A Circular Diffusion Model of Continuous-Outcome Source Memory Retrieval: Contrasting Continuous and Threshold Accounts

Jason Zhou [a], Adam F. Osth [a], Simon D. Lilburn [a], and Philip L. Smith [a]

[a] Melbourne School of Psychological Sciences, The University of Melbourne

Corresponding Author:

Jason Zhou

Melbourne School of Psychological Sciences

The University of Melbourne

Parkville, VIC 3052, AUSTRALIA

jasonz1@student.unimelb.edu.au

Declaration of Interest: none

Funding sources: This research was supported by Australian Research Council Discovery Grant DP180101686, awarded to Philip Smith and Discovery Early Career Researcher Award DE170100106, awarded to Adam Osth.

Data and model code from this article can be found on our Open Science Framework (OSF) page (https://osf.io/p7sxc/). This experiment was not pre-registered.

Abstract

A circular analogue of the diffusion model adapted for continuous response tasks is applied to a continuous-outcome source memory task. In contrast to existing models of source retrieval that attribute all of the variability in responding to memory, the circular diffusion model decomposes noise into variability arising from memory and from decision processes. We compared three models: 1) a single diffusion process with trial-to-trial variability in drift rate, 2) a mixture of two diffusion processes, one with positive drift that does not vary from trial-to-trial, and a second zero-drift process that represents discrete guessing, and 3) a hybrid model that mixed positive and zero-drift processes with trial-to-trial variability in the positive drift process. Comparison of model fits to joint response error and RT data suggest that a memory strength threshold under which no information is retrieved appears to underlie responding in a continuous report source memory task. Additionally, we also conditioned participants' source responding on their confidence in an old/new recognition task, ruling out the possibility that participant guessing was only due to unrecognized items. Overall, our findings support an all-or-none or some-or none view of source memory retrieval and pose a challenge to continuous models of source memory.

Keywords: source memory, continuous-outcome, diffusion model, response time

Episodic memory is memory for particular events or occasions; more generally, it is memory for items or material in context. An important part of episodic memory is source memory. As the name implies, source memory is memory for the source or the origin of material stored in memory: where, in the larger episodic context, the material was first encountered. A number of models have been developed to explain the source memory task, which make distinct assumptions about source memory retrieval. Based on Signal Detection Theory (SDT), continuous models of source memory claim that memory relies on continuous evidence. In signal detection models, retrieved information may be inaccurate but not absent, allowing for a gradual decline in the quality of information retrieved (Banks, 2000; Mickes, Wais & Wixted, 2009). In contrast, threshold or discrete-state models hold that retrieval fails discretely, and so performance is made up of either precise responses driven by memory, or guesses when the memory is below the retrieval threshold (Batchelder & Riefer, 1990; Klauer & Kellen, 2010).

A third class of models can be regarded as hybrids of continuous and threshold models, and are known as dual-process models, in which different kinds of retrieval mechanism support different kinds of memory (Mandler, 1980). Specifically, dual-process models involve the retrieval mechanisms of familiarity, which is continuous and reflects whether or not a particular item was previously encountered, and recollection, which is thresholded and reflects a richer episodic account of the encounter. Dual-process models propose that memory in an item recognition task is supported by a mixture of familiarity and recollection processes, as both types of information retrieved can inform whether an item has been previously encountered or not. Dual-process models and SDT can make similar predictions about item recognition, particularly when the contribution of familiarity in the dual-process model is high (e.g., Glanzer, Kim,

Hilford, & Adams, 1999; Wixted, 2007). However, dual-process models and SDT make different predictions about the source memory task, which is assumed to not rely on familiarity because familiarity is assumed to not contain context information, and thus cannot distinguish between two sources which have both been encountered. Instead, source memory relies exclusively on recollection, and because recollection is a thresholded process, the dual process model makes the same predictions as discrete state models (Yonelinas, 1999).

These competing models of retrieval from source memory have been difficult to distinguish, partly because evidence from accuracy and confidence alone can be inconsistent or, at best, provides only qualified support for a particular account. One more diagnostic method of distinguishing between accounts is to use a continuous-outcome task. Unlike the more traditional two-alternative forced-choice tasks that are widely used in the study of memory, in continuous-outcome tasks responses are made on a continuous scale. Historically, the continuous-outcome task has its origins in the method of adjustment of classical psychophysics (Woodworth & Schlossberg, 1954), in which sensory thresholds were measured by asking participants to adjust the intensity of a variable stimulus to match a standard. It was reintroduced to modern cognitive psychology by Prinzmetal, Amiri, Allen, and Edwards (1998), who used it to study the effects of attention on perceptual variability. It was first applied to the study of memory by Wilken and Ma (2004), who used it to investigate how the representations of items in visual working memory change with the number of items that stored. It has since become the method of choice for many visual working memory researchers because it provides information about the quality of representations in memory that more traditional two-choice tasks do not (Ma, Husain, & Bays, 2014). The first application of the continuous-outcome task to source memory was by Harlow

and Donaldson (2013) to contrast continuous and threshold accounts of the observed response distributions.

A drawback of using continuous response tasks to study memory is that, although decision models for discrete-choice tasks are well developed (e.g., the SDT models used by Egan, 1958; Banks, 1970; DeCarlo, 2003; Wixted, 2007 and the diffusion model of Ratcliff, 1978), until recently there were no formal models of the decision process in continuous-outcome tasks. In other areas of memory research that have used two-choice tasks, such as item recognition and source memory, Ratcliff's diffusion model has provided a unified account of the retrieval processes that underlie response time (RT) and accuracy, and provided a successful account of both choice probabilities and RT distributions for correct and error responses across a range of conditions that affect performance (e.g., Criss, 2010; Fox, Dennis, & Osth, 2020; Osth, Dennis, & Heathcote, 2017; Ratcliff & Smith, 2004; Starns, Ratcliff, & McKoon, 2012; Starns, 2014; White & Poldrack, 2014). This kind of model-based approach, which accounts for both RT and accuracy, has been useful in helping researchers distinguish between alternatives that are difficult to distinguish based on accuracy alone. Here we present a similar approach to the analysis of performance in a continuous-outcome source memory, which is based on the circular diffusion model of Smith and colleagues (Smith, 2016; Smith, Saber, Corbett, & Lilburn, 2020).

The circular diffusion model provides a characterization of the decision processes that are involved in retrieving items from memory in continuous-outcomes tasks and predicts both distributions of decision times and decision outcomes in such tasks. Using this model allows us to distinguish the contributions of memory and decision processes to source memory performance in a precise way. In two-choice tasks, Ratcliff's diffusion model (Ratcliff, 1978;

Ratcliff & McKoon, 2008) provides estimates of the quality of the information in the stimulus, the amount of evidence needed to make a response, and the time for nondecision processes, The circular diffusion model provides a similar decomposition of the components of processing involved in continuous-outcome tasks. Continuous-outcome tasks yield a measure of the *precision* of responding (roughly, the reciprocal of the variance of the distribution of responses). The circular diffusion model analysis shows that precision depends jointly on the quality of the evidence encoded from the stimulus, represented by the *drift rate* of the diffusion process and the quantity of evidence required for a response, represented by the *decision criterion*. (These quantities are defined more formally below). In most applications of continuous-outcome tasks to the study of memory, the quantity of theoretical interest is not the empirically observed precision, but the latent memory strength that gives rise to it. The estimates of drift rate from the circular diffusion model provide a characterization of memory strength that is not confounded with the effects of decision criterion. Like Ratcliff's (1978) diffusion model of two-choice decisions, drift rates and decision criteria can only be independently estimated if decision outcomes and RT are both measured. In our task, described below, we measured both decision times and decision outcomes.

Evidence from Two-choice Tasks

Traditionally, evidence both for and against a threshold in source memory has come from the examination of Receiver Operating Characteristic (ROC) curves (Yonelinas & Parks, 2007; Yonelinas, 1999; Slotnick & Dodson, 2005). In a two-choice paradigm with two possible sources of information, continuous and threshold models make divergent predictions about the shape of

source ROC curves. The continuous model predicts a curvilinear ROC because each of the two sources is associated with a normally distributed memory strength, which overlap with each other. As the response criterion is varied, the ratio of hit rates to false alarms will be such that the resultant shape of the plot is curvilinear (Slotnick & Dodson, 2005). In contrast, in a threshold model, each source is associated with a memory strength threshold, and where the strength of the memory representation fails to meet either response threshold, no information is retrieved and the response is a guess. The threshold model predicts that the ratio of false alarms to hit rates across criterion points is constant, producing a linear ROC (Rouder, Morey, Cowan, Zwilling, Morey & Pratte, 2008). In examining ROCs for a two-choice source memory task, Yonelinas (1999) found that, although the ROCs for recognition memory performance were curvilinear, the ROCs for source memory performance were linear, indicating that source memory can be well described by a threshold process.

The premise that source memory is thresholded was challenged by a reanalysis of the Yonelinas (1999) data by Slotnick and Dodson (2005), in which they conditioned source performance on recognition confidence ratings for each item. This reanalysis demonstrated that if source ROCs were plotted separately for different levels of confidence reported in the item recognition task, the highest confidence source ROCs were in fact curvilinear, contrary to the predictions of the dual-process model. Performance for unrecognized items was at chance; these items were on the diagonal of the ROC. As items rated with lower recognition confidence were included in the original data, the source ROC became increasingly linear, and more consistent with the predictions of the threshold model. The authors argued that only the items that were recognised with high confidence contained diagnostic source information, and that the linearity

of source ROCs observed by Yonelinas (1999) was an artifact of collapsing across all recognition confidence ratings, and was thus not evidence for a recollection threshold.

Yonelinas and Parks (2007) responded to the Slotnick and Dodson (2005) analysis by proposing that source ROCs are typically linear, but become more curvilinear under a number of conditions. One such condition is when an item and a source are treated holistically as one item, known as *unitised familiarity*, which is continuous. We will return to this point in the Discussion. While this proposal represented a concession towards a continuous contribution under certain circumstances, Klauer and Kellen (2010) were later able to account for curvilinear ROCs using only discrete states by allowing for a variable mapping between recognition confidence ratings and source memory thresholds. At present, then, there is a lack of consensus about whether apparently linear or curvilinear ROCs reflect thresholded or continuous retrieval processes.

Continuous-Outcome Tasks

Harlow and Donaldson (2013) addressed the need for more diagnostic data in the source memory literature by using a continuous-outcome task instead of a two-choice task, which yielded a continuous measure of response accuracy. In the Harlow and Donaldson (2013) continuous report paradigm, source information was provided by a point located on the circumference of a circle, which represented the context, and which was paired with a word item. When later cued with that word, participants were required to reproduce the associated location. This procedure allowed for a continuous measure of the error in the angle between the reported and true source locations. The researchers' use of a continuous measure of source memory performance allowed them not only to measure the accuracy of source memory judgments, but also the distribution of response errors. Instead of categorizing responses as either correct or

incorrect as in a two-choice task, their task, which captures an entire distributions of response accuracy, provides a more detailed picture of trial-to-trial variability in retrieval performance. The additional information in such distributions may be more diagnostic than ROC curves of the underlying retrieval processes. Critically, the threshold and continuous models of source memory make divergent predictions about the distributions of response errors in continuous report tasks.

According to the threshold model, items that fall below the recollection threshold will result in guesses, which will be distributed uniformly across all possible response options. Items that exceed the threshold and are successfully retrieved will cluster, with some error, around the true value of the item source. As discussed earlier, this two-process account of continuous report performance parallels similar proposals in the visual working memory literature, like the one of Zhang and Luck (2008) who used a two-component mixture model comprised of a von Mises distribution and a uniform distribution to argue for an item-capacity-limited memory model. The von Mises distribution is a circular analogue of the Gaussian distribution and, like the Gaussian distribution, has a bell-shaped density function. Items in memory are represented with high accuracy and responses to them follow a von Mises distribution; items not in memory lead to guessing and responses to them follow a uniform distribution.

Harlow and Donaldson (2013) took a similar approach in modeling performance in their source memory task, using a wrapped Cauchy distribution to characterize the shape of the marginal distribution of response errors when items exceeded the retrieval threshold. The wrapped Cauchy distribution differs from the von Mises distribution in that its shape is more leptokurtic, with a higher peak and heavier tails. A mixture of a wrapped Cauchy distribution and a uniform distribution produces a high-peaked, heavy-tailed distribution (Harlow & Donaldson,

2013). Harlow and Donaldson (2013) found that source accuracy data were better fit by the threshold model better than by its continuous counterpart, which assumed that all responses follow a single wrapped Cauchy distribution, and which predicts that responses made with moderate memory strength would result in a wider spread of responses around the true location without a uniformly distributed guessing component. Continuous models can account for discrete failures by assuming that a proportion of items are not encoded (e.g., DeCarlo, 2002, 2002). For this reason, the Harlow and Donaldson (2013) paradigm also included a study-test delay manipulation. Because the delay between study and test was not predictable at the encoding stage, an encoding failure account predicts that the proportion of guesses arising from failures must be equivalent across delay conditions. Contrary to this prediction, the authors found that the probability of retrieval was higher for short delays than long delays, ruling out the encoding failure account.

Source Memory for Unrecognized Items

Although Harlow and Donaldson's (2013) method represents an innovative way to characterize the retrieval processes in source memory tasks, a limitation of their approach was that it did not distinguish between recognition failures and source retrieval failures. As mentioned previously, Slotnick and Dodson (2005) showed how source memory ROC shapes depend on recognition confidence in the two-choice paradigm, in that unrecognized items have no source discriminability, and it is possible that continuous source memory judgments are affected in a similar way. Hautus, MacMillan, and Rotello (2008) modeled performance in two-choice source memory tasks using a multivariate signal detection model like that of Banks (2000) and obtained better fits to data by including a process in which source memory retrieval is

not even attempted when an item is not recognized and the person simply guesses. Their findings mirror those of Slotnick and Dodson (2005) who found that source performance was at chance for items recognized with low confidence.

A lack of source discriminability for unrecognized items has been replicated numerous times (Bell, Mieth, & Buchner, 2017; Malejka & Broder, 2016; Onyper, Zhang, & Howard, 2010; but see Fox & Osth, 2020 for exceptions), although these studies often employed a procedure where item and source ratings were obtained in the same test trials. When item recognition and source memory tests were in separate blocks, Osth, Fox, McKague, Heathcote, and Dennis (2018) observed reliable source memory for unrecognized items, but discriminability was still close to chance-level accuracy ($d' \sim 0.1$).

If the lack of source memory for unrecognized items generalizes to continuous report tasks, then the contribution of guesses from the unrecognized items would result in a heavy-tailed error distribution, which would not necessarily reflect a threshold in memory retrieval but might simply reflect a state in which source retrieval was not attempted. In the context of the findings of Harlow and Donaldson (2013), this account of apparent guessing behavior predicts that if unrecognized items are excluded, the heavy tails in the error distribution will disappear, and that a continuous model will be preferred in account for source performance. An aim of our study was therefore to investigate a continuous-report measure of source-memory performance conditional on the accuracy of previous recognition judgments. In order to do this, we must consider the mapping between retrieved information and observed responses through the lens of a decision model.

Insights from Models of Decision-Making

In completing a source or recognition memory task, not only do participants need to retrieve information from memory, they must also make a decision on how to respond based on the information retrieved (Ratcliff, 1978). Much of the existing body of source memory research, particularly in the continuous report paradigm, lacks an explicit account of properties of the decision process. Past research in the recognition memory literature has shown that when the properties of decision processes as well as RTs are taken into account, the kinds of conclusions that can be made about episodic memory differ from those made when decision-making is not explicitly considered (Dube, Starns, Rotello & Ratcliff, 2012; Osth, Bora, Dennis, & Heathcote, 2017; Osth & Farrell, 2019; Ratcliff & Starns, 2013; Starns, Ratcliff, & McKoon, 2012).

Diffusion models have emerged as increasingly influential accounts of decision processes which predict both RT and response accuracy, and which naturally explain well-documented phenomena like the speed-accuracy trade-off (Ratcliff, Smith, Brown & McKoon, 2016). Diffusion models have also been used extensively in the past to model memory retrieval, and more recent research has proposed general theories of memory and decision-making in which decisions about stimuli within recognition memory and visual working memory are made using a diffusion process (e.g., Nosofsky, Little, Donkin, & Fific, 2011; Osth, Jansson, Dennis, & Heathcote, 2018; Smith & Ratcliff, 2009). In the most common form of the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008), the decision process is modeled as noisy evidence accumulation between a pair of absorbing boundaries that represent the decision criteria for the task. Evidence is accumulated until the process reaches one or other boundary: The first boundary reached determines the response and the time to first reach a boundary is the decision time component of RT. The rate of evidence accumulation on any trial is known as the drift rate.

The drift rate often varies across conditions, with high drift rates resulting in high accuracy and fast RTs, while lower drift rates result in slower and less accurate responses (Ratcliff, Smith & McKoon, 2015). The diffusion decision model in shown in Figure 1.

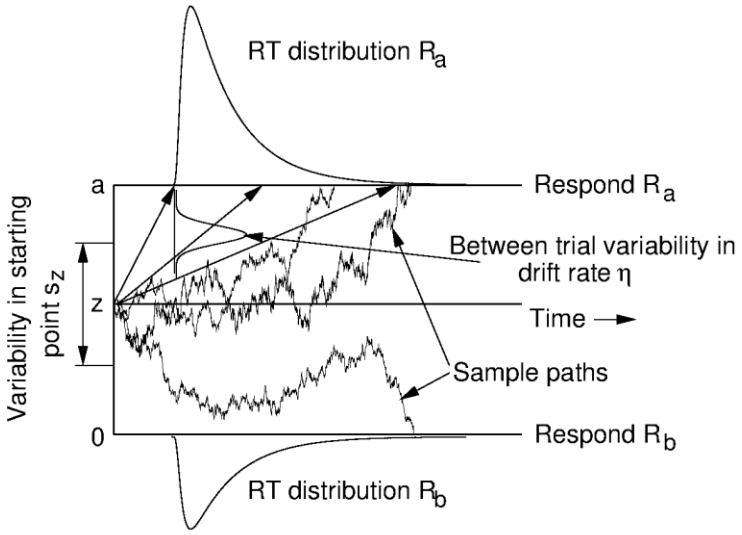

Figure 1. Diffusion decision model. Evidence is accumulated by a Wiener diffusion between a pair of absorbing boundaries that represent the decision criteria for responses R_a or R_b. The starting point is *z* and the boundaries are located at 0 and *a*. The first boundary reached determines the response and the time taken to reach it determines the decision time. The rate at which evidence accumulates is the drift rate, which is normally distributed across trials with standard deviation η.

The diffusion model assumes that multiple sources of variability affect the decision process, including moment-to-moment variability in the accumulation of evidence and trial-to-trial variability in the quality of evidence entering the decision process. The moment-to-moment variability reflects the noisiness of the evidence provided by the retrieval process, while the trial-to-trial variability in drift rate reflects differences in the stimulus information on which the decision is based. When the drift rate varies between trials, then the mean RT for correct responses will be shorter than mean RT for errors. This is because most error responses come from trials with low drift rates, which have long RTs, while most correct responses come from

trials with high drift rates, which have short RTs. Without this variability, correct and error RT distributions will be the same; it is only when drift rates vary between trials that the model predicts this relationship. This phenomenon, known as a *slow error* pattern, has been reliably observed when decision making is difficult (Luce, 1986; Ratcliff et al., 2016) and is frequently observed in recognition memory tasks (Osth et al., 2017; Ratcliff & Smith, 2004).

The circular diffusion model, of Smith (2016) extends the two-choice diffusion model of Ratcliff (1978), which represents decision-making as a one-dimensional evidence accumulation process (diffusion on a line), to account for continuous report tasks. In the circular model, the drift rate is defined as a vector in a two-dimensional (2D) space. As shown in Figure 2, the drift rate vector has a direction, or *phase angle*, that represents the encoded stimulus identity, and a length, or norm, which represents the encoded stimulus quality. When a response is made, the magnitude of the drift vector determines RT in the same way as does the scalar drift rate does in the standard Ratcliff (1978) model, while the point at which the evidence accumulation process exits the circle determines the response outcome.

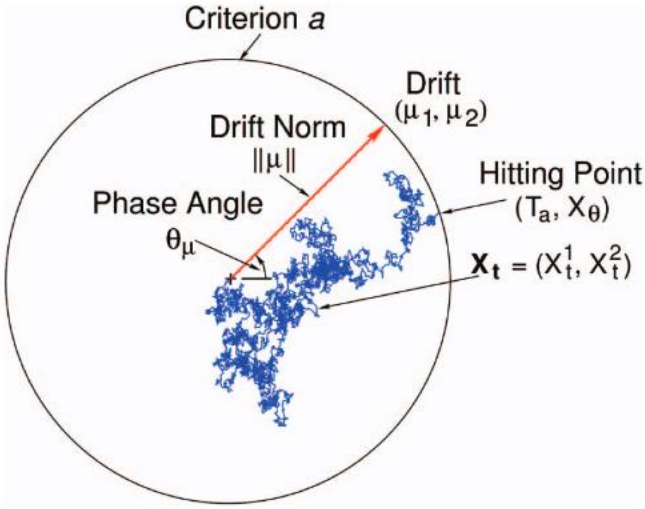

Figure 2. Circular diffusion model of continuous report. Evidence is accumulated by a two-dimensional Wiener diffusion on the interior of a disk, whose bounding circle, of radius *a*, represents the decision criterion. Evidence is accumulated starting at the origin until the process hits the bounding circle. The hitting point, X_θ, is the decision outcome and the hitting time, T_a,is the decision time. The drift rate is vector-valued and consists of two components, (μ_1, μ_2), which jointly specify its magnitude and direction. In polar coordinates the magnitude is represented by the drift norm $\|\mu\|$ and direction is represented by the phase angle θ_μ The noisy sample path represents evidence accumulation on a single experimental trial. From P. L. Smith (2016). "Diffusion theory of decision making in continuous report' *Psychological review*, *123*, 425-451. Figure 2. Copyright American Psychological Association.

The properties of the circular diffusion model closely parallel those of the two-choice diffusion model. When the only source of variability in the model is moment-to-moment variability in the evidence accumulation process the model predicts that decision times will be the same for all decision outcomes. When there is across-trial variability in drift rates, the model predicts that accurate responses will be faster than inaccurate responses. When there is across-trial variability in decision criterion, represented by variability in the diameter of the bounding circle, the model predicts that accurate responses will be slower than inaccurate responses. These properties are continuous counterparts of the slow error and fast error properties predicted by the

two-choice diffusion model with across-trial variability in drift rate and starting point, respectively.

Mathematically, the circular diffusion model predicts that, for a fixed drift rate and decision criterion, the distribution of decision outcomes will follow a von Mises distribution. The spread of decision outcomes predicted by the von Mises distribution depends on a precision parameter, κ. Precision is the inverse of variance: high precision represents low variance and vice versa. The von Mises precision predicted by the circular diffusion model is jointly a function of the drift norm, ||μ||, the decision criterion, *a*, and the noise in the evidence accumulation process, σ^2. Specifically,

$$\kappa = \frac{a\|\mu\|}{\sigma^2}.$$

In words, this equation says that precision is equal to the quality of the information in the stimulus, represented by the drift norm, multiplied by the amount of evidence required for a response, represented by the decision criterion, divided by the noisiness of the evidence accumulation process. This relationship between precision, strength of evidence and the decision criterion is a key feature of the circular diffusion model which motivates our application of the model to the source memory task. As previously discussed, the analytic decomposition of precision in the preceding equation allows to distinguish the effects of evidence quality and decision criterion on precision in fits of the model to data.

Smith (2016) showed that while a single, fixed drift rate results in a von Mises error distribution, trial-to-trial variability in drift rate results in a peaked, high-tailed error distribution, similar to those found by Harlow and Donaldson (2013) and in the visual working memory literature (Zhang & Luck, 2008). The peaked high-tailed distributions are the result of mixing

trials that have high and low drift rate norms. High and low drift rates lead to error distributions with high and low precision, respectively. Mixtures of high and low precision von Mises distributions lead to peaked, heavy-tailed distributions like those found experimentally (van den Berg, Awh, & Ma, 2014). Harlow and Donaldson (2013) interpreted these kinds of distributions as evidence of an underlying memory retrieval threshold. Following van den Berg, Shin, Chou, George and Ma (2012), we consider whether heavy-tailed distributions may instead reflect a mixture of trials with high and low drift rates.

We propose a model which accounts for the leptokurtic source error distribution observed by Harlow and Donaldson (2013) through mixture of high and low drift rates in a single diffusion process, reflecting variability in memory precision, which we refer to, using the terminology of van den Berg et al. (2014), as a *variable-precision* diffusion model. This model is an analogue of the purely continuous SDT models of source memory (e.g., Banks, 2000) in that the only source of evidence in the model is continuously varying drift rates. We compared our variable-precision model to a *threshold diffusion* model, which is a circular diffusion analogue of the threshold model which has two diffusion processes: a memory-driven state with positive drift norm that occurs with probability π, along with a second guessing state that operates if the memory-driven process fails (probability $1 - \pi$). The guessing state differs from the memory-driven state by virtue of it having a drift norm that is zero. When the drift norm is zero, the diffusion process "wanders" randomly around the circle until it terminates at a decision boundary.

Both models are able to account for the heavy-tails in the distributions of response error. However, the models differ in two important respects. First, the variable precision model

specifically predicts that heavy-tailed distributions should be accompanied by slow errors. The threshold model is more flexible than the variable precision model because it does not make *a priori* predictions about the relationship between RTs and the distribution of response errors. However, this comes at the cost of introducing additional parameters, namely the mixing probability parameter π along with different response boundaries for information-driven accumulation (a_1) and guessing accumulation (a_2). For this reason, we carried out model selection using the Bayesian Information Criterion (BIC), which penalizes complexity, where complexity is measured in the number of parameters. The BIC provides a principled way to determine whether the improvement in fit provided by the threshold model over the simpler variable precision model is sufficient to offset the greater penalty associated with its increased flexibility.

The Present Study

Our study had two main aims. The first was to investigate whether the heavy-tailed distributions found by Harlow and Donaldson (2013) could have been the result of guessing about the source of unrecognized items. We did this by investigating source memory performance conditional on confidence in the recognition task. If heavy-tailed distributions of errors are due to source guessing on unrecognized items, then they should be eliminated on trials on which recognition confidence is high. In order to make our results comparable to those of Harlow and Donaldson (2013), we presented the source and item information consecutively, in successive frames of the display, in the same way as they did.

Our second aim was to use the circular diffusion model to implement and compare continuous and threshold models of source memory performance. Because the model predicts

distributions of decision times and decision outcomes, it has the potential to provide a more constrained and more diagnostic comparison of continuous and threshold models of memory than is possible using distributions of decision outcomes alone.

Our experimental task also included a manipulation of the imageability and concreteness of the stimulus words, as rated on the MRC Psycholinguistic Database. Harlow and Donaldson (2013) selected words for low ratings on both metrics to prevent participants from visualizing a concrete object in a source location. We drew our stimuli from pools of words of low and high imageability and concreteness to investigate whether they affected source memory, but we found their effects were negligible, and so we collapsed stimuli into a single pool in our modeling of the data.

Method

Stimuli and apparatus

Stimuli were presented on a 20'' Dell 2009W LDC Monitor with a screen refresh rate of 60 Hz. Software written in MATLAB using PsychToolbox controlled stimulus presentation and recorded responses. Stimuli consisted of words generated from the MRC Psycholinguistic Database, selected/ for low concreteness (minimum 100, maximum 456) and imageability (minimum 100, maximum 481) in the low stimulus set, and high concreteness (minimum 543, maximum 611) and high imageability (minimum 545, maximum 609) in the high stimulus set. Words were displayed in size 24 point "Courier New" white font positioned in the center of a uniform mean luminance field.

Participants

Twenty participants were recruited online through the University of Melbourne SONA system. Each participant was expected to complete four 60-minute sessions, for which they were paid 12 AUD at the completion of each session. One participant who did not complete all four sessions was excluded from analysis ($N = 19$). A small-N design was chosen for this study because by collecting 720 trials over the course of the four sessions, we were able to have a large number of observations for each participant in the sample. Each observation can be thought of as a replication of an effect for that participant (Smith & Little, 2018). In this sense, experimental power was concentrated at the level of the individual participants, rather than at the level of the experimental sample. All participants were provided with plain language statements and consent forms, and gave informed consent prior to data collection.

Procedure

Participants completed the experimental tasks over four sessions, Each of the four sessions consisted of 180 trials, which was broken up into 18 blocks of 10 items each. Blocks were comprised of a study phase, followed by a test phase. In the study phase, participants were presented with a black cross positioned on a randomly generated angle on a dark gray outline of a circle at the start of each trial for 600 ms. The presentation of the cross was followed by the display of a word in the center of the screen for 1500 ms. Locations and words were presented serially in accordance with the experimental design of Harlow and Donaldson (2013) to allow for direct comparison of data. To ensure that participants attended to the source information, they were instructed to indicate the previous location of the cross on the blank target circle using a computer mouse. Responses made within 6 degrees of the true target location were classified as

attended and advanced participants to the next item. Responses further away were deemed unattended and the words "TRY AGAIN" was displayed for 1000 ms, then the location was then re-presented for 250ms, and the verification task was repeated. Participants were then instructed to complete a distractor task, which involved 30 seconds of arithmetic problems, which involved summing two double-digit integers, with participants entering the solution using the keyboard. Following this, they were shown a scrambled list of 10 previously studied items and 10 foils and asked to rate each item on a six-point Old/New confidence scale. Finally, in the source memory retrieval task, participants were cued with the words for 1500 ms, and then indicated the recalled location by a clicking a mouse on the circumference of a gray response circle. Only previously presented items from the study phase were used as cues in the source memory task. There was no time limit on the decision task. A schematic for one trial in each of the phases is shown in Figure 3.

Figure 3. Schematic of display presented to the participant in one trial in each phase of the experiment. There was also an arithmetic distractor task between the encoding and recognition phases of the block, which is not shown. The mouse cursor is shown in the center of the "Source Task" panel to illustrate the procedure. In the actual experiment, the cursor was hidden from the participant and replaced with a red dot four pixels in diameter.

Results

The results are presented in four parts. First, we tested whether individual participants' responses in the source retrieval task were above chance. As responses were made on a

continuous scale, above-chance performance translates into a deviation from uniformity in responding. We did this in order to distinguish participants who were responding at chance from those who showed better-than-chance source memory performance. Second, we investigated source memory judgments conditioned on the prior recognition response and show that conditioning source responding on successful recognition does not fully account for the heavy tails in the distribution of source memory accuracy. Third, we fit a two-component memory-plus-guessing model, like that of Zhang and Luck (2008), to the marginal distributions of response accuracy, conditioned on recognition performance, and show that recognition affects the precision of the source information that is retrieved and not the proportion of guessing responses. Finally, we fit continuous and threshold versions of the circular diffusion model to the joint distributions of RT and accuracy. This analysis generalizes the Zhang and Luck (2008) mixture model analysis to account for both speed and accuracy. The data and model code used are available at https://osf.io/p7sxc/.

Data Screening

Preliminary inspection of the data suggested that some participants performed the source retrieval task with very low accuracy. A Rayleigh test for uniformity, shown in Table 1, identified two participants whose data showed no evidence for a departure from uniformity in at least one condition, which can be interpreted as completely random responding (Fisher, Lewis, & Embleton, 1993). These participants will be referred to as a *low response accuracy* subgroup. Although we retained this subgroup in our analysis, we expected that the data from the remaining

high response accuracy group would be more diagnostic for the purposes of distinguishing between the models.

Table 1
Rayleigh Test for Uniformity for Source Memory Response Error

Participant	$\chi^2(2)$	*p*
1	0.33	.85*
2	559.67	<.01
3	209.21	<.01
4	655.19	<.01
5	361.31	<.01
6	47.33	<.01
7	420.98	<.01
8	944.04	<.01
9	9.84	.01
10	910.91	<.01
11	147.08	<.01
12	526.34	<.01
13	1.88	.39*
15	6.47	.04
16	339.96	<.01
17	12.90	<.01
18	459.85	<.01
19	53.99	<.01
20	40.96	<.01

Note. * *p* values greater than 0.05, indicating no evidence of a departure from uniformity for participants 1 and 13.

Source Memory for Unrecognized Items

The data for each participant were split into three categories on the basis of their confidence in the recognition task. Items rated three and below were deemed unrecognized; successful recognition was defined by ratings of four and above. Recognized items receiving the maximum rating of six were further classified as recognized with high confidence, while items recognized with confidence ratings of four or five were classified as recognized with low confidence. Figure 4 shows the frequency of response errors across all participants grouped according to these categories of confidence in the recognition phase.

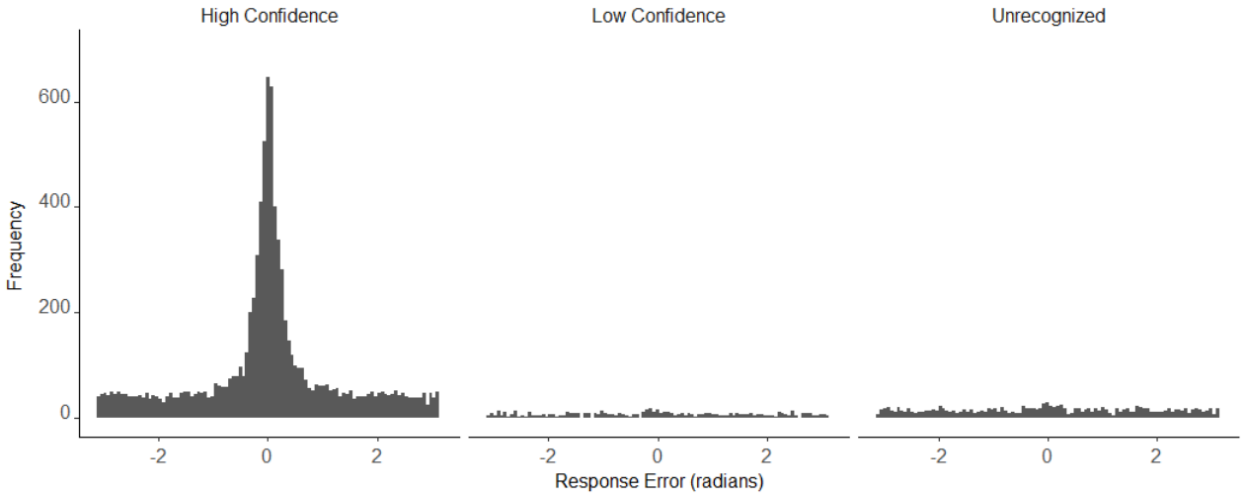

Figure 4. Frequency of angular response error in the source retrieval task, collapsed across participants. The subset of unrecognized items (rated three or below on the six-point confidence scale) yields source responses which are uniform, indicating no source memory for unrecognized items. However, the frequency of unrecognized items was relatively low and the exclusion of these items did not eliminate the heavy tails from the distribution of responses for recognized items.

The results of the Rayleigh tests for unrecognized items are displayed in Table 2. The distributions of these responses were uniform for most participants, indicating that no source memory was present when recognition confidence was low. In the case of Participant 11, where the unrecognized data statistically varied from uniformity, there were few trials classified as unrecognized (8.06% of trials).

Table 2
Rayleigh Test on Source Memory for Unrecognized Items

Participant	$\chi^2(2)$	*p*
1	2.23	0.33
2	2.76	0.25
3	4.47	0.11
4	0.49	0.78
5	0.12	0.94
6	1.74	0.42
7	0.26	0.88
8	1.67	0.43
9	0.21	0.90
10	3.91	0.14
11	11.25	<0.01*
12	3.10	0.21
13	0.98	0.61
15	4.67	0.10
16	4.35	0.11
17	0.29	0.86
18	5.67	0.06
19	4.29	0.12
20	0.22	0.89

Item Recognition Performance

Welch's *t*-test applied to individual-level hit rates for high and low imageability conditions (Table 3) indicated that there was no significant difference in hit rates across the two conditions $t(35.85) = .68$, $p = .503$. Coupled with the Rayleigh test on unrecognized items, this suggests that although source performance for unrecognized items was uniform, the majority of items were successfully recognized, and so guessing due to recognition failure does not fully account for the heavy tails of the error distributions.

Table 3
Item Recognition Hit Rates and False Alarms

Condition	Hit Rate		False Alarms	
	M	*SD*	*M*	*SD*
High Imageability	.87	.14	.14	.10
Low Imageability	.84	.13	.14	.11

Mixture Model

Following Harlow and Donaldson (2013), we used the Zhang and Luck (2008) two-component mixture model to fit the marginal distribution of response error. The model had two free parameters, one for the von Mises precision, κ, which described the spread of responses around the true location, and a mixing parameter π, which described the proportion of trials which were driven by information in a von Mises distribution, as opposed to guesses in a uniform distribution. The best-fitting parameters of the mixture model to the response accuracy data at a group level are shown in Table 4. The parameter estimates for the individual participants are included as supplementary material.

The main difference across the levels of item confidence was that the proportion of guesses increased as confidence decreased, although there was also an increase in the estimated von Mises precision parameters. Changes in precision with confidence are not consistent with a dual process model, which claims that only the probability of remembering a stimulus should change with item confidence. Rather than reflecting a real change in precision, the estimates in Table 4 are likely a reflection of the fact that precision becomes unidentifiable when the proportion of memory-based responses approaches zero because there is little memory information to constrain it.

Table 4

Parameter Values for Best Fits of the Mixture Model to Source Accuracy Conditioned on Recognition Confidence.

Recognition Rating	Low Imageability		High Imageability	
	κ	π	κ	π
High (6)	22.43	0.48	22.89	0.51
Low (4-5)	42.43	0.09	49.73	0.07
Unrecognized (1-3)	51.97	0.06	11.30	0.07

Note. κ refers to the precision of the information-driven retrieval process. π represents proportion of responses driven by information.

Circular Diffusion Models

Unlike the Zhang and Luck mixture model, the circular diffusion model predicts distributions of response error and RT. As discussed earlier, the estimates of drift rate obtained from fits of the model allowed us to decompose precision into two components. One component, the drift norm, characterizes the quality of the information retrieved from memory. The second

component, the decision criterion, characterizes the amount of evidence used to make a response. When there is across-trial variability in drift rate norm, the circular diffusion model can predict heavy-tailed distributions of errors (Smith et al., 2020), like those predicted by the variable-precision model of visual working memory of van den Berg et al. (2014) and one of our aims in fitting the model was to investigate whether such a model could account for the distributions of errors in our data without the addition of a discrete guessing process. As we observed no difference between the summary statistics for the high and low imageability conditions, we combined data between these conditions together and fit the model to the pooled data set. We tested three different versions of the circular diffusion model, which embodied different hypotheses about the process of memory retrieval, as reflected in the evidence entering the decision process. Firstly, the variable-precision model was analogous to the continuous model of source memory presented in Harlow and Donaldson (2013) and was implemented as a circular diffusion model with across-trial variability in drift rates. The model is like a continuous signal-detection model of source memory, in that there is no threshold for memory retrieval, but we have avoided using that terminology to avoid confusion with the continuous nature of the task itself. Mean drift rate was described by the parameter $\boldsymbol{\mu}$, which followed a bivariate normal distribution with independent components (Smith, 2019), with standard deviation $\boldsymbol{\eta} = (\eta_1, \eta_2)$. When predictions for the model are generated in a canonical orientation, in which the drift rate vector points along the positive x-axis and a response at the point $(a, 0)$ is made with zero error, then the horizontal component of drift rate variability, η_1 , represents across-trial variability in stimulus quality and the vertical component, η_2, represents variability in stimulus identity. We considered a number of alternative models of drift-rate variability, including one in which the

two components were equal, but we found the best (most parsimonious) model was one in there was variability in η_1 only and variability in η_2 was negligible. Smith et al. (2020) reported similar results from fits of the circular diffusion model to data from a continuous-outcome perceptual task requiring decisions about the hues of noise-perturbed color patches. We report the fits of this version of the model only and denote the drift rate variability parameter as η without the subscript.

The decision criterion was represented by a. We considered models with and without across-trial variability in criterion. Variability in criterion allows the model to predict a continuous version of the fast-error property of the two-choice diffusion model (Smith, 2016), in which the least accurate responses are also the fastest. We found that criterion variability produced no systematic improvement in fit, so we have omitted it from the models we report here.

Finally, there was a nondecision time parameter, T_{er}, and nondecision time variability s_t. Like the standard diffusion model, the circular model assumes that RT is the sum of the decision time and a time for other (encoding and response) processes. We used the onset of the response circle to begin timing in the source retrieval task, but participants may have started to retrieve information prior to this point, during the display of the cue word immediately prior to the response circle. For this reason, we allowed T_{er} to be negative in the model fits to allow for the premature onset of the retrieval process. In addition, similar to the full diffusion model, there is variability in nondecision time, which is sampled from a uniform distribution with range s_t.

The second model embodied the thresholded retrieval property preferred by Harlow and Donaldson (2013). This threshold diffusion model was implemented as a mixture of two

diffusion processes: one with positive drift rate and no across-trial drift variability, and a second that was modeled as a diffusion process with zero drift rate. The zero-drift process provides a diffusion process implementation of a guessing process, in which the decision process is driven only by noise (Smith et al., 2020). Like other guessing models, the model predicts that responses will be uniformly distributed on the circle, but unlike such models, the zero-drift process predicts both responses and RT. This model had six free parameters. The mean drift rate parameter was shared with the variable-precision model (μ), with the same interpretation, as well as the non-decision time, $T_{er,,}$ and non-decision time variability, s_t. The mixing proportion between information-driven and guessing processes was represented by π. The decision criterion was estimated separately for the information-driven component (a_1) and the guessing component (a_2).

The third model was a combination of the variable-precision and threshold diffusion models. It assumed a mixture of zero-drift and nonzero-drift processes, like the threshold diffusion model, but also allowed for across-trial variability in drift rates, similar to the variable recollection dual process model of Onyper, Zhang, and Howard (2010) which posits that recollection is continuous. This model, which we name the *hybrid diffusion model*, incorporates both the variable-precision and threshold diffusion models in that it had a mixture of information-driven and guessing processes as well as trial-to-trial drift rate variability. This model had seven free parameters, which are shown in Table 5.

Table 5
Symbols and definitions of free parameters estimated in diffusion model variants

		Inclusion in Model		
Symbol	Parameter	Variable-Precision	Threshold	Hybrid
μ	Mean drift	Y	Y	Y
η	Drift variability	Y	N	Y
a_1	Decision criterion, information-driven component	Y	Y	Y
a_2	Decision criterion, guessing component	N	Y	Y
π	Mixing proportion	N	Y	Y
T_{er}	Non-decision time	Y	Y	Y
s_t	Non-decision time variability	Y	Y	Y

Note. Not all parameters were estimated for all three models. The variable-precision diffusion model did not include a mixed guessing process, and therefore lacked *a2*, $\pi 1$ and $\pi 2$. The threshold diffusion model did not have drift variability and lacked η. The hybrid diffusion model included all seven parameters.

The three variants of the circular diffusion model were each fit using maximum likelihood estimation to data on trials that were highly recognized (rated four or higher in the item recognition phase) at an individual level. We excluded trials on which RT exceeded five seconds, and then excluded trials that were extremely fast or slow for each participant, defined as being beyond three standard deviations of the median RT for that participant. These two steps excluded 4.25% of the total number of responses. We compared models using the Bayesian Information Criterion, defined as

$$\text{BIC} = -2LL_{\max} + m \log N,$$

where LL_{max} is the maximized log-likelihood, m is the number of free parameters in the model, and N is the sample size. The BIC values for the three models' fits to each participant is shown in Table 6.

Table 6
Bayesian Information Criterion (BIC) values for Fits of the Models to Individual Data

Participant		Variable-precision	Threshold	Hybrid
High Precision	2	2211.62	1974.04	**1973.43**
	3	1875.35	**1743.92**	1749.93
	4	3724.85	**3158.48**	3162.36
	5	2180.82	**1985.66**	1991.72
	6	1581.47	**1529.58**	1530.76
	7	1448.65	**1275.30**	1278.28
	8	1222.85	**606.70**	620.79
	9	**2060.22**	2065.52	2068.87
	10	1974.49	1857.13	**1832.28**
	11	1821.20	**1638.65**	1648.34
	12	1479.85	**1020.29**	1024.03
	15	**1880.61**	1884.47	1889.55
	16	1965.88	1824.08	**1823.12**
	17	**1978.34**	2001.04	2035.88
	18	2062.39	**1832.53**	1834.55
	19	1620.87	**1545.20**	1551.37
	20	1602.36	1570.42	**1566.27**
Low Precision	1	**885.60**	886.72	892.04
	13	**1566.15**	1573.21	1579.12

Lowest BIC for each participant is indicated in boldface

To assess the extent to which the predictions of these models mimic each other, and thus the diagnosticity of our model selection, we conducted a model-recovery exercise. We simulated data sets using the variable-precision and threshold models with the parameters that provided the best fit to the empirical data for each participant. Each simulated data set was based on the same number of observations as the empirical data. We then fit the variable-precision model and the

threshold model to each simulated data set, using the BIC as the fit statistic. We found that the correct model was recovered in all the cases where the threshold model generated the data, and in all but two of the cases where the continuous model generated the data. These results demonstrate that the models make sufficiently distinct predictions that if the continuous model had generated the empirical data, then use of the BIC to compare the two models would have led to its recovery. A more detailed description of the model-recovery exercise is provided as supplementary material.

Both the threshold and the hybrid models consistently outperformed the continuous model without guessing, particularly for those participants in the high precision subgroup. Table 7 summarizes the number of participants better fit by each model, as well as the summed BIC across participants.

Table 7
Number of Participants Better Fit and Summed BICs for each Model

Model Name	Number of Participants Better Fit	Summed BIC
Variable-Precision	5	35143.57
Threshold	10	31972.94
Hybrid	4	32052.69

Of the five out of nineteen participants whose data was best fit by the variable-precision model, two participants (1 and 13) are in the low precision group, meaning their response error distributions as assessed by the Rayleigh test did not significantly deviate from uniformity, while the remaining three participants (9, 15, and 17) appear close to uniformity as well, as shown in Figure 5, although the hypothesis of uniform responding was rejected, indicating that these

participants are not responding successfully to the source judgement task. Additionally, the difference in BIC for the variable-precision and threshold and hybrid models was moderate for the variable-precision model ($\Delta BIC = 8.00$). In contrast, the evidence for the threshold model ($\Delta BIC = 229.33$) and the hybrid model ($\Delta BIC = 227.53$) for whom those models were the preferred model was very strong. Fits of the models to response error (Figure 5) and RT (Figure 6) data show where the variable-precision model misses the data. For participants who were able to perform the task with moderate-to-high accuracy (as characterized by the presence of a well-defined peak in the error distribution), the variable-precision model underestimates the proportion of responses made with high accuracy and overestimates the proportion of responses made with a moderate level of accuracy. In contrast, the threshold and hybrid variants capture the general structure of both response error and RT data, although both models have some difficulty predicting the extreme peaks in the error distribution and the leading edges of the RT distributions.

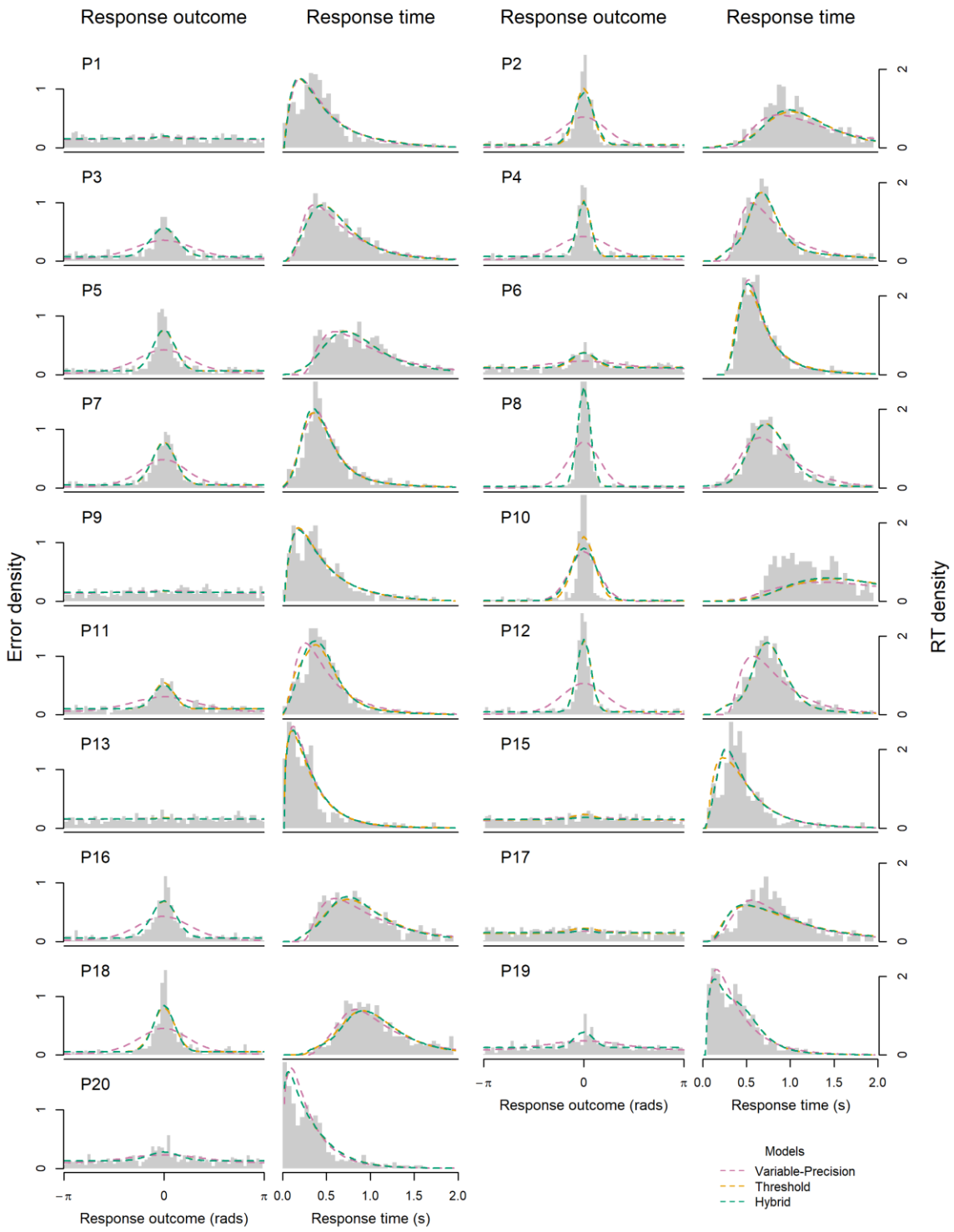

Figure 5. Fits of each of the three models to response outcome and response time data for each participant, presented in histograms.

A more detailed characterization of the performance of the model and points at which it misses the data is shown in Figure 6, which shows joint distributions of errors and RT in the form of a bivariate quantile plot, which depicts how response time (depicted on the y-axis) varies with response accuracy (depicted on the x-axis). The points represent the observed data: response precision is represented by how close the points are to the leftmost axis (with the data grouped into the 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles of response error); the stacks of points represent the speed of the responses conditioned on that level of accuracy (with three response time quantiles: the 10^{th}, 50^{th}, and 90^{th} percentiles).

Like the threshold models, the variable-precision model can predict a leptokurtic distribution of responses, because of trial-to-trial variability in drift norm. Because of this variability, it also predicts slow errors. This slow error pattern would be reflected in the joint quantile plot with RT quantiles becoming gradually slower as response error increases. Visually, this would appear as a bowing upwards as the bivariate quantile plot moves from left to right (cf. the slow errors in Figures 6 and 10 of Smith et al., 2020). However, the data do not systematically exhibit this pattern across participants, and so jointly fitting response error and RT constrains the ability of the variable-precision model to produce leptokurtic distributions of response error. In contrast, the hybrid model does not necessarily rely on trial-to-trial variability in drift norm to capture the shape of the distribution of response error, and can instead produce a leptokurtic shape through the mixture of above-threshold and sub-threshold zero-drift responding, while the threshold model relies only on the latter. Both the hybrid and threshold models are therefore more flexible than the variable-precision model in that their predicted RT distributions are a mixture of both processes.

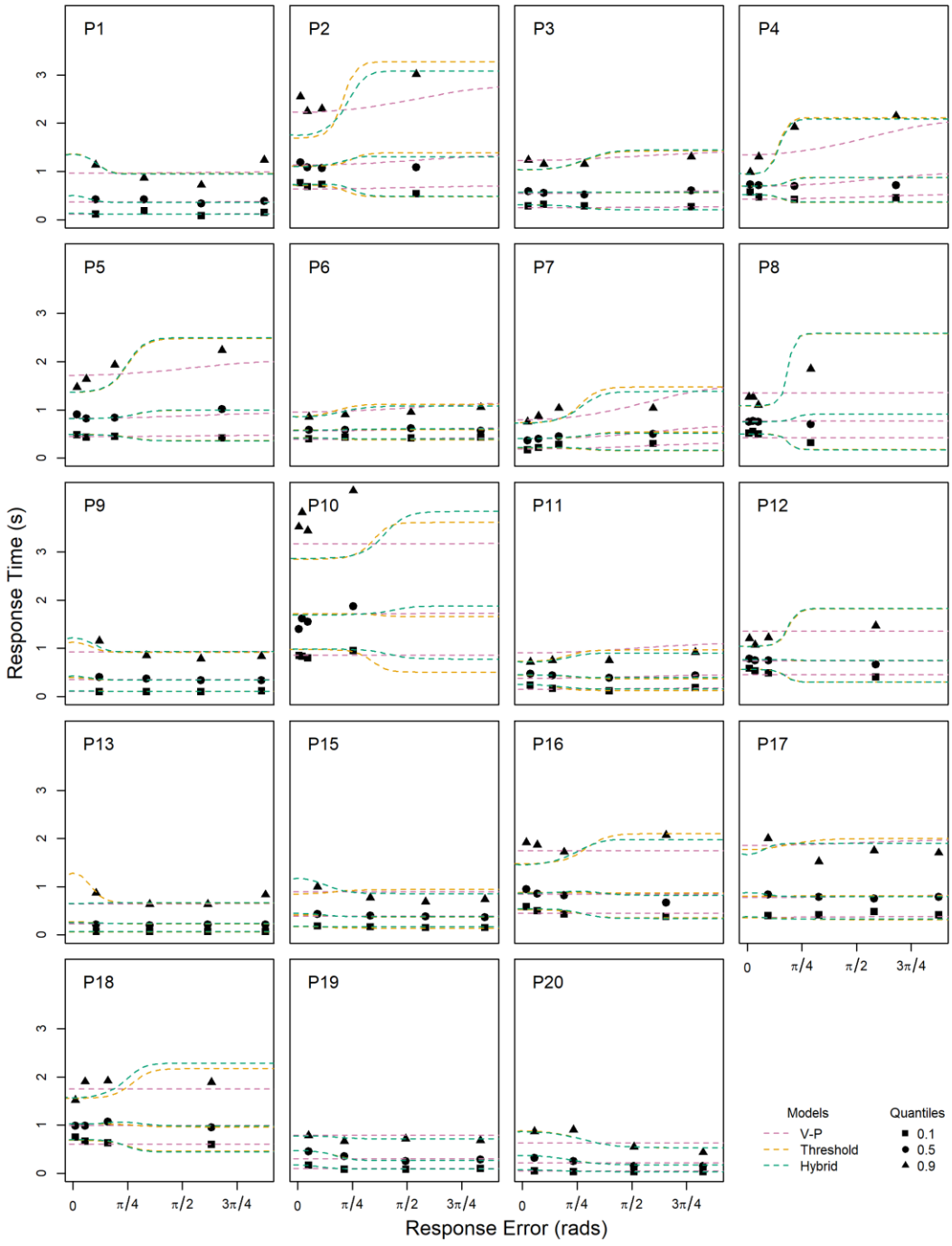

Figure 6. Fits of each of the three models to joint response error and RT quantiles for each participant. Points represent data quantiles, while lines represent the variable-precision (V-P), threshold and hybrid diffusion models.

Two qualifications are worth noting in relation to these data. Firstly, the data for some of the participants show idiosyncratic features that are not captured by any of the models. Participant 10, for example, had long decision times and showed evidence of bimodality in the RT distributions (Figure 8). We carried out an exploratory fit of a mixture model with two different nondecision time parameters and with all of the other model parameters constrained to be the same. The mixture model successfully captured the bimodality in Figure 8 and better characterized the precision of the distribution of decision outcomes than did the model with a single nondecision time. We have not reproduced this fit as we have no theory of why this kind of bimodality might arise and the pattern was only observed for one participant.

Second, our experimental program did not include a trap for very fast responding (cf. Ratcliff, 2018), which meant that participants may have begun to move the mouse prior to the retrieval cue. One of the participants (Participant 20), showed a progressive increase in very fast responses over the four experimental sessions. Most of these responses appear to be anticipations and are not captured by the model, but, rather than excluding the participant, we have chosen to include all data in order to provide a complete picture of the range of the individual differences on the task.

Overall, the advantage of the threshold and hybrid diffusion models over the variable-precision model suggest that participants sometimes do respond in a no-information guessing state, which is mixed with a proportion of responses driven by information which is centered on the target location. In comparing the hybrid and threshold models, the addition of across-trial variability drift rate does not appear to translate into any systematic advantage in fit across participants as compared to a model in which the drift rate is fixed. For the four out of 19

participants for whom the hybrid model fit better than the threshold model, the difference in BIC between the two models was small (ΔBIC = 7.64). The difference was similarly small for the 16 participants for whom the threshold model fit better than the hybrid model (ΔBIC = 6.89). This suggests that the addition of drift rate variability does not improve the fit of the threshold model enough to justify the additional complexity introduced into the model. The average parameter values for each model are displayed in Table 8.

Table 8
Mean Parameter Values Across Participants for Circular Diffusion Models

Model Name	Model Parameters						
	μ	η	a_1	a_2	π	T_{er}	s_t
Continuous	0.92	0.52	1.46	-	-	-0.02	0.14
Threshold	3.64	-	3.00	1.32	0.42	-0.08	0.09
Hybrid	3.66	0.18	3.20	1.30	0.41	-0.07	0.06

Modeling Source Responding Across Recognition Confidence Levels

In addition to the modeling presented above, we were also interested in the change in the source response distributions across recognition confidence levels. Because a pure threshold model assumes that memory exists in discrete states, the threshold variant of the circular diffusion model predicts that changes across recognition confidence only affect the proportion of trials in each of the discrete states and not the quality of evidence retrieved. To test this prediction, we split each participants' source response data into high-confidence and low-confidence sets. We classified items rated six on a six-point confidence scale in the item recognition task as trials recognized with high confidence, while items rated four or five were classified as trials recognized with low confidence.

We fit three versions of the threshold model that differed in flexibility jointly to the low-confidence and high-confidence data to try to find a most parsimonious model for the two conditions. We used the BIC to test whether the improvement in fit for the more flexible models was sufficient to offset the increase in the penalty for increasing the number of parameters. The most constrained model fit the data for the low and high confidence conditions with the same parameters. The second model allowed the mixing proportion of guessing and memory states (π) to vary across confidence conditions. The third model allowed both the mixing proportion and the mean drift rate (μ) to vary across confidence conditions. Participants varied in how they used the confidence scale to express their confidence. Some participants, given that they recognized an item, would typically recognize it with very high confidence (a rating of six), which resulted in too few responses in the low confidence range to characterize the distribution shape. For this reason, we restricted this analysis to those participants for whom there were at least 40 responses in both confidence ranges.

If the change in the source response distribution across recognition confidence is the product of a pure threshold process then we would expect that allowing the mixing proportion to vary as a function of confidence would improve the fit of the threshold model. In contrast, allowing drift rates to vary across recognition confidence should not improve the fit of the model, as the quality of evidence retrieved should not vary as a function of recognition confidence. Table 9 shows the fits of the three threshold models to the high-confidence and low-confidence data. It shows that allowing the mixing proportion to vary as a function of confidence improved the fit for the majority of participants but allowing the mixing proportion and the drift rate to both vary did not.

Table 9
BICs for Joint Fits of Threshold Model to Items Recognised with High and Low Confidence

	Parameters Allowed to Differ		
Participant	None	π	$\pi + \mu$
3	1697.76	**1690.00**	1691.54
4	3054.77	**3029.47**	3034.27
6	1337.23	**1333.50**	1343.53
7	1201.93	**1185.90**	1211.61
9	2028.02	**2033.62**	2045.79
12	808.82	**790.82**	799.57
13	**1507.69**	1540.51	1525.57
15	1645.55	**1641.69**	1653.72
17	1872.12	1882.19	**1871.68**

Lowest BIC for each participant is indicated in boldface

The parameter estimates for the best fit of the three versions of the threshold model are shown in Table 10. In the best fitting version of the threshold model, where only π varied between confidence conditions, the proportion of memory-based trials was higher in the high confidence condition than the low confidence condition. Even in the version of the model where both μ and π were both allowed to vary, the π parameter changes substantially across confidence conditions. Although μ also changes across confidence, this behavior is similar to the Zhang and Luck (2008) mixture model and reflects the same issue of poor identifiability in the low confidence condition. Regardless, model selection does not favor such changes in μ.

Table 10
Mean Threshold Model Parameter Values Across Participants

		Parameters Allowed to Differ		
Parameter	Confidence	None	π	$\pi + \mu$
μ	Low	3.98	3.94	3.40
	High			2.04
π	Low	0.31	0.15	0.21
	High		0.36	0.36

Discussion

In this article, we had two main aims. Our first aim was to attempt to characterize performance on a continuous report source memory task using a mathematical model of the decision process, the circular diffusion model, to ascertain whether it could predict the distributions of decision outcomes and RT from such a task. Specifically, we examined whether the model would allow us to distinguish between the predictions of SDT-like variable precision models and thresholded models of memory in the RT and accuracy data from a continuous outcome retrieval task. Our second aim was to ascertain whether Harlow and Donaldson's (2013) conclusion that source memory is thresholded would be supported for memory when performance was conditioned on recognition confidence.

We found evidence that source memory retrieval is indeed best characterized as a thresholded process. First, we found that even when source responses were conditioned on successful recognition, the marginal distribution of response error was well characterized by a two-component mixture model, consisting of a von Mises and a uniform distribution. This agrees the results of Harlow and Donaldson (2013), who used a wrapped Cauchy to account for the heavy tails of the distribution of errors. Second, we fit the joint distributions of accuracy and RT with the circular diffusion model, and again found that the threshold and hybrid models, both of which assumed a mixture of guessing and memory-based responses, fit the data better than did the variable-precision model, which did not assume such a mixture. Finally, when we analyzed source performance conditioned on recognition confidence, we found that the best threshold model was one in which only the proportion of guesses changed as a function of confidence,

while the drift rates remained constant. This is as predicted by a pure threshold account of source memory retrieval.

Our joint modeling of accuracy and RT contributes to a growing body of work that suggests that source memory retrieval is well characterized by a threshold process. The novelty of our results is that we found evidence for a threshold in retrieval using a continuous-outcome decision task in which we measured both RT and accuracy. Our analysis using the circular diffusion model allowed us to distinguish the contributions of memory strength and decision criterion to the precision of the memory-based component of responding. While Harlow and Donaldson similarly found evidence for a threshold process, their modeling of response-error distributions would not be able to rule out an explanation based on variable precision (e.g., van den Berg et al., 2012). Specifically, a variable precision model has two sources of variability – a variability in the latent strength along with variability in the mapping between strength and the decision outcome.

In the circular diffusion model, variable precision is a natural consequence of variability in the drift norm (the latent memory strength). For each value of the drift rate, there will be trial-to-trial variability in the decision time and decision outcome because of the cumulative effects of moment-to-moment variability in evidence accumulation (Figure 2). When these two sources of variability are combined, the model can produce heavy-tailed distributions like those reported by Harlow and Donaldson (2013), but additionally predicts slow errors. These predictions mirror those of the two-choice diffusion model, which includes variability in drift rate. However, we did not find any evidence for the slow errors predicted by a variable precision account. For this reason, our data provide a highly constrained test of the variable precision account, which, when

implemented in the circular diffusion model, was unable to fit the joint distributions of response error and RT. Instead, the data were well-described by a threshold version of the circular diffusion model, which assumes that a zero-drift guessing process is initiated when retrieval discretely fails.

Source Guessing for Unrecognized Items

One potential recourse for continuous models is to assume that participants do guess, but only guess because they are unable to recognize the items. The model of Hautus et al. (2018) made this assumption on the grounds that when items are not recognized, participants do not even attempt source memory retrieval. For this reason, we collected item recognition confidence ratings in a separate test phase. While we confirmed that unrecognized items do indeed elicit guesses in source memory (consistent with some investigations using two-choice paradigms, e.g., Malejka & Broder, 2016), we conditioned our source memory data in the circular diffusion model fits on recognized items. Thus, these results challenge virtually all continuous models of the source memory task. Nonetheless, it remains unclear from our findings whether a simple threshold model is sufficient or a "some-or-none" hybrid model (Onyper et al., 2010) is better supported. The some-or-none model agrees with a simple threshold model in that retrieval can discretely fail, but when retrieval succeeds it produces continuous evidence, which we modeled using variability in the drift norm. Future work may be able to distinguish between these two possible architectures.

Joint Recognition and Source Modeling

The relationship between recognition and source memory has been a central part of the episodic memory literature, and a complete understanding of this relationship requires modeling of both recognition and source memory with one unified model (Hautus et al., 2008). While we used recognition confidence as a predictor variable on which source responses were conditioned, joint modeling that predicts recognition as well as source memory was beyond the scope of the present study.

An example of a model that provides a joint account of recognition and source confidence in two-choice tasks is the bounded bivariate Gaussian model presented by Starns, Rotello, and Hautus (2014), which is a continuous model that represents evidence in memory as a bivariate Gaussian distribution with recognition and source evidence as its two dimensions. When compared with a bivariate dual-process model that incorporated a threshold recollection process like the threshold variant of the circular diffusion model in the present study, the authors found that the continuous bounded bivariate model successfully predicted a range of qualitative patterns in the joint recognition and source data, while the dual-process model did not. As the data and modeling presented in the current study precludes drawing strong conclusions about the relationship between recognition and source memory, a natural direction for future research in the continuous-outcome domain is to extend the circular diffusion model to joint recognition and source data. To model discrete responses, the 2D decision space of the circular diffusion model can be partitioned into response categories with decision bounds, and also into different confidence regions (Smith, 2016). To reflect the bivariate nature of joint recognition and source evidence, drift rates in the model can be bivariate-normally distributed to connect performance in the two tasks together (Smith, 2019).

Application of the Circular Diffusion Model

In studies of two-choice decision making, Ratcliff and colleagues have argued persuasively for the importance of using models than can account for both accuracy and RT (Ratcliff, 1978; Ratcliff & McKoon, 2008). The ubiquity of the speed-accuracy tradeoff in cognitive psychology has repeatedly shown the limitations of considering RT or accuracy in isolation and highlighted the importance of using models that characterize these two dependent variables as expressions of a single underlying process. The diffusion model represents a theoretical advance over signal detection theory precisely because it provides a process model of the relationship between speed and accuracy. The parameters of the diffusion model estimated from data represent theoretically meaningful components of processing that jointly determine accuracy and RT: drift rate, which represents the quality of the evidence in the stimulus; boundary separation, or decision criterion, which represents the quantity of evidence needed for a response; and nondecision time, which represents the time for processes outside of the decision process. A pragmatic advantage of using the diffusion model rather than SDT in applications is that it is more constrained, because it must account for both RT and accuracy, rather than just accuracy alone. In studies of recognition memory, for example, fits of the diffusion model have allowed researchers to distinguish between alternative models of the recognition process that could not be distinguished using signal detection theory (e.g., Osth, Jansson, et al., 2018; Ratcliff & Starns, 2009; Starns, Ratcliff, & McKoon, 2012). For instance, Ratcliff, Thapar, and McKoon (2004) conducted a diffusion model analysis on the effects of aging on recognition memory, where it has generally been found that older adults exhibited slower RTs than younger adults despite similar accuracy. A counterintuitive result of their analysis was that drift rate changed

negligibly with aging; instead, the slower RTs in older adults were due to higher response boundaries and longer nondecision times. Such insights would not be possible using accuracy data alone.

The circular diffusion model offers the same theoretical advantages in the analysis of continuous-outcome tasks as the standard diffusion model offers for two-choice tasks. Like the two-choice model, the circular diffusion model provides a unified account of RT and accuracy and its estimated parameters characterize the components of processing that give rise to these variables. In models of visual working memory, the distribution of error responses is characterized by the precision of the components of a von Mises mixture model (Bays & Husain, 2008; Oberauer & Lin, 2017; van den Berg et al., 2014; Zhang & Luck, 2008). Such models seek to characterize how precision changes as a function of the number of items in memory – Bays and Husain and van den Berg et al. assumed that precision varies as a power function of display size – but they provide no theoretical account of precision or its cognitive foundations. Many researchers assume that precision is an expression of an underlying neural population code (Bays, 2014) in which it characterizes the variability in the firing rates in a population of tuned detectors. This idea, while plausible (although see Lilburn, Smith, & Sewell, 2019, for contrary evidence), provides no account of RT.

In contrast, the circular diffusion model provides a decomposition of precision into its underlying cognitive components. These components correspond closely to those in the two-choice diffusion model. Rather than viewing the continuous-outcome task as being qualitatively distinct from other kinds of decision tasks, the circular diffusion model views continuous-outcome tasks and two-choice tasks as being expressions of the same underlying psychological

processes. Importantly, as pointed out by Smith (2016), the expression for the von Mises precision parameter in the circular diffusion model closely parallels the sensitivity index for a two-choice, random walk decision model, derived by Link (1975). (A random walk is a discrete-time counterpart of a diffusion process.) The similarity in the sensitivity/precision indexes for the two kinds of decision model highlights their underlying theoretical unity. This unity is not apparent when the experimental tasks are considered at the level of the kinds of data they produce, which appear to be very different, but becomes apparent when performance on them is viewed as an expression of evidence accumulation by a diffusion process.

The novelty of our analysis of a continuous-outcome source memory task using the circular diffusion model, which supports and extends Harlow and Donaldson's (2013) more empirical analysis, is not that it yielded different conclusions than theirs, but rather, that it yielded similar conclusions via a very different route, specifically, via the application of a mathematical model of the decision process that can predict both RT and accuracy. Our study joins the growing body of literature that has shown the theoretical benefits to be gained from characterizing memory retrieval using diffusion decision models.

It is important to note that the source distributional assumptions that provide the best fit in the circular diffusion model may not hold for other models. The circular diffusion model, like Ratcliff's diffusion model for two-choice decisions, assumes the evidence entering the decision process is normally distributed across trials. Unlike his model, the evidence has a bivariate rather than a univariate normal distribution, one component of which represents variability in evidence strength and the other of which represents variability in the retrieved stimulus identity. We found that most of the across-trial variability was in evidence strength and that there was little

variability in retrieved identity. These findings agree with those of Smith et al. (2020) who applied the circular diffusion model to decisions about the hues of noisy color patches. As our models were all versions of the circular diffusion model, we are not able to say whether this finding depends on the properties of this particular decision model or whether it is a feature of source-memory representations more generally.

Our experiment was designed to be similar to the short-delay condition in the Harlow and Donaldson (2013) paradigm, but not the long-delay condition. With longer study-test lags, changes in overall performance and levels of interference in memory may affect the relative performance of the models presented in this study. As such, it is unclear how our results generalize to other continuous-outcome source memory paradigms and making broader conclusions will require future work.

This study represents the first attempt to model both RT and accuracy in a source memory task with continuous-outcome decisions. We used the circular diffusion model, which provides for an elaborated account of decision-making, to fit joint RT and error data. This allowed for more constrained analysis than previous studies which accounted only for response error. The theoretical advantage of analyzing RT and error data using a formal decision model is that it allows identification of independent sources of noise from memory and decision processes, rather than attributing all variability in responding to just memory. In distinguishing these sources of variability, our approach allows for more direct inferences about the nature of memory, and how retrieval from memory drives decision-making.

References

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267-273.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological bulletin*, *74*(2), 81.

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

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*(10), 3632-3645.

Bell, R., Mieth, L., & Buchner, A. (2017). Emotional memory: No source memory without old–new recognition. *Emotion*, *17*(1), 120.

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*(2), 484.

DeCarlo, L. T. (2003). An application of signal detection theory with finite mixture distributions to source discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 767.

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*(3), 389-406.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*.

Fisher, N. I., Lewis, T., & Embleton, B. J. (1993). *Statistical analysis of spherical data*. Cambridge university press.

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, *110*, 104065.

Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review*, *20*(2), 318-325.

Hautus, M. J., Macmillan, N. A., & Rotello, C. B. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review*, *15*(5), 889-905.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review*, *17*(4), 465-478.

Lilburn, S. D., Smith, P. L., & Sewell, D. K. (2019). The separable effects of feature precision and item load in visual short-term memory. *Journal of vision*, *19*(1), 2-2.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, *12*(1), 114-135.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347.

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & cognition*, *44*(1), 63-72.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological review*, *87*(3), 252.

Mickes, L., Wais, P. E., & Wixted, J. T. (2009). Recollection is a continuous process: Implications for dual-process theories of recognition memory. *Psychological science*, *20*(4), 509-515.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological review*, *118*(2), 280.

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological review*, *124*(1), 21.

Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General*, *139*(2), 341.

Osth, A. F., 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., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological review*, *126*(4), 578.

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. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive psychology*, *104*, 106-142.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: I. Color, location, orientation, and spatial frequency. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 261.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychological review*, *125*(6), 888.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

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*(3), 697.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, *111*(2), 333.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260-281.

Ratcliff, R., Smith, P. L., & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current directions in psychological science*, *24*(6), 458-470.

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*(4), 408-424.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975-5979.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & cognition*, *33*(1), 151-170.

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*, *123*(4), 425.

Smith, P. L. (2019). Linking the diffusion model and general recognition theory: Circular diffusion with bivariate-normally distributed drift rates. *Journal of Mathematical Psychology*, *91*, 145-158.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic bulletin & review*, *25*(6), 2083-2101.

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological review*, *116*(2), 283.

Smith, P. L., Saber, S., Corbett, E. A., & Lilburn, S. D. (2020). Modeling continuous outcome color decisions with the circular diffusion model: Metric and categorical properties. *Psychological Review*.

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*(1-2), 1-34.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological review*, *121*(1), 124.

Van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780-8785.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*, *4*(12), 11-11.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152.

Woodworth, R. S., & Schlossberg, H. Experimental psychology. New York: Holt, 1954

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*(6), 1415.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological bulletin*, *133*(5), 800.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233.