Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory

Adam F. Osth
University of Melbourne

Beatrice Bora and Simon Dennis
University of Newcastle

Andrew Heathcote
University of Tasmania

Address correspondence to:
Adam Osth
E-mail: adamosth@gmail.com

Author Note

We would like to thank Jeff Starns and Chad Dube for generously providing their data. This work was supported by two grants from the Australian Research Council, ARC DP150100272, awarded to Simon Dennis and Andrew Heathcote, and DE170100106, awarded to Adam Osth.
Abstract

The relative amount of variability in memory strength for targets vs. lures in recognition memory is commonly measured using the receiver operating characteristic (ROC) procedure, in which participants are given either a bias manipulation or are instructed to give confidence ratings to probe items. A near universal finding is that targets have higher variability than lures. Ratcliff and Starns (2009) questioned the conclusions of the ROC procedure by demonstrating that accounting for decision noise within a response time model yields different conclusions about relative memory evidence than the ROC procedure yields. In an attempt to better understand the source of the discrepancy, we applied models that include different sources of decision noise, including both the diffusion decision model (DDM) and the linear ballistic accumulator (LBA) model, which both include and lack within-trial noise in evidence accumulation, and compared their estimates of the ratio of standard deviations to those from ROC analysis. Each method produced dramatically different estimates of the relative variability of target items, with the LBA even indicating equal variance in some cases. This stands in contrast to prior work suggesting that the DDM and LBA produce largely similar estimates of relevant model parameters, such as drift rate, boundary separation, and nondecision time. Parameter validation using data from Starns’s (2014) numerosity discrimination data demonstrated that only the DDM was able to correctly reproduce the evidence ratios in the data. These results suggest that the DDM may be providing a more accurate account of lure-to-target variability evidence ratios in recognition memory.

Keywords: diffusion decision model; linear ballistic accumulator; receiver operating characteristics; signal detection theory; recognition memory
Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory

Signal detection theory (SDT) is possibly the single most widely used framework of decision making in the field of recognition memory. According to SDT, memory strength is continuously distributed for both targets and lures; decisions are made by placing a decision criterion on the memory strength axis and responding 'yes' if an item's strength exceeds the criterion and 'no' otherwise. SDT models are commonly tested via receiver operating characteristics (ROCs). ROCs are constructed by manipulating bias across several different levels and observing how the hit rate (HR) and false alarm rate (FAR) change across the different levels of bias. Bias can be manipulated in different ways, including target proportion manipulations, payoffs, or via the collection of confidence ratings, although confidence ratings are the dominant method for ROC construction in the literature. SDT models correctly predict the curvilinear shape of the ROC in recognition memory (Egan, 1958; Wixted, 2007).

SDT models make further predictions about the shape of the z-transformed ROC (zROC), specifically that the zROC should be linear with a slope that is equal to the ratio of the standard deviation of the lure distribution to that of the target distribution ($\sigma_{\text{lure}}/\sigma_{\text{target}}$). Models where there is equal variability between targets and lures make the prediction that the slope of the zROC is 1. The results of many ROC experiments conducted over several decades of recognition memory research have almost unilaterally found that the slope of the zROC is less than 1 (Benjamin, Diaz, & Wee, 2009; DeCarlo, 2007; Dube & Rotello, 2012; Glanzer & Adams, 1990; Glanzer, Hilford, Kim, & Adams, 1999; Glanzer, Kim, Hilford, & Adams, 1999; Heathcote, 2003; Hirshman & Hostetter, 2000; Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, McKoon, & Tindall, 1994; Yonelinas, 1994; Wixted, 2007). This has led theorists to reject equal variance (EV) SDT models and in some cases endorse unequal variance (UEV) SDT models, where the target distribution has greater variability than the lure distribution, potentially due to the contribution of
variability in encoding strength (Wixted, 2007).

However, the finding of zROC slopes less than one is one of the only consistently observed regularities in ROC studies of recognition memory. ROC studies initially gained some prominence in the early 1990s to test the predictions of global matching models of recognition memory, as models such as the search of associative memory (SAM: Gillund & Shiffrin, 1984) model and the Minerva 2 model (Hintzman, 1988) made the prediction that the ratio of lure-to-target variability should decrease as performance is increased. This prediction was violated in several influential experiments by Ratcliff and colleagues (Ratcliff et al., 1992, 1994) who found that zROC slopes (which measure lure-to-target variability) were around .8 and were unaffected by strength manipulations, leading them to endorse the constancy of slopes hypothesis. Nonetheless, the constancy of slopes was violated in subsequent investigations: Glanzer, Kim, et al. (1999) and Hirshman and Hostetter (2000) found that the slope of the zROC was reduced by massed repetitions but not by spaced repetitions. Heathcote (2003) found the opposite result, namely reduced zROC slopes for conditions with spaced repetitions rather than massed repetitions. Several investigations also found reduced slopes for words of lower natural language frequency (e.g.: DeCarlo, 2007; Glanzer & Adams, 1990; Ratcliff et al., 1994). Thus, zROC slopes of .8 are no longer considered a regularity of the ROC procedure.

The linear shape of the zROC is also not ubiquitous. Although SDT models make the prediction that zROCs should be linear in shape, alternative measurement models predict a degree of curvilinearity in their shape. For instance, the dual process signal detection model (DPSD: Yonelinas, 1994) posits that recognition memory is a mixture of familiarity, which is an equal variance SDT process, and recollection, which is a high threshold process that only occurs for targets. The contribution of high threshold recollection bends the

---

1Two other regularities in ROC studies are the zROC length effect, where conditions of higher performance have shorter zROC lengths (Stretch & Wixted, 1998; Glanzer, Hilford, & Maloney, 2009), and the variance effect, where distributions that are further from the log likelihood ratio of zero have higher variability (DeCarlo, 2007; Glanzer et al., 2009).
zROC upward, introducing some curvilinearity and is what reduces the zROC slope below one (Yonelinas, 1994; Yonelinas & Parks, 2007). The high threshold model proposes that recognition is solely governed by high threshold "detect" and "non-detect" states, with manipulations of target strength only affecting the probability of entering these states (e.g.; Provence & Rouder, 2012) and predicts linear ROCs (but not zROCs) in circumstances such as when bias is manipulated by target proportions or payoffs (Bröder & Schütz, 2009).

Although aggregate zROCs are often linear in shape, in congruence with SDT models, individual participant zROCs possess a wide variety of shapes that are incompatible with the predictions of any of these models (Ratcliff et al., 1994). More recently, Kellen and Singmann (2016), conducted a thorough analysis of several ROC studies with the dominant models and found that each model had systematic residuals, suggesting that the observed form of the zROC does not conform to the predictions of any of the current models.

**Each of the aforementioned models provides a very different characterization of how evidence informs recognition decisions,** and yet studies using the ROC procedure have not yielded a clear resolution to the debate as to which is the correct model. Inconsistencies among ROC results have contributed to this indecision. Although it is not immediately clear why results across participants and studies are so inconsistent, one possibility is that the assumptions of the ROC procedure may not be valid. A necessary assumption in applying models to ROC data is that performance is constant across each of the bias levels; only the decision criterion varies. This selective influence assumption is a necessity in identifying the parameters of the models. If this assumption is lifted and $d'$ is allowed to vary across bias conditions, however, non-linear zROCs can be produced, even if the SDT model holds. However, given that the models cannot be identified under those circumstances, other procedures or measures are necessary to investigate this possibility.
Evidence Accumulation Models and zROCs

Evidence accumulation models have been used to investigate whether performance is constant across bias conditions. Unlike SDT models, evidence accumulation models of decision making make predictions about both choice and response time (RT). In these models, evidence gradually accumulates toward boundaries corresponding to each of the response options. Once a boundary is reached, the corresponding response is made and the time taken to reach the boundary in part determines RT. The rate of accumulation is called the *drift rate*; increases in the drift rate make the corresponding response more likely and decrease RT. Response boundaries control the speed-accuracy tradeoff: increases in boundaries increase accuracy, because they decrease the likelihood that an incorrect response boundary is reached by chance, but also increase accumulation time, and hence increase RT.

The ability of evidence accumulation models to separate the strength of the evidence from the response threshold is one of their major strengths. In the domain of recognition memory, for example, Ratcliff, Thapar, and McKoon (2004) compared younger and older adults, and found that $d'$ was nearly equivalent between the two age groups, but older adults exhibited slower RTs. Rather than the slower RTs in older adults being reflective of an impairment, an analysis of the results using the diffusion decision model (DDM: Ratcliff, 1978; Ratcliff & McKoon, 2008) revealed that memory strength, as measured by the drift rates, was identical between the two groups. Instead, older adults exhibited wider speed-accuracy thresholds, suggesting that their slower RTs were not due to memory impairment but instead due to a more cautious approach to the task.

If participants employ different speed-accuracy thresholds across bias conditions, performance will not be constant across each point of the ROC even though the strength of the underlying evidence is unchanged by the bias manipulation. Ratcliff and Starins investigated this possibility with their response time confidence model (RTCON: Ratcliff & Starins, 2009) and its successor, RTCON2 (Ratcliff & Starins, 2013). In the RTCON models
each choice and confidence option has its own accumulator; whichever accumulator reaches its threshold first determines the confidence response. Similar to the traditional two choice DDM, drift rates in RTCON are not fixed for a given experimental condition, but vary across trials as samples from a normal distribution. The assumption of cross-trial variability in drift rates was directly inspired by SDT and causes error responses to be slower than correct responses (Ratcliff, 1978; Ratcliff & McKoon, 2008).

The RTCON models were applied to both confidence and RT data from a number of ROC datasets. The model was able to produce a variety of non-linear zROC shapes despite the fact that the evidence distributions, which are analogous to those from SDT using different thresholds across each confidence accumulator. This implies that accuracy is not constant across each confidence level, contrary to the assumptions of SDT. In addition, the estimated ratio of standard deviations (SDs) from the models were much lower than that of the zROC slopes calculated from the data, contrary to SDT models, where the slope of the zROC is equal to the ratio of SDs. Based on these observations, Ratcliff and Starns concluded that the RTCON models seriously question the standard assumptions of the ROC procedure, as neither shape nor slope can be used to draw conclusions about the nature of memory evidence without additionally considering response times.

One possible reason why the RTCON models produce parameter estimates that differ wildly from SDT estimates is that participants may have difficulty adopting consistent speed-accuracy thresholds for each confidence response. This begs the question whether conclusions from the standard yes/no recognition procedure are congruent with the findings of the RTCON models. Starns, Ratcliff, and McKoon (2012) further explored the relation between SDT estimates of the slope of the zROC and those of evidence accumulation models using a binary ROC procedure. In a binary ROC procedure participants produce yes/no responses at test, but bias is manipulated in other ways such as by changing the proportion of targets and lures (e.g., Dube & Rotello, 2012).

Starns et al. (2012) investigated the binary ROC procedure using the traditional
two-choice DDM, which shares several assumptions with the RTCON models. In the DDM, evidence accumulates from the starting point $z$ toward either the upper response boundary, denoted by $a$, or the lower response boundary, which is fixed at zero. These response boundaries correspond to response alternatives in an experiment, which are 'YES' and 'NO' in the case of recognition memory. Drift rates are sampled from a normal distribution with mean $v$ and standard deviation $sv$. Evidence accumulation is noisy, such that a given drift rate can cause termination at different response boundaries (producing errors) at different times (producing variable RTs). To provide a complete account of the relative speeds of correct and error responses, the starting points are sampled from a uniform distribution with width $s_z$. Variability in the starting point allows for the model to predict error responses that are faster than correct responses under speeded conditions (Laming, 1968; Ratcliff, Van Zandt, & McKoon, 1999). RT is the sum of the time for the diffusion process and the time taken for perceptual encoding and motor response output, which is called the nondecision time, denoted by $t_0$, and is estimated as a shift to the RT distribution. Ratcliff, Gomez, and McKoon (2004) later assumed that nondecision time is sampled from a uniform distribution with range $s_t$. The DDM with variability in drift rate, starting point, and nondecision time is commonly referred to as the "full" diffusion model. A diagram of the DDM can be seen in Figure 1 and a list of parameters with descriptions can be found in Table 1.

Starns et al. (2012) applied the diffusion model to the binary ROC paradigm, primarily by assuming that each target proportion condition yields a different value of the starting point $z$ (see also Dube, Starns, Rotello, & Ratcliff, 2012). Conventional applications of the DDM assumed that both targets and lures have the same value of drift rate variability (Arnold, Broder, & Bayen, 2015; Bowen, Spaniol, Patel, & Voss, 2016; Criss, 2010; Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff, Thapar, & McKoon, 2004, 2010, 2011; White & Poldrack, 2014). However, Starns et al. found that such a model was untenable as it predicted zROC slopes that were close to 1, contrary to the data. However, when the $sv$...
parameter of the target distribution was estimated separately from the lure distribution the model was able to reproduce the zROC slopes. Similar to the findings with the RTCON models, the ratio of standard deviations estimated by the DDM was much lower than the zROC slopes, contrary to the assumptions of SDT. The DDM estimated a ratio of standard deviations of about .6, with target distribution about 1.66 times more variable than that of the lure distribution. The zROC slope, in contrast, was around .9, suggesting a target distribution about 1.11 times more variable than the lure distribution in an SDT model.

Table 1

*Parameter names and descriptions for the diffusion decision model (DDM) and linear ballistic accumulator (LBA) model parameters.*

<table>
<thead>
<tr>
<th>Param.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common Parameters</strong></td>
<td></td>
</tr>
<tr>
<td>$v$</td>
<td>Mean of the drift rate distribution.</td>
</tr>
<tr>
<td>$s_v$</td>
<td>Standard deviation of the drift rate distribution.</td>
</tr>
<tr>
<td>$t_0$</td>
<td>Nondecision time for stimulus encoding and response.</td>
</tr>
<tr>
<td><strong>DDM</strong></td>
<td></td>
</tr>
<tr>
<td>$s_t$</td>
<td>Width of the nondecision time distribution.</td>
</tr>
<tr>
<td>$a$</td>
<td>Height of the upper response boundary.</td>
</tr>
<tr>
<td>$z$</td>
<td>Starting point (bias).</td>
</tr>
<tr>
<td>$s_z$</td>
<td>Height of the starting point distribution.</td>
</tr>
<tr>
<td><strong>LBA</strong></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>Height of the starting point distribution.</td>
</tr>
<tr>
<td>$b$</td>
<td>Height of an accumulator's response boundary.</td>
</tr>
</tbody>
</table>

*Notes: Param. = parameter.*
Ratcliff and Starns noted that models such as the DDM and RTCON incorporate several sources of variability not present in SDT, such as variability in speed-accuracy criteria across bias conditions, cross-trial variability in starting points, in addition to within-trial noise. Effects of these sources are consistent with the observation that additional sources of noise in SDT models, such as criterion variability (Benjamin et al., 2009; Wickelgren, 1968), obscure the relationship between evidence variability and zROC slope because as other sources of errors are introduced into the decision process, errors are less likely to be due to evidence variability. As a consequence, targets require even greater variability than lures to produce the same zROC slope that would be observed in a pure SDT model that lacks any additional source of noise.

Nonetheless, the precise role of each of these sources of noise, such as within-trial noise, starting point variability, and differences in speed-accuracy criteria, have yet to be investigated. In addition, there has not been a comparison to the linear ballistic accumulator (LBA: Brown & Heathcote, 2008) model, which is particularly pertinent because it lacks within-trial noise, a crucial component of the DDM (Smith, Ratcliff, & McKoon, 2014). The present article examines parameter estimates from both the DDM and LBA models in order to understand how within-trial noise contributes to the estimate of the ratio of SDs. In the next section, we describe the LBA model, and prior work comparing its parameter estimates to the DDM.

**The Linear Ballistic Accumulator Model and its Relation to the DDM**

Like the RTCON models, the LBA is an accumulator model; each response option receives its own accumulator. Each accumulator has its own drift rate (with mean $v$), boundary ($b$), and a starting point. Unlike the DDM, the accumulation in the LBA is linear (noise free within a trial); all of the variability in the model comes from cross-trial variability in its parameters. As with the DDM, there is cross-trial variability in both drift rate and the starting point; the drift rate is sampled from a normal distribution with
standard deviation $s_v$, while the starting point is sampled from a uniform distribution with range $A$. Starting point variability in the LBA allows for the speed-accuracy tradeoff; if the accumulator with the lower drift rate (which is more likely to be an error response) begins with a high starting point, it can reach its boundary before the accumulator with a higher drift rate and a lower starting point. Bias is implemented in the model by allowing for different boundaries for each accumulator. For example, a low boundary for the "YES" accumulator yields faster yet relatively inaccurate "YES" responses compared to the "NO" accumulator. A diagram of the LBA can be found in Figure 2 and a list of parameters with descriptions can be found in Table 1.

On the surface, the LBA and the DDM appear to be very different models due to the lack of within-trial noise in the LBA. Nonetheless, the LBA has been successful in achieving the same benchmarks, such as right-skewed RT distributions, errors that are faster than correct responses under speed emphasis, and errors that are slower than correct responses under accuracy emphasis (Brown & Heathcote, 2008). Furthermore, their parameter estimates are sufficiently correlated that the two models have endorsed similar conclusions about psychological data. A thorough comparison of the two models was conducted by Donkin, Brown, Heathcote, and Wagenmakers (2011). They simulated data with one model and fit with the other and compared the parameter estimates of the two models. Both models were found to reliably recover changes in drift rate, boundary separation, and nondecision time that were generated from the corresponding model. The points of disagreement among the models were that a) the LBA produces lower nondecision time estimates than the DDM and b) both models interpret changes in response caution in the corresponding model as, in part, a change in nondecision time.

Despite the disagreements of the two models, Donkin et al. (2011) found that in fits to real data both models produced nearly identical conclusions about the effects of manipulations on model parameters. More recently, Rae, Heathcote, Donkin, Averell, and Brown (2014) explored the two models in fits to experiments from several tasks where
speed-accuracy emphasis was manipulated, including recognition memory, lexical decision, and brightness discrimination. Model selection endorsed qualitatively similar models of both the DDM and LBA: in the best models the speed accuracy manipulation influenced not just the response boundaries but additional parameters, including nondecision time and drift rate. This stands in contrast to traditional work assuming that only response boundaries are affected by such manipulations (e.g. Ratcliff & Rouder, 1998).

Although some differences between the DDM and LBA have been found for nondecision time and start point variability in lexical decision data (Heathcote & Hayes, 2012; Heathcote, Loft, & Remington, 2015), the two models strongly agree about drift rates. However, none of the numerous comparisons between the two models have focused on estimates of drift rate variability. Drift rate variability is not typically the focus of evidence accumulation modeling, and estimates of this parameter are often not interpreted. In the LBA model, drift rate variability is sometimes fixed for one accumulator (Rae et al., 2014), as this allows separate estimates of mean drift rate to be identified for each accumulator. However, any accumulator-related parameters in the model can be fixed to enable identification (e.g.: Donkin, Brown, & Heathcote, 2009); one alternative is to estimate the drift rate of only one accumulator and determine the drift rate for the other (Brown & Heathcote, 2008; Donkin et al., 2009; Donkin & Nosofsky, 2012b, 2012a; Donkin et al., 2011; Hawkins, Hayes, & Heit, 2016; Little, Wang, & Nosofsky, 2016; Nosofsky, Cao, Cox, & Shiffrin, 2014), a practice which offers sufficient constraint for the estimation of drift rate variability. In this article, we specifically focus on this parameterization because it is most congruent with the assumptions of SDT models, where each condition receives a single estimate of memory strength; for this reason we refer to this model as the unidimensional LBA. This parameterization is quite similar to that of the DDM because increases in the strength of evidence for one response option are accompanied by decreases in the strength of evidence for the other. We return to implications of alternative parameterizations of rate effects in the General Discussion.
One reason why researchers may have been disinterested in drift rate variability is that in some cognitive tasks there may be no clear psychological interpretation of a change in $sv$ across conditions. This is not the case for recognition memory, where there is obvious theoretical relevance to changes in $sv$, especially given that investigations that have allowed drift rate variability to differ between targets and lures have produced results that are in qualitative (but not quantitative) agreement with the ROC literature (Starns et al., 2012; Starns & Ratcliff, 2014; Starns, 2014). In addition, process models of recognition memory make precise predictions about how variability in memory strength changes across performance conditions in an experiment (Ratcliff et al., 1992, 1994; Shiffrin & Steyvers, 1997), implying that measurement of drift rate variability can be further used to test such models. Thus, any evidence for a divergence between DDM and LBA estimates of the $sv$ parameter brings into question whether it is straightforward to test either process or SDT accounts of recognition memory using evidence accumulation models.

**Divergent Effects of Drift Rate Variability.** Drift rate variability was introduced into evidence accumulation models by Ratcliff (1978) to allow different rates for different memory items, as well as to allow for the prediction of errors that are slower than correct responses. A somewhat counterintuitive prediction of the DDM is that a single fixed drift rate produces equal speeds for correct and error responses when the starting point is equidistant from each response boundary. However, when drift rate variability is introduced, the errors are predominantly composed of low drift rates (which are slow), while the correct responses are predominantly composed of high drift rates (which are fast). When these predictions are averaged together, errors are slower than correct responses (Ratcliff & McKoon, 2008). It is for this reason that drift rate variability can be identified and reliably recovered without requiring an ROC function (Ratcliff & Tuerlinckx, 2002); Starns and Ratcliff (2014) fit the DDM to a large number of recognition memory datasets without ROCs and consistently found greater drift rate variability for targets than for lures, a result which has been replicated (Osth, Dennis, & Heathcote, 2017; Starns, 2014).
Variations in $sv$ can produce quite different behavior in the two models of interest, as can be seen in Figure 3, which depicts both the DDM (top) and LBA (middle) models predictions for correct (left) and error (right) RT across a wide range of the $sv$ parameter for three different values of drift rate. RTs are summarized in terms of the leading edge, as the $10^{th}$ percentile of the RT distribution, the median, and the right tail of the RT distribution, as the $90^{th}$ percentile. For the DDM the leading edge and median hardly vary with $sv$, with most of the changes occurring in the right tail, where differences between correct and error responses get larger as $sv$ is increased. The LBA also shows effects in the right tail of the RT distribution, but a major divergence is that it predicts large effects on the leading edge of the RT distribution for both correct and error responses as $sv$ is increased. This discrepancy between the two models suggest that fits to a given dataset may produce very different estimates of the ratio of SDs ($sv_{lure}/sv_{target}$) across the DDM and LBA.

The differences in the $sv$ predictions between the DDM and LBA might be due to the nature of evidence accumulation (noisy vs. linear) or the architecture of the model (single vs. multiple accumulators). To investigate this issue, the bottom panel of Figure 3 includes predictions from racing noisy accumulators that were simulated with parameters very close to those of the DDM except that there was no variability in nondecision time. One can see that the predictions are qualitatively similar to those of the DDM, in that there is little change in the leading edge of the RT distribution as $sv$ is increased by a very wide range. Thus, it appears that the divergent effects of manipulation of $sv$ on the DDM and LBA, namely the large leading effects in the LBA, are primarily due to the presence vs. absence of within-trial noise. There are analytics available for noisy racing accumulators that include variability in either drift rate (Desmond & Yang, 2010) or starting points (Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014), but not both together, and thus from here we restrict consideration of evidence accumulation models to the DDM and LBA.
Overview

The divergent effects of sv on the leading edge of the RT distribution suggests that the DDM and LBA could produce very different interpretations of zROC slopes. This is an important issue, as the previous investigations finding greater variability for targets for the DDM are not guaranteed to generalize to the LBA. In addition, current global matching models such as the retrieving effectively from memory (REM: Shiffrin & Steyvers, 1997), the bind cue decide model of episodic memory (BCDMEM: Dennis & Humphreys, 2001) and the model of Osth and Dennis (2015) make testable predictions for how lure-to-target variability ratios change across performance conditions. If the DDM and LBA are not found to agree in their estimate of the ratios, this could cast doubt on the ability to test the predictions of memory models. In this article, we compare the DDM, LBA and SDT models in their estimates of the ratio of SDs with experimental procedures that enable estimation of the zROC slope. The first dataset (Experiment 1) comes from a new experiment we conducted that used a two-stage confidence procedure, where participants give an initial "yes" or "no" decision follows by a confidence judgment. Such procedures are common in perception (Baranski & Petrusic, 1998; Vickers, 1979), but are somewhat less common in recognition memory than the single stage confidence procedures, with some exceptions (e.g.: Hirshman & Hostetter, 2000; Van Zandt & Maldonado-Molina, 2004; Weidemann & Kahana, 2016). The two-stage confidence procedure produces a full ROC function, allowing for the estimation of SDT models, and offers an additional advantage because the two choice versions of the DDM and LBA can be applied to the initial yes/no decision. To our knowledge this is the first analysis that has directly compared the estimates of the ratios of lure-to-target variability from two-choice evidence accumulation models and from SDT using confidence ratings.

One difficulty we faced was that the two-stage confidence procedure required the addition of criterion variability in the second stage to achieve a good fit to the zROC functions. For this reason, we additionally fit two archival datasets collected with the
binary ROC procedure (Dube et al., 2012), with bias manipulated using different target proportions. These designs also allow for direct comparison of the evidence accumulation models with SDT.

For each dataset we fit a variety of models that made allowances for the effect of different manipulations – such as word frequency, speed-accuracy emphasis, and study repetitions – on the $sv$ parameter for targets. We subsequently performed model selection to evaluate which model was appropriate for the data. The purpose of this exercise was to evaluate the level of consistency between the models in their interpretation of the effects of these manipulations on variability. Starns and Ratcliff (2014) previously found that virtually none of these variables affected the $sv$ parameter for targets in the DDM, while several investigations have found such effects in SDT (e.g.: Heathcote, 2003; DeCarlo, 2007; Glanzer, Kim, et al., 1999). We know of no such investigations that have been conducted with the LBA. In addition to allowing $sv$ to vary across conditions, we fit DDM and LBA models that lack starting point variability to evaluate its' consequences on the estimated ratio of lure-to-target variability.

To foreshadow our results, in every data set the three models produced very different estimates of the ratio of lure-to-target variability, with the DDM producing the lowest ratios, SDT somewhat higher, and the LBA the highest, with ratios that were in some cases quite close to one, indicating equal variance. We followed up the analyses of recognition memory datasets with DDM and LBA parameter validation exercise based on data from a numerosity discrimination task that allowed trial to trial variability in evidence to be manipulated directly (Starms, 2014). In numerosity discrimination participants are shown a series of asterisks and asked whether the depicted number is 'high' or 'low' relative to a given criterion. From trial to trial the number of asterisks is drawn from distributions whose mean and standard deviation differ between conditions. Manipulations of the standard deviation should produce corresponding effect on the model’s $sv$ estimates.
Experiment 1: Two Stage Confidence Procedure

Participants underwent recognition memory testing with manipulations of speed-accuracy emphasis, word frequency, and concreteness (c.f. Rae et al., 2014) where they gave a yes-no response followed by a high vs. low confidence judgment. This procedure allows for direct comparison between evidence ratios in the DDM and the LBA, which are estimated on the basis of the choice and response times to the initial judgment, and SDT, which is estimated on the basis of the 4-point confidence ROC function.

Participants

36 participants at the University of Newcastle participated in exchange for course credits.

Materials

The experimental item set was 2,069 nouns. This set included items rated for word frequency (min=1, max=314,232, median=469, measured in counts per million) and contextual diversity (min=1, max=8363, median=294), according to the subtitles lexicon of American English (SubtLexUS: Brysbaert & New, 2009) norms. Word concreteness ratings were also measured and controlled (min=183, max=667, median=506) with ratings source from the Medical Research Council psycholinguistic database (MRC2: Coltheart, 1981). All words were between four and seven letters.

Procedure

Participants attended two sessions of one hour each. Each session began with training on the response key mapping followed by 18 study-test cycles, the first two of which were treated as a warm up and not further analysed. Recognition choices ('old', 'new') were indicated with the 'z' and '/' keys. Confidence choices ('probably', 'definitely') were subsequently made with the same keys. The attribution of response keys was
counterbalanced across participants. Reminders of the response keys were located at the bottom of the screen. During the 32 training trials participants were prompted by flashing one of the response labels ('old', 'new', 'probably', or 'definitely') and asked to press the corresponding key as quickly as possible. Participants were also asked to maintain their index fingers positioned on the two recognition keys at all times during the testing phase.

For half of the study-test cycles they were asked to perform very carefully and for the other half to respond as fast as possible while being accurate. These two emphases were administered in an alternating fashion within the session, with half of the participants starting with speed emphasis and the other half with accuracy emphasis. Words were assigned randomly for each participant to study lists with the constraint that there were an equal number of low and high frequency (LF and HF) and low and high concreteness (LC and HC). The cutoff between the 'low' and 'high' categories was the median of the sample.

Each of the two warm-up lists included 12 words to be presented during the study phase and 24 words for the test phase. Each of the 16 experimental lists contained 28 words for study and 50 words for the test. The first and last two items from the study list were used as primacy and recency buffers. Two of the buffer words were randomly selected and tested to keep participants from guessing the control strategy, although the results of these buffer words were not included in analyses.

During the study phase, words were presented at the center of the screen at a rate of one per second. Participants were encouraged to make use of any mnemonic strategy to help them remember the words. During the test phase, participants were instructed that when each word appeared on the screen they had to make an old/new decision followed by a confidence decision. The confidence response was followed by feedback. For accuracy blocks the feedback indicated whether the response was correct or not. For speed blocks it it indicated if it was necessary to respond faster, with the message "TOO SLOW" for responses made in less than 0.75s, and also when responses were made in less than 0.25s, with the message "TOO FAST". Positive and negative feedback were presented in green
and red, respectively. Participants were instructed that they had up to six seconds to respond, and if they exceeded that limit the following feedback was given: "TIME LIMIT EXCEEDED! NO RESPONSE RECORDED!"

At the completion of each study-test cycle a feedback page displayed percentage of correct responses and encouraged to maintain greater than 60% accuracy at all times. To aid participants with fatigue and in maintaining high levels of concentration, minimum breaks of 30 seconds were enforced between cycles and participants were encouraged to take longer breaks if needed, pressing the space-bar on the keyboard when they felt ready to progress.

Hierarchical Bayesian Models

Parameters for both models were estimated using hierarchical Bayesian analysis. We briefly highlight three advantages of hierarchical Bayesian analysis over conventional methods (for more see Lee, 2011; Lee & Wagenmakers, 2014; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011). First, it provides estimates of both group and participant level parameters, avoiding distortions caused by fitting to group data (Brown & Heathcote, 2003; Estes & Maddox, 2005; Heathcote, Brown, & Mewhort, 2000). Second, the produce posterior distribution that directly quantify estimation uncertainty. This is especially useful for the present purposes, as drift rate variability is notoriously difficult to estimate (Ratcliff & Tuerlinckx, 2002). Third, evidence accumulation models canonically require large numbers of trials per participant to reliably estimate parameters (Wagenmakers, 2009) and the same is true for target variability in SDT (Macmillan, Rotello, & Miller, 2004; Yonelinas & Parks, 2007). Given a sufficient number of participants, hierarchical methods enable better estimation of individual participant’s parameter because they pulled toward the more certain group estimate, a phenomenon referred to as “shrinkage”.

Estimation of posterior distributions requires Markov chain Monte Carlo (MCMC)
algorithms. However, in models such as the DDM and LBA, parameter estimates can be highly correlated with each other (Ratcliff & Tuerlinckx, 2002; Turner, Sederberg, Brown, & Steyvers, 2013), which is problematic for conventional MCMC algorithms. For this reason, we used differential evolution Markov chain Monte Carlo (DE-MCMC: Turner et al., 2013), a method of posterior sampling that is robust to parameter correlations. On each MCMC iteration a proposal for a vector of parameters is generated by randomly sampling two other parameter vectors from other MCMC chains, taking a scaled difference between those parameter vectors, and then adding that scaled difference to the current parameter vector. This makes enables the proposal to take account of correlations (for a detailed and technical description see Turner et al., 2013).

For each class of models, we estimate both equal variance (EV) and unequal variance (UEV) versions. Although the UEV-SDT models have been extensively supported in the ROC literature, no investigations have examined whether UEV assumptions apply to the LBA model. In addition, for the LBA and DDM models, we also applied versions of the model that lack starting point variability to investigate how it contributes to the estimation of the ratio of SDs; we refer to these as the fixed start point (FSP) variants. Note that when starting point variability is removed from the LBA it reduces to the linear approach to threshold with ergodic rate (LATER) model (Carpenter & Williams, 1995).

In SDT, it is conventional to fix the $\sigma$ parameter of the lure distribution at 1 and estimate the variability of the target distribution, which is sometimes referred to as $\tau$ (e.g.: Pratte, Rouder, & Morey, 2010). We do so here and further adopt it for the evidence accumulation by modelling the relative variability of the target distribution $s_{\text{target}}/s_{\text{lure}}$ as $\tau$. We sometimes report $z$ROC slopes for SDT models, for which the appropriate comparison is the inverse of $\tau$ ($s_{\text{lure}}/s_{\text{target}}$). SDT models fits have found that $\tau$ can vary across factors such as word frequency (DeCarlo, 2007; Glanzer & Adams, 1990; Ratcliff et al., 1994) and speed-accuracy emphasis (Starns et al., 2012). Nonetheless, for the DDM, factors such as word frequency and item strength did not produce changes in the
estimate of $\tau$ (Starns & Ratcliff, 2014; Starns, 2014). Thus, for each UEV model we evaluate variants where $\tau$ varies across the factors manipulated in our experiment. These models are described using factor notation. For instance, we denote the word frequency factor as $F$ and speed-accuracy emphasis factor as $E$, and a model where $\tau$ is allowed to vary across word frequency and speed-accuracy emphasis is denoted $\tau \sim E, F$. To decide between each variant, we use model selection approaches which balance the goodness of a model’s fit with a quantitative measure of the model’s complexity.

**Parameterizations**

As mentioned previously, we did not implement an independent race version of the LBA where each decision alternative receives its own drift rate (e.g.: Heathcote, Loft, & Remington, 2015; Rae et al., 2014). Such a model requires drift rates for each condition: a $v_{YES}$ and $v_{NO}$ parameter. This is somewhat contrary to the assumptions of SDT and the DDM where a single estimate of memory strength drives the decision. For this reason, we used a unidimensional LBA where each condition receives its own drift rate and constrain the drift rates of each accumulator to sum to 1. This parameterization makes it such that increases in $v$ produce increases in strength for the "YES" option and corresponding decreases in strength for the 'NO' option, which is similar to the DDM.

**Model Formulations**

In this subsection, we describe which parameters vary over which factors for each of the models. The data indicated no significant differences in HR and FAR for the concreteness factor, and model selection scores did not improve by inclusion of a concreteness factor. For this reason, all models collapsed across high concreteness and low concreteness items.

We assumed a hierarchical structure where each model parameter is sampled from a group level distribution with mean $M$ and standard deviation $\varsigma$. To avoid placing too much constraint on each model, prior distributions on $M$ and $\varsigma$ were only very weakly
informative. Specific values of each prior distribution’s parameters along with details of the hierarchical fitting procedure can be found in Appendix A. Model formulations for each variant along with their number of parameters can be found in Table 2.

SDT. In SDT models, both a $\mu$ and $\sigma$ parameter need to be fixed for the model to be identifiable. We fixed the parameters of the HF lure distribution $\mu_{HF_{lure}}$ and $\sigma_{lure}$ to 0 and 1, respectively. The only mechanism present in SDT for capturing a speed-accuracy emphasis manipulation (E) is to reduce discrimination; thus we allow the $\mu$ parameters to vary over the E factor.

The two stage confidence procedure can be modeled by placing three ordered criteria, $c_1$, $c_2$, and $c_3$ on the memory strength axis. $c_2$ corresponds to the initial yes/no decision, while $c_3$ and $c_1$ correspond to high/low confidence judgments after a "yes" or "no" decision, respectively. Criteria were allowed to vary over the E factor but were fixed across the word frequency factor (F). The $\sigma_c$ was also allowed to vary across the E factor.

We found that the standard SDT model was unable to account for the shapes of the zROCs in the data. Specifically, zROCs were concave down in all conditions, which is contrary to the standard SDT model, which predicts linear zROCs. This is also contrary to both the dual process signal detection model (DPSD: Yonelinas, 1994) and the mixture signal detection (MSD) model (DeCarlo, 2002), both of which predict slightly concave up zROCs.

For this reason, we fit models with variability in the second decision ($c_1$ and $c_3$). When criterion variability, which we denote as $\sigma_c$, is included, the evidence variability $m$ at $c_1$ and $c_3$ is distributed as:

$$m - c \sim N(\mu_m - c, \sqrt{\sigma_m^2 + \sigma_c^2})$$

(1)

where $m$ refers to memory evidence. Criterion variability and memory variability are additive, effectively increasing the total variability of the decision variable. This decreases
performance on the outer two zROC points, producing convex zROCs. Fixed criterion and
criterion variability variants are referred to as FC and CV variants, respectively.

Criterion variability is usually indistinguishable from evidence variability without a
selective influence assumption. Prior work employing criterion variability has employed
selective assumptions similar to ours, such as a manipulation that selectively influences
stimulus variability (Benjamin et al., 2009; Kellen, Klauer, & Singmann, 2012) or a similar
two-stage decision process (e.g., remember-know judgments; Wixted & Stretch, 2004). The
psychological interpretation of the criterion variability variants is that criterion setting for
the subsequent confidence decision is more difficult for participants to maintain than the
initial yes/no decision.

**DDM.** Conventionally in the DDM, only the response boundary parameter $a$ is
allowed to vary across a speed-accuracy emphasis manipulation. However, Rae et al. (2014)
found that several additional parameters were affected by the manipulation, the most
critical of these being the drift rate. We follow suit and allow drift rates to vary across E.
For the EV model, the $sv$ parameter applies to both targets and lures, but in the UEV
variants, $sv$ only applies to lures and $\tau$ reflects the relative variability for targets.

Due to the manipulation of $a$ in our experiment, a fixed value of $sz/a$ would imply
larger absolute levels of starting point variability in the accuracy emphasis condition
relative to the speed emphasis condition. For this reason, we parameterized $sz/a$ relative to
the speeded condition. Model variants with a fixed start point (FSP, $sz/a = 0$) were also fit.

DDM likelihoods were estimated using the partial differential equation method of
Voss and Voss (2008) and a combination of analytic and numerical integration methods of
accounting for parameter variability as implemented in fast-dm-30 (Voss, Voss, & Lerche,
2015).

**LBA.** Similar to the DDM, Rae et al. (2014) found that parameters besides the
response threshold $B$ were affected by the E manipulation, including drift rate. We follow
suit and use their endorsed model. Allowing $B$ to vary across response types accounts for
response bias. Like with the DDM, the \( sv \) parameter corresponds to both targets and lures in the EV model, and applies only to lures in the UEV variants, but was fixed across both accumulators.

LBA likelihoods were calculated using the analytics provided by Brown and Heathcote (2008), with the extension to the truncated normal rate case given in Footnote 1 of Heathcote and Love (2012).

Table 2

*All model variants of the SDT, DDM, and LBA models along with the total number of parameters.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SDT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC-EV</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E )</td>
<td>9</td>
</tr>
<tr>
<td>FC-UEV</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \tau \sim 1 )</td>
<td>10</td>
</tr>
<tr>
<td>FC-( \tau \sim E )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \tau \sim E )</td>
<td>11</td>
</tr>
<tr>
<td>FC-( \tau \sim F )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \tau \sim F )</td>
<td>11</td>
</tr>
<tr>
<td>FC-( \tau \sim E,F )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \tau \sim E,F )</td>
<td>13</td>
</tr>
<tr>
<td>CV-EV</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \sigma_e \sim E )</td>
<td>11</td>
</tr>
<tr>
<td>CV-UEV</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \sigma_e \sim E; \tau \sim 1 )</td>
<td>12</td>
</tr>
<tr>
<td>CV-( \tau \sim F )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \sigma_e \sim E; \tau \sim E )</td>
<td>13</td>
</tr>
<tr>
<td>CV-( \tau \sim E )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \sigma_e \sim E; \tau \sim F )</td>
<td>13</td>
</tr>
<tr>
<td>CV-( \tau \sim E,F )</td>
<td>( \mu_{\text{target}} \sim E,F; \mu_{L\text{Flure}} \sim E; c_1, c_2, c_3 \sim E; \sigma_e \sim E; \tau \sim E,F )</td>
<td>15</td>
</tr>
<tr>
<td><strong>DDM</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSP-EV</td>
<td>( z/a \sim E; a \sim E; v \sim T,E,F; sv \sim 1; t_0 \sim E; s_t \sim E )</td>
<td>19</td>
</tr>
<tr>
<td>FSP-UEV</td>
<td>( z/a \sim E; a \sim E; v \sim T,E,F; sv \sim 1; t_0 \sim E; s_t \sim E; \tau \sim 1 )</td>
<td>20</td>
</tr>
<tr>
<td>FSP-( \tau \sim E )</td>
<td>( z/a \sim E; a \sim E; v \sim T,E,F; sv \sim 1; t_0 \sim E; s_t \sim E; \tau \sim E )</td>
<td>21</td>
</tr>
<tr>
<td>FSP-( \tau \sim F )</td>
<td>( z/a \sim E; a \sim E; v \sim T,E,F; sv \sim 1; t_0 \sim E; s_t \sim E; \tau \sim F )</td>
<td>21</td>
</tr>
</tbody>
</table>

*Continued on next page*
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSP-τ ~ E,F</td>
<td>z/a ~ E; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E; τ ~ E,F</td>
<td>23</td>
</tr>
<tr>
<td>SPV-EV</td>
<td>z/a ~ E; sᵣ/a ~ 1; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E</td>
<td>19</td>
</tr>
<tr>
<td>SPV-UEV</td>
<td>z/a ~ E; sᵣ/a ~ 1; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E; τ ~ 1</td>
<td>20</td>
</tr>
<tr>
<td>SPV-τ ~ E</td>
<td>z/a ~ E; sᵣ/a ~ 1; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E; τ ~ E</td>
<td>21</td>
</tr>
<tr>
<td>SPV-τ ~ F</td>
<td>z/a ~ E; sᵣ/a ~ 1; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E; τ ~ F</td>
<td>21</td>
</tr>
<tr>
<td>SPV-τ ~ E,F</td>
<td>z/a ~ E; sᵣ/a ~ 1; a ~ E; v ~ T,E,F; sv ~ 1; t₀ ~ E; sᵣ ~ E; τ ~ E,F</td>
<td>23</td>
</tr>
<tr>
<td>LBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSP-EV</td>
<td>B ~ R,E; t₀ ~ E; v ~ T,E,F; sv ~ 1</td>
<td>17</td>
</tr>
<tr>
<td>FSP-UEV</td>
<td>B ~ R,E; B ~ R,E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ 1</td>
<td>18</td>
</tr>
<tr>
<td>FSP-τ ~ E</td>
<td>B ~ R,E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ E</td>
<td>19</td>
</tr>
<tr>
<td>FSP-τ ~ F</td>
<td>B ~ R,E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ F</td>
<td>19</td>
</tr>
<tr>
<td>FSP-τ ~ E,F</td>
<td>B ~ R,E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ E,F</td>
<td>21</td>
</tr>
<tr>
<td>SPV-EV</td>
<td>B ~ R,E; A ~ E; t₀ ~ E; v ~ T,E,F; sv ~ 1</td>
<td>17</td>
</tr>
<tr>
<td>SPV-UEV</td>
<td>B ~ R,E; A ~ E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ 1</td>
<td>18</td>
</tr>
<tr>
<td>SPV-τ ~ E</td>
<td>B ~ R,E; A ~ E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ E</td>
<td>19</td>
</tr>
<tr>
<td>SPV-τ ~ F</td>
<td>B ~ R,E; A ~ E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ F</td>
<td>19</td>
</tr>
<tr>
<td>SPV-τ ~ E,F</td>
<td>B ~ R,E; A ~ E; t₀ ~ E; v ~ T,E,F; sv ~ 1, τ ~ E,F</td>
<td>21</td>
</tr>
</tbody>
</table>

Notes: FC = fixed criterion, CV = criterion variability, FSP = fixed start point, SPV = starting point variability, EV = equal variance, UEV = unequal variance, N = number of parameters per participant, “~” = varies as a function of, 1 = one parameter for all conditions, E = speed/accuracy emphasis, F = word frequency, R = response options (yes vs. no), T = item type (target vs. lure).
Model Selection

Prior to fitting the models, all responses faster than .2 or slower than 2.5 seconds were removed from the analysis. Additionally, only trials where participants gave responses to both decisions were included. This resulted in the exclusion of only 1.01% of trials.

Models cannot be assessed on goodness of fit alone, as more complex models will invariably fit better than less complex ones. For this reason, we used a model selection technique that quantifies the complexity of the model and subtract that from a measure of its ability to fit the data, the widely applicable information criterion (WAIC: Watanabe, 2010). In WAIC, model complexity is measured by the variability in the likelihood of a data point across posterior samples summed across all data points. We preferred WAIC to DIC (Spiegelhalter, Best, Carlin, & van der Linde, 2002) because it is more numerically stable than DIC (see Gelman et al., 2014). WAIC is an approximation to leave-out-one cross validation, and so chooses a model on the basis of its ability to predict new data. Smaller values of WAIC mean that a model gives better out-of-sample predictions by striking a better balance between goodness-of-fit and simplicity.

Table 3

*Model selection results with WAIC for the SDT models of Experiment 1. All scores are reported relative to the best performing model (WAIC = 6409), which is highlighted in bold.*

<table>
<thead>
<tr>
<th>SDT</th>
<th>FC</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>2002</td>
<td>1018</td>
</tr>
<tr>
<td>UEV</td>
<td>1017</td>
<td>35</td>
</tr>
<tr>
<td>$\tau \sim E$</td>
<td>983</td>
<td>0</td>
</tr>
<tr>
<td>$\tau \sim F$</td>
<td>1007</td>
<td>31</td>
</tr>
<tr>
<td>$\tau \sim E,F$</td>
<td>964</td>
<td>39</td>
</tr>
</tbody>
</table>

*Continued on next page*
Notes: EV = equal variance, UEV = unequal variance, FC = fixed criterion variants, CV = criterion variability variants, E = speed/accuracy emphasis factor, F = word frequency factor.

Model selection results for the SDT models can be seen in Table 3, and DDM and LBA in Table 4, with WAIC specified relative to the best performing model, which received a score of zero. Since the SDT and DDM/LBA models were evaluated based on different data, their model selection results are not comparable. Similar to other model selection measures like DIC, WAIC is on a log-likelihood scale, meaning that differences between models that are greater than 10 points can be considered large. We will highlight the broad trends here. For the SDT models, the criterion variability (CV) models outperform each of the fixed criterion (FC) models by an very large margin ($\Delta$ WAIC $\sim$ 1000), indicating that the extra variability for the confidence criteria is needed. For the DDM, several of the fixed start point ($s_z/a = 0$) models outperform their counterparts with starting point variability (SPV), indicating that there is no need for starting point variation, but with the exception of the $\tau^{-}$ E,F models. This is interesting considering that the SPV variants are considered the standard in the literature. For the LBA, the SPV variants outperform the FSP variants by a very large margin ($\Delta$ WAIC $\sim$ 2500), indicating a superiority of the LBA over the LATER model. Finally, the winning DDM variant outperformed the winning LBA variant.

Table 4

Model selection results with WAIC for the DDM and LBA models of Experiment 1. All scores are reported relative to the best performing model of the two classes (WAIC = 42765). The best performing model within each class (LBA or DDM) is highlighted in bold.

Continued on next page
The model classes disagree on the appropriate factoring of the $\tau$ parameter. For both the LBA and SDT models, the $\tau \sim E$ model wins, whereas for the DDM, the standard UEV model wins. The SDT results stand in contrast to previous results showing that word frequency decreases zROC slope (DeCarlo, 2007; Glanzer & Adams, 1990; Ratcliff et al., 1994). However, it should be noted that those studies not only used single stage confidence measures, but also employed six confidence options instead of four. The DDM results are
not altogether surprising given prior results showing that manipulations of performance do not generally affect estimates of $\tau$ (Starns & Ratcliff, 2014). Our results demonstrate that this is not the case for the LBA. However, both the DDM and SDT results are inconsistent with current global matching models such as the REM, BCDMEM, and Osth and Dennis (2015) models, which predict lower ratios of lure-to-target variability for LF than HF words.

zROCs

**zROC Shapes.** Empirical zROCs, along with predictions from the winning SDT models, can be seen in Figure 4. The model predictions include both the fixed criterion (FC: left panel) and criterion variability (CV: right panel) variants of the $\tau \sim \mathcal{E}$ model. For both the data and the models, zROCs were calculated for each participant and subsequently averaged together. For the SDT models, predictions for the HR and FAR were derived analytically and subsequently $z$-transformed. Since uncertainty in the predicted $z(\text{HR})$ and $z(\text{FAR})$ was quite small, we depicted only the mean of the posterior predictive distribution for the SDT models.

Inspection of the data reveals non-linear zROCs with inverted V-shapes for all conditions. This is contrary to the standard SDT models which predict linear zROCs, and the DPSD (Yonelinas, 1994) and mixture signal detection (MSD: DeCarlo, 2002) models, which predict U-shaped zROCs. Nonetheless, the CV variant of the model was able to provide an excellent account of the data in all conditions and is able to capture the inverted-V shape. This is because the criterion variability only applies to the second decision, which functions to push down the performance of the first and third zROC points and allows the model to deviate from the linear shape. As mentioned previously, the psychological interpretation is that participants may encounter relatively more difficulty in setting a criterion for the subsequent confidence judgment relative to the initial yes-no decision. This result may also be challenging to evidence accumulation models of the two stage confidence procedure in which evidence continues to accumulate in the confidence
decision, predicting better performance for the second decision relative to the first decision (Pleskac & Busemeyer, 2010; Moran, Teodorescu, & Usher, 2015). However, such models have not been tested using ROC functions, and thus it remains to be seen what kinds of ROC functions they produce and how parameters corresponding to the second decision might alter the function. It remains possible that the better performance effects predicted by these models were occurring but were masked by additional criterion variability in the second decision.

When evaluating ROC shapes it’s important to consider the behavior of individual participants to ensure that the observed group average is not artifactual. Figure 5 shows three individual participants along with model predictions from the selected SDT model. Uncertainty in the posterior predictive distribution is depicted using contour plots, where the thickest lines show the highest posterior density. One can see that the inverted-V shape is evident in some of the individual participant data (S7 and S19). Nonetheless, some participants did not show this pattern, as evident from the linearity in participant S32’s data. Each of these participants are well accounted for by the model.

**Ratio of Lure-to-Target Variability Estimates.** Figure 6 depicts posterior estimates of the ratio of SDs \(1/\tau\) for each of the winning models averaged across participants. Corresponding decision noise variants are also depicted to illustrate how variability in the starting point or criterion affect the ratio estimates. Specifically, we show both the fixed criterion (FC) and criterion variability (CV) variants of the \(\tau \sim E\) SDT model, along with the fixed start point (FSP) and starting point variability (SPV) variants of the UEV DDM and the \(\tau \sim E\) LBA models.

The criterion variability (CV) SDT model yielded mean ratio estimates of .63 and .7 in the speed and accuracy conditions, which is in the ballpark of previous investigations of zROC slopes in recognition memory (e.g.: Heathcote, 2003; Ratcliff et al., 1992, 1994). These ratio estimates were lower than those of the fixed criterion (FC) SDT model (speed \(M = .8\), accuracy \(M = .73\)), which is consistent with previous work showing that models
DIFFUSION VS. LBA

with criterion variability have lower ratios than fixed criterion models (Benjamin et al., 2009). Prior work has shown that LF words produce lower zROC slopes than HF words with LF words showing around a .1 reduction in slope (Glanzer & Adams, 1990; Ratcliff et al., 1994). However, in our study the $\tau \sim E,F$ model produced virtually identical slopes for LF and HF words ($\sim .71$).2

Similar to previous work with the DDM in the binary ROC paradigm (Starns et al., 2012), the DDM produced estimates of the ratio of SDs that were more extreme than those of the SDT models. Contrary to the supposition that this discrepancy is due to the presence of starting point variability, the FSP and SPV variants produced nearly identical ratios ($\sim .57$).

The LBA models produced ratio estimates that were much higher than both other classes of models, with values above 1 in the speeded condition. This stands in stark contrast to previous work showing that models taking account of RT distributions show strong evidence for greater target variability. The LBA models did have greater target variability in the accuracy conditions, but the estimates were still quite close to 1 (SPV $M = .9$, FSP $M = .95$). Similar to the DDMs, the presence or absence of starting point variability did not produce substantial differences in the ratio estimates, with the only difference appearing in the accuracy condition.

Posterior Predictives: Choice Probabilities and Response Times. To evaluate how well the models match the data, posterior predictive distributions were generated by taking $3.33\%$ (1 in every 30) of total posterior samples and simulating the experimental conditions with each parameter set. Predictions were generated for each participant and were subsequently averaged together. Due to space considerations, we restrict depiction of the posterior predictives to winning models, namely the fixed start point (FSP) unequal variance (UEV) DDM and the $\tau \sim E$ LBA model with starting point

2It should be noted that the $\tau \sim E,F$ model produced lower zROC slopes for LF words in the speeded condition ($LF = .54$, $HF = .6$), but lower zROC slopes for HF words in the accuracy condition ($LF = .7$, $HF = .64$).
variability (SPV). To demonstrate the consequences of unequal variance on fits to RT distributions, we additionally show equal variance (EV) counterparts. Posterior predictive distributions can be seen in Figure 7 for the DDM (top) and LBA (bottom) models. Predicted HR and FAR (left panel) are shown along with the correct and error RT distributions (middle and right panels) summarised by 10\textsuperscript{th}, 50\textsuperscript{th} and 90\textsuperscript{th} percentiles.

It is apparent that the difference in fit between the EV and UEV models is extremely subtle. Recall from Figure 3 that changes in $sv$ produce diverging predictions from the DDM and LBA; the LBA predicts a large decrease in the leading edge while the DDM mostly predicts effects in the right tails of the RT distribution. One should note when inspecting Figure 7, however, that there are other differences between targets and lures, such as differences in drift rates and bias, that can prevent a clear qualitative difference between targets and lures being visible. Nonetheless, there appears to be a trend in the data that distinguishes the UEV and EV variants of the DDM: the right tail (90\textsuperscript{th} percentile) of the correct RT distribution is faster for targets in every condition. The UEV DDM appears able to capture these differences and provides a better overall account of the right tails, while the EV model produces tails that are nearly equivalent for targets and lures.

Both the EV and UEV variants of the LBA appear unable to account for the differences between targets and lures. In order for the LBA to produce substantially larger $sv$ estimates for targets, there must be a faster leading edge for targets coinciding with either faster or slower right tails. Inspection of Figure 7 reveals no such consistent trend across the conditions in the experiment. Correct RTs in the speeded condition reveal a contrary trend, namely a greater spread for lures than for targets, which may explain why the LBA produced slightly higher $sv$ estimates for lures than for targets in this condition. The general lack of differences in spread in the accuracy condition likely prevented the LBA from estimating large values of $\tau$ relative to the DDM.
Discussion

We conducted an experiment with a two-stage confidence procedure, where participants give an initial yes/no judgment followed by a confidence judgment. An advantage of this procedure is that it allows for the estimation of the ratio of lure-to-target variability for three classes of models: SDT, DDM, and the LBA. Each of these models produced very different estimates of the ratio of SDs. Consistent with previous evidence from Starns et al. (2012), the DDM produced estimates of the ratio of SDs that were lower than those from SDT. We attempted to evaluate the source of the divergence by applying models with and without contributions of starting point variability, in addition to applying the LBA, a model that lacks within-trial noise. The LBA, in contrast, estimated a ratio of SDs that was much higher than both the DDM and SDT, with roughly equal variability between targets and lures in the speed emphasis condition. This stands in stark contrast to the work of Starns and Ratcliff (2014), who argued that model fits taking into account RT distributions provide strong evidence for greater target variability.

The LBA diverges from the DDM in that it requires a large leading edge difference between targets and lures in order to estimate a high value of \( \tau \), whereas analyses of the .1 quantile in our experiments suggested that these were quite similar between targets and lures across a range of conditions. If the LBA’s account of the data is correct, the evidence for greater target variability may not be as widespread as was previously believed, with speeded responding particularly providing evidence for EV models.

There was also disagreement amongst the models on how the experimental manipulations affect the \( \tau \) parameter, with the LBA and SDT models favoring an effect of speed vs. accuracy emphasis on \( \tau \) and the DDM favoring a model where \( \tau \) is unaffected by experimental manipulations. The LBA and DDM did agree that starting point variability plays only a very minor role in the estimated ratio of SDs.

A potential drawback of the two stage confidence procedure is that each model class’s estimate of the ratio of SDs is constrained by different data. The SDT models are based on
the choices in both decisions, whereas the DDM and LBA models are based on the RT of
the first decision. To address this limitation, we next examine the binary ROC paradigm,
where each model uses the choices in each of the bias conditions to constrain the model
parameters, with the DDM and LBA models gaining the additional constraint of RT
distributions. This may produce substantial constraint on the LBA, as it requires a large
leading edge difference between targets and lures in order to produce a low zROC slope.
An additional advantage is that previous investigations of the binary ROC procedure have
generally not required criterion variability within the SDT framework to achieve a good fit
to the ROC function (Dube & Rotello, 2012; Dube et al., 2012; Starns et al., 2012),
potentially enabling us to use a simpler model. For these reasons we fit two binary ROC
datasets from Dube et al. (2012) to see whether we obtain results consistent with the two
stage confidence procedure. To pre-empt the results of that present analysis, we observed a
similar pattern with the LBA, but somewhat different results for the DDM.

Binary ROC: Dube et al. (2012) Datasets

In two experiments, Dube et al. (2012) manipulated target proportions across five
levels (25%, 33%, 50%, 67%, and 75%), testing both singly and multiply presented items (5
presentations in Experiment 1, 10 presentations in Experiment 2). A summary of each of
the experiments can be found in Table 5. Henceforth, we refer to Dube et al.’s Experiments
1 and 2 as D-DS1 and D-DS2 to distinguish them from our Experiment 1. Due to the
repetition manipulation, we consider EV models, UEV models with a single value of $\tau$
across both strength conditions, and a model where $\tau$ varies across the strength factor ($\tau \sim
S$), as strength has been found to produce larger estimates of $\tau$ in SDT models (Heathcote,
2003).

A common assumption in DDM modeling of response proportion manipulations that
the manipulation affects only the bias or the drift criterion with which decisions are made
(Criss, 2010; Dube et al., 2012; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; White &
Poldrack, 2014). A thorough examination of response proportion effects was carried out by White and Poldrack (2014), who found that the drift criterion did not vary significantly with this manipulation, whereas a manipulation of the criterial evidence required to make a decision – participants were instructed to respond relative to a 'strong' or 'weak' memory criterion – produced changes in the drift criterion. Hence, we did not consider models where the drift criterion varies.

Nonetheless, it remains a possibility that participants do not employ the same speed-accuracy threshold across each response proportion manipulation, producing different levels of accuracy in each condition, and thus a violation of the assumptions of SDT. As mentioned previously, investigations with the RTCON models found strong evidence of participants using different speed/accuracy thresholds for each confidence accumulator, with variation in such speed/accuracy thresholds being responsible for non-linear confidence zROCs (Ratcliff & Starns, 2009, 2013; Voskuilen & Ratcliff, 2016). Thus, for these datasets, in addition to investigating whether EV or UEV models are appropriate for the DDM and LBA, we additionally investigate models where decision boundaries are varied across each target proportion condition. We refer to these variants as varied boundary (VB) models, and we contrast them with fixed boundary (FB) models, which are the standard models for manipulations of response proportion in the DDM. We additionally investigated both fixed start point (FSP) and starting point variability (SPV) variants.

Table 5

Summary of the experiments analysed from Dube et al. (2012).

<table>
<thead>
<tr>
<th>N</th>
<th>Obs.</th>
<th>Cond.</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>480</td>
<td>1x Targets, 5x Targets, Lures</td>
</tr>
<tr>
<td>26</td>
<td>480</td>
<td>1x Targets, 10x Targets, Lures</td>
</tr>
</tbody>
</table>

Notes: N = number of participants, Obs. = mean number of observations per participant, Cond. = conditions
Model Formulations

Model formulations and the number of parameters can be found in Table 5. Here, we briefly expand on details necessary for implementing the models.

**SDT.** For SDT, we investigated a model with one criterion corresponding to each of the proportion conditions, yielding a total of five criteria ($c_{25}$, $c_{33}$, $c_{50}$, $c_{67}$, and $c_{75}$). The model was allowed different $\mu_{\text{target}}$ parameters for the weak and strong conditions ($\mu_{\text{weak}}$ and $\mu_{\text{strong}}$) while $\mu_{\text{ture}}$ and $\sigma_{\text{ture}}$ were fixed to 0. There were only three models under consideration: the EV model, the UEV model, and the $\tau \sim S$ model.

**DDM.** In the FB models, only the response bias $z$ varied across target proportion conditions, yielding five $z$ parameters ($z_{25}$, $z_{33}$, $z_{50}$, $z_{67}$, and $z_{75}$) and a single $a$ parameter. The VB model also has 5 $a$ parameters ($a_{25}$, $a_{33}$, $a_{50}$, $a_{67}$, and $a_{75}$). An additional complication for the VB models is the $s_z/a$ parameterization. Given that $a$ varies across bias conditions in this model, this would entail different magnitudes of $s_z$ in each bias condition. For this reason, we used the threshold for the 50% condition in the denominator ($s_z/a_{50}$) to ensure that $s_z$ is constant across proportion conditions.

**LBA.** Unlike the DDM, the LBA does not have a response bias parameter but can have different thresholds for the accumulators corresponding to each response. Thus, an analog of the FB DDM can be constructed by preserving the average of the threshold parameters ($B_{\text{mean}}$) for each response across the target proportion conditions. We accomplished this by first allowing different thresholds for "yes" and "no" responses for the 50% condition ($B_{\text{OLD50}}$ and $B_{\text{NEW50}}$). For the other target proportion conditions, we estimated a bias parameter $\beta$ ($\beta_{25}$, $\beta_{33}$, $\beta_{67}$, and $\beta_{75}$). The $B$ parameters for target proportion condition $i$ were then parameterized as follows:


\[ B_{OLD_i} = B_{OLD50} + c\beta_i \]  
\[ B_{NEW_i} = B_{NEW50} - c\beta_i \]

where \( i \) is the target proportion condition (excluding the 50% condition) and \( c \) is an indicator variable that is 1 if \( i \) is 25 or 33 and is -1 when \( i \) is 67 or 75. Given that \( \beta \) is positive, this produces high values of \( B_{OLD} \) for conservative conditions (25% and 33% old) and low values of \( B_{OLD} \) for liberal conditions (67% and 75% old). Most importantly, \( B_{mean} \) is preserved across the target proportion conditions, making this analogous to the DDM where only \( z \) varies across bias conditions. For the varied boundary (VB) class of models, we allowed \( B_{OLD} \) and \( B_{NEW} \) to vary freely across each target proportion condition.

Table 6

All model variants of the SDT, DDM, and LBA models along with the total number of parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDT</td>
<td>( \mu_{target} \sim S; \ c \sim P )</td>
<td>7</td>
</tr>
<tr>
<td>EV</td>
<td>( \mu_{target} \sim S; \ c \sim P; \ \tau \sim 1 )</td>
<td>8</td>
</tr>
<tr>
<td>UEV</td>
<td>( \mu_{target} \sim S; \ c \sim P; \ \tau \sim S )</td>
<td>9</td>
</tr>
<tr>
<td>DDM</td>
<td>( z/a \sim P; \ a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1 )</td>
<td>12</td>
</tr>
<tr>
<td>FSP-FB-EV</td>
<td>( z/a \sim P; \ a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim 1 )</td>
<td>13</td>
</tr>
<tr>
<td>FSP-FB-UEV</td>
<td>( z/a \sim P; \ a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim S )</td>
<td>14</td>
</tr>
<tr>
<td>FSP-FB-( \tau \sim P )</td>
<td>( z/a \sim P; \ a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1 )</td>
<td>13</td>
</tr>
<tr>
<td>SPV-FB-EV</td>
<td>( z/a \sim P; \ s_z/a \sim 1; a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1 )</td>
<td>14</td>
</tr>
<tr>
<td>SPV-FB-UEV</td>
<td>( z/a \sim P; \ s_z/a \sim 1; a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim 1 )</td>
<td>14</td>
</tr>
<tr>
<td>SPV-FB-( \tau \sim P )</td>
<td>( z/a \sim P; \ s_z/a \sim 1; a \sim 1; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim S )</td>
<td>15</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 5: Model Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSP-VB-EV</td>
<td>$z/a \sim P; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1$</td>
<td>16</td>
</tr>
<tr>
<td>FSP-VB-UEV</td>
<td>$z/a \sim P; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim 1$</td>
<td>17</td>
</tr>
<tr>
<td>FSP-VB-$\tau \sim S$</td>
<td>$z/a \sim P; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim S$</td>
<td>18</td>
</tr>
<tr>
<td>SPV-VB-EV</td>
<td>$z/a \sim P; s_z/a \sim 1; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1$</td>
<td>17</td>
</tr>
<tr>
<td>SPV-VB-UEV</td>
<td>$z/a \sim P; s_z/a \sim 1; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim 1$</td>
<td>18</td>
</tr>
<tr>
<td>SPV-VB-$\tau \sim S$</td>
<td>$z/a \sim P; s_z/a \sim 1; a \sim P; v \sim T,S; sv \sim 1; t_0 \sim 1; s_t \sim 1; \tau \sim S$</td>
<td>19</td>
</tr>
<tr>
<td>LBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSP-FB-EV</td>
<td>$B \sim R; \beta_{25,33,67,75}; t_0 \sim 1; v \sim T,S; sv \sim 1$</td>
<td>11</td>
</tr>
<tr>
<td>FSP-FB-UEV</td>
<td>$B \sim R; \beta_{25,33,67,75}; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim 1$</td>
<td>12</td>
</tr>
<tr>
<td>FSP-FB-$\tau \sim S$</td>
<td>$B \sim R; \beta_{25,33,67,75}; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim S$</td>
<td>13</td>
</tr>
<tr>
<td>SPV-FB-EV</td>
<td>$B \sim R; \beta_{25,33,67,75}; A \sim 1; t_0 \sim 1; v \sim T,S; sv \sim 1$</td>
<td>11</td>
</tr>
<tr>
<td>SPV-FB-UEV</td>
<td>$B \sim R; \beta_{25,33,67,75}; A \sim 1; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim 1$</td>
<td>12</td>
</tr>
<tr>
<td>SPV-FB-$\tau \sim S$</td>
<td>$B \sim R; \beta_{25,33,67,75}; A \sim 1, t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim S$</td>
<td>13</td>
</tr>
<tr>
<td>FSP-VB-EV</td>
<td>$B \sim R,P; t_0 \sim 1; v \sim T,S; sv \sim 1$</td>
<td>15</td>
</tr>
<tr>
<td>FSP-VB-UEV</td>
<td>$B \sim R,P; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim 1$</td>
<td>16</td>
</tr>
<tr>
<td>FSP-VB-$\tau \sim S$</td>
<td>$B \sim R,P; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim S$</td>
<td>17</td>
</tr>
<tr>
<td>SPV-VB-EV</td>
<td>$B \sim R,P; A \sim 1; t_0 \sim 1; v \sim T,S; sv \sim 1$</td>
<td>16</td>
</tr>
<tr>
<td>SPV-VB-UEV</td>
<td>$B \sim R,P; A \sim 1; t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim 1$</td>
<td>17</td>
</tr>
<tr>
<td>SPV-VB-$\tau \sim S$</td>
<td>$B \sim R,P; A \sim 1, t_0 \sim 1; v \sim T,S; sv \sim 1, \tau \sim S$</td>
<td>18</td>
</tr>
</tbody>
</table>

Notes: FSP = fixed start point, SPV = starting point variability, EV = equal variance, UEV = unequal variance, $N$ = number of parameters per participant, "$\sim"$ = varies as a function of, 1 = one parameter for all conditions, $P$ = proportions of targets and lures (5 levels), $S$ = strength (weak vs. strong), $R$ = response options (yes vs. no), $T$ = item type (target vs. lure).
Model Selection

Prior to fitting the models, all responses that were faster than .25 or slower than 3.0 seconds were removed from the analysis (the same criteria used by Dube et al., 2012, in their DDM fits), excluding 1.13% of trials. The sampling process and assessment of convergence were identical to the previous analysis.

Table 7 provides model selection results for the SDT models and Table 8 for the DDM and LBA models. WAIC results are reported relative to the winning model, which is given a zero value. In SDT, the \( \tau \sim S \) model is preferred for Dube et al.’s Experiment 2 but not 1. This may be because the repetition manipulation is stronger in the second experiment (10 presentations vs. 5 presentations in the first) and because there was a ceiling effect present in their first experiment for the strong items. In addition, the target proportion manipulation appeared to produce more widely spaced ROC points in their Experiment 2, and Dube et al. (2012) noted that this dataset was much more diagnostic for model selection.

For the LBA and DDM models, several trends are present. First, model variants with varied boundaries (VB) are favored over models with fixed boundaries (FB) by a very large margin, with VB models winning by several hundred for both the DDM and LBA models. This finding severely undermines the notion that response proportions selectively influence the response bias parameters, which is surprising when one considers that the VB models are considerably more complex. Second, EV models lose over either UEV or \( \tau \sim S \) models in each case. Nonetheless, there are disagreements among each model type as to how the repetition manipulation affects the \( \tau \) parameter. For the LBA the \( \tau \sim S \) parameter is preferred only in Dube et al.’s Experiment 1 (contrary to SDT, which prefers it only for their Experiment 2), and for the DDM it is preferred for both datasets. However, for each case the preference for the \( \tau \sim S \) over the UEV model is relatively small (\( \Delta \) WAIC < 10) and so the disagreement is only minor.
Table 7
Model selection results with WAIC for the SDT models for the Dube et al.’s (2012) Experiment 1 (D-DS1) and 2 (D-DS2). Results are reported relative to the winning model, which is in boldface (absolute WAIC: D-DS1 = 1657, D-DS2 = 2164).

<table>
<thead>
<tr>
<th></th>
<th>EV</th>
<th>UEV</th>
<th>τ ~ S</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-DS1</td>
<td>213</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>D-DS2</td>
<td>79</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: UEV = unequal variance, S = strength (strong vs. weak).

For the LBA, similar to previous results, the variants with starting point variability (SPV) win over the fixed start point (FSP) variants in every comparison by fairly substantial margins, suggesting it’s a crucial component of the model. The DDM showed the opposite results, with the winning model having a fixed start point in both datasets. Finally, the best LBA variants outperform the best DDM variants in both datasets. This result is contrary to our Experiment 1, where the winning DDM outperformed the winning LBA model.

Table 8
Model selection results with WAIC for the DDM and LBA models for the Dube et al.’s (2012) Experiment 1 (D-DS1) and 2 (D-DS2). Results are reported relative to the winning model (absolute WAIC: D-DS1 = 6372, D-DS2 = 9823). The best model in each class (DDM or LBA) is depicted in boldface.
### DDM

<table>
<thead>
<tr>
<th></th>
<th>FSP</th>
<th>SPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>VB</td>
<td>FB</td>
</tr>
<tr>
<td>EV</td>
<td>712</td>
<td>284</td>
</tr>
<tr>
<td>UEV</td>
<td>643</td>
<td>146</td>
</tr>
<tr>
<td>$\tau \sim S$</td>
<td>551</td>
<td>140</td>
</tr>
</tbody>
</table>

### LBA

<table>
<thead>
<tr>
<th></th>
<th>FSP</th>
<th>SPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>VB</td>
<td>FB</td>
</tr>
<tr>
<td>EV</td>
<td>1019</td>
<td>140</td>
</tr>
<tr>
<td>UEV</td>
<td>992</td>
<td>132</td>
</tr>
<tr>
<td>$\tau \sim S$</td>
<td>961</td>
<td>104</td>
</tr>
</tbody>
</table>

### D-DS2

<table>
<thead>
<tr>
<th></th>
<th>FSP</th>
<th>SPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>VB</td>
<td>FB</td>
</tr>
<tr>
<td>EV</td>
<td>636</td>
<td>91</td>
</tr>
<tr>
<td>UEV</td>
<td>471</td>
<td>40</td>
</tr>
<tr>
<td>$\tau \sim S$</td>
<td>474</td>
<td>31</td>
</tr>
</tbody>
</table>

### DDM

<table>
<thead>
<tr>
<th></th>
<th>FSP</th>
<th>SPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>VB</td>
<td>FB</td>
</tr>
<tr>
<td>EV</td>
<td>1026</td>
<td>288</td>
</tr>
<tr>
<td>UEV</td>
<td>905</td>
<td>217</td>
</tr>
<tr>
<td>$\tau \sim S$</td>
<td>902</td>
<td>213</td>
</tr>
</tbody>
</table>
Notes: EV = equal variance, UEV = unequal variance, FSP = fixed start point variants, SPV = starting point variability variants, FB = fixed boundary variants, VB = varied boundary variants, S = strength (strong vs. weak).

**zROC Shapes.** zROCs calculated from the data along with predictions from the winning SDT, DDM, and LBA models can be seen in Figure 8 using the same methods described previously. To achieve stable estimates of the zROCs from the DDM and LBA, 100,000 simulations were used to generate HR and FAR for each parameter set. Uncertainty in the posterior predictive distribution was portrayed using contour plots; the size of the contours conveys the total uncertainty in the z(HR) and z(FAR).

Inspection of the zROCs reveals that the data are somewhat irregular, and that all models have some difficulty in capturing the shapes of the zROCs. **Note that binary ROCs are inherently more noisy than confidence ROCs.** Because cumulative confidence ratings are used to construct confidence ROC function it must increase monotonically. In binary ROCs, each point on the ROC is independent of the others, which allows for a considerably wider array of possible shapes. Even though the depicted fits are for variants with varied boundaries, which does allow them to capture different levels of accuracy at each point of the zROC function, they are still unable to capture some of the deviations from linear zROC shapes, such as the sharply decreased accuracy of the left-most zROC point in Dube et al.’s (2012) second experiment. This implies that there are constraints in the RT distributions that prevent the models from being able to capture any pattern of zROC
(Voskuilen and Ratcliff (2016) provided a similar demonstration of constrained predictions for RTCON2 in associative recognition).

**Ratio of Lure-to-Target Variability Estimates and Predicted zROC Slopes.**

Figure 9 shows ratios of lure-to-target variability estimates along with the predicted zROC slopes averaged across participants for the winning SDT, DDM, and LBA models. To ensure stable estimates of zROC slopes from the DDM and LBA models, predictions for each data point were generated using 100,000 simulations for each participant’s set of parameters. For both datasets, SDT estimates a relatively low ratio of SDs, .48 for Dube et al.’s Experiment 1 and .63 for their second experiment. Although the $\tau \sim S$ model was the preferred model for their Experiment 2, the mean ratio estimates for weak (.625) and strong (.635) items were extremely similar to each other.

Contrary to the results from our experiment using the two-stage confidence procedure, the $\tau \sim S$ DDMs produced zROC slopes that were somewhat similar to the models underlying SD ratio in each case. Nonetheless, they are considerably higher than those estimated from SDT, averaging about .61 and .73 for weak and strong items in Dube et al.’s Experiment 1, and .77 and .84 for their second experiment. Across the weak and strong conditions, the DDM estimated considerably higher ratios in the strong condition for Dube et al.’s Experiments 1 and 2 ($M = .68$, and $M = .8$, respectively) than the weak condition ($M = .63$ and $M = .6$, respectively).

The LBA variants produced zROC slopes quite similar to the model’s underlying ratio of SDs, but but even higher than those from the DDM. In both datasets, the slopes were around .89 and .95 for weak and strong items for the UEV models. Interestingly, in Dube et al.’s Experiment 1 the estimates of the ratio were very close to 1 for strong items. In addition, EV models were able to produce zROC slopes that were around .95 in each condition.

Once again for both DDM and LBA starting point variability has very little effect on zROC slope. Both the fixed start point (FSP) and starting pointvariability (SPV) variants
of the models estimate very similar ratio of SDs and predict very similar zROC slopes in each condition. This is corroborated by an analysis of the correlations between model parameters and predicted zROC slopes for the DDM and LBA models in Appendix B which found that the magnitudes of starting point variability exhibited very little correlation with the predicted zROC slopes. An additional finding in this analysis was that $\tau$ is much more predictive of the zROC slope for the LBA than the DDM, which suggests that within-trial noise is a major factor that obscures the relationship between the evidence ratio and observed zROC slope.

**Posterior Predictives: Choice Probabilities and Response Times.** Posterior predictive distributions were generated as described previously and we again restrict consideration to the winning models. Due to the substantial noise in the data and model predictions, differences between the EV and UEV variants were mostly noticeable in the choice probabilities. For this reason, we directly contrast the DDM and LBA here; their posterior predictive distributions can be seen in Figure 10 for Dube et al.’s Experiment 1 (top row) and 2 (bottom two row).

Although the RT data are noisy, the DDM appears to be producing a somewhat poorer fit than the LBA, overpredicting the tail of the RT distribution in several of the conditions. This likely explains why the DDM performs more poorly than the LBA in model selection despite the fact that it predicts zROC slopes that are closer to those estimated from SDT than does the LBA. Constraints from the RT distribution may have prevented the LBA from properly capturing the zROC function. Recall that, for the LBA, as $sv$ is increased, there are strong effects on both the leading edge and tail of the RT distribution, while for the DDM the effect of $sv$ is largely in the upper tail (the 90th percentile). Although the 10th percentiles of correct RTs to targets were somewhat faster for targets than for lures ($M_{ lure} = .6, M_{ weak} = .55, M_{ strong} = .55$, for Experiment 1, and $M_{ lure} = .58, M_{ weak} = .53, M_{ strong} = .53$, for Experiment 2), which is consistent with the predictions of the UEV model, the 90th percentiles were fairly consistent across each
category ($M_{lure} = 1.21$, $M_{weak} = 1.25$, $M_{strong} = .99$, for Experiment 1, $M_{lure} = 1.25$, $M_{weak} = 1.25$, $M_{strong} = 1.15$, for Experiment 2). The similar slow tails of the RT distributions for targets and lures may be why the DDM produced excessively long tails for targets.

**Discussion**

We sought to test the generality of our results with the two-stage confidence procedure by applying the SDT, DDM, and LBA models to data from another procedure which allows their ratio of lure-to-target variability estimates to be compared, the binary ROC procedure. In Dube et al.’s (2012) two experiments, bias was manipulated across five target proportion levels and participants responded using the two choice procedure. Results of these analyses largely converged with our Experiment 1 using the two-stage confidence procedure, in that there were large divergences between the estimates of the ratio of SDs ($1/\tau$) among models in each of the datasets, with SDT estimating the lowest ratios, DDM estimating somewhat higher ratios, and the LBA estimating ratios that were close to 1. However, they differ in that the ordering of DDM and SDT ratios is reversed in the results for binary and two-stage procedures, whereas the LBA ratio was consistent.

A potential reason for the contrasting results for the DDM is that the RT distributions for targets and lures were extremely similar in both of Dube et al.’s (2012) experiments. Thus, the DDM’s parameter estimates may reflect a compromise between fitting the zROC slopes and the RT distributions. The LBA predicted slopes that were very close to 1, were nearly equivalent to the evidence ratios, and were much higher than those of SDT. Thus, it was somewhat surprising that the winning LBA variants outperformed the winning DDM variants in model selection, as the LBA models predicted zROC slopes that were much higher than those of SDT. We additionally performed a correlational analysis between model parameters and zROC slopes that found that in the LBA, target variability ($\tau$) was highly predictive of zROC slope, whereas the correlations
were more moderate in the DDM. This suggests that the within-trial noise in the DDM may confound the relationship between target variability and the zROC slope.  

One possible reason why model selection favored the LBA is that the manipulations of choice proportions may not provide a great deal of additional constraint on the estimation of $\tau$ in evidence accumulation models. Evidence for this conjecture comes from our investigation of whether manipulations of target proportion can selectively influence response bias by comparing models with fixed boundaries (FB) across proportion conditions against models with varied boundaries (VB). VB variants greatly outperformed the FB variants for both datasets and for both classes of models. This failure of selective influence is similar to the observations of Ratcliff and Starns (2009) about the confidence ROC procedure, specifically that it cannot be assumed that performance is constant across each bias condition on an ROC, which further obscures the relationship between the ratio of SD estimates in evidence accumulation models relative to those of SDT. Failure of selective influence in the ROC procedure has far reaching implications for measurement that we elaborate further on in the General Discussion.

Parameter Validation: Numerosity Discrimination Data from Starns (2014)

Our results thus far demonstrated that the LBA deviates considerably from the DDM in its estimates of lure-to-target variability, with some cases actually showing evidence for equal variability. This is contrary to the assertions of Starns and Ratcliff (2014), who argued that RT distributions show strong evidence for greater target variability, and contrary to the analyses of Donkin et al. (2011) who demonstrated that the two models yield largely similar psychological conclusions from RT data. We also found that the DDM has an inconsistent ordinal relationship with SDT, estimating lower ratios with the two-choice confidence procedure, but larger ratios with the binary ROC procedure. The LBA, in contrast, consistently estimated ratios that were closer to one than the DDM or SDT for both procedures. We also found a stronger relationship between evidence ratios
and zROC slopes for the LBA than the DDM, suggesting within trial noise in the latter may obscure this relationship. Given these contrasting results, the question arises: as a measurement model which one is providing a better account of the ratio of evidence variability, the LBA or the DDM?

In order to provide independent evidence to address this question we fit the DDM and LBA to three perceptual choice data sets reported by Starns (2014). His experiments explicitly manipulated the variability of stimulus characteristics that provided evidence about the correct response. Participants were presented with between 1 and 100 asterisks and were asked to respond "high" if the number was greater than a criterion of 50 and "low" if it was less. Performance tends to increase as the mean of the numerosity distribution is further from the response criterion and if the variance is lower, which is isomorphic to the SDT account of recognition memory. Starns explicitly manipulated both the across trial mean and variance of the number of asterisks, creating conditions in which means further from the criterion were associated with either more variability, or less, or the same variability as the conditions that were closer to the criterion.

Starns (2014) constructed five numerosity discrimination analogs of memory experiments in his Experiment 1 along with a supplementary experiment; each condition we denote with a letter. The stimulus generating distributions for each are shown in Figure 11. Two recognition-memory experiment analogs used three stimulus generating distributions: one "low" distribution with a mean less than 50 (analogous to an unstudied or lure condition) and two "high" distributions with means greater than 50 (analogous to a studied or target condition), one (denoted "weak") with a mean closer to the criterion and another with a mean further from the criterion (denoted "strong"). In one condition (A) all distributions had roughly the same standard deviation. In a second condition (B) extra variance was added to the strong "high" distribution, consistent with recognition memory findings that targets have higher evidence variability.

Three further numerosity conditions were analogous to a source memory design,
where zROC slopes have indicated higher variability for sources studied with high strength (Starns, Pazzaglia, Rotello, Hautus, & Macmillan, 2013). There were four stimulus generating distributions: two low and two high. In the equal variance (C) experiment, all distributions had the same variability. In an unequal variance (D) experiment, the strongest low and high distributions had the highest variability. To ensure that strong conditions don’t invariably show higher estimates of $sv$, Starns ran an additional analog of the source experiment with unequal variance (E), where the weak distributions had higher variance than the strong distributions. Details of these datasets are described in Table 9.

Table 9

<table>
<thead>
<tr>
<th>N</th>
<th>Obs.</th>
<th>Cond.</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>1336.35</td>
<td>A,B,C,D</td>
</tr>
<tr>
<td>21</td>
<td>1086.05</td>
<td>E</td>
</tr>
</tbody>
</table>

Notes: $N = \text{number of participants}$, $\text{Obs.} = \text{mean number of observations per participant}$, $\text{Cond.} = \text{conditions}$

Starns (2014) found that the DDM was able to recover the relative orderings of the generating distributions in the $sv$ estimates. We applied both the DDM and the LBA to his data (details of our modeling procedures for both the DDM and LBA can be found in Appendix C). Our results are depicted in Figure 12. To ease comparison with the results reported so far, we constructed an analog of lure-to-target variability by taking the ratio of
for a given condition relative to the $sv$ from the low condition in each experiment. In most cases, this was a low-to-high variability estimate, with the exception of the source memory cases where there was a low-to-low variability comparison due to the manipulation of variability of one of the "low" distributions. For the low-to-low variability estimates, the numerator always had the lowest variability (weak low distributions in D and E, strong low distributions in E).

The cognitive representation of evidence about numerosity may be no more than montonically related to actual numerosity, and so only ordinal rather than quantitative correspondence might be expected. Nonetheless, it’s impressive how well the DDM is able to recover the actual ratio in each of the datasets considered except the equal-variance recognition-memory experiment analog (A) where it underestimates the ratio in the strong high case. The LBA, in contrast, behaves in a manner quite similar to our fits to recognition memory data in that it produces ratios that are quite close to 1 in each case. In two experiments, the LBA reproduces the unequal variance pattern. This is clearly so in (E), where the weak distributions have high variance, for both the weak low and high distributions (95% credible intervals 0.85-0.9 and 0.82-0.87 respectively), and just in (B), where the strong distribution has high variance (0.94-0.99). However, in (D), where low and high strong distributions have high variance, they are clearly in the wrong direction for both the strong low and high distributions (1.07-1.15 and 1.02-1.1, respectively). This would be analogous to estimating higher variability for targets than lures in a recognition memory experiment.

Why does the LBA reproduce the correctly ordered ratios in some cases but not in others? It’s important to recall that for the LBA, a high value of $sv$ implies a fast leading edge of the RT distribution. The strong conditions, which yield high performance, require strong values of drift rates, which also produce faster leading edges. The data (not depicted here for space considerations), in contrast, do not show radically fast leading edges in strong conditions with high variance. Thus, the model is forced to compromise between an
accurate estimation of sv, performance, and the leading edge. In condition E, the strong conditions have low variability, and the model is able to sensibly recover the sv parameter because the mean drift rate and sv parameters are not placed in conflict with one another.

Absolute values of the estimates of sv can be found in Appendix C. Analysis of these estimates found some unusual trends; for distributions that were unchanged across experimental conditions, neither model produced identical estimates of sv across each condition. This suggests that the DDM may only be able to recover the relative differences in variability across conditions, whereas a direct measurement of evidence variability may not be tenable from either model.

General Discussion

ROC functions have been widely used to test models of memory. Nonetheless, results from the ROC procedure have often yielded inconsistent shapes and results that sometimes deviate radically from the predictions of models, leading researchers to consider other means of constraint such as response times (Ratcliff & Starns, 2009, 2013; Starns et al., 2012). Our work extends this line of research by comparing three major models of decision making, namely signal detection theory (SDT), the diffusion decision model (DDM), and the linear ballistic accumulator (LBA), to evaluate how each estimates the ratio of lure-to-target variability in recognition memory. We did this using two ROC measurement methods, a two-stage confidence procedure from a new experiment and a bias manipulation in a binary ROC procedure (using two datasets from Dube et al., 2012). Consistent with the analyses of Starns et al. (2012), in our experiment using the two-stage confidence procedure we found major deviations between the SDT estimates of the evidence ratio and those derived from the DDM. Deviations were also observed between the two models’ estimates in the analyses of the datasets of Dube et al. (2012), but in the opposite direction, with the DDM estimating higher ratio of SDs than SDT.

Ratcliff and Starns (2009) conjectured that observed deviations between the DDM
evidence ratio and the zROC slopes could be due to the fact that evidence accumulation models consider other sources of variability, such as within-trial noise and variability in the starting point of accumulation. We evaluated this conjecture using two methods: a) applying the LBA, a model which lacks within-trial noise, and b) fitting models that lacked starting point variability. We found that starting point variability had very little role in the estimation of either the evidence ratio or on predicted in zROC slopes. For the DDM, model selection preferred models that lacked starting point variability in each dataset considered. The primary role of starting point variability in the DDM is to accommodate errors that are faster than correct responses (Laming, 1968; Ratcliff et al., 1999). However, such fast error patterns are not generally observed in recognition memory tasks (Ratcliff & Smith, 2004), presumably because recognition memory is dominated by evidence variability, which produces errors that are slower than correct responses (Ratcliff, 1978). These analyses suggest that the DDM with no starting point variability is a viable model to apply to recognition memory, which has some practical significance for researchers as this reduces an additional integration term in the model which makes model fitting an order of magnitude faster. For the LBA, in contrast, models with starting point variability were always selected over models without. However, as was the case for the DDM, starting point variability had very little role in the estimation of either the evidence ratio or on predicted zROC slopes. Thus, both our DDM-based and LBA-based results refute Ratcliff and Starns’ conjecture with respect to start-point variability.

Although prior work with the DDM has consistently shown higher variability for targets, analyses of this type have not been conducted with the LBA. We tested a version of the LBA designed to be analogous to the DDM in assuming there is a single source of evidence that drives both accumulators, such that each condition receives a single drift rate and the mean drift rate of both accumulators sums to one. Although Donkin et al. (2011) demonstrated strong correlations between the DDM and LBA on their primary parameters (i.e., non-decision time, thresholds and mean drift rates), we showed that changes in the
standard deviation of the drift rate distribution ($sv$) yields very different effects in the two models, with our version of the LBA showing a major speed-up in the leading edge of the RT distribution as $sv$ is increased while the DDM showed very little change. These differences imply that the two models could yield very different estimates of the ratio of lure-to-target variability, and this is what we found; the LBA producing consistently higher evidence-ratio estimates than the DDM, and in one case (the speed emphasis condition of Experiment 1) the LBA estimates of target and lure variability were indistinguishable.

The analyses of the binary ROC datasets of Dube et al. (2012) allowed for direct predictions of zROC slopes from the models. Both the DDM and the LBA differ from SDT in that simple SDT models have a one-to-one relationship between the evidence ratio and the zROC slope, whereas the evidence accumulation models have a more complex relationship. Indeed, in Appendix B we found that: a) numerous parameters affected zROC slope and b) the relative variability of targets ($\tau$) for the DDM was only moderately predictive of zROC slope ($r \sim -.45$), whereas the LBA showed much stronger correlations ($r \sim -.8$). These results suggest that the within-trial noise in the DDM obscures the relationship between the evidence ratio and the predicted zROC slope, affirming the other conjecture of Ratcliff and Starns.

The divergence between the DDM and the LBA is problematic because it casts doubt on the ability of evidence accumulation models to test predictions from memory models, and motivated us to replicate and extend the parameter validation exercise conducted by Starns (2014). He fit the DDM to data from several experiments using a numerosity discrimination task, and showed that the DDM’s $sv$ parameter estimates were affected in the same direction as explicit manipulations of the trial-to-trial variability of numerosity across experimental conditions. We replicated this result for the DDM using our hierarchical Bayesian methods and additionally fit the LBA using the same methods. The DDM was successful in recovering whether the generating case was equal or unequal in variance in a majority of the cases. The LBA was successful when the generating distributions were
equal in variance, and when the high variance condition was associated with worse performance. However, it failed in the cases where the better performing "high" and "low" distributions had high variance, likely because the high performing drift rates and the high required value of $sv$ were placed in opposition to each other in capturing the leading edge of the RT distribution.

These findings shed light on why the LBA estimated equal variance in some of the recognition studies we examined, such as the speeded condition of Experiment 1 and the strong items in the first dataset of Dube et al. (D-DS1). If targets have stronger drift rates than lures, it’s possible that the LBA has to compromise by underestimating the variability of targets in order to capture the leading edge in the RT distributions. This yields the hypothesis that our conditions that estimated equal variance should have stronger drift rates for targets than for lures, while our conditions that yielded greater target variability should have either comparable drift rates between targets and lures or stronger drift rates for lures. In order to test this hypothesis, drift rates for targets and lures have to be placed on a comparable scale. In the LBA model we applied, where drift rates sum to 1, the drift rate for the 'yes' accumulator is $v$ and the 'no' accumulator is $1 - v$. Targets perform better with a high value of $v$ while lures perform better with a high value of $1 - v$. Thus, to place targets and lures on a scale where higher values indicate better performance in each case, we compared $v_{\text{target}}$ against $1 - v_{\text{lure}}$. The results for each of the datasets can be seen in Figure 13.

The drift rates for targets and lures were consistent with our hypothesis. In Experiment 1 (left panel), the speeded condition, which resulted in roughly equal variability between targets and lures in the LBA, shows higher drift rates for targets for both HF and LF items. The accuracy condition, which resulted in larger target variability, shows similar drift rates for targets and lures. For the binary ROC datasets of Dube et al. (2012), the first dataset (D-DS1) showed higher variability for strong targets than weak targets, with strong targets exhibiting only slightly higher variability than lures. Consistent
with our hypothesis the mean drift rates for strong targets were roughly equal to those for lures, whereas weak drift rates were much lower. For the second dataset (D-DS2), the selected model only used a single \( s_v \) parameter for targets that was shared across weak and strong items, and this was higher than the variability for lures, again consistent with the hypothesis.

The deviations between the DDM and the LBA highlight a crucial role for within-trial noise in estimating evidence ratios. Brown and Heathcote (2008) made a case for linear (noise-free) accumulation on the grounds of computational convenience, as it allows for easily computable analytical expressions without sacrificing the ability to capture RT distributions and the relative speeds of correct and error responses under speed and accuracy emphasis. The strong correlations between the DDM and the LBA in estimates of drift rate, boundary separation, and nondecision time found by Donkin et al. (2009) suggested that within-trial noise may be unnecessary for researchers who are merely interested in measuring these quantities. However, for cases where measurement of the evidence ratio is necessary, the constrained one-dimensional version of the LBA we considered here may not be appropriate, although that conclusion may not generalise to other parameterisations of the LBA or different types of racing accumulator models. In any case, of the models considered here, the DDM seems preferable for this purpose. Our results also clearly indicate that the evidence ratio estimates provided by SDT without taking account of RT may be problematic, as they can be not only greater than those provided by the DDM in the two-stage procedure, but also less than the DDM in the binary ROC procedure.

Failure of Selective Influence in the ROC Procedure

Ratcliff and Starns (2009, 2013) demonstrated failures of selective influence with confidence based ROCs by demonstrating that participants often adopt different speed-accuracy thresholds for each confidence response, which produced a wide variety of
ROC shapes. Our results generalize this finding to the binary ROC procedure, in which participants respond using only two choices and bias is manipulated by varying the proportions of targets and lures. Despite the disagreement between the LBA and DDM in the ratio of lure-to-target variability estimates, both gave strong preference for the target proportion manipulation affecting the speed-accuracy threshold as well as the bias of responding, implying that there are different levels of accuracy across each bias condition.

These results stand in contrast to the assumptions of measurement models such as SDT or high threshold models, which require the assumption that sensitivity is constant for each point on the ROC. A recent debate in the literature concerns whether the form of the binary ROC conforms to the predictions of SDT models, which predict curvilinear ROCs (Dube & Rotello, 2012; Dube et al., 2012; Starns et al., 2012) or high threshold models, which predict linear ROCs (Bröder & Schütz, 2009; Kellen, Klauer, & Bröder, 2013). It is possible that the failure of selective influence in the procedure has contributed to a difficulty in resolving this debate. Indeed, the forms of the binary ROCs observed across the diverse datasets collected in Figures 5 and 6 of Kellen et al. (2013)’s article shows a wide variety of ROC shapes that do not easily conform to the predictions of either of the model classes.

The failure of selective influence of the ROC procedure has implications that extend beyond model testing. The ROC procedure is used to separate sensitivity from bias in a wide range of psychological tasks, including medical diagnostics and child welfare referrals (Rotello, Heit, & Dube, 2015). Within the domain of episodic memory, there has been an increasing reliance on ROCs for measurement of eyewitness memory. This has led to some major changes in the literature, including a preference for simultaneous over sequential lineups, which is the opposite of what is recommended when performance is calculated using the diagnosticity ratio (Clark, Benjamin, Wixted, Mickes, & Gronlund, 2015; Mickes, Flowe, & Wixted, 2012; Wixted & Mickes, 2014). Estimation of sensitivity may be contaminated by different levels of accuracy in each ROC point and the conclusions drawn
from choice alone using these procedures may, therefore, be erroneous.

Independent Race Models

Following Donkin et al. (2011), we have focused on the “sum-to-one” variant of the LBA because it is most analogous to the DDM in that a single source of evidence can be used to drive a decision. This "unidimensional" LBA is analogous not only to SDT models, but also to the Exemplar Based LBA model (EB-LBA), where the summed similarity between the cues and the exemplars present is the evidence driving a decision (Brown & Heathcote, 2008; Donkin et al., 2009; Donkin & Nosofsky, 2012b, 2012a; Donkin et al., 2011; Hawkins et al., 2016; Nosofsky et al., 2014). Up to now we have focused mainly on within-trial noise in the DDM as enabling it to better able than the LBA to measure evidence variability. However, only one LBA parameter needs to be fixed to make the model identifiable (Donkin et al., 2009). Hence, a second possibility is that the unidimensional LBA is over-constrained, which occurs because it fixes two parameters, one in assuming that $sv$ is the same for both accumulators and another through the sum-to-one constraint on the rates for the two accumulators.

In many LBA applications only one parameter is fixed (Heathcote & Love, 2012; Rae et al., 2014; Heathcote, Loft, & Remington, 2015; Provost & Heathcote, 2015; Heathcote, Suraev, Curley, Gong, & Love, 2015), with separate mean rates estimated for each accumulator and only one $sv$ fixed, typically for the accumulator that mismatches the stimulus (e.g., in recognition memory the “yes” accumulator for a lure or the “no” accumulator for target). In every such case model selection favours a higher $sv$ for the mismatching than matching accumulator, and this version of the LBA is also usually selected over the DDM. Hence, it appears that the assumption of equal variance across accumulators and/or sum-to-one mean rates may not be appropriate for the LBA.

These considerations suggest it may not be straightforward to combine the LBA, and perhaps other racing accumulator models, with the majority of process models of
recognition memory, which produce a single evidence value. In this sense the DDM provides a more natural “back-end” decision process for these process models. When a race model provides the back end, as in the case of the EB-LBA model, the single evidence value is transformed into positive (i.e., target supporting) and negative (i.e., lure supporting) evidence. However, in short-term recognition memory there is evidence that suggests positive and negative evidence are not a simple transformation of the same underlying evidence (E. E. Johns & Mewhort, 2002, 2003; D. Mewhort & Johns, 2000). Further, Mewhort and Johns (Experiment 7) showed that different types of positive and negative evidence can also be necessary in long-term study-test list paradigm of the sort examined in this paper. Hence, it may be inappropriate use unidimensional evidence, or a simple transformation thereof, as input to a racing accumulator decision process. Instead, process models may need to generate separate types of positive and negative evidence, as in the iterative resonance (D. J. K. Mewhort & Johns, 2005) and the recognition through semantic synchronization (B. T. Johns, Jones, & Mewhort, 2012) models.

Conclusion

We applied three prominent models of decision-making, the SDT, DDM, and LBA, to two different paradigms that allow for the construction of ROCs. Each model produced dramatically different estimates of the ratio of SDs in addition to different conclusions about how manipulations that affect performance, such as word frequency, speed-accuracy emphasis, and repetitions, affect the relative variability of the lure and target distribution. A parameter validation exercise using data from a numerosity discrimination task suggested that neither the LBA nor DDM provided direct estimates of the absolute magnitude of trial to trial stimulus variability across experiments, but that the DDM appeared better able to recover ratios of variability within an experiment. This suggests that the DDM might provide a better model for recognition memory decisions when measurement of the evidence ratio is crucial. Consideration of the unidimensional nature of
evidence assumed by process models of recognition memory also suggests that the DDM may provide a more compatible “back-end” decision-process model.
Figure 1. Basic description of the Ratcliff diffusion model (DDM). The drift rate, $v$, for a given trial is a sample from a normal distribution with standard deviation $sv$ relative to a drift criterion $d_c$. This drift rate is then used to drive a diffusion process (bottom panel). During the diffusion process, evidence accumulates through the trial, beginning at the starting point $z$, and continues until either the upper response boundary ($a$) or the lower response boundary ($0$) is reached. The boundary that is reached is the choice, and the time taken to reach the boundary is the response time (RT). Evidence accumulation is noisy, such that diffusion processes with the same drift rate will often reach different boundaries and produce different RTs; three sample trajectories with the same drift rate are depicted.
Figure 2. The linear ballistic accumulator (LBA) model. Each response option has its own accumulator; in the case of two choice recognition memory, these correspond to "YES" (left accumulator) and "NO" (right accumulator) responses. Each accumulator has its own boundary at height $b$. Accumulation is driven by the drift rate, which is a sample from a normal distribution with mean $v$ and standard deviation $sv$. The starting point is sampled from a uniform distribution with range $A$. $B$ denotes the distance from the height of the starting point distribution to the response boundary.
Figure 3. Illustration of the manipulation of the $sv$ parameter on both the DDM (top), LBA (middle), and racing noisy accumulators (bottom). Predictions are depicted for the .1, .5, and .9 quantile summaries of the RT distribution for both correct (left panel) and error (right panel) RTs. DDM parameters are: $v = .5, 1.5, 3.0$, $a = 1.5$, $z/a = .5$, $sz/a = .1$, $s_t = .1$, $t_0 = .4$. LBA parameters: $v = .6, .7, .8$, $A = .2$, $b_{YES} = .6$, $b_{NO} = .6$, $t_0 = .4$. Noisy accumulator model parameters (using LBA parameter labels) are $v = .5, 1.5, 3.0$ (constrained to sum to 3.5), $b_{YES} = 1.5$, $b_{NO} = 1.5$, $A = .15$, and $t_0 = .4$. 
**Figure 4.** Average zROCs in the data along with the averaged predictions from both the fixed criterion (left panel) and criterion variability (right panel) variants of the $\tau^{-} E$ SDT model variants. Notes: spd. = speed emphasis condition, acc. = accuracy emphasis condition.
Figure 5. zROCs from the data and the selected SDT model for three individual participants. Notes: spd. = speed emphasis condition, acc. = accuracy emphasis condition.
Figure 6. Ratio of lure-to-target standard deviations for the winning SDT, DDM, and LBA models (see text for details) averaged across participants. Notes: FSP = fixed start point, SPV = starting point variability, FC = fixed criterion, CV = criterion variability.
Figure 7. Posterior predictive distributions for both the DDM (top row) and LBA (bottom row) along with the data for the HR and FAR (left panel) and correct and error RT (middle and right panels). In the RT plots lower lines correspond to the .1 quantile, middle lines to the .5 quantile and upper lines to the .9 quantile.
Figure 8. Average zROCs in the data along with the averaged predictions from the SDT (left), DDM (middle), and LBA (right) models for the data from Dube et al.’s first (D-DS1, top row) and second datasets (D-DS2, bottom row). For the DDM the depicted model was the fixed start point (FSP) variant with varied boundaries (VB). For the LBA model, the depicted model was the starting point variability (SPV) variant with varied boundaries (VB).
Figure 9. Ratio of SD estimates along with predicted zROC slopes for the SDT, DDM, and LBA models for the two datasets of Dube et al. (2012, D-DS1 and D-DS2). Notes: FSP = fixed start point, SPV = starting point variability.
Figure 10. Posterior predictive distributions for both the DDM and LBA along with the data for the HR and FAR (left column) and correct and error RT (middle and right columns) for D-DS1 (top row) and D-DS2 (bottom row).
Figure 11. Stimulus generating distributions for Starns’s (2014) numerosity discrimination experiments.
Figure 12. Ratio of variability estimates averaged over participants for the five numerosity discrimination cases of Starns (2014) for the LBA and DDM models along with the ratios of the generating distributions. Conditions A and B are recognition-memory experiment analogs and C-E are source-memory experiment analogs. For C and D the numerator of the ratio was the weak low distribution while for E the numerator was the strong low distribution.
Figure 13. LBA drift rates for targets (v) and lures (1 − v) averaged over participants for each condition in Experiment 1 (left panel) and the two datasets of Dube et al. (right panel).
Appendix A
Details of Hierarchical Bayesian Modeling

Prior Distributions on Model Parameters for Recognition Memory Data

Participant parameters are sampled from group level mean and standard deviation parameters $M$ and $\varsigma$. We begin with discussion of the DDM:

\[ z/a \sim TN(M_z, \varsigma_z, 0, 1) \] (4)
\[ s_z/a \sim TN(M_{sz}, \varsigma_{sz}, 0, 1) \] (5)
\[ a \sim TN(M_a, \varsigma_a, 0, \infty) \] (6)
\[ t0 \sim TN(M_{t0}, \varsigma_{t0}, 0, \infty) \] (7)
\[ s_t \sim TN(M_{st}, \varsigma_{st}, 0, \infty) \] (8)

Because the $z/a$ and $s_z/a$ parameters are proportions and their constituents are positive these parameters fall between zero and one, so were truncated to the $(0, 1)$ interval. The remaining parameters are bounded below at 0 but unbounded on the right.

Participants’ drift rate variability parameters are treated as samples from zero truncated normal distributions, and $\nu$ parameters are sampled from normal distributions because drift rates can be negative or positive.

\[ \nu_{target} \sim Normal(M_{\nu_{target}}, \varsigma_{old}) \] (9)
\[ \nu_{lure} \sim Normal(M_{\nu_{lure}}, \varsigma_{new}) \] (10)
\[ s\nu_{lure} \sim TN(M_{s\nu_{lure}}, \varsigma_{s\nu_{lure}}, 0, \infty) \] (11)
\[ \tau \sim TN(M_{\tau}, \varsigma_{\tau}, 0, \infty) \] (12)

The EV variants of the conventional DDM lack a $\tau$ parameter and instead have only a single $s\nu$ parameter that corresponds to both targets and lures.
For the group level mean \((M)\) parameters, we used mildly informative priors:

\[
M_{z,sz} \sim \text{TN}(0.5, 0.5, 0, 1) \quad (13)
\]
\[
M_{st} \sim \text{TN}(0.25, 0.25, 0, \infty) \quad (14)
\]
\[
M_{t0} \sim \text{TN}(0.5, 0.5, 0, \infty) \quad (15)
\]
\[
M_a \sim \text{TN}(2, 2, 0, \infty) \quad (16)
\]
\[
M_{r,sr} \sim \text{TN}(1, 1, 0, \infty) \quad (17)
\]
\[
M_{\text{vtarg}et} \sim \text{Normal}(2, 2) \quad (18)
\]
\[
M_{vlure} \sim \text{Normal}(-2, 2) \quad (19)
\]

For the group level standard deviation \((\varsigma)\) parameters we used the following mildly informative priors:

\[
\varsigma_a, \varsigma_s, \varsigma_v, \varsigma_{vold}, \varsigma_{vnew}, \varsigma_{d}, \varsigma_{dc} \sim \Gamma(1, 1) \quad (20)
\]
\[
\varsigma_{z, sz, st, ter} \sim \Gamma(1, 3) \quad (21)
\]

For the LBA, participant level parameters are sampled from the following group level distributions:

\[
A \sim \text{TN}(M_A, \varsigma_A, 0, \infty) \quad (22)
\]
\[
B \sim \text{TN}(M_B, \varsigma_B, 0, \infty) \quad (23)
\]
\[
s_{vlure} \sim \text{TN}(M_{svlure}, \varsigma_{svlure}, 0, \infty) \quad (24)
\]
\[
\tau \sim \text{TN}(M_{r}, \varsigma_{r}, 0, \infty) \quad (25)
\]
\[
t0 \sim \text{TN}(M_{t0}, \varsigma_{t0}, 0, \infty) \quad (26)
\]
\[
s_t \sim \text{TN}(M_{st}, \varsigma_{st}, 0, \infty) \quad (27)
\]
For the group level mean parameters, we used mildly informative priors that were nearly identical to those employed by (Hawkins et al., 2016):

\[ M_{A,B,svlure} \sim TN(5, 5, 0, \infty) \]  
\[ M_{vtarget} \sim TN(.75, .75, 0, \infty) \]  
\[ M_{vlure} \sim TN(.25, .75, 0, \infty) \]  
\[ \tau \sim TN(1.0, 1.0, 0, \infty) \]

For the group level standard deviation parameters we used the following mildly informative priors:

\[ \varsigma_{A,B,vtarget,vlure,svlure} \sim \Gamma(1, 3) \]  
\[ \varsigma_{svlure,\tau} \sim \Gamma(1, 1) \]

For the SDT models, we use the following parameterizations for participant level parameters:
For the group level mean parameters, we use the following mildly informative priors:

\[
\begin{align*}
\mu_{\text{target}} & \sim \text{Normal}(M_{\mu_{\text{target}}}, \varsigma_{\mu_{\text{target}}}) & (34) \\
\mu_{\text{lure}} & \sim \text{Normal}(M_{\mu_{\text{lure}}}, \varsigma_{\mu_{\text{lure}}}) & (35) \\
\mu_{c1} & \sim \text{Normal}(M_{c1}, \varsigma_{c1}) & (36) \\
\mu_{c2} & \sim \text{Normal}(M_{c2}, \varsigma_{c2}) & (37) \\
\mu_{c3} & \sim \text{Normal}(M_{c3}, \varsigma_{c3}) & (38) \\
\mu_{c5} & \sim \text{Normal}(M_{c5}, \varsigma_{c5}) & (39) \\
\mu_{c33} & \sim \text{Normal}(M_{c33}, \varsigma_{c33}) & (40) \\
\mu_{c50} & \sim \text{Normal}(M_{c50}, \varsigma_{c50}) & (41) \\
\mu_{c67} & \sim \text{Normal}(M_{c67}, \varsigma_{c67}) & (42) \\
\mu_{c75} & \sim \text{Normal}(M_{c75}, \varsigma_{c75}) & (43) \\
\tau & \sim \text{TN}(M_{\tau}, \varsigma_{\tau}, 0, \infty) & (44) \\
\sigma_c & \sim \text{TN}(M_{\sigma_c}, \varsigma_{\sigma_c}, 0, \infty) & (45)
\end{align*}
\]
Given that a majority of the $\mu_{lure}$ parameters were fixed at 0, we similarly placed the mean of the prior for the $M_{\mu_{lure}}$ parameter (which only correspond to LF lures) at 0.

Priors on the group means for the criteria ($M_c$) were selected to be in the approximate regions as their selected values. For the group level standard deviation we similarly use mildly informative priors:

$$\varsigma_{\mu_{target}, \mu_{lure}, c_1, c_2, c_3, c_{25}, c_{33}, c_{50}, c_{67}, c_{75}, \sigma_c} \sim \Gamma(1, 1)$$

**Details on MCMC Procedure for Recognition Memory Data**

For each model, the number of chains was set equal to three times the number of parameters. After 2,500 burn-in iterations were discarded, the MCMC chains were thinned by only accepting 1 sample every 10th iteration which continued until 1,500 MCMC iterations.
samples were collected for each chain. Convergence was assessed using the Gelman-Rubin statistic; a model was considered converged if this statistic was less than 1.1 for all parameters at both the subject and group level. This criterion was satisfied for all models. Chains were also visually assessed for convergence.
Appendix B

Correlations between Model Parameters and Predicted zROC Slopes

To evaluate how well each of the model parameters’ roles in generating zROC slopes, all parameters were pooled across posterior samples and participants and correlations were calculated between the predicted zROC slope and each parameter. These results can be found in Figure C1 for each of the winning models along with their variants that lacked or contained starting point variability. What is first apparent from the figure is that a number of parameters outside the $\tau$ parameter are predictive of zROC slope, which is contrary to the assumptions of SDT where slope $= 1/\tau$. What is quite surprising is that drift rates are quite predictive of zROC slopes in each of the models, with $v_{lure}$ exhibiting correlations as strong as .45 in the DDM and .25 in the LBA. In addition, consistent with the findings that estimates of $1/\tau$ were relatively unaffected by the presence of starting point variability, estimates of $s_z/a$ in the DDM exhibit only weak correlations with the zROC slope, with the largest correlation reaching .14 for D-DS2. Finally, $\tau$ exhibited much stronger correlations with the predicted slopes in the LBA ($r \sim -.8$) than with the DDM ($r \sim .45$).
Appendix C

DDM and LBA Estimates of sv in the Numerosity Discrimination Experiments of Starns (2014)

Prior to applying any models to the data, response times faster than .25 and slower than 3.5 seconds were removed from the data. For both the DDM and LBA we fit hierarchical Bayesian models estimating a single parameter of each type for all conditions except v and sv, where we allowed separate estimates for the three conditions in the item-recognition analogue experiments (lure and strong and weak target conditions) and the four conditions in the source-memory analogue experiments (strong vs. weak crossed with male vs. female). This is resulted in 11 and 13 parameters per participant for the DDM and 10 and 12 parameters per participant for the LBA respectively for the two types of experiment. Each model parameter was assumed to be drawn from a truncated Gaussian population distributions. For the DDM (assuming a diffusion constant of 1), the truncation range for a was (0,5), for v (-2,7), for sv, (0,3), for z and sz (as proportions of a) (0,1), and for the two non-decision time parameters (0.1,1) seconds. For the LBA the only upper truncation was for non-decision time, at 1s, and the only lower truncations were at zero for the all the v parameters, and 0.1s for non-decision time. For both the LBA and DDM the priors for the population standard deviation parameters were exponential was a scale parameter of one and the the population mean priors were normal with the same truncations as their corresponding parameters. For the DDM the these priors had means of two for the a, v and sv parameters, 0.5 for z, 0.2 for sz, and 0.4s and 0.1s for the non-decision time mean and range. Prior standard deviations were 2 for all except the z and sz parameters, and for the non-decision time mean and range (in each case 1 and 0.5 respectively). For the LBA these priors had means of one for the start-point noise and threshold and sv parameters, two for the v parameters, and 0.3s for non-decision time. Their standard deviations were one for all but the v parameters, which had standard deviations of two.

Sampling was performed using the same methods as elsewhere in the paper, with thinning set at 15 or the LBA and 20 for the DDM. All results are based on 3300 and 3900
thinned samples for the DDM and 3000 and 3600 thinned LBA samples for the recognition and source analogue experiments respectively (i.e., chains of length 100 and three times as many chains as participant parameters). The DDM and LBA generally performed comparably according to WAIC (we give DDM - LBA WAIC differences, so positive values favour the DDM, SE = standard error): in the item-recognition analogue experiments the LBA was clearly preferred in the equal variance experiment (-238, SE = 71) but results were equivocal in the unequal variance experiment (56, SE = 57). In the source-recognition analogue experiments selection was equivocal for the equal variance case (2.6, SE = 58), and for the unequal variance experiments it favored the DDM when the stronger condition was more variable (132, SE = 53) and was equivocal when the weaker condition was more variable (105, SE = 96).

Table C1

Posterior median \( sv \) estimates for Starns’s (2014) recognition memory analog numerosity experiments with (A) equal variance and unequal variance (B), with greater strong high variance. Unequal variance entries in bold should be greater than corresponding equal variance entries, and entries in italics should be equal. In order to facilitate comparison with Starns’s results we scaled our results to his assumption that the diffusion coefficient is 0.1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>LBA</th>
<th>DDM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equal (A)</td>
<td>Unequal (B)</td>
</tr>
<tr>
<td>Strong High</td>
<td>0.186</td>
<td><strong>0.211</strong></td>
</tr>
<tr>
<td>Weak High</td>
<td>0.186</td>
<td>0.208</td>
</tr>
<tr>
<td>Low</td>
<td>0.187</td>
<td>0.203</td>
</tr>
</tbody>
</table>

The underlying estimates of \( sv \) are given in Tables C1 and C2. Comparison of our results with those of (Starns, 2014) shows a good correspondence between our different
Table C2

Posterior median sv estimates for Starns’s (2014) source memory analog numerosity experiments with (C) equal variance and unequal variance, with either (D) greater strong high variance or (E) greater weak variance. Unequal variance entries in bold should be greater than corresponding equal variance entries, and entries in italics should be equal. In order to facilitate comparison with Starns’s results we scaled our results to his assumption that the diffusion coefficient is 0.1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>LBA</th>
<th>DDM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equal (C)</td>
<td>Unequal (D)</td>
</tr>
<tr>
<td>Strong Low</td>
<td>0.186</td>
<td><strong>0.204</strong></td>
</tr>
<tr>
<td>Weak Low</td>
<td>0.188</td>
<td><strong>0.228</strong></td>
</tr>
<tr>
<td>Weak High</td>
<td>0.195</td>
<td><strong>0.218</strong></td>
</tr>
<tr>
<td>Strong High</td>
<td>0.187</td>
<td><strong>0.216</strong></td>
</tr>
</tbody>
</table>

estimation methods. The DDM always estimated sv to be greater for the distribution with greater variance (bold entries in the tables). The LBA produced the same pattern for all conditions except D, where the high variance distributions are strong. Comparison across the different conditions reveals marked differences between the two models. Experiments with distributions having the same generating variance (e.g., the weak high generating distribution was identical for conditions A through D) should yield the same estimate of sv. However, not only did these estimates of sv change across conditions for each model, they also changed in different ways. For the LBA, the equal variance conditions yielded lower estimates of sv than the unequal variance conditions, whereas for the DDM estimates of sv were higher in equal variance conditions.

These results clearly illustrate that a direct relationship between numerosity and evidence variability is untenable for either model. Instead it appears the experimental
context in which numerosity variability is experienced scales the way in which it affects evidence variability. Importantly, the scaling for the DDM happens in a way that largely preserves the ratios between conditions, whereas for the LBA it does not, so that where the ratio is of interest the DDM appears to provide a better measurement model. Further, comparison of estimates for conditions in the unequal variance experiments that have the same numerosity variance as the corresponding equal variance experiment show opposing trends (italic entries in the table); the former are always greater than the latter for the DDM, but always less for the LBA. The same underestimate by the DDM is also reported by (Starns, 2014). However, it should also be noted that given that all parameters varied across blocks, it is possible that differences across blocks in these other parameters affected the estimates of $sv$ across each of the conditions.
References


Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and...


DIFFUSION VS. LBA


**Figure C1.** Correlations between the DDM (left) and LBA’s (right) parameters and the predicted zROC slope, collapsed across conditions, participants, and posterior samples, for D-DS1 (top) and D-DS2 (bottom). Depicted are the starting point variability (SPV) and fixed start point (FSP) variants of the winning DDM and LBA models (described in the text).