Correlated Racing Evidence Accumulator Models

Angus Reynolds
University of Tasmania

Peter D. Kvam
University of Florida

Adam F. Osth
University of Melbourne

Andrew Heathcote
University of Tasmania

Many models of response time that base choices on the first evidence accumulator to win a race to threshold rely on statistical independence between accumulators to achieve mathematical tractability (e.g., Brown & Heathcote, 2008; Logan et al., 2014; Van Zandt et al., 2000). However, it is psychologically plausible that trial-to-trial fluctuations can cause both positive correlations (e.g., variability in arousal, attention or response caution that affect accumulators in the same way) and negative correlations (e.g., when evidence for each accumulator is computed relative to a criterion). We examine the effects of such correlations in a racing accumulator model that remains tractable when they are present, the log-normal race (LNR Heathcote & Love, 2012). We first show that correlations are hard to estimate in binary choice data, and that their presence does not noticeably improve model fit to lexical-decision data (Wagenmakers et al., 2008) that is well fit by an independent LNR model. Poor estimation is attributable to the fact that estimation of correlation requires information about the relationship between accumulator states but only the state of the winning accumulator is directly observed in binary choice. We then show that this problem is remedied when discrete confidence judgments are modelled by an extension of Vickers' (1979) "balance-of-evidence" hypothesis proposed by Reynolds et al. (submitted). In this "multiple-threshold race" model confidence is based on the state of the losing accumulator judged relative to one or more extra thresholds. We show that not only is correlation well estimated in a multiple-threshold log-normal race (MTLNR) model with as few as two confidence levels, but that it also resulted in clearly better fits to Ratcliff et al.'s (1994) recognition memory data than an independent mode. We conclude that the MTLNR provides a mathematically tractable tool that is useful both for investigating correlations between accumulators and for modelling confidence judgments.

Racing evidence accumulator models are widely used to model dynamic decision processes (Donkin & Brown, 2018). In these models, a decision process is represented by multiple racing accumulators, one corresponding to each possible choice. For example, a binary decision has two accumulators, each to their own threshold, with the first finishing accumulator determining both the choice and the corresponding decision time. Response time (RT) is the sum of decision time and the time taken to encode the stimulus and produce a response (t_{er}).

Several such models have been proposed that differ in the distributional assumptions they make. For example, the Linear Ballistic accumulator (LBA) assumes two types of trial-to-trial variability that is independent over accumulators, a uniform distribution of distances from the start-point of accumulation to threshold, and normally distributed rates of evidence accumulation (Brown & Heathcote, 2008). Other models assume an important role for moment-to-moment noise during accumulation (e.g., Heathcote, 2004; Logan et al., 2014; Van Zandt et al., 2000), but the LBA assumes accumulation to be well-approximated by a constant rate for the duration of each race, hence the term 'ballistic' (although 'deterministic' is probably more appropriate, as a change in input can change the trajectory of accumulation, see Brown & Heathcote 2005). The log-normal race (LNR) is another deterministic accumulation model that makes an even simpler distribution assumption; that the finishing times of each accumulator have a log-normal distribution, corresponding to rates and/or the start-point to threshold distance having a log-normal distribution (Heathcote & Love, 2012).

Both the LBA and LNR assume independence between trials (the outcome of one trial does not affect the outcome of following trials) and independence between accumulators within each trial (e.g., the rate of the left accumulator is independent of the rate of the right accumulator). In this paper we will focus on the second, within-trial independence, assumption. Heathcote & Love (2012) showed that the likelihood for a correlated LNR model has a simple analytic form, but did not attempt to fit such a model to data. Given that only the finishing time of the winning accumulator is ever observed, it may seem questionable whether correlation is identifiable given that the only thing known about the the losing accumulator is that its evidence total is less than that of the winner when it reached threshold.

In first section of this paper we define and apply the correlated LNR model to the same binary lexical-decision (word vs. nonword) data that was fit by Heathcote & Love (2012) with an independent LNR. We demonstrate that the correlation parameter is indeed difficult to identify in such binary choice data. Heathcote & Love demonstrated that in order to model errors that are as fast or faster than correct responses, variability in the finishing time of the the accumulator that matches the stimulus (i.e., the accumulator that corresponding to the correct choice for a given stimulus) must be smaller than for the mismatching accumulator (i.e., the accumulator corresponding to the wrong choice). We show that estimating a correlation is particularly difficult in this case, because its effects can be closely mimicked by unequal variance in an independent model.

In the second section of this paper we show the parameters of the correlated LNR are identifiable when participants give confidence ratings that are modeled using an extension of Reynolds et al.'s (submitted) multiple-threshold instantiation of "Balance of Evidence" theory (Vickers, 1979). In this extension, the choice is determined, as in a standard choice race model, by the first accumulator to cross its choice threshold. Confidence is determined by the evidence total of the losing accumulator relative to thresholds placed below its choice threshold. We show that a correlated LNR model of this type provides a good fit to recognition-memory data with confidence ratings (Ratcliff et al., 1994). We also show that in this setup correlations are identifiable even in the difficult unequal-variance case.

First, however, we motivate our analysis by discussing how different evidence-accumulation models are affected by within-trial correlations and what psychological mechanisms might cause either positive or negative correlations.

Psychological Mechanisms Causing Correlation

The issue of correlation has perhaps not been of much concern to evidence-accumulation modellers because it is not relevant for the most popular account, the diffusion (DDM, Ratcliff, 1978). The diffusion has only a single accumulation process and the input to that process can be conceptualized as the difference in evidence for the two alternatives. This means that the effect of differences in evidence correlations can be fully accommodated by differences in the diffusion model's trial-to-trial rate variability parameter. Indeed, there is a large class of models with correlated parameters that are equivalent to the diffusion model (Bogacz et al., 2006). In contrast, racing accumulator model like the LBA and LNR, and non-deterministic versions such as racing single-barrier diffusion models that assume trial-to-trial variations in parameters (Leite & Ratcliff, 2010; Logan et al., 2014; A. Osth & Farrell, in press), are sensitive to the effects of such correlations, and hence to the effects of the psychological mechanisms causing correlations that we now discuss.

In racing evidence accumulation models one of the chief motivation for independence in accumulation rates, start points, or thresholds across accumulators is mathematical simplicity. Assuming independence results in likelihood functions that are easy to derive and that can be quickly computed analytically. Additionally, even in the simple case of binary choice, which will focus on in this paper, being able to separately consider the matching and mismatching accumulator makes interpreting parameter estimates much simpler. The matching accumulator is only affected by evidence for a correct response and its starting point and threshold, and the mismatching accumulator only by evidence for the incorrect response and own starting point and threshold. Further, even if an independence assumption does not represent the data generating process it can be the case that parameter estimates for the independent model remain useful, and they may even be preferable if there is a sufficient payoff in terms of estimation properties (see van Ravenzwaaij et al. 2017 for a similar argument with respect to trial-to-trial variability in the diffusion model).

However, there are a range of mechanisms that could potentially cause the rates of both accumulators to rise and fall together (i.e., a positive correlation), or for one to rise when the other falls, or vice versa (i.e., a negative correlation). Negatively correlated rates can arise when judging a uni-dimensional stimulus (x) against a stimulus criterion (c), such as classifying all perceptual stimuli above the criterion as "big" and all those below as "small" (Leite & Ratcliff, 2011; White & Poldrack, 2014). If the input to the "big" accumulator is $x - c$, and to the "small" accumulator $c - x$, the two inputs will be perfect negatively correlated. A less than perfect, but still negative, correlation would ensure if there were also an independent addition of noise to each accumulator (see Brown et al., 2008, for a related example).

Positive correlations can result if trial-to-trial fluctuations in attention or arousal scale evidence-accumulation rates by the same amount in all accumulators. In the cognitive literature the idea that attention has a variable "sensitivity" or "gain" is part of both older theories, such as the spotlight theory of visual attention (e.g., Posner & Boies, 1971) and more recent work, such the Prospective Memory Decision Control theory of dual-task costs in prospective-memory paradigms (Boag et al., 2019). The normalization framework that is prominent within the neurosciences (Carandini & Heeger, 2011) also features the idea of a broadly-tuned gain or signal-boost mechanism. Often, there is also overlapping activation between neural populations coding evidence for different alternatives in a choice set, such as when the alternatives are have some similarity because they share some common features or fall along an ordinal scale. Such similarity has been suggested to translate into positive correlations in rate across accumulators (see Figure 8 of Kvam, 2019).

Factors inducing positive and negative correlations could be simultaneously present, so that the resulting overall correlation between rates for each accumulator will represent their aggregate effect, which could be positive, negative, or even approximate independence. Some factors that potentially cause positive correlations in rates, such as trial-to-trial fluctuations in arousal, could also cause positive correlations in evidence thresholds. For example, a less cautious setting (i.e., a higher start point and/or lower threshold for both accumulators) when arousal is higher. Positive correlations could also be associated with post-error slowing, which has been explained as occurring due to increased response caution in order to increase accuracy (e.g., Dutilh et al., 2013). In the LNR separate effects of rates and the distance between accumulation start-points and thresholds are not identifiable. Hence, the aggregate effect of all of the various potential factors causing correlations in variables that affect accumulation is reflected in the binary choice case in a single correlation. Our aim in this paper is to explore the properties of this single correlation in the LNR, leaving the more difficult investigation of the different underlying causes to later research.

The Log-normal Race

The LNR assumes log-normal distributions for evidence accumulation rates and for start-point to threshold differences. The ratio of two log-normal distributions is also log-normal, so these two components are non-separable and we are left with a log-normal distribution of threshold-crossing times for each accumulator. Hence, for simplicity we will refer to the rates with the understanding that they can also reflect effects on the distance between start points and thresholds. In particular, we will address the log-rate, as the logarithm is normally distributed, and so it is fully characterized by its mean and variance (i.e., the log-mean,μ and log-variance, σ^2). Further, when we refer to accumulators as having equal or unequal variances we mean the log-variance.

The outcome of a log-normal race is the minimum of two or more log-normal distributions, one driven by each alternative. The minimum or maximum of multivariate normal distributions has been extensively studied and we can use these results to implement correlated accumulator models. The minimum of correlated bivariate normals can be described analytically as a function of normal CDFs and PDFs. Nadarajah & Kotz (2008) describes various useful results for the maxima and minima of bivariate normal distributions. Since the logarithmic function is monotonic, these results apply directly to the log of $RT - t_{er}$, which we will refer to as decision time.

Consider a bivariate normal distribution: $[X_1, X_2] \sim MVN(\mu, \Sigma)$ with vector of μ= $[\mu_1, \mu_2]$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. If $Z = min(X_1, X_2)$ (i.e., the winning time) then the probability density function of Z is

$$g(Z) = g_1(Z) + g_2(Z) \tag{1}$$

where each function on the right hand side are defective densities for either minimum. For example, the integral of $g_1(Z)$ over the real number line gives the proportion of the time X_1 is the minimum.

To use as an LNR model we take the above formula and make Z the log of decision time. Sometimes the minimum of the bivariate normal could be non-identified, such as when know decision time but not which response is the minimum. When we do know this, we do not need to take the sum of each equation, but can instead focus on the the equation that relates to the appropriate response. We can also plot each equation separately to represent correct and error RT densities.

$$g_1(Z) = \frac{1}{\sigma_1}\phi\Big(\frac{Z-\mu_1}{\sigma_1}\Big) * \Phi\Big(\frac{\rho(Z-\mu_1)}{\sigma_1\sqrt{1-\rho^2}} - \frac{Z-\mu_2}{\sigma_2\sqrt{1-\rho^2}}\Big) \tag{2}$$

$$g_2(Z) = \frac{1}{\sigma_2}\phi\Big(\frac{Z-\mu_2}{\sigma_2}\Big) * \Phi\Big(\frac{\rho(Z-\mu_2)}{\sigma_2\sqrt{1-\rho^2}} - \frac{Z-\mu_1}{\sigma_1\sqrt{1-\rho^2}}\Big) \tag{3}$$

The functions ϕ and Φ represent the PDF and the CDF of the standard normal distribution, respectively.

Basu & Ghosh (1978) showed that the distribution of the minimum uniquely determines the bivariate distribution, except for the trivial case where the first and second elements of the bivariate distribution switch places. Since our RT data is identified with a choice, this trivial case can be ignored. This means that there is no exact mapping from the pair of defective densities in the independent case to the correlated case, although as we show very close approximations are possible.

As described in Heathcote & Love (2012), the likelihood corresponding to each defective density can also be written as a product of log-normal densities and survival functions. If f and S are the probability density function and survival function (1-CDF) of log-normal distributions respectively, the likelihood of the first accumulator winning with decision time (dt) is:

$$L_1(dt) = f(dt|\mu_1, \sigma_1^2) \times S(dt|\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(ln(dt) - \mu_1), (1-\rho^2)\sigma_2^2). \tag{4}$$

The equation for an independent LNR simplifies to.

$$L_1(dt) = f(dt|\mu_1, \sigma_1^2) \times S(dt|\mu_2, \sigma_2^2). \tag{5}$$

We fit the LNR model using these likelihoods in a hierarchical Bayesian framework using Differential Evolution-MCMC sampling (Ter Braak, 2006; Turner et al., 2013) as implemented in the R software DMC (Heathcote et al., 2019). We used sampled a linear transformation of the logistic function ($\lambda(x) = 1/(1 + e^{-x})$) of the correlation parameter so that samples could range over the real line, then transformed them back to (-1,1). To allow for unequal variances and enforce diagonal dominance of the covariance matrix (i.e., covariance cannot exceed the variance of either accumulator), the minimum of the ratio of the two σ values, $c = min(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})$, is used as a scale factor, so the linear transformation is $2c(\lambda(\rho) - c)$. Other details of the fitting methods and priors are provided in supplementary materials. The data sets and code to perform the fits is available at osf.io/4hn7p/

An LNR model with equal match and mismatch accumulator variance can be motivated from the negative correlation mechanism described previously. Suppose that the two stimulus types to be classified produce subjective values distributed on a log-scale of $S_1 \sim N(\mu_1, \sigma_1^2)$ and $S_2 \sim N(\mu_1, \sigma_1^2)$ that are transformed into the evidence values that determine log-rates by comparison to a criterion on the log scale, c, where typically $\mu_1 < c < \mu_2$, so that $E_1 \sim c - S_1$ and $E_2 \sim S_2 - c$. Hence, for stimulus 1 the log-rate for the matching accumulator is $r_{match|1} \sim N(c - mu_1, \sigma_1^2)$ and for the mismatching accumulator $r_{mismatch|1} \sim N(\mu_1 - c, \sigma_1^2)$. Similarly for stimulus 2, $r_{match|2} \sim N(\mu_2 - c, \sigma_2^2)$ and for the mismatching accumulator $r_{mismatch|2} \sim N(c - \mu_2, \sigma_2^2)$. In both cases the matching and mismatching accumulators have equal variance and a perfect negative correlation. This will also be the case if the criterion has a normal distribution, but the magnitude of the negative correlation will approach zero if independent criterion samples apply to each accumulator. If the same criterion sample applies then it will also reduce, and can even become positive if the criterion variance is large enough.

Unequal variance between matching and mismatching accumulators can occur if variability in log-rates differs as a function of the average degree of match. For example, Ratcliff et al. (2018) suggested trial-to-trial rate variability increases with the mean, which would result in greater variance for the matching than mismatching accumulator (assuming above chance accuracy). Heathcote & Love (2012) suggested that the mismatching accumulator might have more variance—the pattern they found in their fit of the independent LNR to the data from Experiment 1 of Wagenmakers et al. 2008)—if a template-matching process produces evidence (and hence the log-rate) for each response, such that a poorer match producing outputs that are not only weaker but also more variable.

Using the equation for the minimum of a bi-variate normal, we can predict the accuracy of a response given the decision time and how this varies with the relative variances of matching and mismatching accumulators. If we make $g_1(Z)$ the correct response and $g_2(Z)$ the error, the estimated accuracy at time t is $\frac{g_1(t)}{g_1(t)+g_2(t)}$. An equal variance model predicts that accuracy is highest for short decision time and decreases with increasing decision time, and so that error tend to be slower than correct responses, as is usually observed when accuracy is emphasized over speed. If the matching accumulator is more variable, initially same decrease in accuracy occur, but this can then reverse at longer decision times. Finally, if the mismatching accumulator has a higher variance, then it is possible to have accuracy increase with decision time and hence to predict that errors are faster than correct response, as can occur when speed is emphasized over accuracy.

We note one final point about the correlated LNR germane to modeling response confidence. The balance-of-evidence hypothesis states that confidence is proportional to the difference in the evidence totals (and hence the log-rates) of the two accumulators. If the log-rates are perfectly negatively correlated then there is a perfect negative relationship between decision time and the evidence-total difference (and hence confidence). That is, there is no overlap in decision time for different levels of confidence. However, as the magnitude of the correlation decreases so does the the link between confidence and decision time, so when there is a perfect positive correlation there is no relationship. Before addressing confidence data, we first compare the fit of the independent and correlated LNR models to Wagenmakers et al.'s (2008) binary-choice lexical-decision data (. For brevity, we call the independent LNR the ILNR and the correlated LNR the CLNR.

Fitting Binary Choice Data

Wagenmakers et al.'s (2008) lexical-decision task required participants to classify high, low and very-low frequency words and non-words. As shown in Figure 1, instructions emphasizing the speed of responding produced faster errors than instructions that emphasized the accuracy of responding, particularly for the easier conditions (i.e., high-frequency words and non-words) and for faster responses (i.e., the 10^{th} percentile of the RT distribution).

We first compared the best unequal-variance ILNR model from (Heathcote & Love, 2012) with an equal-variance CLNR model with the same number (34) of parameters.The latter model had half of its σ parameters replaced by correlation parameters. Additionally a simpler CLNR model with only one correlation parameter was fit. According to the DIC (Deviance Information Criterion) model-selection criterion Spiegelhalter et al. (2002) the ILNR model provided the best fit (DIC = -20535), the most complex CLNR model fared worst (DIC = -19378), and its less complex version was intermediate but still much worse (DIC = -19764). Figures 1 and 2 graphically illustrate the fits; the correlated model is clearly unable to fit the fast errors that occur in speed-emphasis conditions.

We performed a cross-fit simulation study (i.e., fitting one model to data simulated from another como to exam-

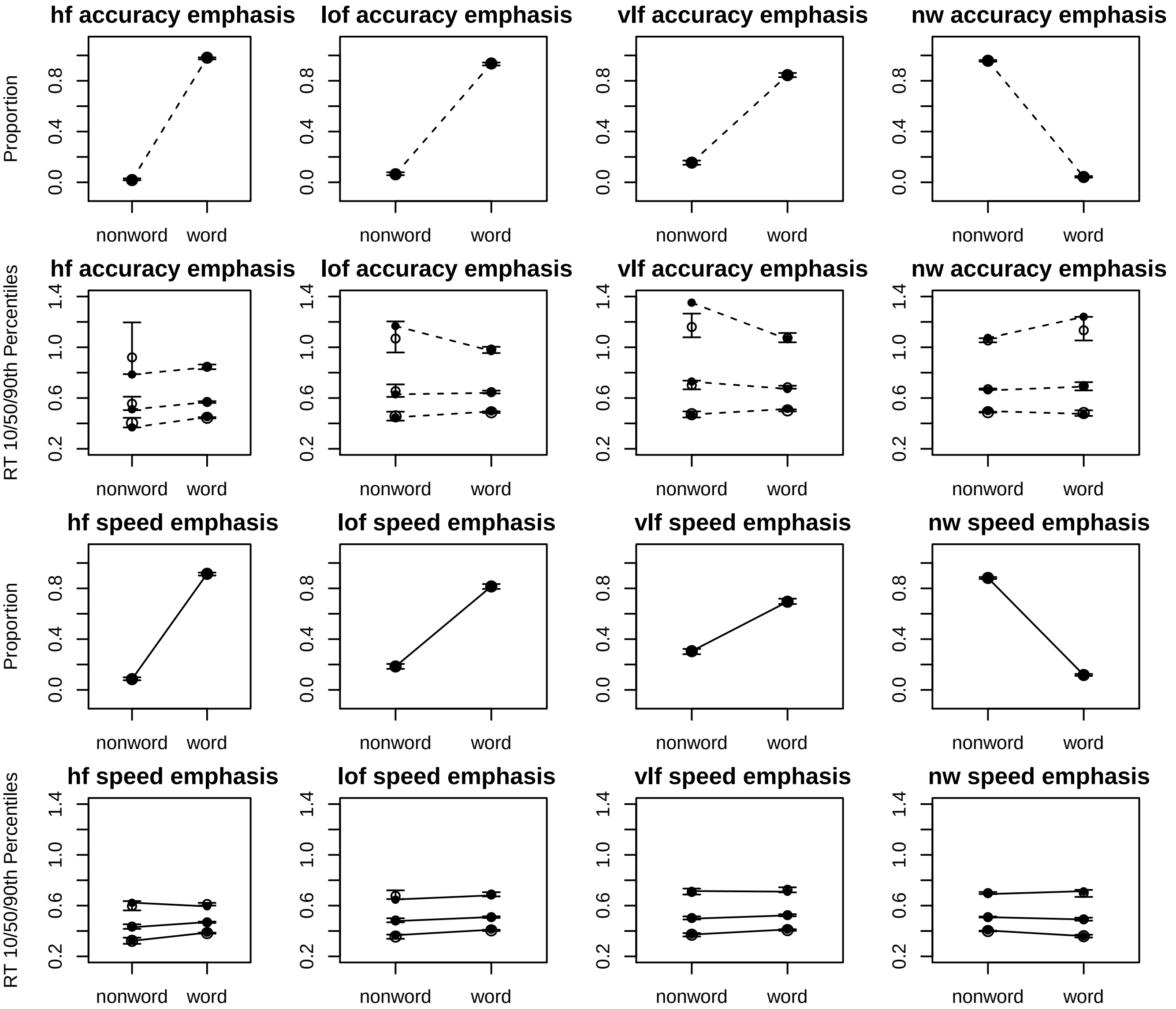

Figure 1. Binary choice fits with 95% credible intervals for the ILNR model. hf = high-frequency word, lf = low-frequency word, vlf = very-low frequency word, nw = non-word.

ine del)mparinckryg model mimithe ILNR model with the CLNR model with the same number of parameters. Based on the posterior means from the fits of each model to the original data for each of the 17 participant, we simulated a large number of trials (19200, 10 times as many as in the original data) in order to investigate approximately asymptotic estimation performance. We then fit both models to these simulations to compare the ability of one model to mimic the other's data. Results are presented in supplementary materials in the form of post-predictive plots comparing defective RT CDFs of the simulated data to fits. The ILNR model was able to match the simulated CLNR model almost exactly, but the CLNR model could not so closely match the simulated ILNR data. This suggests that with the same number of parameters, an unequal-variance ILNR is more flexible than an equal-variance CLNR model, at least when it must account for faster errors.

Finally, we fit a model with both unequal variance and a single correlation parameter. This model had a better DIC than any previous model, although the improvement was not large (DIC = -20553) and visual inspection of the fit (see supplementary materials) revealed no noticeable improvement. Estimates of the difference between match and mismatch accumulator variance were greatly reduced relative to the ILNR model. However, there was very little updating between the prior and posterior density for the correlation

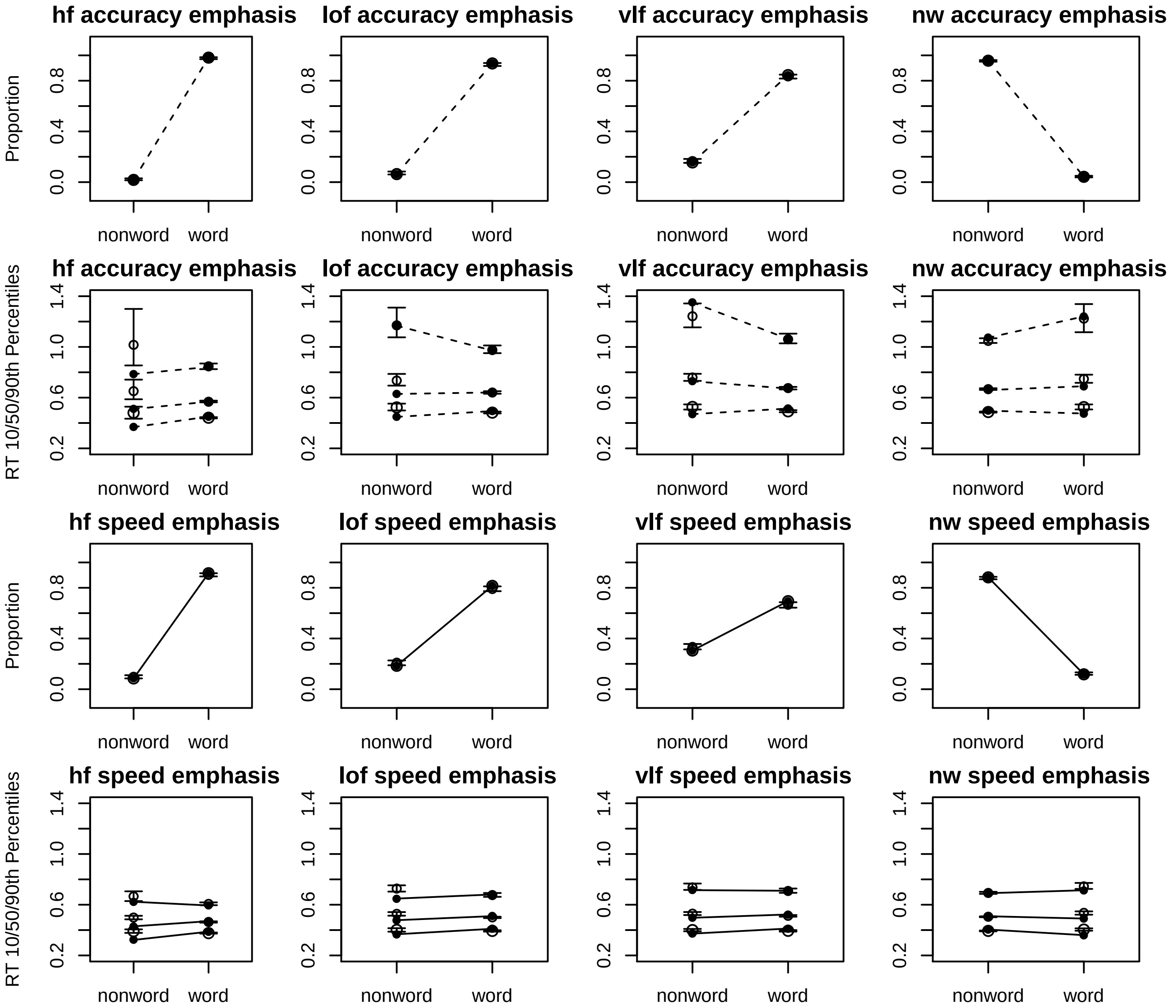

Figure 2. Binary choice data and fits with 95% credible intervals for the CLNR model with the same number of parameters as the ILNR model.hf = high-frequency word, lf = low-frequency word, vlf = very-low frequency word, nw = non-word.

parameter. There was also a large increase in the correlations among posterior parameter estimates, suggesting this model is subject to parameter identification difficulties. Indeed, even the fits of the equal-variance CLNR model to the large simulated data sets displayed similar difficulties. This suggests that, although the unequal-variance CLNR appears to provide a good description, it is unlikely to produce useful parameter estimates based on realistic choice RT data. In the next section we show that not only are these estimation difficulties ameliorated when fitting correlated models to confidence data, but that there can then also be a qualitative increase in goodness-of-fit over independent models.

The Multiple Threshold Race

It is difficult to identify the correlation parameter because it depends on the relationship between the state of the winning accumulator, which is observed at the moment of choice, and the state of the losing accumulator, which is unknown except that it is less than the winning accumulator. It seems likely that if we knew where the losing accumulator was relative to the winning accumulator when it hits threshold (i.e., the "balance of evidence") this could help to identify the correlation parameter. Vickers (1979) suggested that confidence is proportional to the balance of evidence, but did not specify a mechanism by which the balance of evidence could be translated into the discrete confidence judgments that are typically collected in experiments. Reynolds et al. (submitted) proposed a general mechanism that can solve this problem, and applied it to modeling the addition of an intermediate "don't-know" judgment to binary choice.

The extension of Reynolds et al.'s (submitted) approach to any number of discrete confidence judgments is straightforward. For a high vs. low confidence judgment, for example, an extra threshold is added to each accumulator. A decision is made in the usual way in favor of the accumulator that first hits its upper threshold. Confidence in that decision is determined by the state of the losing accumulator. If it is below its lower threshold confidence is high (as the balance of evidence is larger); if it is above, confidence is low (as the balance of evidence is smaller).

Suppose a high confidence response occurs when the eventual finishing time of the losing accumulator is at least twice that of the decision time (*dt*). As the LNR is deterministic this is equivalent to saying the loser has accumulated less than half the amount of evidence needed to trigger a response on that trial and so the the lower threshold, expressed as a proportion of the upper threshold, is $d = 0.5$. The likelihood of a high confidence response is the density of the winning choice at *dt* multiplied by the survival function of the losing accumulator finishing after $\frac{dt}{d}$.

$$L_1(dt) = f(dt|\mu_1, \sigma_1^2) \cdot S(\frac{dt}{d}|\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(ln(dt) - \mu_1), (1-\rho^2)\sigma_2^2). \tag{6}$$

The losing accumulator is in the low-confidence state if it has not finished before *dt* (so it is the loser) but has finished after $\frac{dt}{d}$ (so it is above the lower threshold). The corresponding probability is the survivor function of the losing accumulator at *dt* minus the survivor function at dt/d. When multiplied by the density of the winning choice at *dt* this gives the likelihood of a low confidence response.

$$\begin{aligned} L_1(dt) = {} & f(dt|\mu_1, \sigma_1^2) \cdot \Big(S(dt\big|\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(ln(dt) - \mu_1) \\ & - S(\frac{dt}{d}\big|\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(ln(dt) - \mu_1), (1-\rho^2)\sigma_2^2)), (1-\rho^2)\sigma_2^2)\Big). \end{aligned} \tag{7}$$

Adding more thresholds to allow a greater number of confidence levels is simply a case of taking differences of survival functions evaluated to adjacent thresholds. When fitting such models the additional thresholds must be parameterised so that their order is preserved. In the next section we fit model with two lower thresholds to accommodate data with three (low/medium/high) confidence levels.

Fitting Confidence Data

In this section, we use recognition memory data from Ratcliff et al. (1994) to evaluate the performance of the multiple threshold log-normal race (MTLNR) model. Participants studied a list of words, and then performed test trials with words that had either been previously studied (old) or not (new). They pressed one of 6 buttons to simultaneously choose if the test word was old or new and to indicate if they had low, medium or high confidence in their choice. Words could be either high or low frequency, and some words were studied more than others, so overall there were 6 conditions in a 3 (new, weak old, strong old) by 2 (high vs. low word frequency) design.

In order to compare the correlated and independent MTLNR models on an equal footing we first tested versions with equal numbers of parameters (29). We again again fit hierarchical Bayesian models using the DMC software. Details of the fitting methods and priors are provided in supplementary materials, and the data sets and code to perform the fits are available at osf.io/4hn7p/. The models have 12 log-mean parameters (i.e., two for each condition, one for the matching and one for the mismatching accumulator). In the unequal-variance uncorrelated model there are 12 corresponding log-variance parameters. In the correlated equal-variance model there are 6 log-variance parameters that are the same for matching and mismatching accumulators and 6 correlation parameters, one for each condition. Finally there are 4 confidence thresholds, two for each accumulator (splitting them into three regions corresponding the three confidence responses) and a single non-decision time parameter.

In contrast to the choice data, the correlated model provided a qualitatively better fit than the independent model (see Figures 3 and 4). Both models capture the bow shape of the RT distributions over the confidence levels, with slower RTs for low confidence responses. Neither fit is perfect, but the correlated model is generally better, particularly for slower responses (i.e., the 90^{th} percentile). The correlated model has an even clearer advantage in fitting response proportions, with independent model particularly struggling with high-confidence responses. As Table 1 shows, DIC was better by a very large margin (7811) for the correlated than the independent model. The advantage was greater in terms of the best fit (minimum posterior deviance 238449 vs. 246286, a difference of 7837) indicating that the correlated model is slightly more complex.

We also fit two models with more parameters, both with unequal variance, one with a separate correlation parameter for each condition, and the other with just one correlation. The former model was better (by 580) and the latter model worse (by 561) than the equal-variance correlated model. We also fit a simpler equal-variance correlated model with just one correlation parameter, but its fit was substantially worse (by 2629) than the equal-variance model with separate correlations for each condition, although it was still substantially better than the independent unequal-variance model (by 5182). Overall, these results suggest that allowing

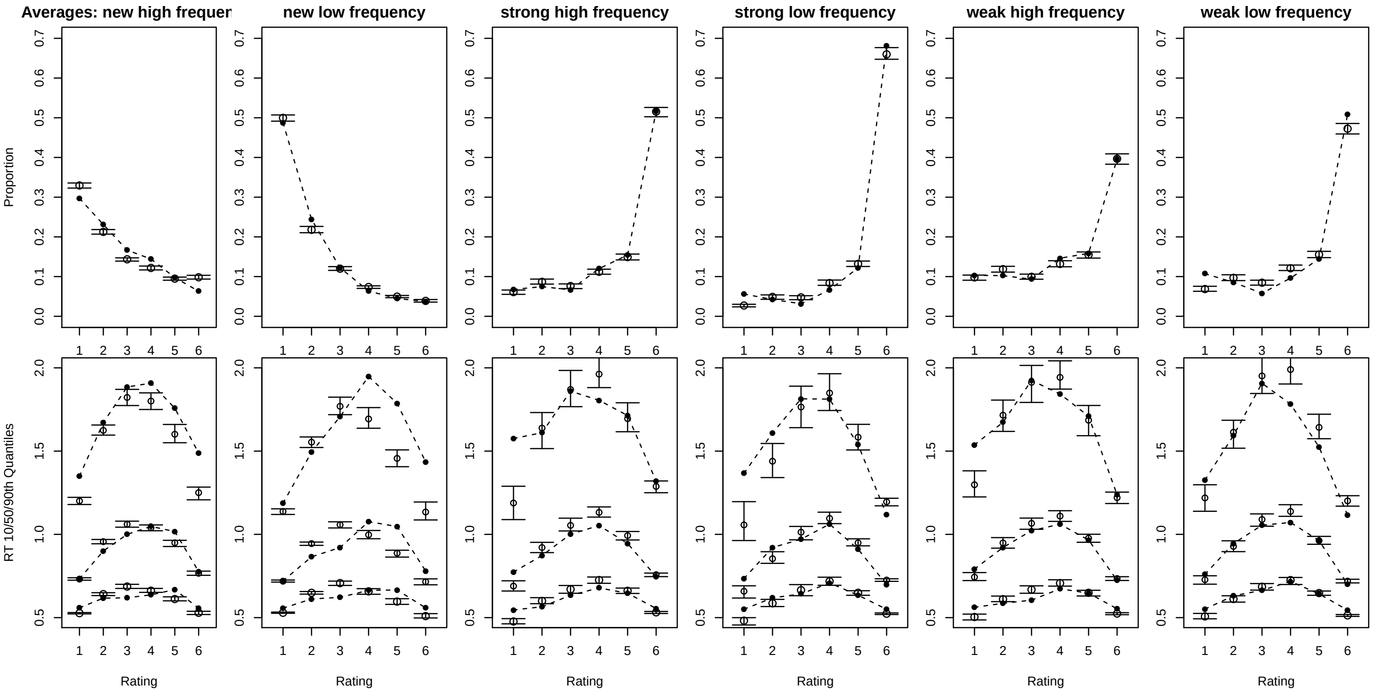

Figure 3. Independent MTLNR fits to the confidence data, averaging over all subjects. Responses 1-6 represent a continuum from high confidence new to high confidence old.

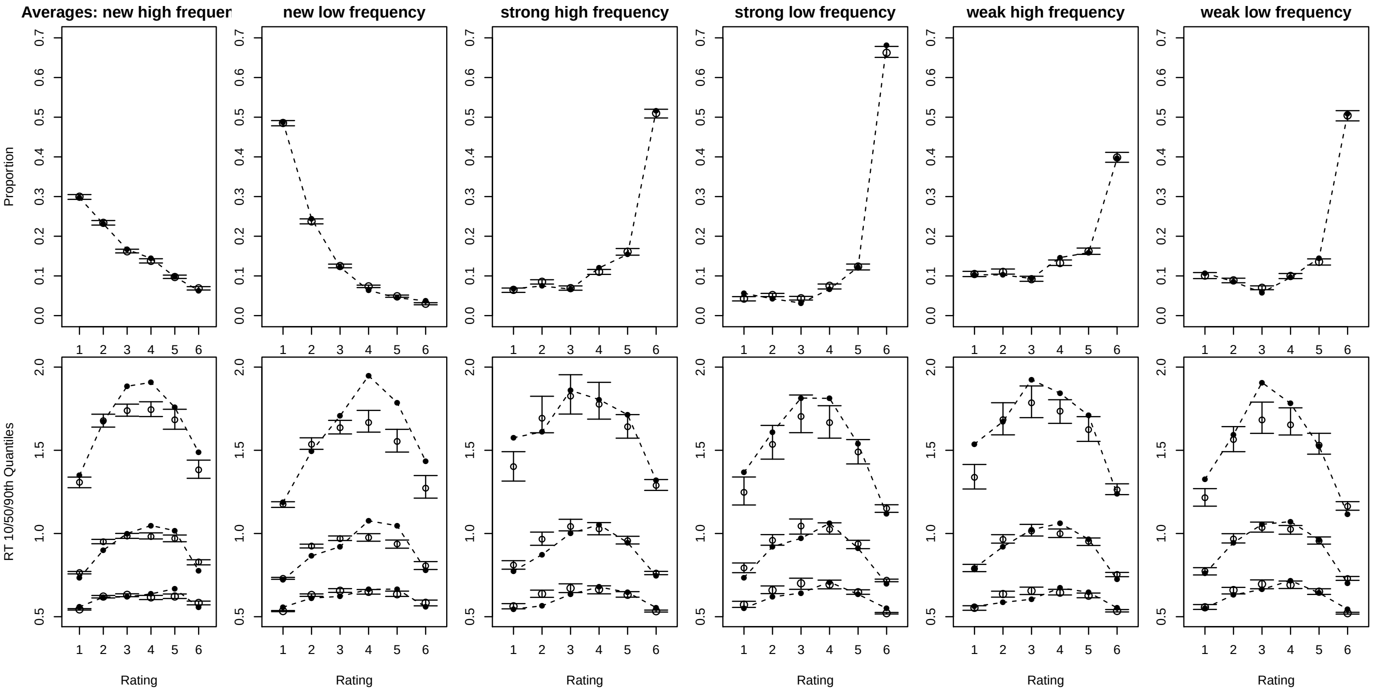

Figure 4. Correlated MTLNR fits to the confidence memory data, averaging over all subjects. Responses 1-6 represent a continuum from high confidence new to high confidence old.

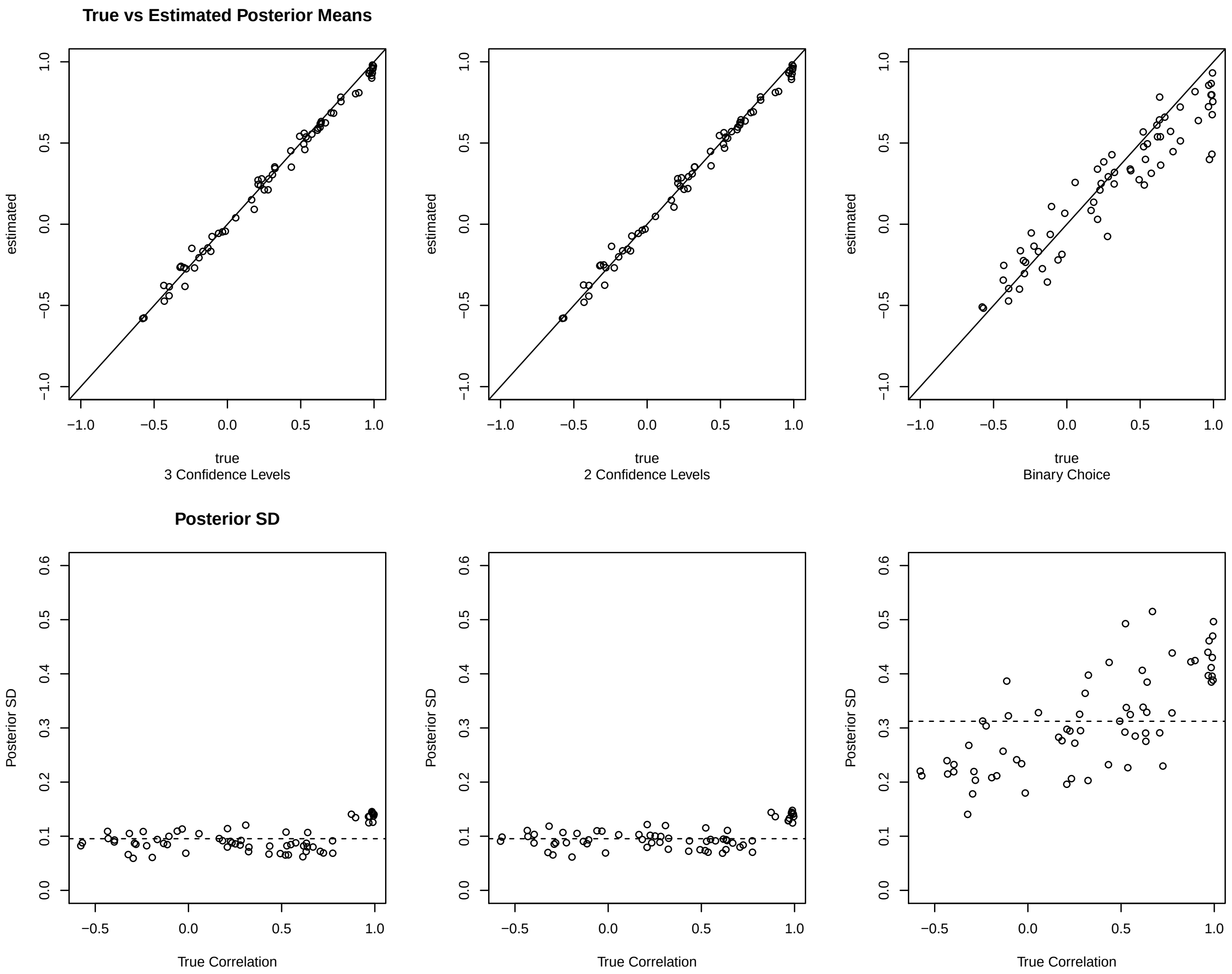

Figure 5. Correlation parameter recovery across decreasing confidence levels. The first row shows, for each participant and condition, the true vs estimated (posterior) means and the second row shows uncertainty of these estimates in the form of the standard deviations of the posterior samples. Each model is fit to simulated data from the equal variance correlated LNR model, with approximately the same number of trials per simulation as there is data per subject (8960 trials and 11 subjects).

Table 1
Numbers of variance and correlation parameters, total number of parameters and DIC values for five models fit to the confidence-rating data. All models had 12 log-mean (μ) parameters, 4 intermediate thresholds (2 per response) and a single non-decision time parameter. Models are ordered according to DIC, from the best at the top to the worst at the bottom.

no. σ	no. ρ	no. pars (total)	DIC
12	6	35	238381
6	6	29	238961
12	1	30	239522
6	1	24	241590
12	0	29	246772

for correlation is more important than unequal-variance, but that the two can trade off. Although the model with both was preferred by DIC, and it was accompanied by little visible improvement in fit over the simpler 29 parameter equal-variance model (see supplementary materials).

In order to focus on the correlations that are the most novel aspect of our investigation, from here we address the equal-variance model with separate correlations for each condition. However, we acknowledge that further investigation with a wider array of data sets will be required to better asses the case for also including unequal variance. Given the results from the fits of the correlated MTLNR to the binary data in the speed-emphasis condition it seems likely that greater variance for the mismatching than matching accumu-

Table 2

Correlation (ρ), standard-deviation (σ) and threshold (d) parameter posterior means for each participant, and group level means, for the 29 parameter correlated MTLNR model. “Low“ and “High“ refer to word frequency.

Participant	1	2	3	4	5	6	7	8	9	10	11	Group
ρ_{NewLow}	0.32	0.63	0.97	0.24	0.63	0.99	-0.01	0.43	0.52	0.72	-0.32	0.47
$\rho_{WeakLow}$	-0.17	0.21	0.90	-0.24	0.28	0.98	-0.58	-0.23	0.28	0.55	-0.57	0.13
$\rho_{StrongLow}$	-0.06	0.31	0.87	-0.03	0.18	0.98	-0.43	-0.40	0.25	0.62	-0.32	0.18
$\rho_{NewHigh}$	0.49	0.77	0.99	0.21	0.77	1.00	-0.19	0.54	0.61	0.71	-0.30	0.51
$\rho_{WeakHigh}$	0.16	0.52	0.97	-0.10	0.67	0.99	-0.29	-0.11	0.43	0.53	-0.40	0.31
$\rho_{StrongHigh}$	0.23	0.64	0.97	0.05	0.58	0.99	-0.28	-0.13	0.32	0.64	-0.43	0.32
σ_{NewLow}	0.75	0.65	0.57	0.85	0.55	0.53	0.81	0.80	0.61	0.75	0.89	0.71
$\sigma_{WeakLow}$	0.93	0.71	0.58	1.02	0.58	0.49	0.95	0.93	0.65	0.71	1.14	0.79
$\sigma_{StrongLow}$	0.86	0.70	0.56	0.95	0.62	0.52	0.84	0.91	0.60	0.67	0.93	0.74
$\sigma_{NewHigh}$	0.75	0.75	0.70	0.92	0.56	0.57	0.84	0.85	0.64	0.78	0.92	0.75
$\sigma_{WeakHigh}$	0.84	0.79	0.66	0.99	0.58	0.56	0.83	0.94	0.70	0.79	1.02	0.79
$\sigma_{StrongHigh}$	0.83	0.75	0.67	0.99	0.59	0.56	0.81	1.05	0.64	0.74	1.01	0.79
$d2_{New}$	0.82	0.95	0.92	0.85	0.74	0.93	0.81	0.72	0.78	0.76	0.86	0.83
$d1_{New}$	0.57	0.76	0.83	0.61	0.51	0.91	0.46	0.44	0.59	0.42	0.45	0.60
$d2_{Old}$	0.91	0.96	0.94	0.83	0.74	0.92	0.88	0.73	0.72	0.77	0.91	0.85
$d1_{Old}$	0.47	0.69	0.87	0.53	0.50	0.89	0.43	0.39	0.46	0.38	0.43	0.55

lator will be particularly required to fit data with fast errors. In the present data errors were generally slower than correct responses and so the equal-variance model performed quite well.

Parameter Estimation

We performed a simulation study to investigate the estimation properties of the MTLNR models. We used the same number of trials as in Ratcliff et al.'s (1994) design, as it was quite large, 8960 trials per participant (2240 each for high and low frequency new condition, and 1120 for each of the four old conditions). Otherwise we followed the same procedure as in the previous cross-fit study, simulating new data from the mean posterior parameter estimates of the fits to the confidence data. The pattern of results was the opposite to the lexical experiment; the equal-variance correlated model was able to match independent unequal-variance simulations, but the independent model was not able to match the correlated simulations. These results are shown in supplementary materials and reflect the results in Figure 3 in terms of the causes of the misfit.

We also used the simulated data from the correlated model fit to investigate how the number of confidence ratings influenced the quality of estimation. As shown in Figure 5, we compared the fits with 3 levels (as in the data) to fits with 2 levels (by collapsing the less numerous low and medium confidence responses) and 1 level (i.e., binary choice). The figure shows excellent recovery of the correlation parameter with all three confidence levels and little if any degradation when collapsing to two levels. In contrast, recovery is poor for binary choice. Figure 5 also shows similar pattern in terms of the standard deviation of the posterior estimates of the correlation parameters. Hence, not only do the correlation parameters improve the model fit to the confidence data, but the fitting confidence ratings helps to identify the correlation parameters.

Given that these findings indicates parameters are well recovered in this design, we now discuss the values estimated from the data, focusing on the correlation, variability and threshold estimates shown in Table 2 (see supplementary materials for the remaining parameters).

Correlations in every condition were on average positive, with an overall mean of 0.48 and a range of 0.29 - 0.73. Individual-participant values were quite variable, with participants 3 and 6 having estimates mostly close to 1, whereas participant 7 and 11 had all negative estimates and participants 1, 4 and 8 had two to four negative estimates. The heterogeneous nature of these vales suggests that mechanisms causing both negative and positive correlations are present, and that the balance of the effects of these mechanisms can differ markedly across, and even within, participants. However, there are also patterns that are quite consistent within participants, with larger correlations for for new than old items and for high than low frequency items. In particular, for the average over participants, the 95% credible interval for new items (.475 - .499) was much greater than for either weak (.199 - .237) or strong (.233 - .271) old items, it was for high frequency (.398 - .426) compared to low frequency (.296 - .324).

In contrast to the ρ estimates, the σ estimates varied much less over individuals, with σ estimates being greater for old (weak: 0.783-0.801, strong: 0.775 - 0.772) than new (0.724 - 0.735) items. The greater old than new σ estimates are also consistent with findings from receiver-operating characteristic (ROC) (Heathcote, 2003; Wixted, 2007; A. F. Osth et al., 2017) and diffusion model analyses of recognition-memory data Starns et al. (2012). It is possible that this difference is related to the correlation difference between new and old, in that that greater old variability would produce a greater influence of the stimulus-criterion mechanism that reduces correlations.

Threshold estimates were fairly high, mostly being in the upper half of the range, so relatively small differences in the balance-of-evidence differentiate low and medium confidence responses. This reflects the generally positive correlations, which mandate that both the winning and losing accumulators will tend to have large evidence totals when a decision is made. Reflecting this fact, thresholds were particularly high for the participants with large positive correlations and generally were placed lower as correlations decreased. Threshold placement was fairly consistent across accumulators. For most participants the distance between the lower (d_1) and upper (d_2) thresholds, which delineates medium-confidence responses, was smaller for the new than old accumulator.

Correlation, RT and the Balance of Evidence.

Given the prominent role played by correlations in explaining in Ratcliff et al.'s (1994) data, in this section we further explore the effects of correlation on the predictions of the MTLNR model. We first show that the balance of evidence in the MTLNR model produces a representation of response probabilities that is directly analogous to that of Gaussian signal-detection theory. This type of signal-detection model has been used extensively to model response probabilities in recognition memory research using confidence-based ROCs (e.g., Heathcote, 2003; Wixted, 2007). We then explore the relationship between the balance of evidence and RT.

Because of the presence of the logarithmic transformation, in the MTLNR confidence is related to the ratio of decision times (DT). Since DT for both accumulators is log-normal, their ratio, and hence the distribution of the balance of evidence (BoE), is log-normal, and so log(BoE) has a normal distribution. As the logarithmic transformation is monotonic, the values of integrals of BoE distributions between boundaries (which correspond to response probabilities) are preserved on the log(BoE) scale. Hence, log(BoE) = log(NewDT)-log(OldDT) ~ $N(\mu_New-\mu_Old,\sigma^2_{New}+\sigma^2_{Old}-2\rho\sigma_{New}\sigma_{Old})$, is analogous to the normal distributions of memory strength assumed by signal-detection theory, with the criterion dividing new and old responses placed at zero. In the MTRLNR version, criteria demarcating different confidence responses are arrayed around zero (see Pratte et al., 2010, for an analogous representation), with values equal to

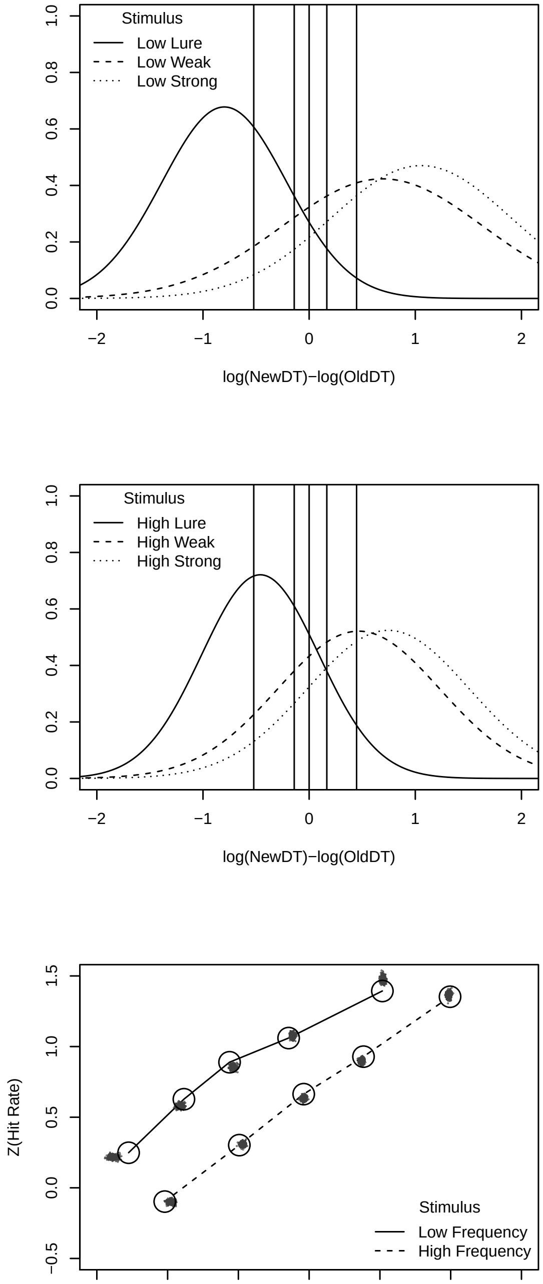

Figure 6. Signal-detection theory plot and zROC curves derived from group parameters of fits of the correlated MTLNR model for low-frequency (top panel) and high-frequency (middle panel) items in the confidence data. The lower panel shows zROCs averaged over participants (lines and open points) with model fits based on 500 samples from the posterior per participant shown as clouds of small grey points.

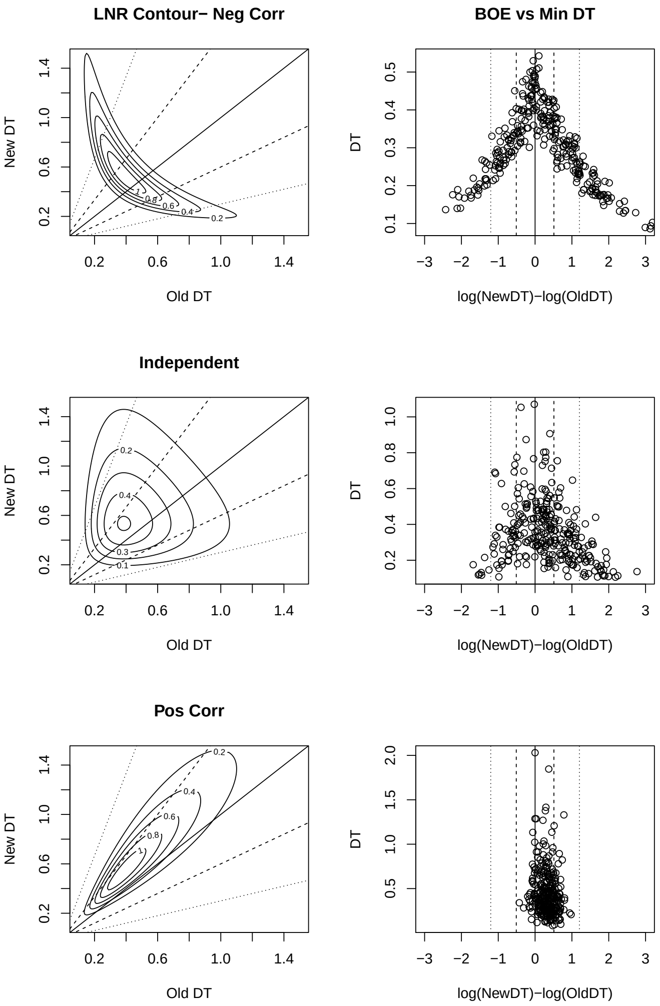

Figure 7. Top, middle and bottom rows show representation of MTLNR models with negative (-0.9), zero and positive (0.9) correlations respectively with and old stimulus ($\mu_{old} = -0.95, \mu_{new} = -0.63, \sigma = .55$, d1=.3 and d2=.6 (both old and new response)). Left Column: bi-variate densities of decision time (DT) for the new vs. old accumulators as contour plots (density values are marked on the curved contours), with sloping lines from the origin denoting decision criteria (solid lines dividing new and old, dashed lines dividing high and medium confidence, and dotted lines medium and low confidence). Right Column: Samples of the minimum of the new and old accumulator DT (i.e., the winning DT, corresponding to $RT - t_0$) vs. BoE values with decision criteria (vertical lines).

the logarithm of the relative thresholds (d) for the old accumulator and the negative logarithm of the relative thresholds for the new accumulator.

Figure 6 shows zROC plots (i.e., plots of the z-transformed of hits against the z-transformed of false alarms). Lines joining open circle represent the data averaged over participants, with similarly averaged MTLNR model predictions for 500 randomly drawn posterior samples drawn as clouds of grey points. Mean and standard deviation parameters for the normal distributions averaged over participants are given in Table 3. In agreement with patterns typically found in ROC analysis (Glanzer et al., 2009), standard deviations for low-frequency items were greater than high-frequency items both when new and old, and the means displayed a word-frequency mirror effect, with the smallest values for low-frequency new items and the largest values for low-frequency old items, with high frequency means intermediate. High vs. medium confidence criteria were more extreme for new (median -0.647; 95% credible interval -0.66 - -0.633) than old (0.55; 0.538 - 0.562) but the opposite pattern led for the medium vs. low confidence criteria for new (-0.172; -0.178 - -0.167) compared to old (0.55; 0.538 - 0.562).

Models based on Gaussian distributions produce perfectly linear zROC plots for fits to a single participant. The slight deviations from linearity in fits in Figure 6, particularly for the low-frequency condition, occur because of the averaging over participants. The fit is good for high-frequency but there are slight misses for high confidence error conditions for the low-frequency condition. Plots of individual-participant fits in supplementary materials show this is mainly due to one individual (Participant 10) who has a much stronger version of the concavity evident in Figure 6, and which the MTLNR model (which produces strictly linear zROCs at the individual level) is unable to capture. It has been proposed that this pattern requires requires either an additional mechanism or a more complex model. Ratcliff et al. (1994) attributed it to a guessing process whose effects are most evident for rare responses like high-confidence errors. Ratcliff & Starns (2013) ascribed it to unusual settings of the threshold parameters in the RTCON2 model. In the next section we explore whether variability in MTLNR thresholds can also address this pattern. However, we first explore the effect of correlation on the signal-detection representation that provides a good account for the majority of participants.

Figure 7 represents the bivariate decision-time distributions underlying the MTLNR signal detection representation. It shows the case of an old stimulus that differs across the three rows of the figure only in the correlation parameter. In all cases the bi-variate distributions in the left-hand column are shifted above the main diagonal (solid line) so the old accumulator is most often faster and so wins most often. Relative to the middle row, where the correlation is zero, in the top row, where there is a strong negative correlation, the old ac-

Table 3

Signal Detection means, SDs and criteria pertaining to Figure 6, generated by the equal variance, correlated MTLNR fit to the confidence data. Criterion 3 is fixed to 0

		Low Frequency			High Frequency		
		2.5%	50%	97.5%	2.5%	50%	97.5%
New	mean	-0.817	-0.799	-0.781	-0.476	-0.460	-0.447
	sd	0.693	0.707	0.721	0.685	0.700	0.716
Strong	mean	0.805	0.839	00.869	0.684	0.716	0.747
	sd	0.919	0.942	0.968	0.866	0.890	0.915
Weak	mean	0.778	0.808	0.838	0.753	0.784	0.813
	sd	1.028	1.055	1.084	0.883	0.908	0.933

cumulator wins less often, and higher-confidence responses are more frequent. In the bottom row, where there is a strong positive correlation, the old accumulator wins less often and higher-confidence responses are less frequent. These effects occur because the increase from negative to positive correlation reduces the difference in decision time between accumulators, and hence the spread of the BoE, as illustrated on the x-axis of the right hand column of Figure 7.

The right hand column of Figure 7 illustrates that an increase in correlation also reduces the overlap of the RT distributions for adjacent confidence responses. High negative correlations are implausible because there is typically substantial overlap in the RT distributions of different confidence categories (e.g., the spread of the percentiles in Figure 4 is much larger than the differences between response). At the other extreme, high positive correlations mean that RT carries little information about the strength of the evidence favoring one or other choice (i.e., there is little difference between the distributions of the minimum decision time in the lower right hand panel of Figure 7).

Threshold Variability

It has been suggested that signal detection theory decision criteria are subject to trial-to-trial variability (Benjamin et al., 2009; Mueller & Weidemann, 2008). However, the magnitude of this variability can be difficult to estimate, and more recent work has suggested it has only a small effect (Kellen et al., 2012). In this section, we investigate the estimation of variability in MTLNR thresholds, and whether allowing for such variability enables the model to address Participant 10's non-linear zROCs.

There are at least two ways this type of variability can be introduced in the MTLR: letting let each threshold vary independently, or varying all thresholds together in a lockstep. The second method requires fewer parameters and allows for a wider range of variation in the regions that each threshold can vary over while maintaining the appropriate order. Order changes greatly complicate the derivation of likelihood functions, as well as raising interpretation problems, and so we only consider models here in which the order is preserved. We also assume that the choice threshold does not vary, only the confidence threshold(s).

On the assumption that confidence-threshold variability is uniformly distributed, results reported in Terry et al. (2015) enable the derivation of analytic likelihoods. The RT density part of the likelihood (corresponding to the choice) remains a simple log-normal as in the constant-threshold model, but the survival function (corresponding to the state of the losing accumulator) is now composed of CDFs for a shifted uniform random variable divided by a log-normal random variable. The CDF of the latter ratio provided by Terry et al. is $F_{UonL}(dt, A, b, v, sv)$, where A is the range of uniform start-point noise, b is a threshold, v is the mean rate of accumulation and sv its standard deviation. A equates to MTLNR threshold noise, and so we will use the same notation here; b equates to our confidence threshold (d) and sv to our σ parameter. The accumulation rate parameter v corresponds to $-\mu$ (as μ is negative of accumulation rate).

Hence, the likelihood for a high confidence for the first accumulator (i.e., the second accumulator's evidence total is below its lowest threshold, $d1$) in an MTLNR model with threshold variability is:

$$f(dt|\mu_1, \sigma_1) \times (1 - F_{UonL}(dt|A_2, d1_2, -\mu_2, \sigma_2)) \qquad (8)$$

The likelihood equations for lower confidence responses are analogous to Equation 7 using the survivor functions for the uniform on log-normal random variable (i.e., $1 - F_{UonL}$). It is also straightforward to extend the likelihood equations to allow variability in the lower threshold to extend below zero. For trials on which this occurs a high confidence response cannot be produced. The extension simply involves adjusting the likelihoods for each confidence level based on the probability that the lower threshold falls below zero, and we employ it in the fits reported below (see supplementary materials for details). Finally, including correlations is exactly the same as before, only now rather than passing the conditional log-mean parameter into a log-normal CDF function, we pass the negative of the conditional to the uniform-on-log-normal CDF.

Figure 8 shows zROCs for Participant 10 with fits of the MTLNR model both without (top panels) and with (bottom

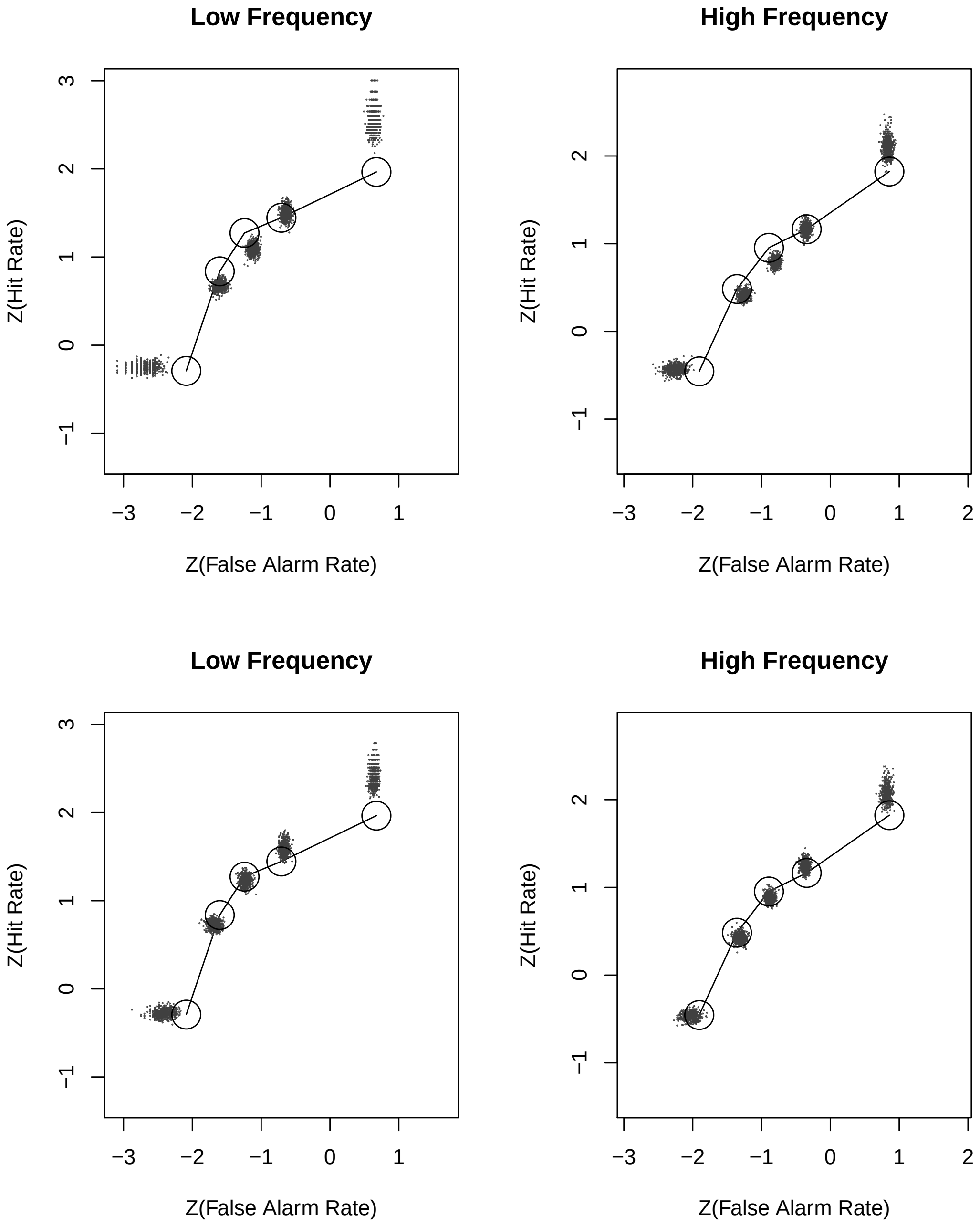

Figure 8. zROC plots for participant 10. The upper two panels show fits of the MTLNR model without threshold variability and the lower panels include threshold variability.

panels) threshold variability. Threshold variability enables the model to better capture the concave pattern, with good fits to the middle points on the curves, although there is still some misfit to the probability for high confidence errors, particularly for low frequency items where accuracy is higher and hence error are rarer. Details of how the variable threshold MTLNR model was fit, parameter estimates, and zROCs for the remaining participants, can be found in the supplementary materials. Separate values of *A* were estimated for each accumulator and estimates were generally larger for the accumulator corresponding to old responses, indicating a higher level of noise in the calibration confidence for new responses. Six of the eleven participants had negligible levels of threshold variability, but the remaining 5 had substantial threshold variability in at least one accumulator and DIC summed over participants improved substantially, by 1250, when threshold variability was added. Larger values of threshold variability were associated with substantial reductions in correlations estimates for participants 8 and 10. However, for participant 6, who also had substantial threshold variability, the reduction was much less.

Discussion

There are many plausible reasons to believe that the inputs to, and parameters of, racing accumulators might be correlated with each other. In the introduction to this paper, we described some mechanisms that cause both positive and negative correlations, but it is likely that there are many others. For example, all of the models we have addressed here assume independence during the accumulation process (i.e., the evidence total in one accumulator does not affect the evidence total in other accumulators), but other models, such as the leaky competing accumulator (LCA Usher & McClelland, 2001), assume an interaction. If that interaction is competitive it would appear as a negative correlation to the models we address here, whereas if it were excitatory it would appear as a positive correlation (see Teodorescu & Usher, 2013, or a wide-ranging discussion of different types of dependence).

However, investigating inter-accumulator correlation is challenging because the assumption of independence brings with it benefits in terms of conceptual and mathematical simplicity that are commonly lost when correlation is present. It also seems likely that correlations, which are fundamentally about the relationship between accumulators, will be challenging to estimate in binary choice data because for each choice only the state of one accumulator, the winner, is observed. In this paper we attempted to address these challenges using an evidence-accumulation model, the lognormal race (LNR Heathcote & Love, 2012), that remains tractable in the face of correlation, and using an equally tractable extension of that model combined with the multiple threshold architecture proposed by Reynolds et al. (submitted), the MTLNR, which uses confidence judgments to provide information about the state of the losing accumulator.

Our initial exploration of the correlated LNR model applied to binary-choice data (Wagenmakers et al., 2008) confirmed it is poorly identified, and so if little practical use for investigating the characteristics (e.g., sign and magnitude) of inter-accumulator correlation. We also found that an independent model where the matching accumulator (which corresponds to the stimulus) and mismatching accumulator (which does not) differ in their variance can very closely mimic a correlated model with equal variance. Indeed, as is explored in more detail in supplementary material, we found a consistent pattern whereby positive correlation is mimicked by greater mismatching than matching variance, whereas negative correlation is mimicked by greater matching than mismatching variance. The former pattern is almost always found in fits of the independent LNR model and a related independent deterministic accumulator model, the LBA (Brown & Heathcote, 2008), suggesting the possibility that it might arise because it is mimicking the presence of an underlying positive correlation.

However, Heathcote & Love (2012) pointed out that greater mismatch than match variance is the only mechanism that can enable the independent LNR to accommodate error responses that are as fast or faster then correct responses, as is commonly seen when the speed of responding is emphasized over accuracy. We fit Wagenmakers et al.'s data because their design manipulated the speed vs. accuracy of responding and confirmed that the same mechanism was still required by the correlated LNR to model fast errors, although the estimated difference in variance was attenuated. It is an interesting topic for future research whether a correlated LBA, which has an additional mechanism for accommodating fast errors (start-point noise), will be able to accommodate speed vs. accuracy manipulations without requiring unequal variance.

We then confirmed that the MTLNR not only enables good estimation of correlation but also that a correlated version of this model provided a clearly better fit than an independent version to Ratcliff et al.'s (1994) data where three (high/medium/low) confidence ratings were made simultaneously with a binary (new vs. old) recognition memory choice. A good overall fit was achieved by an equal-variance correlated MTLNR model, and although model selection also favored the additional flexibility afforded by unequal variance, there was little noticeable improvement in fit, likely because errors were relatively slow in this data. As we show in supplementary materials, parameter estimation was still excellent for the unequal variance correlated model, so it can readily be used in applications where fast errors are present.

Interestingly, a simulation study revealed that the dramatic reduction in the uncertainty of parameter estimates relative to the binary-choice case was equally good for two- and three-level confidence ratings. A two-level (e.g., higher vs. lower) rating is easier to elicit both in a manual sense, in that it is easier to equate response production times across ratings and so avoid the need to estimate extra parameters, and in terms of compliance with instructions to utilize all rating levels. Hence, these results support the wider adoption of a simultaneous binary choice and two-level confidence rating procedure, not only in investigations of inter-accumulator correlation but also potentially in investigations that are more focused on confidence itself.

Good fits to Ratcliff et al.'s (1994) data previously required a complex and analytically intractable model (RTCON2 Ratcliff & Starns, 2013). Although the fits of our much simpler and analytically tractable correlated MTLNR were quite good for most participants, they clearly failed for

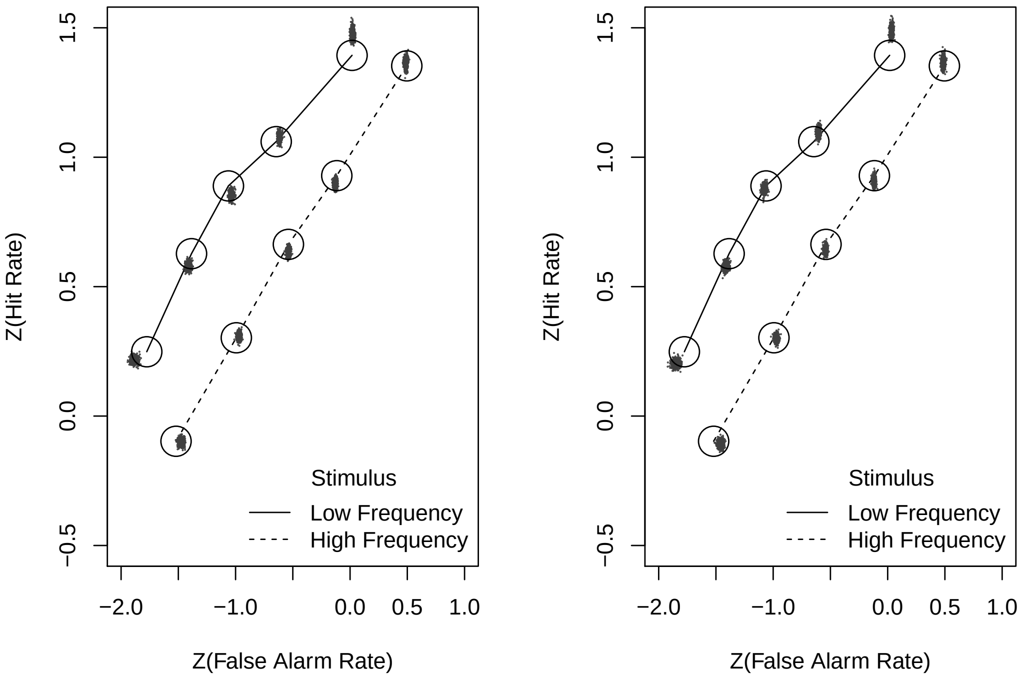

Figure 9. zROC plots for data averaged over participants (lines and open circles) with similarly averaged model fits based on 500 samples from the posterior per participant shown as clouds of small grey points. The left panel show fits for the without threshold variability and the right panel with threshold variability.

one participant. However, this misfit was largely ameliorated by allowing confidence thresholds to vary from trial to trial. Fortunately, this more flexible model remained analytically tractable. Overall, our results suggest that a correlated MTLNR model allowing for both unequal variance between accumulators and threshold variability has the potential to provide a comprehensive and practical new approach to investigations of confidence.

Our results supporting the necessity of correlation to model Ratcliff et al.'s (1994) data are consistent with those of Ratcliff & Starns (2013). In order to fit this data, they had to augment their RTCON model (Ratcliff & Starns, 2009), which had been successful in fitting a range of other confidence data sets, to allow for competitive interactions during accumulation. However, on average we found evidence for positive correlations, which are more consistent with an excitatory interaction rather than the competitive mechanism of RTCON2. That said, there were considerable individual differences, with a few participants estimated to have negative correlations in all conditions, a few to have mix of positive and negative and the remainder all positive, including a few with very strongly positive correlations. Such diversity is consistent with the LCA, which has been found to display behavior ranging from competitive to mutually excitatory depending on individual differences in parameter estimates (Usher & McClelland, 2001). As discussed earlier, these individual differences are also consistent with differences in the balance of the effects of mechanisms that produce either negative or positive correlations when there is no interaction during accumulation.

In a recent investigation of a different type of correlation in the inputs to an evidence-accumulation model—correlations between stimulus dimensions as characterized the multivariate generalization of signal-detection theory (GRT Ashby & Townsend, 1986)—Smith (2019) noted the practical advantages of models with analytic equations in terms of being able to efficiently fit and evaluate sets candidate models, as we have done here. He also noted the theoretical advantages in the insights into the way a model works that are often afforded by analytic prediction equations. We found this was the case with the MTLNR model without threshold variability, as we were able to show that its response-probability predictions, while also being constrained by RT measurements, match those of the simple unidimensional Gaussian signal-detection theory that has been widely applied to response-probability data.

This relationship not only provides insights into the way the model works, but also shows that its predictions are quite constrained. Such constraint makes the model highly testable, and provides a reference for understanding the effects of additional mechanisms. This reference allowed us to understand how threshold variability is important in modeling non-linear zROCs. We note, however, that further mechanisms may be required to accommodate unusual zROC patterns displayed by some participants, such as random responding as originally suggested by Ratcliff et al. (1994). Further mechanisms may also be required to accommodate unusual patterns in the relationship between confidence and RT. For example, RT decreases as confidence increases for most participants, but a few participants in Ratcliff & Starns (2009) had fast low confidence responses. Complex and flexible models such as RTCON are able to accommodate such exceptions, but at the cost of not making any clear predictions about the pattern displayed by the majority of participants.

Related to the latter point, a further insight that was afforded by analysis of the MTLNR model is in regard to the effect of correlation on the relationship between RT and confidence. A negative correlation is associated with a strong inverse relationship between RT and confidence (i.e., RT decreases as confidence increases). Indeed, when the correlation is near negative one there is little overlap in RT distributions for adjacent confidence ratings. In practice, overlap is usually quite marked, so this is likely the reason for an absence of very strong negative correlation estimates in our fits. When independence holds, there is still a negative relationship, but as the correlation becomes strongly positive the relationship disappears entirely. Overall, this pattern is consistent with the variable but mostly inverse relationship observed between confidence and RT. Correlation cannot, however, reverse the relationship between confidence and RT, suggesting that it cannot explain fast low-confidence responses, and so an extra mechanism will be required for such cases. However, this also means that the MTLNR model is constrained to generally predict an inverse relationship between confidence and RT.

These considerations also shed light on ROC analyses based on RT rather than confidence, which suggest that RT carries information about memory strength in recognition paradigms (Weidemann & Kahana, 2016). Perhaps surprisingly then, two of the 11 participants we fit had very strong positive correlations, suggesting that their RT carried little or no information about memory strength. However, for the remainder this was not the case, although consistent with Weidemann & Kahana's results, the average positive correlations indicate that RT usually carries less information than confidence ratings. In any case, strong individual differences in correlation suggest that a model-based approach, such as the one used here, may be necessary if RT information is to be useful in augmenting confidence ratings when assessing the quality of memory on an individual basis. Including relative threshold variability into the model also weakens the relationship between decision time and confidence, causing estimates of correlation to decrease for some subjects.

Although the simplicity of the LNR affords advantages, it comes with some limitations. Foremost is its inability to differentiate accumulation rate effects from the effects of the distance between the start and end points of accumulation. Although present in the model conceptually, the parameters governing these different processes interact linearly and so are not separately identified without imposing further constraints. Hence, any attempt to investigate differences in correlations between these processes will require another approach. One possibility is to specify and estimate the parameters of structural relationships that imply correlations, such as those corresponding to criterion referenced inputs that specify rates, or global fluctuations in rates or thresholds. Identification of the parameters of such mechanisms will depend on appropriate design constraints in terms of rate mechanisms that will most likely possible in perceptual-choice paradigms (i.e., where it is possible to specify the mapping between objective stimulus values and subjective magnitudes, see van Ravenzwaaij et al., 2019).

A second approach to this issue is to implement a multiple-threshold balance-of-evidence mechanism within the a model such as the LBA that intrinsically provides separate estimates of rates and other parameters. Reynolds et al. (submitted) took this approach, which could be extended from independent to dependent racing accumulators like those used here. Fortunately, the correlated case of the LBA-MTR requires only one-dimensional numerical integration.

Although slow, such integration is usually quite stable, and so that approach may provide a fruitful avenue for future research. Another issue likely requiring numerical integration is variability in non-decision time. If such variability affects both accumulators equally, as seems most likely, it will make correlation estimates more positive if its effects are not explicitly modeled, but the degree to which this occurs depends on the magnitude of the non-decision variability relative to variability in the decision process. When Heathcote & Hayes (2012) estimated the range of uniform non-decision time variability for the LBA in a lexical-decision data set it was a relatively modest 0.1s, suggesting that any positive inflation of correlation would be modest. We report results in supplementary materials that support only modest inflation of MTLNR correlation estimates for non-decision magnitudes of a similar size. However, inflation can be larger for more substantial non-decision time variability. Clearly, more work is required to investigate this, and other potential influence on correlation estimates. Although the work presented represents only a first step in this enterprise, we hope that the tractable models that we have developed will provide both benchmarks and useful tools for future investigations.

In closing, we note that, even without augmentations such as correlations and threshold variability, it seems likely that the MTLNR model, and indeed the general multiple-threshold approach to confidence ratings, has more flexibility than the balance-of-evidence hypothesis on which it is based. This occurs through the threshold parameters by which it produces discrete responses. Pleskac & Busemeyer (2010) argued that a balance-of-evidence model instantiated with a counter or accumulator model such as the LBA may not be able to account for the the effects of speed emphasis, because a reduction in decision thresholds also reduces the maximum possible difference in evidence, predicting a reduction in both overall confidence and the variability of confidence ratings. Multiple threshold models are not so constrained; if confidence threshold settings can also be modified by speed emphasis, which seems quite plausible. In future research, we plan to investigate this issue by applying the model to data with both a speed vs. accuracy manipulation like the first data set we examined here, but also with confidence ratings like the second data set so that model parameters are estimable.

References

Ashby, G., F, & Townsend, T., J. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Basu, A., & Ghosh, J. (1978). Identifiability of multinormal and other distributions under competing risk models. *Journal of Multivariate Analysis*, *8*, 413-429.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115.

Boag, R. J., Strickland, L., Loft, S., & Heathcote, A. (2019). Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition*, *191*.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Brown, S. D., & Heathcote, A. (2005). Practice Increases the Efficiency of Evidence Accumulation in Perceptual Choice. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(2), 289–298.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*(2), 396–425.

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Donkin, C., & Brown, S. D. (2018, June). Response Times and Decision-Making. In E.-J. Wagenmakers (Ed.), *The stevens handbook of experimental psychology and cognitive neuroscience.*

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging*, *28*(1), 64–76.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*(3), 431–455.

Heathcote, A. (2003). Item Recognition Memory and the Receiver Operating Characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1210–1230.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods*, *36*, 678–694.

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology*, *66*(2), 125–136.

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior research methods*, *51*(2), 961–985.

Heathcote, A., & Love, J. (2012, August). Linear deterministic accumulator models of simple choice. *Frontiers in Pschology*, *3*(292).

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457–479.

Kvam, P. D. (2019). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology*, *91*, 14–37.

Leite, F. P., & Ratcliff, R. (2010, January). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, *72*(1), 246–273.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651–687.

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

Mueller, S. T., & Weidemann, C. T. (2008, June). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–494.

Nadarajah, S., & Kotz, S. (2008). Exact distribution of the max/min of two gaussian random variables. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, *16*(2), 210–212.

Osth, A., & Farrell, S. (in press). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*, *66*(2), 125–136.

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

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901.

Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review*, *78*, 391–408.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 224–232.

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719.

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

Reynolds, A., Garton, R., Kvam, P., Griffin, V., Sauer, J., Osth, A., & Heathcote, A. (submitted). A dynamic model of deciding not to choose. *Manuscript submitted for publication*.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *B64*(4), 583–639.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012, February). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1-2), 1–34.

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review*, *120*(1), 1–38.

Ter Braak, C. J. (2006). A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

Terry, A., Marley, A., Barnwal, A., Wagenmakers, E.-J., Heathcote, A., & Brown, S. D. (2015). Generalising the drift rate distribution for linear ballistic accumulators. *Journal of Mathematical Psychology*, *68*, 49–58.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

van Ravenzwaaij, D., Brown, D., Scott, Marley, A. J., A, & Heathcote, A. (2019). Accumulating advantages: A new approach to multialternative forced choice tasks. *Psychological Review*.

van Ravenzwaaij, D., Donkin, C., & Joachim. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*, 547–556.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208–256.

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

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159.

Weidemann, C. T., & Kahana, M. J. (2016, April). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*(4), 150670.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385–398.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.