Journal of Mathematical Psychology 98 (2020) 102431

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A note on decomposition of sources of variability in perceptual decision-making

Inhan Kang [a], Roger Ratcliff [a,*], Chelsea Voskuilen [b]

[a] The Ohio State University, United States of America
[b] Ford Motor Company, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 22 October 2019
Received in revised form 1 July 2020
Accepted 16 July 2020
Available online 10 August 2020

Keywords:
Double-pass procedure
Linear ballistic accumulator model
Response time and accuracy
Sources of variability

ABSTRACT

Information processing underlying human perceptual decision-making is inherently noisy and identifying sources of this noise is important to understand processing. Ratcliff, Voskuilen, and McKoon (2018) examined results from five experiments using a double-pass procedure in which stimuli were repeated typically a hundred trials later. Greater than chance agreement between repeated tests provided evidence for trial-to-trial variability from external sources of noise. They applied the diffusion model to estimate the quality of evidence driving the decision process (drift rate) and the variability (standard deviation) in drift rate across trials. This variability can be decomposed into random (internal) and systematic (external) components by comparing the double-pass accuracy and agreement with the model predictions. In this note, we provide an additional analysis of the double-pass experiments using the linear ballistic accumulator (LBA) model. The LBA model does not have within-trial variability and thus it captures all variabilities in processing with its across-trial variability parameters. The LBA analysis of the double-pass data provides model-based evidence of external variability in a decision process, which is consistent with Ratcliff et al.'s result. This demonstrates that across-trial variability is required to model perceptual decision-making. The LBA model provides measures of systematic and random variability as the diffusion model did. But due to the lack of within-trial variability, the LBA model estimated the random component as a larger proportion of across-trial total variability than did the diffusion model.

1. Introduction

In mathematical modeling of human choice and reaction time (RT) in perceptual and cognitive tasks, most models of the decision-making process involve the accumulation of noisy evidence up to a decision criterion. Some models assume that this within-trial variability is a sufficient source to explain behavioral patterns and that other sources of variability are not needed (Churchland, Kiani, & Shadlen, 2008; Deneve, 2012; Ditterich, 2006; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Hanks, Mazurek, Kiani, Hopp, & Shadlen, 2011; Kiani, Corthell, & Shadlen, 2014; Palmer, Huk, & Shadlen, 2005; Usher & McClelland, 2001; Zhang, Lee, Vandekerckhove, Maris, & Wagenmakers, 2014). In contrast, the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2008; Wiecki, Sofer, & Frank, 2013) and the linear ballistic accumulator model (LBA model; Brown & Heathcote, 2008) assume that it is impossible to produce exactly the same process settings from trial to trial and this is represented by assuming that some of the parameters in the decision process vary from trial to trial. This across-trial variability enables the models to account for complicated relationships between correct and error RTs. Although within- and across-trial variabilities are needed to fit correct and error RTs, experimental manipulations that identify sources of across-trial variability in the evidence accumulation rate (drift rate) had not been developed until recently.

Ratcliff, Voskuilen, and McKoon (2018) provided direct evidence for across-trial variability in evidence driving the decision process using experimental data from a double-pass paradigm (Burgess & Colborne, 1988; Cabrera, Lu, & Dosher, 2015; Gold, Bennett, & Sekuler, 1999; Green, 1964; Hasan, Joosten, & Neri, 2012; Lu & Dosher, 2008, 2014; Swets, Shipley, McKey, & Green, 1959). In this paradigm, the exact same stimulus is presented twice with some moderately large interval between the two presentations (we call these repeated trials 'double-pass trials'). If evidence extracted from a stimulus varies from trial to trial and some of the variability is systematic and item-based, then the agreement between repetitions will be greater than chance. Results showed this to be the case thus providing behavioral evidence for the existence of external noise in the decision process.

* Correspondence to: The Ohio State University, 291 Psychology Building, 1835 Neil Avenue Columbus, OH 43210, United States of America.
E-mail address: ratcliff.22@osu.edu (R. Ratcliff).

https://doi.org/10.1016/j.jmp.2020.102431
0022-2496/

Furthermore, in several perceptual and cognitive tasks examined in Ratcliff et al. the amount of across-trial variability in drift rate was able to be decomposed into internal (random) and external (systematic) components.

In this note, we aim to provide additional model-based evidence of the necessity for across-trial variability in processing using the linear ballistic accumulator (LBA) model. To this end, we fit the double-pass data in Ratcliff et al. (2018) with the LBA model and examine if the conclusions are consistent with Ratcliff et al.'s. If so, this analysis would provide additional evidence for across-trial variability in drift rate because the LBA model has different architectural assumptions. The LBA model has to capture what is explained by within-trial variability in the diffusion model analysis by the combination of across-trial variability parameters in drift rate and starting point. Across-trial variability in drift rate in the LBA models can be decomposed into variability due to internal and external sources using the double-pass data as is done for the diffusion model in Ratcliff et al. (2018). However, because there is no within-trial variability in the LBA model, the decomposition will be different from that of the diffusion model. Results show that the LBA model analysis produces an account of the behavioral data and is consistent with the diffusion model analysis.

2. The linear ballistic accumulator model

The goal of this note is to investigate if the LBA model can account for the double-pass data in the same way as the diffusion model in Ratcliff et al. (2018) and to find additional model-based evidence for across-trial variability with the LBA model. In contrast to the diffusion model, the LBA model has only across-trial variability and no within-trial variability, and so all variability in the data needs to be captured by across-trial variability components in the model. Fig. 1 illustrates the LBA model for a two-choice case (Brown & Heathcote, 2008).

The LBA model assumes separate accumulators that represent the two choice options (Responses 1 and 2 in the figure). Evidence accumulation starts at a starting point k with a rate d (i.e., the slope of the 'linear' trend of evidence accumulation without noise) until the accumulated evidence reaches a decision threshold b. Thus, the decision time (DT), the time to reach the threshold for the winning accumulator, is $DT = (b - k)/d$. It is assumed that both the starting point and the drift rate vary across trials. The starting point k is assumed to be uniformly distributed between 0 and A ($k \sim U(0, A)$). The drift rate d is assumed to be normally distributed with a standard deviation of s but with different means for the two accumulators. For a two-choice case, the original parameterization of the LBA model assumed that the two mean drift rates are v and $1 - v$ so that they sum to one (Brown & Heathcote, 2008). But sampled drift rates are not constrained to sum to one; $d_1 \sim N(v, s)$ and $d_2 \sim N(1 - v, s)$. Because the two trial-wise drift rates are assumed to be normally distributed, both of them may have a negative value with a small probability $\Phi(-\frac{v}{s}) \cdot \Phi(-\frac{(1-v)}{s})$ where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. In this case, the process does not terminate properly. To address this issue, different distributional assumptions of drift rate have been proposed (Terry et al., 2015) but we do not employ these versions of the model. The model also assumes a nondecision time (t_0) that corresponds to the minimum possible RT. This represents the time for processes other than the decision process, such as encoding stimuli or producing output. Thus, the RT is determined as the sum of decision time and nondecision time: $RT = DT + t_0$.

3. Sources of variability assumed by the models

There are a number of different possible sources of variability that map into model parameters in somewhat different ways. Fig. 2 shows these sources and how they are related to variability parameters implemented in the diffusion model (left) and the LBA model (right). In the diffusion model, this includes moment-to-moment fluctuations in the decision process within each trial, which is captured by within-trial variability, and changes in attention, vigilance, or sequential effects, which are captured as a part of across-trial variability in drift rate ('across-trial internal random variability'). In the LBA model, there is no within-trial variability unlike the diffusion model and so the model must account for all variabilities as variability across-trials.

In both models, some of the across-trial variability ('across-trial external systematic variability') may come from experimental manipulations, item configurations, and/or some applications (not examined here) from experimental external noise added to stimuli (Lu & Dosher, 1998; Swets et al., 1959).[1,2] Also, some sources of variability may fit in more than one category. For example, differences in encoding may be a function of both configurations of previous trials and attention (which may represent random across-trial variability) as well as differences in the stimuli (which may represent systematic across-trial variability). In this note, we focus on the external noise due to different item configurations used in Ratcliff et al. (2018)'s experiments. In each experimental condition, nominally equivalent stimuli (for the numerosity judgment task, for example, an image with 45 asterisks placed randomly in a 10 × 10 array) were used with differences in their configurations, but the repetitions (e.g., 100 trials apart) used the exact same stimuli.

4. Experiments 1–5

Ratcliff et al. (2018) used the double-pass procedure in five perceptual decision-making tasks (as did Ratcliff & McKoon, 2018, in their Experiment 11). In all the tasks, there were two choice options and subjects were asked to make a decision after a stimulus was presented.

In the numerosity discrimination task, a single array that contained a number of asterisks was presented. The task was to decide if the number of asterisks was greater or less than a criterion number, 50. The number of asterisks was between 36–65 and data were grouped into three conditions with different numbers of asterisks, closer or further from the criterion (correct responses for 36–40 asterisks were grouped with correct responses for 61–65 asterisks and errors were grouped in the same way, 41–45 were grouped with 56–60, and 46–50 were grouped with 51–55).

In the letter discrimination tasks, subjects were presented with two different letters on the left and the right side of the screen, which were the response options, along with a fixation point ('+') in the center. After a short period of time, one of the two letters appeared in the center and then was masked. The task was to decide which of the two letters was presented in the center. There were three presentation durations of the target letter (10, 20, or 30 ms).

In the motion discrimination task, a single array containing dots in a circular area was presented for 400 ms. Some proportion of the dots moved coherently to the left or the right while the

[1] The experimental external noise may be better described as random because it is 'noise'. However, in most double-pass experiments, it is systematic because the exact same visual noise (e.g., the same pattern of random pixels) is added to both presentations of the stimulus.

[2] Experimental external noise was not added to the stimuli in the tasks we examined in this note, and thus our focus was on the item effects, variability from different configurations of nominally equivalent items.

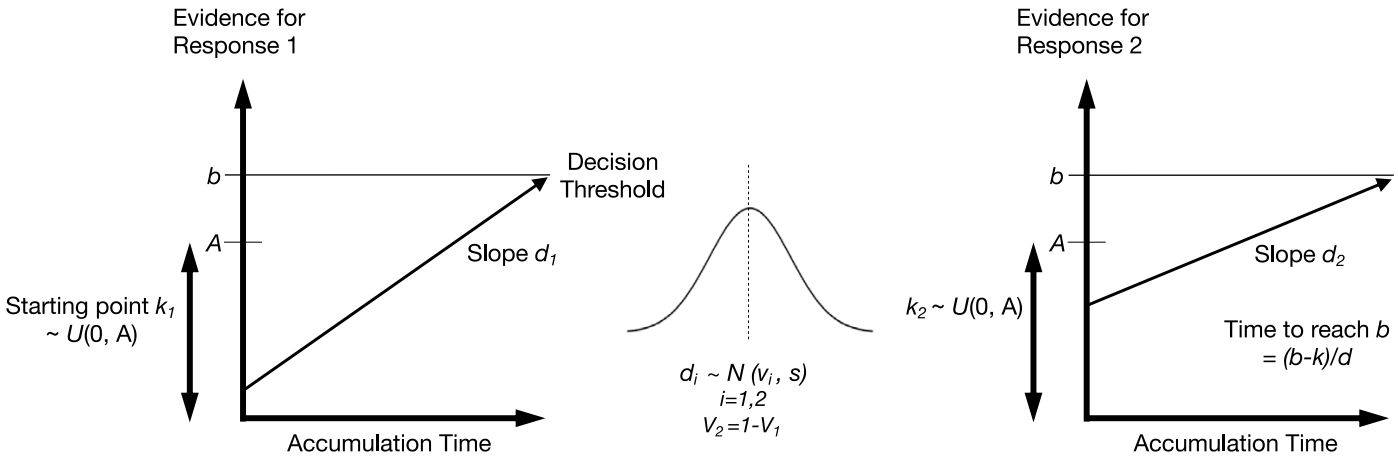

Fig. 1. An illustration of the linear ballistic accumulator model (a two-choice case). There are two accumulators each of which represents one of the two choice options (1 and 2). Each accumulator starts evidence accumulation at a starting point k with a rate d until one of the accumulators reaches a decision threshold b. For each accumulator, starting point k is assumed to be uniformly distributed between 0 and A ($k \sim U(0, A)$) and drift rate d is assumed to be normally distributed with a standard deviation of s but with different means for the two accumulators. For a two-choice case, the original parameterization of the LBA model assumed that the two mean drift rates are $v_1 = v$ and $v_2 = 1 - v$ so that they sum to one (Brown & Heathcote, 2008). But sampled drift rates d_1 and d_2 are not constrained to sum to one; $d_1 \sim N(v, s)$ and $d_2 \sim N(1 - v, s)$.

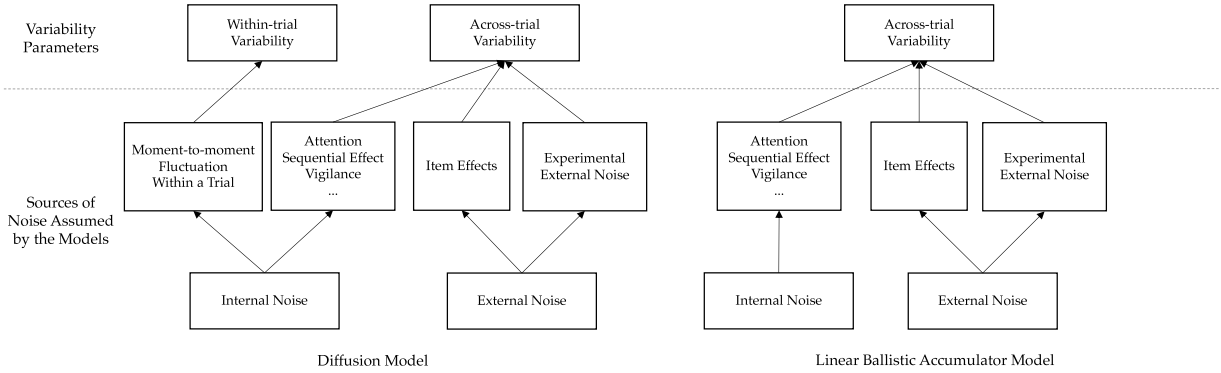

Fig. 2. Internal and external sources of noise assumed by the models (left: the diffusion model; right: the LBA model) and their effect on variability parameters.

other dots moved randomly. The task was to decide the direction of the dots moving in the same direction. There were three proportions of the dots that moved coherently (10%, 15%, or 20%).

In the static brightness discrimination task, black and white pixels were randomly placed in a 64 × 64 array. They were presented for 100 ms and then masked. The task was to decide whether there were more white pixels than black ones. There were two different proportions of white versus black or black versus white pixels (43% vs. 57% or 46% vs. 54%).

The dynamic brightness discrimination task was similar to the static task except the 60 × 60 pixel array of the display changed to another random selection of pixels (with the same probability of white versus black) every 16.67 ms frame of the display. There were two conditions in which the proportion of pixels were 46% vs. 54% or 48% vs. 52%.

In each of the tasks, each block consisted of 90 or 96 trials. The stimuli presented in the second block were identical to those in the first block. The order of the stimuli within a block was the same in the repeated block. The numerosity discrimination task was replicated with the stimuli presented in random order in the repeated block (which did not change the results). Details of the experiments, procedures, and prior studies using these tasks can be found in Ratcliff et al. (2018).

5. Method: Accuracy–agreement plot

Agreement between two responses on double-pass trials provides a measure of how much variability is from random (internal) versus systematic (external) sources. If there is no systematic variability from item configurations and all variability is from random internal sources, the repeated tests of the same stimuli are independent and the probability of agreement can be calculated as $q^2 + (1 - q)^2$ ('baseline') where q is the probability of choosing the correct response. If there is systematic variability, this produces a positive correlation over trials between the two responses on the double-pass trials and thus the two responses are more likely to agree with each other. In this case, the probability of agreement should be higher than this baseline.

Fig. 3 shows the example of accuracy–agreement functions generated from bivariate binomial random variables (Burgess & Colborne, 1988; Lu & Dosher, 2008; Ratcliff et al., 2018). The probability of agreement between two elements of a bivariate binomial sample is plotted against accuracy (a success probability of the bivariate binomial distribution, which is the same for both elements). Different curves represent accuracy–agreement functions with different correlations (from 0 to 0.8 by 0.2) and dots on each of the curves represent different accuracy values (from 0.5 to 0.95 by 0.05). If the correlation is zero (the leftmost function), two binomial samples are independent and the agreement probability should be the baseline level ($q^2 + (1 - q)^2$). For example, $q = 0.5$ gives $0.5^2 + (1-0.5)^2 = 0.5$ and $q = 0.8$ gives $0.8^2 + (1-0.8)^2 = 0.68$. If the correlation is positive, it increases the agreement probability and moves the functions further to the right.

The accuracy–agreement functions can be simulated using the LBA model instead of the bivariate binomial distribution used in

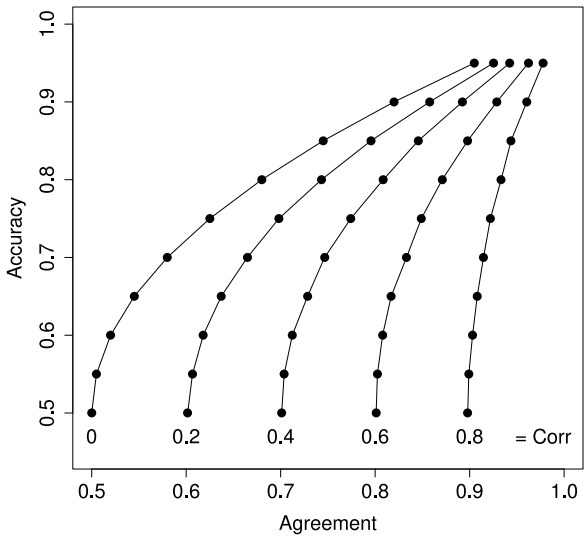

Fig. 3. An example of the accuracy–agreement plot generated from bivariate binomial random variables. The probability of agreement between two elements of a bivariate binomial sample is plotted on the *x*-axis against accuracy (a success probability of the bivariate binomial distribution, which is the same for both elements) on the *y*-axis. The different curves represent different correlations and different dots on each curve represent different accuracy values.

the example above. Before simulating the functions, the model needs to be fit to the choice proportion and RT distribution data to obtain parameter values for use in the simulations. Different values of drift rate v (and $1-v$ for the competing accumulator) and across-trial total variability in drift rate (s) are used in the simulation. The other parameters are fixed to the best-fitting parameters. Given v and s, the shapes of the accuracy–agreement functions are determined by the decision threshold b and the upper bound of the starting point A. The simulation procedure is as follows:

1. Given the mean drift rate v for a condition and across-trial external variability s, item drift rates for the two accumulators are sampled from $d_{1i} \sim N(v, s)$ and $d_{2i} \sim N(1-v, s)$, $i = 1, \ldots, I$ where i is the subscript for the ith stimulus (item) and I is the number of stimuli used in the condition. Each item is assumed to be presented twice, and so the total number of trials in the condition is $T = 2I$.
2. Set drift rates for the double-pass of the same item i: $d_{1i,1} = d_{1i,2} = d_{1i}$ and $d_{2i,1} = d_{2i,2} = d_{2i}$.
3. For each trial, sample trial-wise starting points k_1 and k_2 from $U(0, A)$ and simulate choice and RT data given the drift rates, the starting points above and b from the model fit.
4. Calculate the agreement of two responses on the double-pass trials and accuracy for the condition. Plot these values.
5. Repeat the above with different values of v. This produces the accuracy–agreement function for a single s.
6. Change s, repeat the above, and plot the next function.

In step 2 of the simulation procedure, it is assumed that drift rates for two presentations of the same item are exactly equal. This means that all across-trial total variability is systematic, i.e., $s = s_E$ where s_E indicates across-trial systematic/external variability (standard deviation) and that no across-trial internal variability is considered in the simulated accuracy–agreement functions. Across-trial internal variability can easily be implemented in the simulation procedure by adding random noise to the double pass drift rates: for one accumulator, $d_{1i,j} = d_{1i} + \epsilon_{1i,j}$ where $j = 1, 2$ represents the two presentations and $\epsilon_{1i,j}$ is a random number from a normal distribution with mean zero and standard deviation s_I that represents across-trial internal variability. In this case, $s > s_E$ due to added random noise and s_I is determined by $s_I^2 = s^2 - s_E^2$. For the other accumulator, a separate random number is added to $d_{2i,j}, j = 1, 2$ with the same sampling procedure. For simulations using this procedure, drift rates for double-pass trials are not identical, because they differ by across-trial internal variability. Ratcliff et al. (2018) performed both analyses for the diffusion model, with and without across-trial internal variability, and found that adding across-trial internal variability did not change their conclusions. It turned out across-trial internal variability reduced accuracy but it did not change the shapes and locations of accuracy–agreement functions.

Given a fixed level of across-trial total variability, increasing across-trial external variability s_E is the same as increasing the correlation between double-pass trials because the proportion of systematic variability relative to the total across-trial variability increases. Thus, as s_E increases, the accuracy–agreement function would be placed further to the right, as it does as the correlation increases in Fig. 3 and so different values of s_E produce different accuracy–agreement functions (as different curves in Fig. 3). Also, as the mean drift rate v increases ($1 - v$ for the competing accumulator decreases), item drift rates would tend to be higher and so both accuracy and agreement increase. Thus, increasing v produces the same effect on accuracy–agreement functions as increasing q in Fig. 3.

To examine how much variability is from internal and external sources, the model predictions of accuracy and agreement simulated as above are compared to the observations from double-pass experiments. For the data, the observed accuracy is calculated as the number of correct responses divided by the number of trials for each condition. Then, the proportion of double-pass trials on which the two responses agree is computed. These accuracy and agreement values are calculated for each condition and for each subject and then averaged over subjects. The results are then plotted on the predicted accuracy–agreement plot from the simulation procedure.

There are two important reference curves for this comparison. The one is the leftmost curve in the accuracy–agreement plot (baseline). This curve represents accuracy and agreement when there is no across-trial external variability ($s_E = 0$) and double-pass trials are independent (all variability in the model predictions comes from variability in starting point). If the observed accuracy and agreement fall to the right of the leftmost curve, the observed agreement is higher than the baseline-level agreement obtained when the double-pass trials are independent. This implies the existence of systematic external variability that leads to a higher agreement. The other reference curve is the predicted accuracy–agreement curve that matches the estimated across-trial total variability s. This estimate is obtained from the LBA model fit to the experimental data. This curve represents accuracy and agreement when all variability is from external systematic sources ($s = s_E$). If the observed accuracy and agreement fall to the left side of this curve, the difference between the observed values and this curve represents across-trial internal variability.

Ratcliff et al. (2018) performed the same analysis but with the diffusion model. Except for the motion discrimination task, the observed accuracy and agreement were located between the two reference lines, showing the proportions of across-trial internal variability and external variability. For the motion discrimination task, the observed and predicted double-pass agreement were very close, showing that the across-trial internal variability was small relative to the amount of systematic variability in this task.

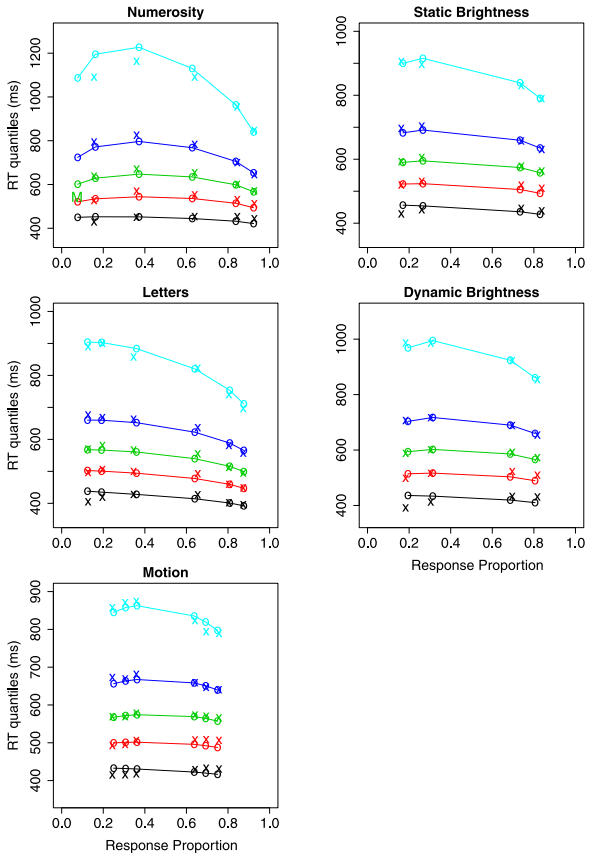

Fig. 4. Quantile probability functions for the data (x) and the LBA model predictions (o and lines joining them) from the five experiments. Quantile RTs for the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles (stacked vertically, with the 0.1 quantile at the bottom) are plotted against the response proportions. Correct responses are plotted on the right while the errors are plotted on the left. The LBA model predictions were generated for each subject and then averaged. The data were averaged in the same way.

6. Results

We fitted the LBA model to the data from all five tasks by a quantile-based method (G^2 method; Ratcliff & Childers, 2015; Ratcliff & Smith, 2004). The model was fit to each subject separately and the average parameter estimates are shown in Table 1. Fig. 4 shows the quantile-probability plot of the five tasks in which the five RT quantiles (0.1, 0.3, 0.5, 0.7, and 0.9) are plotted against the response proportions. Correct responses are plotted on the right while the errors are plotted on the left. In the plot, the model prediction (the circles and the lines joining them) matches the data (x's) well. This demonstrates that the LBA model fits the behavioral data about as well as the diffusion model.

To simulate accuracy–agreement functions, we used sixteen levels of mean drift rate (v), from 0.5 to 1.25 by a difference of 0.05, and ten levels of across-trial external (systematic) standard deviation in drift rate (s_E), from 0 to 0.50 by a difference of 0.05. No across-trial internal (random) variability was used in the simulation. For the other parameters involved in the simulation procedure, namely the decision threshold (b) and the upper bound (and range) of the distribution of starting points (A), the mean parameter estimates from fits to data were used (Table 1). For each combination of v and s_E, it was assumed that drift rates were the same for the two repeated presentations of the same stimulus (i.e., item drift rates d_{1i} and d_{2i} for the two accumulators), while the starting points differed randomly for the two presentations. The model produced simulated choices and RTs with these parameter values and this allowed choice accuracy and agreement between the two repeated presentations to be computed. The simulated accuracy–agreement functions are shown in Fig. 5. Different levels of v (the dots on the curves) produced different levels of accuracy along the curves while different levels of s_E produced the different curves.

On the simulated accuracy–agreement plot, the red squares represent the observed accuracy and the degree of agreement between the double-pass trials for the different experimental conditions. Despite the differences in the model structures, the LBA model produces similar results as those in Ratcliff et al. (2018)'s diffusion model analysis. For all five tasks, the red squares deviate from the leftmost (baseline) function and lie on or close to one of the other simulated functions. For the numerosity task, the red squares fall closely on the function for $s_E = 0.2$, for the letter task, they fall between the functions for $s_E = 0.05$ and $s_E = 0.1$, for the random dot motion task, they fall between the functions for $s_E = 0.1$ and $s_E = 0.15$, for the static brightness task, they fall closely on the function for $s_E = 0.05$, and for the dynamic brightness task, they fall closely on the function for $s_E = 0.1$. The deviation of the red squares from the leftmost function provides evidence for external (systematic) noise for all five tasks within the framework of the LBA model. The location of the data accuracy–agreement on the simulated accuracy–agreement plot can provide an interval measure of across-trial systematic variability. For example, $0.05 < s_E < 0.1$ for the letter discrimination task and $0.1 < s_E < 0.15$ for the random dot motion task.

On the same accuracy–agreement plot, the thick black curves represent the accuracy–agreement functions that most closely correspond to the across-trial variability estimates from the model fit (e.g., for the motion discrimination task, the estimated standard deviation for across-trial variability in drift rate is 0.261 and the thick black curve indicates the accuracy–agreement function simulated with variability of 0.25). In all five tasks, the red

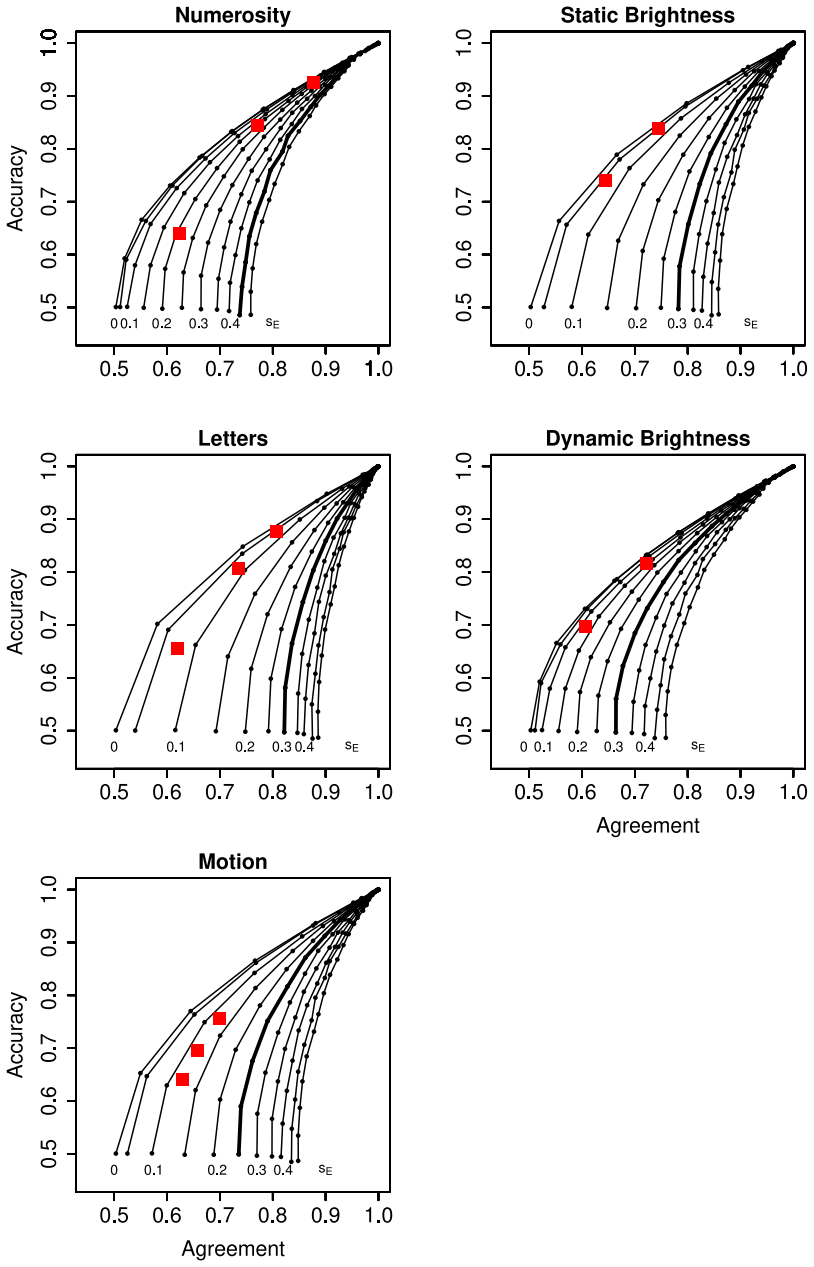

Fig. 5. Accuracy–agreement functions simulated using the LBA model for the five experiments. Sixteen levels of drift rate (v) and eleven levels of across-trial external (systematic) standard deviation in drift rate (s_E) were used to produce the functions. Different curves represent different levels of s_E and different dots on each curve represent different levels of v. For the other model parameters, the best parameter estimates were used (Table 1). The thick curve in each panel represents the accuracy–agreement function that best matches the estimated across-trial total variability from the LBA model fit to the data. The red squares represent the observed accuracy–agreement calculated from the double-pass data and the values were calculated for each condition and for each subject and then averaged over subjects.

Table 1
The LBA parameter estimates from fits to data. The model was fit to the data from each subject separately and then parameter estimates were averaged across subjects.

Discrimination task	b	A	t_0	s	v_1	v_2	v_3	G^2
Numerosity	0.475	0.407	0.334	0.473	1.180	0.942	0.645	98.6
Letter	0.306	0.176	0.240	0.308	0.831	0.743	0.597	73.6
Motion	0.353	0.267	0.263	0.261	0.695	0.641	0.598	75.6
Static brightness	0.344	0.241	0.278	0.297	0.784	0.678		101.6
Dynamic brightness	0.415	0.298	0.231	0.310	0.765	0.644		83.6

Note. The parameters are decision threshold b; upper bound of the starting point A; the minimum nondecision component of response time, t_0; across-trial (total) variability in drift rate s; drift rates v_1, v_2, and v_3 for the most difficult, easier, and easiest conditions, respectively.

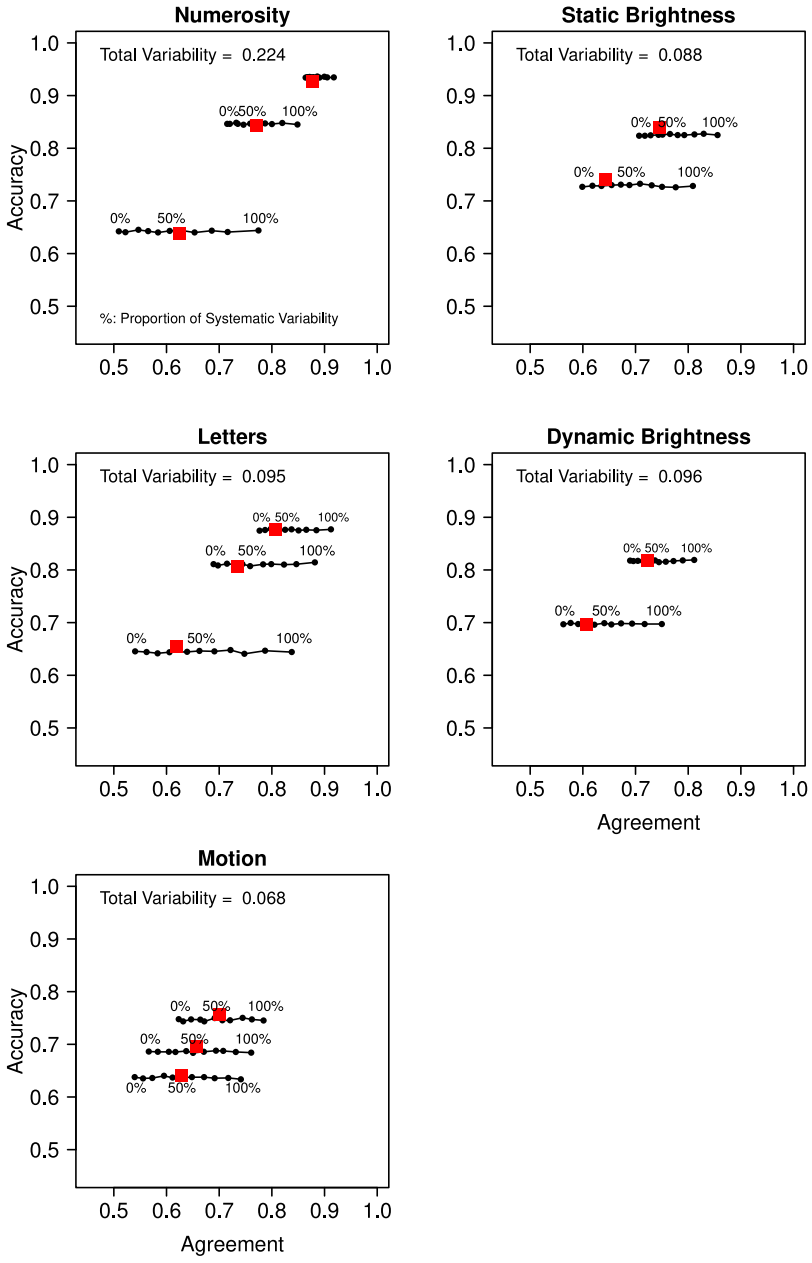

Fig. 6. Decomposition of the estimated across-trial total variability. On top of each panel, the estimated total variability (s^2) is presented, which is the squared value of the across-trial standard deviation s in Table 1. The black dots indicate the simulated accuracy–agreement values as a function of p (the proportion of systematic noises among the total variability) and the red squares represent the observed accuracy–agreement values calculated from the double-pass data.

squares fall to the left of the thick black curves. The distance between them, which is a measure of across-trial internal variability, is estimated to be a larger proportion of total across-trial variability than that in Ratcliff et al. In particular, for the motion discrimination task, the red squares fall almost exactly on the thick black curve in the diffusion model result (Figure 4-C in Ratcliff et al., 2018) while they deviate substantially from the thick black curve in the LBA model result.

With the LBA model, we can further examine how much of the variability in processing is from random internal and systematic external sources given the across-trial total variability estimated from the data. Fig. 6 shows another way of displaying the decomposition of the total variability in processing. We first assumed that $p \times 100\%$ of the across-trial total variability is systematic and generated model predictions of accuracy and agreement with different p's, from 0.0 to 1.0 by 0.1. This simulation was based on the following representation of the drift rates for the double-pass trials: for one accumulator,

$$d_{i,j} = d_i + \epsilon_{i,j} = v + \gamma_i + \epsilon_{i,j}, \quad j = 1, 2$$
$$\text{var}(\gamma_i) = s_E^2, \quad \text{var}(\epsilon_{i,j}) = s_I^2$$
$$\text{cov}(\gamma_i, \epsilon_{i,j}) = 0$$

where v is the mean drift rate for one accumulator, γ_i is the item effect of the stimulus i on drift rates (increase or decrease of drift rate due to the item configuration) and $\epsilon_{i,j}$ is a residual for each presentation. Thus, s_E^2 represents across-trial external (systematic) variability while s_I^2 represents across-trial internal (random) variability. From the equation above, $s^2 = \text{var}(v_{i,j}) = s_E^2 + s_I^2$. While we estimated s^2 from the LBA model fit to the data, s_E^2 and s_I^2 were not estimated separately. However, accuracy and agreement can be simulated given the assumption that $s_E^2 = ps^2$ and $s_I^2 = (1-p)s^2$.

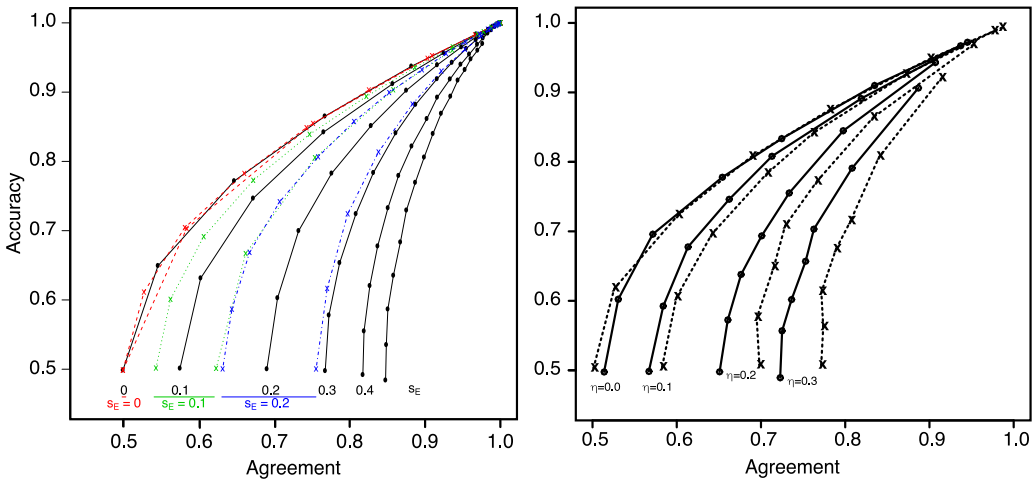

Fig. 7. The precision of accuracy–agreement functions simulated using the LBA model (left) and the diffusion model (right; from Ratcliff & Smith, 2020) for the motion discrimination task. Left: The same simulation procedure as in Fig. 5 was used to produce the functions with the LBA model, but with $s_E = 0.0 - 0.5$ by increments of 0.1. For the first three s_E values, a rough confidence interval of predicted accuracy–agreement was obtained by taking $A \pm 2SE(A)$ in the simulation procedure. The x's with the red-dashed, green-dotted, and blue-dot-dashed curves represent these intervals for $s_E = 0.0$, 0.1, and 0.2, respectively, and the horizontal lines at the bottom represent the length of the corresponding intervals. Right: The functions were simulated by the diffusion model with the boundary separation $a = 0.098$ and across-trial variability in starting point $s_z = 0.062$ (the solid curves) and $a = 0.098$ and $s_z = 0$ (the dashed curves). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The black dots in Fig. 6 display these predictions as a function of p. The leftmost (rightmost) dot represents that all variability is from random internal (systematic external) sources. The dots in the middle indicate the simulated accuracy–agreement values with 0.1 increments of p. The red squares represent the observed accuracy and agreement values as in Fig. 5. The model predictions (black dots) closest to the red squares represent the approximate proportions of systematic (due to the item effects in this study) across-trial variability. The results show that, for all three conditions, about 60% and 40–45% of the total variability are from the external sources in the numerosity judgment task and the letter discrimination task, respectively. In the motion discrimination task, about 50–55% of the total variability is from the external noise for all three conditions. In the static brightness task, the proportions of the external variability are 30% and 25%, and in the dynamic brightness task, the proportions are 40% and 30%, for the easier and harder conditions, respectively. The remaining amount of total variability can be attributed to random internal sources.

Figs. 5 and 6 provide evidence for the existence of systematic external sources (the distance between the red squares and the leftmost reference) and furthermore, a measure of how much variability is from internal and external sources. Precision of this measure depends on (1) the precision of the data accuracy and agreement values and (2) the precision of predicted functions. The data accuracy and agreement values have high precision. In all five experiments, the standard errors of accuracy (a standard deviation of the binomial proportion) are between 0.004 and 0.006. The standard errors of the data agreement can be obtained using a bootstrap resampling method and they are between 0.022 and 0.028.

The shape of the predicted accuracy–agreement functions is determined by the decision threshold b and the upper bound of the starting point A, and so the precision of the predicted functions can be examined based on the standard errors of the parameter estimates. Ratcliff and Smith (2020) examined the precision of the functions for the motion discrimination task predicted by the diffusion model. They compared the functions simulated with across-trial variability in starting point $s_z = 0.062$ (the estimate from the model fit) and $s_z = 0$, with the boundary separation estimate $a = 0.098$, and found that the systematic component can be reasonably precisely measured until the across-trial variability in drift rate η becomes larger than 0.1 (Fig. 7, the right panel). In the LBA model, A (which best matches s_z in the diffusion model) cannot be set to zero because this parameter is the only source of variability in the model given the same drift rates. Thus, if $A = 0$, there is no variability and the model cannot properly simulate choice and RT data. Instead, we performed a parameter recovery study to obtain $SE(A)$, the standard error of A, and simulated accuracy–agreement functions with $A \pm 2SE(A)$. A larger value of A produces faster errors and so the estimation of A largely depends on the shift and spread of the error RT distributions relative to the correct RT distributions. This means that $SE(A)$ is determined in part by the value of true A. To examine this, the recovery study was carried out with different values of A. We let Ar be the ratio of A relative to b. In the simulations, we used $Ar = 0 - 0.9$ with increments of 0.1. The other parameter values were $b = 0.4$, $t_0 = 0.25$, $s = 0.3$, $v_1 = 0.9$, $v_2 = 0.75$, and $v_3 = 0.6$. These values are close to the average of the parameter estimates (across the five tasks) from the model fits in Table 1. For each value of Ar, choices and RTs were produced by a simulation for 100 subjects, three conditions, and 500 observations per condition. Then the LBA model was fit to the data using the G^2 method.

The estimates and standard errors of b, A, and s obtained from the parameter recovery are shown in Table 2. The recovery was reasonably good except that A was overestimated when the value of true A was 0.00 or 0.04 ($Ar = 0.0 - 0.1$). The standard errors largely depended on the value of true A. The value of $SE(A)$ when $A = 0$ was smaller than that when $A = 0.04 - 0.12$ ($Ar = 0.1 - 0.3$) but this was because A is bounded to be positive. Except for this case, the values of $SE(A)$ decreased as a function of Ar. For $A = 0.28 - 0.32$ or $Ar = 0.7 - 0.8$ (which match four of the five tasks in the current study), the values of $SE(A)$ were about 1/10 of the mean value of A. Then the precision of the accuracy–agreement functions was examined with these standard errors. Fig. 7 shows the predicted accuracy–agreement functions for the motion discrimination task with six values of s_E: $s_E = 0.0 - 0.5$ with increments of 0.1. For the first three values of s_E, a rough confidence interval of the function was obtained by simulating the curve with $A \pm 2SE(A)$. The x's with the red-dashed, green-dotted, and blue-dot-dashed curves represent these intervals for $s_E = 0.0$, 0.1, and 0.2, respectively, and the horizontal lines at

Table 2
Estimates (the first three rows) and standard errors (the next three rows) of the parameter estimates of the upper bound of the starting point A and across-trial variability in drift rate s (the LBA model). The values of true A in the leading row correspond to $Ar = 0.0 - 0.9$ with increments of 0.1 given $b = 0.4$ where Ar is the ratio of A relative to b. The value of true s was $s = 0.3$.

True A	0.00	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36
b	0.437	0.418	0.403	0.399	0.391	0.400	0.397	0.400	0.397	0.395
A	0.081	0.086	0.091	0.121	0.144	0.195	0.233	0.276	0.314	0.353
s	0.312	0.315	0.308	0.312	0.304	0.305	0.303	0.294	0.297	0.299
$SE(b)$	0.037	0.036	0.034	0.032	0.027	0.021	0.017	0.013	0.011	0.013
$SE(A)$	0.080	0.083	0.084	0.082	0.079	0.065	0.052	0.033	0.020	0.015
$SE(s)$	0.043	0.043	0.040	0.039	0.029	0.043	0.036	0.031	0.030	0.027

the bottom represent the length of the corresponding intervals. It turned out the standard errors of the accuracy–agreement function are too large to allow precise measurement of s_E. In contrast, in the diffusion model framework, because within-trial variability accounts for some of the variability that the LBA model has to capture by s, the model produces a less variable estimate of the systematic across-trial variability (Fig. 7, the right panel; Ratcliff & Smith, 2020). However, the leftmost function in Fig. 7 left panel was obtained precisely from the LBA model, and so the deviation between this function and the double-pass accuracy and agreement from the data plotted in Fig. 5 provides reliable evidence for the existence of systematic external sources.

7. Discussion

In this note, we have used the LBA model to interpret the experimental data produced from the double-pass experiments presented in Ratcliff et al. (2018). The estimate of the across-trial variability parameter in the LBA model was decomposed into across-trial internal (random) variability and external (systematic) variability based on the observed accuracy and agreement calculated from the double-pass trials. In this decomposition, the double-pass agreement was higher than the baseline agreement predicted with the assumption that the two responses on the double-pass trials were independent. Also, the double-pass agreement was lower than the agreement that would be expected if all across-trial total variability was systematic, i.e., the drift rates were identical on both passes. This result showed that there were substantial proportions of systematic sources of variability (but not all of the total variability) across all five tasks, which is consistent with the result from the diffusion model analysis in Ratcliff et al. This provided model-based evidence for external variability and demonstrated the necessity of across-trial variability to model human choice and RT data.

The decomposition result can also be used to estimate how much variability in processing is from internal and external sources (Ratcliff & Smith, 2020). The distance between the double-pass accuracy and agreement from the data and the baseline curve can be used as a measure of systematic across-trial variability. Furthermore, the distance between the double-pass accuracy and agreement from the data and the predicted function assuming that the estimated across-trial total variability is systematic can be used as a measure of random across-trial variability. In our result, the LBA model estimated the random component as a larger proportion of the total across-trial variability than the diffusion model result in Ratcliff et al. (2018). This was because the LBA model had to account for what is captured by within-trial variability in the diffusion model solely by across-trial variability parameters.

The measures of random and systematic sources of variability that the LBA model provided had much larger standard errors than those from the diffusion model. The locations of the accuracy–agreement functions predicted from the LBA model were affected to a large degree by the precision of the range of the starting points A, except for those of the baseline curve (Fig. 7). This is because across-trial variability in starting point is the only source of variability in the model predictions. Thus, the LBA model was not able to provide precise measures of random and systematic sources of variability in processing. In contrast, the diffusion model predicts accuracy–agreement functions that are stable and do not vary much with across-trial variability in starting point, until the total across-trial variability in drift rate becomes large (Ratcliff & Smith, 2020). Despite the limitation of the LBA model as a measurement tool for different sources of variability in drift rate, the evidence from the LBA model for the existence of the systematic sources of variability was reliable because the baseline curve was precisely predicted and the double-pass accuracy and agreement from the data deviated from this curve.

One of the reviewers raised an alternative explanation: within-trial noise could be correlated between repetitions of an item. This could account for the greater than chance agreement in double-pass trials without across-trial systematic variability. However, this means that double-pass trials should produce similar paths of evidence accumulation with similar rates for the two passes. This results in systematic differences in drift rate from the mean for these two items and such differences result in across-trial systematic/external variability in drift rate. Also, this correlated within-trial variability explanation is less plausible because it would require that all sources of within-trial noise such as neural spikes, fluctuations in attention, sequential effects, vigilance, etc., would be set to similar values across widely separated presentations of a test stimulus.

Recently, the decomposition method used in this note was called into question by Evans, Tillman, and Wagenmakers (2020). They argued that Ratcliff et al. (2018) conflated different (systematic and random) sources of across-trial variability in drift rate, and failed to provide evidence for their central claim. In fact, Evans et al. began their discussion by misinterpreting how internal and external sources of variability were defined in Ratcliff et al. (we followed the convention in Ratcliff et al. and earlier studies using the double-pass procedure in this note). Evans et al. stated that "internal noise refers to random within-trial variability in drift rate and external noise refers to random between-trial variability in drift rate". Ratcliff et al. explicitly discriminated within-trial (internal) variability and across-trial internal variability. Evans et al. claimed that Ratcliff et al. attempted to show the evidence for random across-trial variability in drift rate but failed to distinguish systematic sources of variability in the across-trial total variability from the random sources. However, providing the evidence for the systematic sources was the central aim of Ratcliff et al. (as stated in their abstract) and this note.

Evans et al. also attempted to show that systematic and random sources of across-trial variability cannot be distinguished from each other using the double-pass paradigm. The main argument made by them was that there is a trade-off between the across-trial total variability in drift rate and the correlation between the two responses in double-pass trials and so different

combinations of these two quantities can produce nearly the same accuracy–agreement functions. They further argued that the correlation cannot be identified due to this trade-off. However, the trade-off and the identification problem appear only when both quantities have to be identified simultaneously with no other constraint. If one of the quantities can be constrained, then the other can be identified. In Ratcliff et al. and the current note, the across-trial variability in drift rate is constrained from the model fit to the accuracy and RT data. From this, the correlation could be measured by comparing the double-pass accuracy and agreement calculated from the data and the accuracy–agreement functions predicted by the models. With an accurate estimate of across-trial variability in drift rate, this two-step approach can minimize the trade-off between the two quantities.

Evans et al. also admitted that the problematic trade-off can be solved by the two-step approach. But they pointed out that this two-step approach requires an unbiased and precise estimate of the across-trial total variability, which they argued could not be obtained. They based this argument on their simulation study in which parameter recovery of the across-trial variability in drift rate (η) in the diffusion model was examined. They showed that when η is small, it is poorly recovered, but when η and drift rate v (for a single condition) have reasonably large values that correspond to typical diffusion model estimates including those in Ratcliff et al., parameter recovery is quite good. In fact, Evans et al.'s simulation study demonstrated that the estimates of across-trial total variability in drift rate were good enough to provide evidence for the systematic components of the variability and show the necessity of modeling across-trial variability in drift rate (which is the central claim of Ratcliff et al. and the current paper).

Parameter recovery of the across-trial variability in drift rate parameter s in the LBA model has been well documented by previous studies (Donkin, Averell, Brown, & Heathcote, 2009; Donkin, Brown, Heathcote, & Wagenmakers, 2011; Visser & Poessé, 2017) and also by our parameter recovery study (Table 2). However, as shown in Fig. 7, the LBA model was not able to produce reasonably precise measures of random and systematic components of the across-trial variability in drift rate. This is because the accuracy–agreement functions predicted by the model, except for the leftmost baseline function, are highly dependent on the range of the starting points and that even a quite small standard error in the estimate produces accuracy–agreement functions that differ substantially. Our results show, for the LBA model, the existence of systematic and random components of the across-trial total variability in drift rate, but the relative proportions of systematic and random components cannot be precisely measured.

The double-pass paradigm, along with appropriate mathematical models, provides a way of studying internal and external sources of variability in perceptual and cognitive decision-making. The evidence for systematic across-trial variability we provided suggests that it is necessary for mathematical models to represent such variability. Implementing across-trial variability parameters as in the LBA model (and the diffusion model used in Ratcliff et al.) is a direct way to achieve this aim. Models that do not explicitly consider the contributions of sources of variability to data are unable to properly explain variability in data (e.g., the relative speeds of correct and error responses) and they would likely attribute this variability to the wrong cognitive components, thus failing to provide an accurate account of the cognitive processes underlying perceptual and cognitive decision-making.

Acknowledgment

This work was supported by the National Institutes of Health grants R01-AG041176 and R01-AG057841.

References

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

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A*, *5*, 617–627.

Cabrera, C. A., Lu, Z. L., & Dosher, B. A. (2015). Separating decision and encoding noise in signal detection tasks. *Psychological Review*, *122*, 429–460.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*, 693–702.

Deneve, S. (2012). Making decisions with unknown sensory reliability. *Frontiers in Neuroscience*, *6*, 75.

Ditterich, J. (2006). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks*, *19*, 981–1012.

Donkin, C., Averell, L., Brown, S. D., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods*, *41*(4), 1095-1110.

Donkin, C., Brown, S. D., Heathcote, A., & Wagenmakers, E. J. (2011). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Psychonomic Bulletin & Review*, *55*, 140–151.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, *32*, 3612–3628.

Evans, N. J., Tillman, G., & Wagenmakers, E. (2020). Systematic and random sources of variability in perceptual decision-making: Comment on Ratcliff, Voskuilen, and McKoon (2018). *Psychological Review*, Available online at https://doi.org/10.31234/osf.io/j98qd.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, *402*, 176–178.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, *71*, 392–407.

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience*, *31*, 6339–6352.

Hasan, B. A. S., Joosten, E., & Neri, P. (2012). Estimation of internal noise using double passes: Does it matter how the second pass is delivered? *Vision Research*, *69*, 1–9.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*, 1329–1342.

Lu, Z. L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, *38*(9), 1183–1198.

Lu, Z. L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*, *115*, 44–82.

Lu, Z. L., & Dosher, B. A. (2014). *Visual psychophysics: From laboratory to theory*. Cambridge, MA: MIT Press.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*, 376–404.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279.

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

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review*, *125*(2), 183–217.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367.

Ratcliff, R., & Smith, P. L. (2020). *Estimating systematic and random sources of variability in perceptual decision-making: A reply to evans*. Tillman, & Wagenmakers, (In Review).

Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review*, *125*(1), 33–46.

Swets, J. A., Shipley, E. F., McKey, M. J., & Green, D. M. (1959). Multiple observations of signals in noise. *Journal of the Acoustical Society of America*, *31*, 514–521.

Terry, A., Marley, A. A. J., Barnwal, A., Wagenmakers, E.-J., Heathcote, A., & Brown, S. D. (2015). Generalising the drift rate distribution for linear ballistic accumulators. *Journal of Mathematical Psychology*, *68–69*, 49–58.

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

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*, 61–72.

Visser, I., & Poessé, R. (2017). Parameter recovery, bias and standard errors in the linear ballistic accumulator model. *British Journal of Mathematical and Statistical Psychology*, *70*, 280–296.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*, 1–9.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*(14), 1–10.

Zhang, S., Lee, M. D., Vandekerckhove, J., Maris, G., & Wagenmakers, E. J. (2014). Time-varying boundaries for diffusion models of decision making and response time. *Frontiers in Psychology*, *5*, 1364.