrete-state models: A comment on Heck and Erdfelder (2016). Psychonomic Bulletin & Review, 25, 2406-2416.

Starns, J. J., Dubé, C., & Frelinger, M. E. (2018). The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models. Cognitive Psychology, 102, 21-40.

Starns, J. J., Pazzaglia, A. M., Rotello, C. M., Hautus, M. J., & Macmillan, N. A. (2013). Unequal-strength source zROC slopes reflect criteria placement and not (necessarily) memory processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1377-1392.

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.

Starns, J. J., & Ratcliff, R. (2012). Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. Psychonomic Bulletin and Review, 19, 139-145.

Starns, J. J., & Ratcliff, R. (2014). Validating the Unequal-Variance Assumption in Recognition Memory Using Response Time Distributions instead of ROCs: A Diffusion Model Analysis. Journal of Memory And Language, 70, 36-52.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. Cognitive Psychology, 64, 1-34.

Starns, J. J., Ratliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 1137-1151.

Starns, J. J., Rotello, C. M**.,** & Hautus, M. J. (2014). Recognition memory zROC slopes for items with correct versus incorrect source decisions discriminate the dual process and

unequal variance signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1205-1225.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

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

Thomas, E., Myers, J. L., 1972. (1972). Implications of latency data for threshold and nonthreshold models of signal detection. *Journal of Mathematical Psychology*, *9*(3), 253–285. http://doi.org/10.1016/0022-2496(72)90018-1

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550-592.

Tanner, W. P. Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

Tillman, G., Van Zandt, T., & Logan, G. D. (in press). Sequential sampling models without random between-trial variability: The racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*.

Verde, M. F., & Rotello, C. M. (2003). Does Familiarity Change in the Revelation Effect? Journal of Experimental Psychology. Learning, Memory & Cognition, 29, 739-747.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 1147–1166.

van Ravenzwaaij, D., Brown, S.D., Marley, A.J. & Heathcote, A. (invited revision). Accumulating advantages: A new approach to multialternative forced choice tasks, *Psychological Review*.

Vickers, D. (1979). *Decision Processes in Visual Perception*. Academic Press.

Wagenmakers, E. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*(5), 641-671.

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

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.