Systematic and random sources of variability in perceptual decision-making

Comment on Ratcliff, Voskuilen, and McKoon (2018)

Evans, N.J.; Tillman, G.; Wagenmakers, E.-J.

DOI
10.31234/osf.io/j98qd

Publication date
2019

Document Version
Submitted manuscript

Systematic and Random Sources of Variability in Perceptual Decision-Making: Comment on Ratcliff, Voskuilen, and McKoon (2018)

Nathan J. Evans[a], Gabriel Tillman[b], and Eric-Jan Wagenmakers[a]

[a] Department of Psychology, University of Amsterdam, The Netherlands

[b] Australian College of Applied Psychology, Australia

Word count: 6,355

Correspondence concerning this article may be addressed to: Nathan Evans (nathan.j.evans@uon.edu.au). Associated code is available at `https://osf.io/d8xfb/`.

Abstract

The existence of variability in information processing is a key assumption in models of human cognition, with variability being required to successfully account for a range of human behaviour. Evidence accumulation models (EAMs) commonly assume two broad variabilities in information processing: within-trial variability, which is thought to reflect moment-to-moment fluctuations in perceptual processes, and between-trial variability, which is thought to reflect variability in slower-changing processes like attention, or systematic variability between the stimuli on different trials. Recently, Ratcliff, Voskuilen, and McKoon (2018) claimed to "provide direct evidence that external noise is, in fact, required to explain the data from five simple two-choice decision tasks" (p. 33), suggesting that at least some portion of the variability in between-trial information processing is due to "noise". We agree that Ratcliff et al. (2018) provided evidence that drift rate varies – in at least some manner – between trials in an experiment. However, we argue that Ratcliff et al. (2018) conflated two different potential sources of between-trial variability: random (i.e., "external noise") and systematic (e.g., item effects). We also argue that their analyses failed to distinguish between these sources, meaning that their results could have been due to "external noise" and/or item effects. Furthermore, we contend that the concept of "noise" within cognitive models merely serves as a convenience parameter for sources of variability that we know exist, but are unable to account for. Therefore, we question the usefulness of experiments aimed at testing the general existence of "random" variability, and instead suggest that future research should attempt to replace the random variability terms within cognitive models with actual explanations of the process.

Keywords:

diffusion model — between-trial variability — random variability — systematic variability

The existence of variability in information processing is a key assumption in models of human cognition, with variability being required to successfully account for a range of human behaviour, such as in memory (Shiffrin & Steyvers, 1997; Osth & Dennis, 2015), practice and learning (Wagenmakers & Brown, 2007; Evans, Brown, Mewhort, & Heathcote, 2018), automaticity (Logan, 1988, 1992), and decision-making (Ratcliff, 1978; Usher & McClelland, 2001; Brown & Heathcote, 2005, 2008), among many others. Within the decision-making literature, evidence accumulation models (EAMs; Stone, 1960) propose that evidence (i.e., processed information) is accumulated for each decision alternative at some rate (known as the "drift rate"), until the evidence for one alternative reaches some threshold level, which triggers an overt response for that alternative (see Ratcliff, Philiastides, & Sajda, 2009; Ratcliff & Smith, 2004; Tillman, Osth, van Ravenzwaaij, & Heathcote, 2017; Brown, Marley, Donkin, & Heathcote, 2008; Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Ratcliff, Van Zandt, & McKoon, 1999; Evans & Brown, 2017; Ratcliff, Thapar, & McKoon, 2001; Evans, Rae, Bushmakin, Rubin, & Brown, 2017; Ratcliff, Thapar, & McKoon, 2011; Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017; Ratcliff, 2006; Evans, Steyvers, & Brown, 2018; Ratcliff & Starns, 2013; Ratcliff & McKoon, 2008; Forstmann, Dutilh, Brown, Neumann, & von Cramon, 2008; Tillman, Benders, Brown, & van Ravenzwaaij, 2017; Forstmann et al., 2011 for applications). Most EAMs include two sources of variability in information processing (though see Usher & McClelland, 2001; Brown & Heathcote, 2008). The first is within-trial variability, which represents moment-to-moment fluctuations in our perceptual processing (Ratcliff, 1978). Within-trial variability is ongoing during a single decision, and is implemented in EAMs as a temporal sequence of random draws from a normal distribution with a mean of zero. The second source of variability in information processing is between-trials, which may reflect variability in processing for different items from the same category across trials,

fluctuations in processes like attention, or sequential and/or time-based effects. Between-trial variability is also commonly implemented in EAMs as random draws from a normal distribution (related to the concept of signal detection theory, Ratcliff, 1978), though these draws occur from one decision to the next, rather than within a single decision. The inclusion of between-trial variability in information processing allows EAMs to account for key qualitative benchmarks observed in decision-making (Ratcliff, 1978; Brown & Heathcote, 2008; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), and well-known behavioural findings in related areas, such as sequential effects (Ratcliff et al., 1999) and differences in electroencephalography (EEG) signals (Ratcliff et al., 2009).

A recent article by Ratcliff et al. (2018) assessed "whether current models can explain accuracy and RT data with only internal noise or whether the external noise, or variation between stimulus exemplars, is also required" (p.33). Here internal noise refers to random within-trial variability in drift rate and external noise refers to random between-trial variability in drift rate, where "random variability" is variability that cannot be modelled deterministically and instead must be modelled as a random variable from a probability distribution. Ratcliff et al. (2018) conducted five different *double-pass* experiments, where a double-pass meant that the exact same stimulus was presented on two different trials of the experiment. Ratcliff et al. (2018) assessed the level of agreement between these identical trials as well as the accuracy over the entire experiment, and tested whether a diffusion model (Ratcliff, 1978) could explain these data. Their findings indicated that only a diffusion model with a non-zero η parameter – the standard deviation of a normal distribution for between-trial variability in drift rate – could generate the trends observed in agreement and accuracy. Based on these findings, Ratcliff et al. (2018) claimed to "provide direct evidence that external noise is, in fact, required to explain the data from five simple two-choice decision tasks" (p. 33), which provided validation for the random between-trial

variability parameter in drift rate within the diffusion model (i.e., η) proposed by Ratcliff (1978).

There are several key points made by Ratcliff et al. (2018) that we agree with. First and foremost, for reasons that will become clear later, we agree that Ratcliff et al. (2018) provided evidence that drift rate varies – in at least some manner – between trials in an experiment. This stands in contrast to recent neuroscientific proposals that drift rate remains identical across decisions in an experiment (e.g., O'Connell, Shadlen, Wong-Lin, & Kelly, 2018; Ditterich, 2006a, 2006b; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Churchland, Kiani, & Shadlen, 2008). We also agree that understanding which sources of random variability are necessary to explain empirical data is an important question for all fields that use computational models, which has been discussed for several other classes of models within the decision-making literature (Regenwetter & Robinson, 2017, 2019; Kellen, Klauer, & Singmann, 2012, 2013; Bhatia & Loomes, 2017). For the diffusion model, these questions involve determining whether or not random between-trial variability parameters are necessary for explaining empirical trends in choice and response time data, and whether they are useful for improving our understanding of decision-making (see Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002 for proposals of random between-trial variability in different parts of the process; though see van Ravenzwaaij & Oberauer, 2009; van Ravenzwaaij, Donkin, & Vandekerckhove, 2017 for issues regarding parameter recovery and statistical power). However, we believe it is important to distinguish between the multiple types of between-trial variability, and that the lack of distinction in Ratcliff et al. (2018) led to misleading conclusions. We argue that although Ratcliff et al. (2018) did show that drift rate varies between trials, they did not provide evidence that it was due to "external noise", or in other words, random between-trial variability.

Our article aims to address three key points relating to the study of Ratcliff et al. (2018). Firstly, we attempt to clarify the exact definition of between-trial variability, and argue that two separate factors could cause this change: random factors (i.e., noise) and systematic factors (i.e., effects of specific variables). Secondly, we question whether these potential sources of variability can be distinguished in the double-pass paradigm, and what conclusions can actually be drawn from the findings of Ratcliff et al. (2018). We also explore a range of additional analyses and an alternative paradigm, though each appears to fail at distinguishing between systematic and random sources of variability in simulations. Lastly, we provide a discussion about what "random" variability means, when we should attempt to model variability as systematic vs. random, what implications these decisions have for cognitive modelling and our understanding of decision-making, and how researchers might go about replacing random variability parameters with systematic explanations in future research.

What Factors Make Up "Variability"?

Ratcliff et al. (2018) showed that simulated data generated from the diffusion model without between-trial variability in drift rate (i.e., $\eta = 0$) was qualitatively inconsistent with their empirical data. However, this conclusion is not novel, with Ratcliff (1978) showing that drift-rate variability in the diffusion model predicts error response times to be slower than correct response times, which is often observed in empirical data (though see O'Connell et al., 2018, who suggest that this trend is also predicted by thresholds that decrease over the course of a trial). The novel claim of Ratcliff et al. (2018) was that "external noise" was necessary to explain the data from the double-pass experiments. They assumed that *any* between-trial variability is evidence for the presence of external noise. However, is external noise the only factor underlying between-trial variability?

Initially, Ratcliff et al. (2018) suggested that they aimed to assess "whether the external noise, or variation between stimulus exemplars" was required to explain their data, which implies there are multiple factors that could contribute to drift rate varying between trials. However, from that point onwards, Ratcliff et al. (2018) use the words "between-trial variability" and "external noise" interchangeably, suggesting that any type of between-trial variability reflects external noise. We believe that Ratcliff et al. (2018) conflated two different sources of potential between-trial variability throughout most of their article: *random* sources of variability and *systematic* sources of variability. In general, "variability" means that something changes from one moment to the next, and in the context of between-trial variability in EAMs, from one trial to the next. For example, if the drift rate on trial N is different to the drift rate on trial N+1 or N+2, then there is between-trial variability in drift rate. However, the reason that drift rate changes from one trial to the next could be due to a systematic source or a random source of variability.

Systematic variability is caused by factors that are known, such as experimental manipulations, and these factors could be explicitly modelled with different drift rates across the levels of the factor. For example, an experimenter may make some stimuli in a task more difficult than others. Although this manipulation would result in a difference in drift rate between the items of different difficulty, the differences would not be considered "random" or "noisy", but would instead be measurable, systematic differences. In such a case, researchers would expect the drift rate to systematically decrease as task difficulty increased. In contrast, random variability is caused by factors that are unknown or are known but not easily modelled. For example, rather than try to explicitly model a participant's fluctuations in attention or mind wandering from trial-to-trial, a researcher would model the associated fluctuations in drift rate as draws from a probability distribution, such as the normal distribution (Ratcliff, 1978).

Although systematic and random sources of variability are theoretically different, they can be easy to conflate, as systematic factors are often modelled as random factors out of convenience. For example, in situations where there are a large number of factors and data are relatively sparse, attempting to model all factors may compromise the properties of the model (e.g., generalizability and identifiability; see our Appendix C for an example). In addition, these types of variability are not mutually exclusive (i.e., both can occur in a given task), making them even easier to conflate.

However, based on our definitions, and assuming that process models are intended to provide explanations of cognitive phenomena, we believe that these sources of variability should *not* be conflated. Attributing the variability to random sources provides *less* of an explanation than modelling the variability as a function of systematic sources. Including a random source of variability acknowledges that variability occurs, and assumptions about the probability distribution of the random variables can even be guided by theory. However, apart from potential distributional assumptions, attributing the variability to random sources provides no precise explanation of how and why the variability occurs. In contrast, attempting to model the variability as a function of a systematic factor provides an explicit, precise explanation of how and why the variability occurs, which can be directly compared to other explanations of why the variability occurs. Therefore, systematic sources of variability are more theoretically meaningful, and do more to further our understanding of the cognitive processes that we aim to explain. Ratcliff et al. (2018) found that variability in drift rate between trials is necessary, but this is not sufficient to attribute the variability to external noise. Their findings failed to show, as they claimed, "direct evidence that external noise is, in fact, required", because their results could have been due to systematic variability in drift rate, which we argue is more theoretically meaningful, and should be separated from random variability.

The Double-Pass Paradigm Does Not Allow Systematic and Random Sources of Variability To Be Distinguished

As mentioned earlier, Ratcliff et al. (2018) provided evidence for variability in drift rate between trials through an elegant "double-pass" paradigm (Green, 1964; Burgess & Colborne, 1988; Gold, Bennett, & Sekuler, 1999; Lu & Dosher, 2008; Cabrera, Lu, & Dosher, 2015). The double-pass paradigm involves a participant viewing a large number of unique stimuli, and having each unique stimuli repeated once at a different point in the experiment. Specifically, the five experiments of Ratcliff et al. (2018) each contained 8-9 unique blocks of 90-96 stimuli (differing between experiments), where each unique block was repeated once. This allowed Ratcliff et al. (2018) to assess the "agreement" in the responses made on identical trials (i.e., how often the same response was made to the same stimulus), which was used in combination with the overall accuracy to infer whether there was variability between trials in drift rate.

Although Ratcliff et al. (2018) provided details on how they simulated from the diffusion model, they provided no formal definitions of how the standard, univariate (i.e., single unique stimulus) diffusion model extends to the double-pass paradigm. However, based on the descriptions of their simulation, we can infer that they defined the relationship between the drift rates of the double-pass trials as a bivariate normal distribution. Formally, this can be written as:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & \rho_v \\ \rho_v & 1 \end{bmatrix} \right) \tag{1}$$

where v_i is the drift rate for a specific stimulus i, $v_{i,1}$ is the drift rate on the first presentation of the stimulus (i.e., the first "pass") and $v_{i,2}$ is the drift rate on the second presentation,

"$\sim$" means "distributed as", BN is the bivariate normal distribution, μ_v is the mean drift rate for all trials, η_v^2 is the between-trial variance in drift rate, and ρ_v is the correlation in drift rate between double-pass trials. However, it should be noted that other formal definitions could be used to represent the extension of the diffusion model to the double-pass paradigm, such as:

$$v_{i,j} \sim \mathrm{N}(\mu_{v_i}, \eta_v^2) \tag{2}$$

where $v_{i,j}$ is the drift rate for a specific stimulus i on presentation (i.e., "pass") j, μ_{v_i} is the mean drift rate for specific stimulus i, and N is the univariate normal distribution. Under this definition ρ_v is no longer required, as the relationship between the two presentations is reflected in the mean drift rate for each specific stimulus. However, throughout our article we choose to use the formal definition in Equation 1, as we believe this provides the clearest formalization for 1) understanding the analyses performed by Ratcliff et al. (2018), and 2) contrasting systematic and random sources of variability. Using our definition in Equation 1, if there were no between-trial variability in drift rate, then every trial (regardless of whether the stimulus was identical or not) would have an identical drift rate. Formally, this would involve setting η_v^2 to 0, meaning that the bivariate normal definition could be simplified to:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} = \begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix} \tag{3}$$

as now the drift rate of every trial is simply the bivariate normal mean.

If there were only a single source of variability in drift rate between trials, which was

systematically based on the precise identity of the stimulus, then the two presentations of each unique stimulus would have identical drift rates, though this drift rate would differ from other trials with different unique stimuli. Formally, this would involve setting ρ_v to 1, meaning that the bivariate normal definition could be simplified to:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right) \tag{4}$$

as the drift rate on each trial can be represented as a deterministic function of the stimulus presented (i.e., $v_i = f(i)$).

If there were only random sources of variability in drift rate between trials (or systematic sources that were not related to the stimulus presented), then the drift rates on each trial would be independent of one another. Formally, this would involve setting ρ_v to 0, meaning that the bivariate normal definition could be simplified to:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \tag{5}$$

as the drift rate on each trial is an independent random draw from the bivariate normal distribution. This model represents pure random between-trial variability, and is the definition of between-trial variability used within the standard univariate diffusion model (i.e., $v_{i,j} \sim \text{N}(\mu_v, \eta_v^2)$).

Based on these formal definitions, when $\eta_v^2 = 0$ there is no variability between trials in drift rate. When $\eta_v^2 > 0$, then the source of variability is determined by ρ_v: $\rho_v = 0$ means the variability is all random, $\rho_v = 1$ mean the variability is all systematically based

on the stimulus, and $0 < \rho_v < 1$ mean both sources of variability exists. Therefore, for the bivariate normal drift rate distribution in the double-pass paradigm, the η_v^2 parameter determines whether or not between-trial variability in drift rate exists, and the ρ_v parameter that determines what type of between-trial variability exists. Note that this differs from the univariate normal drift rate distribution in standard paradigms – as shown in our definitions above – where the η_v^2 parameter determines whether or not random between-trial variability exists.

Based on these formal definitions of systematic and random variability, what did the analysis of Ratcliff et al. (2018) assess? As discussed above, Ratcliff et al. (2018) calculated the agreement (how often participants made the same response to the two identical double-pass trials) and accuracy (how often participants made the correct response across the entire experiment) from the empirical data, and compared these to what would be predicted by the diffusion model, with different potential parameters for μ_v and η_v^2. Specifically, Ratcliff et al. (2018) found that $\eta_v^2 = 0$ made a strong prediction about these agreement-accuracy functions, where the agreement would always be lower for a fixed level of accuracy than when $\eta_v^2 > 0$, unless the mean drift rate was extremely high and both variables began to asymptote (see Figure 1, the bottom-right panel). Importantly, this trend was not supported by the empirical data, which were consistent with non-zero values of η_v^2, resulting in a claim that there was external noise (i.e., random variability) between trials in drift rate in their tasks. However, according to our definitions above, a non-zero η_v^2 only determines that *some* amount of between-trial variability exists, and the value of ρ_v determines whether it is random, systematic, or both. Importantly, Ratcliff et al. (2018) simulated their data with identical drift rates for the double-pass trials with identical stimuli, which is formally equivalent to $\rho_v = 1$. Therefore, Ratcliff et al. (2018) did not attempt to distinguish between the type of between-trial variability, which is of crucial importance for their main

claim.

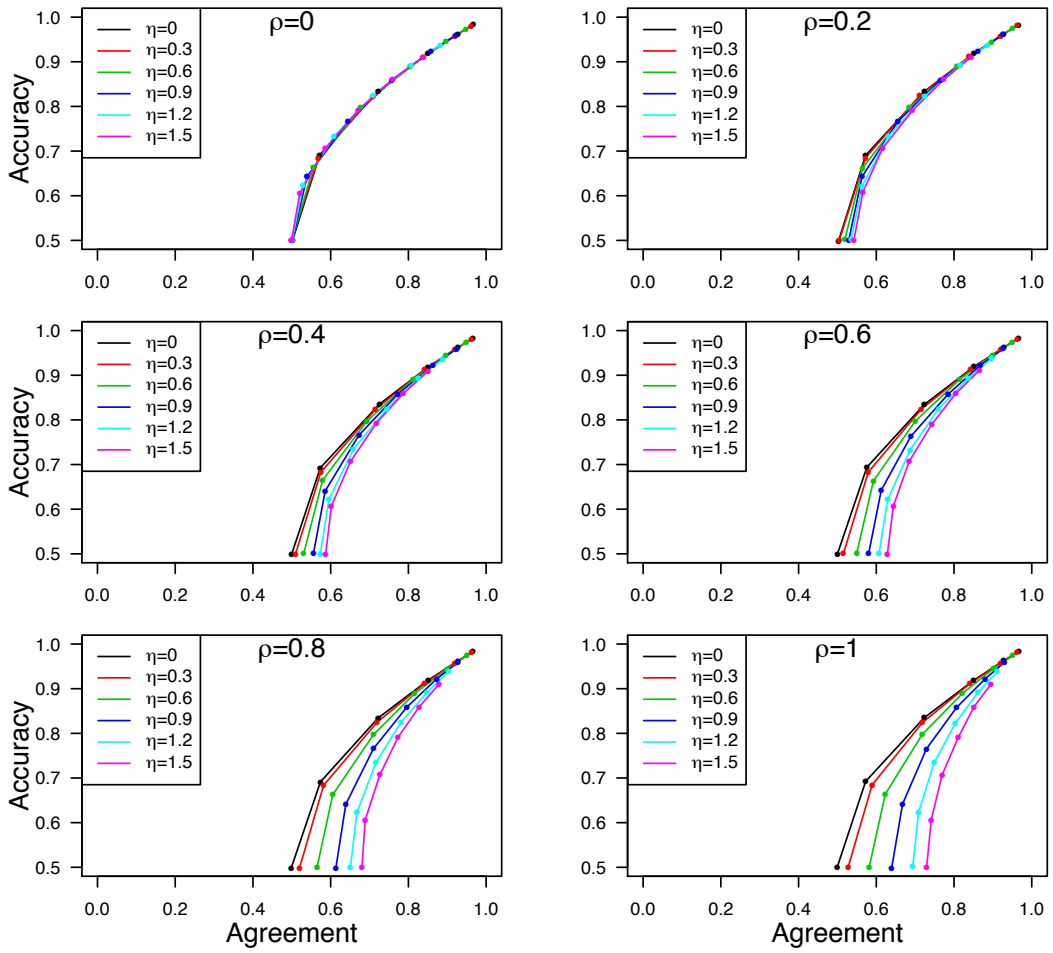

Figure 1. The agreement (x-axis) and accuracy (y-axis) functions for different mean drift rates (points on each line), variability in drift rates (different lines) and correlation in drift rates for identical items (different panels). The bottom-right panel displays the predicted functions when the correlation is fixed to 1, as in Ratcliff et al. (2018). The different mean drift rates used to generate the different points on each line were 0 (lowest accuracy), 0.5, 1, 1.5, 2, and 2.5 (highest accuracy). These simulations used the following fixed parameters: a (threshold) = 1.5; z (starting point) = 0.75 (i.e., unbiased starting evidence); ter (non-decision time) = 0.3; s (the diffusion coefficient) = 1. All simulations used the method of Evans (2019).

Although we believe the analysis of Ratcliff et al. (2018) failed to distinguish between systematic and random sources of between-trial variability in drift rate – as they fixed ρ_v to 1 in all simulations – their general analysis method might still be able to distinguish between these sources of variability by performing simulations where ρ_v is allowed to take on different values. However, our simulations (Figure 1) appear to suggest that this is not the case. The bottom right panel displays simulations with $\rho_v = 1$, as in the simulations of Ratcliff et al. (2018). Interestingly, the strong prediction shown by Ratcliff et al. (2018) for $\eta_v^2 = 0$ (where the agreement would always be lower for a fixed level of accuracy than when $\eta_v^2 > 0$), which was inconsistent with the empirical data, also occurs in one other situation: when $\rho_v = 0$. As shown in the top-left panel of Figure 1, when $\rho_v = 0$ the agreement-accuracy function all resemble that of $\eta_v^2 = 0$, with small agreement relative to accuracy, regardless of the actual value of $\eta_v^2 = 0$. Therefore, two conclusions follow from the data and analysis method of Ratcliff et al. (2018): that between-trial variability exists in drift-rate (i.e., $\eta_v^2 > 0$), and that *at least some* of this variability is due to the systematic source of item (i.e., $\rho_v > 0$; the between-trial variability cannot be purely "noise"). However, when both η_v^2 and ρ_v are non-zero, the parameters appear to have the same qualitative impact on the agreement-accuracy functions, where increasing either parameter increases the level of agreement. For example, Figure 2 (left panel) shows four different agreement-accuracy functions, which look very similar to one another – in fact, each form pairs that appear to overlap perfectly. However, these functions were generated with four very different combinations of η_v^2 and ρ_v values, with η_v^2 ranging from 0.6 (green line) to 1.5 (purple line), and ρ_v ranging from 0.4 (purple line) to 1 (green line). Importantly, this means that we are unable to determine from the agreement-accuracy functions whether the between-trial variability is purely systematic (i.e., $\rho_v = 1$) or a mixture of systematic and random ($0 < \rho_v < 1$). Therefore, it appears virtually impossible to determine whether random

sources of between-trial variability (i.e., "external noise") exist by using the agreement and accuracy of a double-pass paradigm, which is in direct contrast to the claim made by Ratcliff et al. (2018).

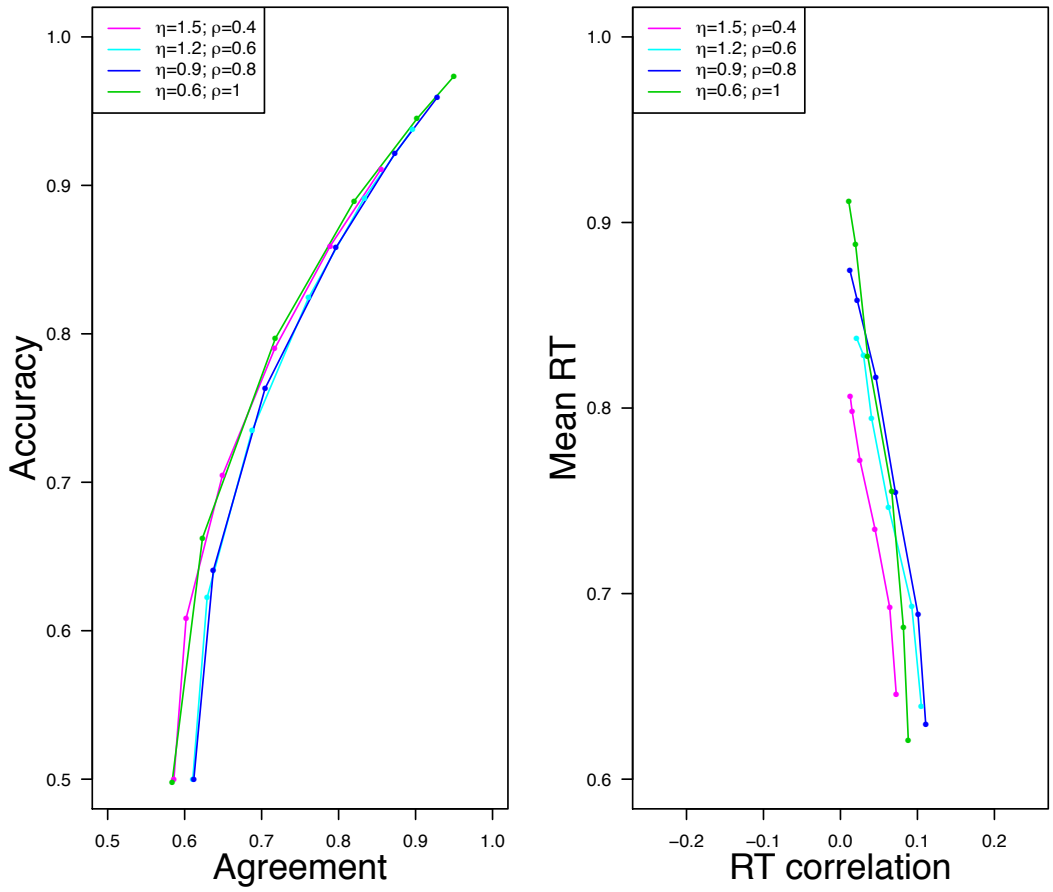

Figure 2. **Left:** The agreement (x-axis) and accuracy (y-axis) functions for different mean drift rates (points on each line), for four selected different combinations of variability and correlation that closely mimic one another, suggesting that the assessment of these summary statistics is of limited value. **Right:** The correlation in (x-axis) and mean (y-axis) response time functions for different mean drift rates (points on each line), for the same for combinations. Although there appears to be one function ($\eta = 1.5$, $\rho = 0.4$) clearly distinguished from the others, this area of the parameter space was still unidentifiable (see Appendix A).

Are there other ways to make inferences about systematic and random sources of variability?

We also tested two other potential analysis methods through simulation, which we believed could have allowed an assessment of whether or not ρ_v was less than 1. Firstly, although agreement and accuracy were unable to identify the values of η_v^2 and ρ_v (other than that they were non-zero), the equivalent analyses with response time may be able to distinguish their predictions. The "agreement" (i.e., correlation) between the log response time of the identical double-pass trials, and the overall mean response time, is shown in Figure 3. Unfortunately, this analysis seems to suffer from the same issue as its choice-based counterpart: increases in the response time correlation can be created by either increasing η_v^2, or increasing ρ_v.

However, it is also possible that the joint information from both of these types of data may provide adequate constraint to separately identify η_v^2 and ρ_v; something that is difficult to accurately assess from the separate functions. For example, in Figure 2 (right panel), it appears that the functions do not perfectly overlap in all cases, though the imperfect overlap may be due to poor hand-tuning of parameter values. To more rigorously assess the identifiability of η_v^2 and ρ_v when constrained to jointly account for response agreement, response accuracy, response time correlation, and response time mean, we fit the bivariate diffusion model to these variables through pseudo-likelihood methods (e.g., Turner & Sederberg, 2014; Holmes, 2015), where the simulated response choices and response times from the bivariate diffusion model provided the parameters for (1) a multinomial distribution governing the response choice combinations across both presentations, and (2) a bivariate normal distribution governing the natural logarithm of the response times across both presentations (see Appendix A for a more detailed description and formal definition of the analysis). However, the parameters remained unidentifiable (see Appendix A), and formal comparison using the Bayesian Information Criterion (BIC; Schwarz, 1978) showed

evidence in favour of a model with ρ_v fixed at 1 over a model with a freely estimated ρ_v parameters, as both models produced near-identical likelihoods with very different parameter sets, meaning that ρ_v fixed at 1 model was slightly preferred based on parsimony (see Appendix B). It should also be noted that these analyses appear to show that other parameters may also be involved in the tradeoff, such as μ_v, which is incorrectly and inconsistency estimated in all cases.

Although these problematic tradeoffs could theoretically be solved by first fitting the standard univariate diffusion model to estimate the μ_v and η_v^2 parameters, and then performing the agreement-accuracy analyses with the μ_v and η_v^2 values constrained, the robustness of this assessment would be completely reliant on the correct point estimate being found for η_v^2, which is difficult to reliably obtain (Ratcliff & Tuerlinckx, 2002; Lerche & Voss, 2016) regardless of method (Boehm et al., 2018). Overall, these analyses suggest that summary statistics contain inadequate information for inferences on the parameters of the drift rate distribution.

Secondly, we attempted to fit a diffusion model with a different drift rate estimated for each trial, and with these drift rates constrained to follow the bivariate normal distribution defined in Equation 1 (i.e., a hierarchical model). Although it seems unlikely that the drift rate used to generate each trial could be recovered, it is possible that the small pieces of information about the overall drift rate distribution contained within each trial would allow the bivariate normal parameters, and most importantly ρ_v, to be recovered. However, we were unable to recover the values of ρ_v (the full details can be seen in Appendix C), suggesting that there are pragmatic difficulties in making inferences about random sources of between-trial variability with the double-pass paradigm.

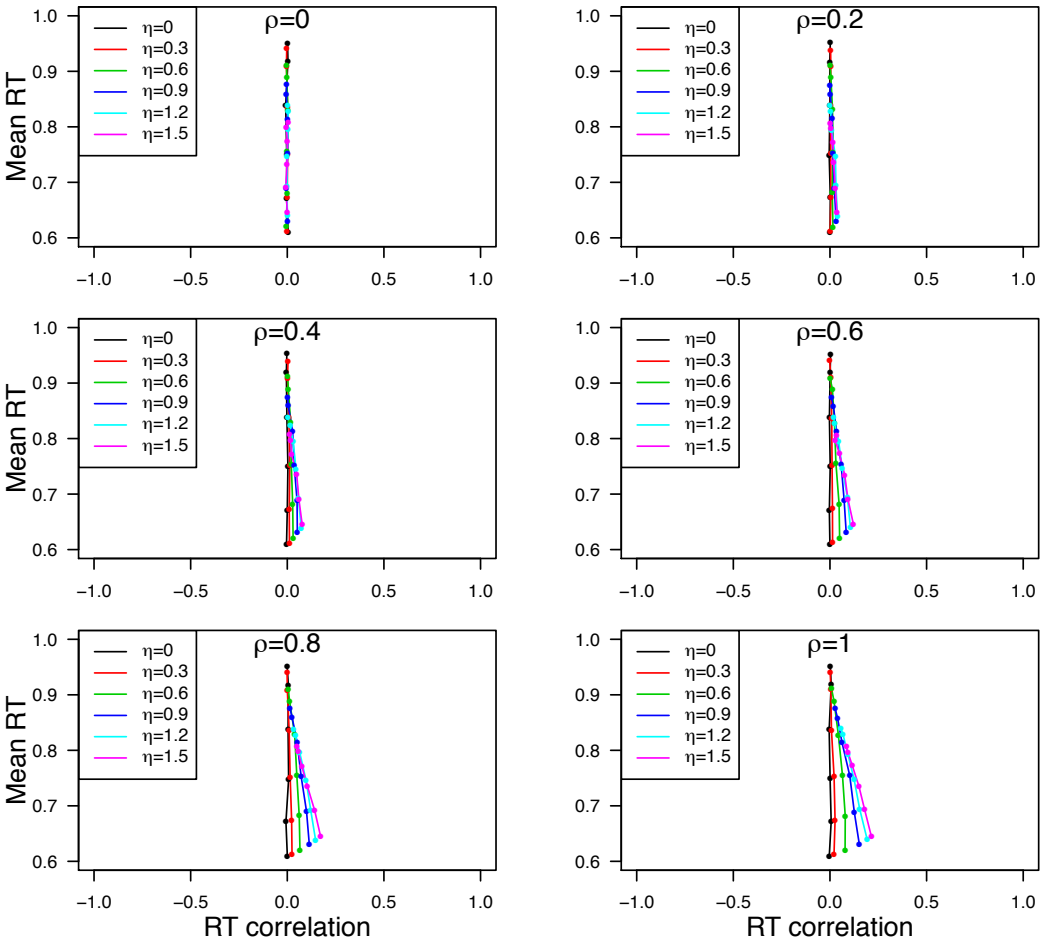

Figure 3. The correlation in (x-axis) and mean (y-axis) response time functions for different mean drift rates (points on each line), variability in drift rates (different lines) and correlation in drift rates for identical items (different panels).

As discussed previously, Ratcliff et al. (2018) used simulations where only *systematical* variability was present ($\rho = 1$) to argue for the presence of *random* variability ($\rho < 1$). Based on our definitions and analyses, it appears that Ratcliff et al. (2018) could not logically conclude in favour of the existence of random between-trial variability using

the double-pass paradigm that they implemented. However, we also considered whether the concept of the double-pass experiment could still be useful for distinguishing between systematic and random variability in the form of a "multi-pass" extension (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008; Cabrera et al., 2015). For example, the double-pass concept could be integrated into a standard experimental paradigm, where the majority of stimuli presented are standard trials, which are not repeated, and mostly serve as "filler" trials. However, some specified, relatively small (e.g., 5 or 6) number of stimuli are repeated a large number of times (e.g., 30+), in order to allow adequate estimation of drift rates for these specific stimuli. Within this paradigm, the different potential sources of between-trial variability in drift rate could be represented in four discrete models: a no-variability model, where the drift rate of all trials is fixed to the same value; a random-only variability model, which is identical to the no-variability model, but with the Ratcliff (1978) η parameter added to the model; a systematic-only variability model, which is identical to the no-variability model, but with different drift rates for each of the different multi-pass stimuli (i.e.., different identical stimuli have different identical drift rates); and a systematic-and-random variability model, which is identical to the systematic-only variability model, but with the Ratcliff (1978) η parameter added to the model. These models could then be compared using formal model comparison (e.g., Bayes factors; Gronau et al., 2017; Evans & Brown, 2018; Annis, Evans, Miller, & Palmeri, 2019; Evans & Annis, 2019; see Evans, Bennett, & Brown, 2018 for an application of Bayes factors using the diffusion model) to determine the sources of between-trial variability in drift rate.

However, the multi-pass approach has two key limitations in identifying sources of between-trial variability in drift rate. Firstly, although this approach breaks down the previous parameter identification problem into a comparison between four discrete models, these models still may not be recoverable; that is, even when the true data generating model

is one of these four models, the correct model may not be able to be identified. Indeed, our recovery simulation (see Appendix D) found that even with 6 multi-pass stimuli that each had 50 trials (i.e., 300 trials dedicated to non-filler trials), the Bayes factor between the true data generating model and the closest competing model was often small (i.e., the evidence was ambiguous), and the true generating model was not always correctly identified. This situation also becomes more complicated in situations with multiple participants, as these item effects should not be thought of as *item random-effects* (e.g., in the context of mixed-effect modelling methods), meaning that hierarchical modelling techniques are required to capture differences between participants in item effects, with hierarchical modelling often increasing the approximation error in Bayes factor estimation (Annis et al., 2019; Evans & Annis, 2019). Therefore, the multi-pass paradigm may not provide a robust solution to the limitations of the double-pass paradigm, especially within a reasonable number of trials. Finally, and perhaps most importantly, there may generally be limited usefulness in experiments aimed at testing the existence of "random" variability. As we argue in the next section, the concept of "noise" within cognitive models merely serves as a convenience parameter for sources of variability that we know exist, but are unable to account for.

What Is "Random" Variability, and When Is It Useful?

Although Ratcliff et al. (2018) emphasized the importance of "external noise" (i.e., random variability), there was little discussion regarding what exactly is meant by "noise", whether the concept is general or context dependent, or what the exact purpose of noise is. We believe that "noise" (or "randomness"), especially in the case of psychological models, simply serves as a convenient parameter that allows the absorption of variability from sources that are difficult to explicitly incorporate into the model. These sources might be completely unknown, difficult to measure, or their inclusion might negatively influence

the usefulness of the model (e.g., making the model computationally taxing, reducing generalizability, or causing problems with parameter identifiability). This same concept of "randomness" is used within statistical models, where the "noise" (i.e., ϵ) is simply the unexplained variance within the model (De Boeck & Wilson, 2004).

Given the definition above, the concept of "noise" appears to be highly context dependent, rather than a general phenomenon that does/does not exist. Specifically, we believe that this random variability is highly dependent both on the specific experiment used, as different experiments may have different sources and amounts of variability, and the model defined, as with complete information about all sources of variability the randomness could be completely eliminated. For example, in experimental designs where there are greater chances of variability being caused by factors that are difficult to measure, such as attention (e.g., online experiments), there is greater chances of "random" variability occurring. In addition, when factors that are known to cause variability are ignored and not explicitly included within the model, such as different items, then there is likely to be random variability. Therefore, we are uncertain how useful general questions about the existence of random variability in specific parameters – such as those posed by Ratcliff et al. (2018) – actually are, as these question may not be answerable at this general level, and instead may be completely dependent on each experimental data set and model defined.

However, we also believe that the incorporation of random variability within models can be extremely useful, regardless of whether the variability is truly "noise", or just unaccounted for as we suggest above. Previous research has shown that cognitive models – most notably the diffusion model – are able to capture key patterns of data with the incorporation of random variability parameters, without the difficulty of having to try and identify the many potential sources of systematic variability (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). In fact, a similar logic was used by Ratcliff

(1978) to justify to incorporation of random between-trial variability in drift rate (i.e., the "external noise" discussed within this paper) into the diffusion model, suggesting that different items within a memory task have different drift rates, though attempting to model each of these different drift rates may prove impossible. Although it is clear that item effects will causes trial-to-trial variability in drift rate, and that this variability is not from a "random" source as theoretically each item could be modelled with a different drift rate, the use of the random variability parameter allows the model to still provide a good account of the data while avoiding the difficulty of modelling the drift rate for each item.

Having said this, adding too much unnecessary random variability into cognitive models can dilute our understanding of the cognitive processes they aim to explain. As mentioned previously, we believe that the goal of cognitive modelling should be to understand as many causes of variability as possible, meaning that we should be attempting to find factors that causes systematic variability in the parameters within our models, and explicitly including them within our model definitions. Merely labelling all variability as random and including random between-trial variability for all parameters does not allow us to gain any additional understanding of the underlying process causing these variabilities, and in some cases can result in models becoming unfalsifiable (Jones & Dzhafarov, 2014); though it should be noted that practically these models can remain quite constrained with reasonable distributional assumptions (Ratcliff, 2002; Smith, Ratcliff, & McKoon, 2014; Heathcote, Wagenmakers, & Brown, 2014). Furthermore, in the separate issue of using cognitive models as *measurement tools* for estimating latent parameters, recent research has drawn into question the measurement properties of the diffusion model with variability parameters, as the addition of these variability parameters can lead to poorer parameter recovery (Lerche & Voss, 2016) and lower power in detecting effects of interest (van Ravenzwaaij et al., 2017). Therefore, although the random between-trial variability parameters

within EAMs can serve as a convenient placeholders, which allow the models to account for phenomena that it otherwise could not, future research should aim to identify these sources of between-trial variability, and explicitly model them.

How can we replace random distributions with systematic explanations?

Throughout our article we have argued for the importance of identifying and modelling systematic sources of variability, as they provide more complete explanations of psychological processes than attributing the variability to random draws from a probability distribution. However, random variability is a fundamental basis of several cognitive models (though see Regenwetter & Robinson, 2017, 2019; Kellen et al., 2012, 2013; Bhatia & Loomes, 2017 for assessments of which sources of variability are necessary in different cognitive models) – particularly the "full" diffusion model – and replacing these random distributions with systematic explanations may seem like a difficult task to researchers. Here, we attempt to provide some recommendations on *how* researchers could begin identifying and modelling systematic sources of variability, which may be able to replace some of the random sources of variability contained within some models.

One method for identifying different sources of variability, or comparing models that make different proposals about the psychological process underlying the variability, is to use experimental design to tightly constrain the predictions of the models (e.g., Ludwig & Davies, 2011; Teodorescu & Usher, 2013; Servant, White, Montagnini, & Burle, 2015; Evans, Dutilh, Wagenmakers, & van der Maas, 2019). This method is in line with the efforts of Ratcliff et al. (2018), though as we argue within the current article, their experimental design did not provide adequate constraint to test their claims. However, the double-pass paradigm can be combined with other experimental manipulations to provide valuable insight about how evidence accumulation may operate, and types of noise may be present within drift rate. For example, Ludwig and Davies (2011) combined the

double-pass paradigm with manipulations of stimulus strength and the time of a response signal, which allowed them to discriminate between different observer models that proposed different processes for how accumulated evidence changes over time. Additionally, recent studies have attempted to constrain EAMs with additional sources of data instead of experimental manipulations. For example, Evans et al. (2019) was able to compare a range of different EAM variants, which strongly mimic one another in choice response time data, by constraining the models to also account for the secondary responses made by participants (something they termed "double responding"). Another example is the study of Servant et al. (2015), who provided further constraint on the diffusion model by making it account for electromyography (EMG) recordings (specifically, partial EMG bursts). These studies showcase that further insight into systematic sources of variability can be provided by further constraining models through either experimental manipulations, additional behavioural data, or psychophysiological data.

Another method for integrating systematic sources of variability into existing models is the use of *front-end* models (e.g., Nosofsky & Palmeri, 1997; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2004; Brown et al., 2008; Purcell et al., 2010; Ratcliff & Starns, 2013; Trueblood, Brown, & Heathcote, 2014; Servant, Tillman, Schall, Logan, & Palmeri, 2019). In the context of EAMs, front-end models directly constrain the drift rate in each experimental condition, where a *front-end* function transforms the stimulus information into a drift rate, which is then fed into the *back-end* EAM. Front-end models have been a useful method for creating detailed models of specific tasks, as researchers only need to design the task-specific function to transform the stimulus information into drift rates, rather than building a complete model from the ground up. For example, in the area of multi-attribute choice, where choices are made between alternatives that each have different values of the same explicit attributes, several of the dominant models are direct

front-end extensions of common EAMs: multi-alternative decision field theory (MDFT; Roe et al., 2001) is a front-end extension of the Ornstein-Uhlenbeck process (Busemeyer & Townsend, 1992, 1993), the multi-attribute leaky competing accumulator (mLCA; Usher & McClelland, 2004) is a front-end extension of the LCA (Usher & McClelland, 2001), and the multi-attribute linear ballistic accumulator (mLBA; Trueblood et al., 2014) is a front-end extension of the LBA (Brown & Heathcote, 2008). These models each provide a specific account of multi-attribute choice tasks, where the attribute values for all alternatives are transformed into drift rates for each alternative, meaning that a single set of parameter values can generalize to all possible combinations of attribute values. Front-end models can also involve constraining the drift rate to be a function of some source of data that is thought the reflect incoming sensory information. For example, Purcell et al. (2010) placed a neural front-end on an extension of the LCA, constraining the drift rate at each point in time to be a function of the current visual neuron activity, which was measured in non-human primates via single-cell recordings. These studies showcase that systematic explanations of variability in drift rate can be created with front-end models, either as task-specific models with front-end functions that constrain the drift rate based on the stimulus information, or general models with front-end functions that constrain the drift rate based on another source of data thought to reflect incoming sensory information.

One final method for providing systematic explanations for random variability is through creating new models, where the seemingly random variability is a natural by-product of the underlying process. This is probably the most difficult method to practically implement, as it requires a complete model development process, but it can also provide the greatest theoretical insights, as more detailed explanations of the underlying process can be compared. One inspiration for developing several EAMs – and their specific underlying processes – has been the concept of *neural plausibility*: that is, designing models with pro-

cesses that are high-level reflections of actual neural processes (Usher & McClelland, 2001, 2004; Verdonck & Tuerlinckx, 2014). For example, Usher and McClelland (2001) used the concept of neural plausibility to develop the LCA, a model that contains several components that reflect the behaviour of neurons. The LCA contains components such as lateral inhibition, where decision alternatives inhibit the future accumulation of one another based on their current accumulated evidence, leaky accumulation, where accumulated evidence decays over time, and an evidence truncation at 0, where evidence for an alternative cannot be negative. By including these additional, neurally plausible components, the LCA only requires a single sources of random variability, which is within-trial (i.e., "internal noise") variability in drift rate, while still being able to meet several of the key qualitative benchmarks in choice and response time data. Furthermore, previous research has found models with lateral inhibition to out-perform other variants of EAMs when constraining the models with experimental design (Teodorescu & Usher, 2013) or additional data (Evans et al., 2019), suggesting that the development of models through neurally plausible mechanisms can lead to superior explanations. In a more recent example, the Ising decision maker (IDM; Verdonck & Tuerlinckx, 2014) uses the dynamic Ising model to represent pools of binary neurons that feed into an accumulation process, which involves within-pool excitation, between-pool inhibition, and external (i.e., stimulus-based) excitation. Interestingly, the IDM does not require between-trial in drift rate, and the neurally plausible dynamics of the model naturally produce between-trial variability in starting point, meaning that the model replaces one sources of random variability with a systematic explanation, and provides a clear process for how the other occurs. These studies showcase how new models can be developed that replace random distributions with systematic explanations, or have the random distributions naturally fall out of their dynamics, and that neural plausibility can be a useful principle for developing these more explanation-focused models.

Conclusion

Based on a double-pass paradigm for five different tasks, Ratcliff et al. (2018) attempted to assess "whether current models can explain accuracy and RT data with only internal noise or whether the external noise, or variation between stimulus exemplars, is also required" (p.33). From the assessment of agreement-accuracy functions, Ratcliff et al. (2018) claimed that they "provide direct evidence that external noise is, in fact, required to explain the data from five simple two-choice decision tasks" (p. 33). However, we argue that Ratcliff et al. (2018) conflated two different types of external, between-trial variability – systematic variability and random (i.e., noise) variability – and further show that their analysis methods were insufficient to determine the existence of random between-trial variability in drift rate, suggesting that they failed to provide any evidence for "external noise" in drift rate. Furthermore, we contend that "noise" terms within cognitive models are not necessarily indicative of true "randomness", and that instead noise terms represent variance in the process that exists, but that we cannot currently explain. Therefore, although we agree that noise terms, such as the between-trial variability parameters in the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002), provide a useful placeholder within cognitive models that have greatly aided the ability of the models to explain empirical data trends, we believe that future research should aim to eventually discard these terms, and replace them with actual explanations of the process.

References

Annis, J., Evans, N. J., Miller, B. J., & Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. *Journal of Mathematical Psychology*, *89*, 67–86.

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review*, *124*(5), 678–687.

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the diffusion decision model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75.

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

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

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*, 396–425.

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

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*, 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

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.

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In *Explanatory item response models* (pp. 3–41). Springer.

Ditterich, J. (2006a). Evidence for time-variant decision making. *European Journal of Neuroscience*,

24, 3628–3641.

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

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

Evans, N. J. (2019). A method, framework, and tutorial for efficiently simulating models of decision-making. *Behavior Research Methods*, 1–15.

Evans, N. J., & Annis, J. (2019). Thermodynamic integration via differential evolution: A method for estimating marginal likelihoods. *Behavior Research Methods*, 1–18.

Evans, N. J., Bennett, A. J., & Brown, S. D. (2018). Optimal or not; depends on the task. *Psychonomic Bulletin & Review*, 1–8.

Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, *24*, 597–606.

Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*, 589–603.

Evans, N. J., Brown, S. D., Mewhort, D. J., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*, 592.

Evans, N. J., Dutilh, G., Wagenmakers, E.-J., & van der Maas, H. L. (2019). Double responding: A new constraint for models of speeded decision making. *PsyArXiv*.

Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports*, *7*, 16433.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & Cognition*, *45*, 1193–1205.

Evans, N. J., Steyvers, M., & Brown, S. D. (2018). Modeling the covariance structure of complex datasets using cognitive models: An application to individual differences and the heritability of cognitive ability. *Cognitive Science*, *42*, 1925–1944.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., & von Cramon, D. Y. (2008). Striatum and

pre-sma facilitate decision making under time pressure. *Proceedings of the National Academy of Sciences*, *105*, 17538-17542.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derffuss, J., Imperati, D., & Brown, S. D. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *The Journal of Neuroscience*, *34(47)*, 17242–17249.

Gold, J., Bennett, P., & Sekuler, A. (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.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision–making. *Journal of Neuroscience*, *35*, 2476–2484.

Heathcote, A., Wagenmakers, E.-J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review*, *121*(4), 676–678.

Holmes, W. R. (2015). A practical guide to the probability density approximation (pda) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68*, 13–24.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*, 1–32.

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.

Kellen, D., Klauer, K. C., & Singmann, H. (2013). On the measurement of criterion noise in signal detection theory: Reply to benjamin (2013). *Psychological Review*, *120*(3), 727-730.

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, *7*, 1324.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Logan, G. D. (1992). Shapes of reaction–time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914.

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.

Ludwig, C. J., & Davies, J. R. (2011). Estimating the growth of internal evidence guiding perceptual decisions. *Cognitive Psychology*, *63*(2), 61–92.

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

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K., & Kelly, S. P. (2018). Bridging neural and computational viewpoints on perceptual decision-making. *Trends in Neurosciences*.

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

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143.

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

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*, 278–291.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*, 195ñ-237.

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

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial–to–trial variability of the EEG. *Proceedings of the National Academy of Science*, *106*, 6539–6544.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological*

Science, *9*, 347–356.

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

Ratcliff, R., & Starns, J. J. (2013). Modeling response times, choices, and confidence judgments in decision making. *Psychological Review*, *120*, 697–719.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and aging*, *16*(2), 323.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and iq on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464.

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

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *102*, 261–300.

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

Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*(5), 533–550.

Regenwetter, M., & Robinson, M. M. (2019). The construct-behavior gap revisited: Reply to hertwig and pleskac (2018). *Psychological Review*, *126*(3), 451–454.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multi–alternative decision field theory: A dynamic artificial neural network model of decision–making. *Psychological Review*, *108*, 370–392.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Servant, M., Tillman, G., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2019). Neurally constrained modeling of speed-accuracy tradeoff during visual search: gated accumulation of modulated evidence. *Journal of neurophysiology*, *121*(4), 1300–1314.

Servant, M., White, C., Montagnini, A., & Burle, B. (2015). Using covert response activation to test latent assumptions of formal decision-making models in humans. *Journal of Neuroscience*,

35(28), 10371–10385.

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

Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review*, *121*, 679–688.

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

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*, 239–249.

Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2017). An evidence accumulation model of acoustic cue weighting in vowel perception. *Journal of Phonetics*, *61*, 1–12.

Tillman, G., Osth, A. F., van Ravenzwaaij, D., & Heathcote, A. (2017). A diffusion decision model analysis of evidence variability in the lexical decision task. *Psychonomic bulletin & review*, *24*(6), 1949–1956.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multi-attribute linear ballistic accumulator model of context effects in multi–alternative choice. *Psychological Review*, *121*, 179–205.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review*, *21*(2), 227–250.

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*, 368–384.

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

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*, 757–769.

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

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*, 463–473.

Verdonck, S., & Tuerlinckx, F. (2014). The Ising decision maker: A binary stochastic network for choice response time. *Psychological Review*, *121*(3), 422.

Wagenmakers, E.-J., & Brown, S. D. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830–841.

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*, 140–159.

Appendix A

Parameter Recovery of the Bivariate Diffusion Model Through Jointly Fitting To Response Choice and Response Time

In order to provide a more rigorous assessment (i.e., beyond visual inspection of the functions in Figures 1, 2, and 3) of whether the η_v^2 and ρ_v parameters are identifiable when jointly constrained by the choice agreement-accuracy and response time correlation-mean functions, we estimated the bivariate diffusion model parameters by directly fitting the model to these functions. Specifically, these fits involved the use of pseudo-likelihoods (e.g., Turner & Sederberg, 2014; Holmes, 2015), where the likelihood function for response choice and response time were created using simulations from the bivariate diffusion model. For simplicity, we assumed that response choice combinations across both presentations were governed by a multinomial distribution, which can be formally expressed as:

$$N_{c,c}, N_{c,e}, N_{e,c}, N_{e,e} \quad \sim \quad M(p_{c,c}, p_{c,e}, p_{e,c}, p_{e,e}) \tag{A1}$$

where M is the multinomial distribution, N is the number of response combinations that fall into the category, the subscripts c and e reflect correct and error responses respectively, the first subscript refers to the response for the first presentation, and the second subscript refers to the response for the second presentation (e.g., $N_{c,e}$ in the number of trials where a correct response was given on the first presentation, and an error response was given on the second presentation). The p parameters of the multinomial distribution are the probability of these events occurring, and were obtained through simulating the bivariate diffusion model. We also assumed that the natural logarithm of response time (as response time distributions are typically positively skewed) across both presentations were governed

by a bivariate normal distribution, which can be formally expressed as:

$$\begin{bmatrix} t_{i,1} \\ t_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_t \\ \mu_t \end{bmatrix}, \eta_t^2 \begin{bmatrix} 1 & \rho_t \\ \rho_t & 1 \end{bmatrix} \right) \tag{A2}$$

where $t_{i,j}$ is the natural logarithm of the response time for item i on presentation j. The μ_t, η_t^2, and ρ_t parameters of the bivariate normal distribution are the mean, variance, and correlation of the natural logarithm of response times, and were obtained through simulating the bivariate diffusion model. We estimated 5 parameters from the bivariate diffusion model – μ_v, η_v^2, ρ_v, a (threshold), and ter (non-decision time) – with the starting point (z) fixed to $\frac{a}{2}$, in order to try and make the model as simple as possible to maximize the chances of recovery.

We performed three parameters recoveries with three different generating values of ρ_v: 0, 0.5, and 1. The synthetic data generated for each recovery contained 1,000 pairs of simulated trials, each generated using the following parameters: $a = 1.5$, $ter = 0.3$, $\mu_v = 3$, $\eta_v^2 = 0.5$. We fit the models with Differential Evolution Markov chain Monte Carlo (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, 2013) with $3k$ (where k is equal to the number of free parameters in the model) parallel chains and 1,000 iterations of sampling, where the maximum likelihood estimate was the sampled parameter set with the highest likelihood. We also used two estimation runs per model to 1) help ensure that we had reached the global maxima, and 2) assess whether the estimated parameters appeared to be consistent, or alternatively, show signs of unidentifiable trade-offs.

The results of the recovery can be seen in Table A1. In all cases the generated ρ_v values are not recovered, and in several cases the estimated values greatly vary from the generated values. Furthermore, although the maximum log-likelihood from each estimation

run is near-identical, the estimated parameter values differ substantially, suggesting that 1) both η_v^2 and ρ_v are unidentifiable from the joint agreement-accuracy and correlation-mean functions, and 2) other parameters – most noticeably μ_v, but also a to some extent – also appear to be involved in this trade-off, varying greatly between estimation runs.

Table A1: Displays the parameter values and log-likelihood for the maximum likelihood estimate of the first recovery assessment in Appendix A. "Data" refers to the data set, defined by the generating ρ_v value, and "Run" refers to the estimation run (either the first or second of two total runs).

Data	Run	a	ter	μ_v	η_v^2	ρ_v	$log[p(D\|\theta)]$
0	1	1.49	0.31	3.53	1.21	0.08	-279.21
0	2	1.36	0.3	2.84	0.44	0.2	-279.35
0.5	1	1.44	0.31	3.42	1.1	0.34	-245.83
0.5	2	1.47	0.3	3.24	0.9	0.41	-245.67
1	1	1.5	0.31	3.72	1.32	0.53	-228.54
1	2	1.4	0.31	3.22	0.87	0.47	-228.3

One limitation of our previous parameter recovery is that it only used a single set of parameter values (outside of varying ρ_v), and it could be argued that the parameters may be identifiable in different regions of the parameter space. Specifically, in Figure 2 one function appears as though it may be distinguishable from the others when jointly assessing the agreement-accuracy and correlation-mean functions (μ_v = [0, 0.5, 1, 1.5, 2, 2.5]; $\eta_v^2 = 1.5^2$; $\rho_v = 0.4$), suggesting that this may be a region of the parameter space where the η_v^2 and ρ_v parameters could be identifiable. To ensure that the unidentifiability of η_v^2 and ρ_v still held for this region of the parameter space, we generated 6 data sets with these parameter values, each using a different μ_v parameter from the function. The

results of this recovery can be seen in Table A2. As in the previous recovery, the generated ρ_v values are not recovered, and the estimated parameter values different substantially between estimation runs despite the near-identical maximum log-likelihoods. This further suggests that the η_v^2 and ρ_v parameter values of the bivariate drift rate distribution are not recoverable based on these summary statistics.

Table A2: Displays the parameter values and log-likelihood for the maximum likelihood estimate of the second recovery assessment in Appendix A. "Data" refers to the data set, defined by the generating μ_v value, and "Run" refers to the estimation run (either the first or second of two total runs).

Data	Run	a	ter	μ_v	η_v^2	ρ_v	$p(D\|\theta)$
0	1	1.31	0.29	0.09	1.11	0.49	-1052.38
0	2	1.3	0.3	0.13	1.16	0.67	-1056.64
0.5	1	1.25	0.31	0.33	0.75	0.71	-1008.96
0.5	2	1.38	0.32	0.49	1.84	0.23	-1009.54
1	1	1.5	0.3	1.16	1.93	0.41	-1080.48
1	2	1.7	0.34	1.7	3.31	0.3	-1080.8
1.5	1	1.56	0.33	2.29	2.87	0.25	-967.47
1.5	2	1.48	0.32	1.85	2.22	0.29	-967.45
2	1	1.29	0.29	1.52	0.78	0.63	-882.76
2	2	1.4	0.31	2.12	1.71	0.15	-883.15
2.5	1	1.31	0.3	2.13	1.07	0.82	-738.88
2.5	2	1.62	0.33	3.86	2.63	0.28	-740.58

Appendix B

Model Recovery in the Bivariate Diffusion Model Through Jointly Fitting To Response Choice and Response Time

In order to provide a more statistically robust test between the different between-trial variability hypotheses, we used BIC to compare the three formal models that reflect these hypotheses: random-only variability ($\rho_v = 0$), systematic-only variability ($\rho_v = 1$), and systematic and random variability ($0 < \rho_v < 1$). The fitting method was identical to that in Appendix A, and we performed the model recovery using the three simulated data sets from Appendix A with different ρ_v values; $a = 1.5$, $ter = 0.3$, $\mu_v = 3$, $\eta_v^2 = 0.5$, $\rho_v = [0, 0.5, 1]$.

The results of the model recovery can be seen in Table B1. Firstly, and most importantly, for all three data sets the systematic-only variability (ρ_v fixed at 1) and systematic and random variability (ρ_v freely estimated) models have near-identical log-likelihoods for the maximum likelihood estimates, despite the systematic-only variability model being largely misspecified in two of these cases, showing the perfectly mimicry between the models based on the unidentifiability of η_v^2 and ρ_v. Importantly, this results in the systematic-only variability model being preferred on BIC, due to the log-likelihoods being identical between models, and the systematic-only variability model being more parsimonious. Secondly, the random-only variability model shows near-identical log-likelihoods for the maximum likelihood estimates to the other models when the data are generated with a ρ_v of 0 or 0.5, though this is not the case when the data are generated with a ρ_v of 1 (i.e., systematic-only variability). These results again show the inability to distinguish between systematic and random between-trial variability in drift rate in the double-pass paradigm of Ratcliff et al. (2018), and that the findings of Ratcliff et al. (2018) only rule out a random-only variability model, and show no evidence for random variability being required in addition

to systematic variability.

Table B1: Displays the log-likelihood for the maximum likelihood estimate and the associated BIC values of the model recovery assessment in Appendix B. "Data" refers to the data set, defined by the generating ρ_v value, and "Model" refers to the model fit to the data, defined by the ρ_v constraints in the model, being either freely estimated, fixed at 0 or fixed at 1.

Data	Model	$log[p(D\|\theta)]$	BIC
0	Free	-279.21	596.42
0	0	-279.87	590.15
0	1	-279.35	589.11
0.5	Free	-245.83	529.66
0.5	0	-247.72	525.84
0.5	1	-245.69	521.78
1	Free	-228.54	495.08
1	0	-263	556.4
1	1	-227.52	485.45

Appendix C

Parameter Recovery of the Bivariate Drift Rate Distribution Through Direct Fitting

Due to the inability to distinguish between the effects of the η_v^2 and ρ_v bivariate diffusion model parameters based purely on summary statistics, we attempted to provide a model-based estimation of these parameters from the full response time distributions of the data. Importantly, we could not directly estimate the ρ_v parameter (as is commonly done for the η parameter in the univariate diffusion model), as no analytic solution currently exists for the likelihood function of a bivariate diffusion model, and attempting to numerically integrate or simulate this likelihood function proved too computationally taxing to implement. Specifically, we attempted to estimate a drift rate for each trial, and then constrained the estimated trial drift rates to follow a bivariate normal distribution:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & \rho_v \\ \rho_v & 1 \end{bmatrix} \right) \tag{C1}$$

with the same definition as Equation 1 in the main text. The threshold (a) and non-decision time (ter) were fixed to have the same values for all trials, and the starting point (z) was fixed to $\frac{a}{2}$, in order to try and make the model as simple as possible to maximize the chances of recovery.

Although it seems unlikely that a single trial – even with all other parameters constrained – would provide adequate information to properly constrain a drift rate parameter and allow it to be recoverable, it is still possible that the pieces of information contained within the drift rate estimated for each trial could provide adequate constraint for estimating the hierarchical parameters of the bivariate drift rate distribution.

The simulated data sets were identical to those in Appendix A and Appendix B (i.e., three simulated data sets with different ρ_v values; $a = 1.5$, $ter = 0.3$, $\mu_v = 3$, $\eta_v^2 = 0.5$, $\rho_v = [0, 0.5, 1]$). The model was estimated using a Bayesian framework with a hierarchical structure for the drift rate distribution, using DE-MCMC with 12 chains and 3,000 sampling iterations, with the first 1,000 iterations discarded as burn-in. However, the information from the drift rates of each trial appeared to do little to inform the hierarchical structure, with the model estimating poorly (i.e., non-converging posterior distributions), and the average posterior values for all three generating ρ_v values being close to 0.5 (0.42-0.66).

Appendix D

Model Recovery in the Multi-Pass Paradigm

Due to the inability to distinguish between the effects of the η_v^2 and ρ_v bivariate diffusion model parameters within the double-pass paradigm, we proposed that a multi-pass paradigm – where some relatively small number of stimuli are repeated a large number of times, in amongst a larger number of "filler" trials – may be able to distinguish between the effects of systematic and random variability. Specifically, we proposed that systematic and random effects could be distinguished using formal model comparison between four discrete models: a no-variability model, a random-only variability model, a systematic-only variability model, and a systematic-and-random variability model. Formally, the no-variability model can be defined as:

$$v_i = \mu_v \tag{D1}$$

where v_i is the drift rate for stimulus i, and μ_v is the mean drift rate for all trials. The random-only variability model can be defined as:

$$v_i \sim \mathrm{N}(\mu_v, \eta_v^2) \tag{D2}$$

where η_v^2 is the variance in drift rate for all trials. The systematic-only variability model can be defined as:

$$v_i = \mu_{v_i} \tag{D3}$$

where μ_{v_i} is the mean drift rate for stimulus i. The systematic-and-random variability model can be defined as:

$$v_i \sim \mathrm{N}(\mu_{v_i}, \eta_v^2) \tag{D4}$$

To gain some insight into whether these different discrete models – representing different explanations for the sources of between-trial variability in drift rate – can be distinguished within a multi-pass paradigm with a reasonable number of trials, we performed a model recovery simulation. Specifically, we compared each of these four models on four data sets – one data set generated by each of the four models – using Bayes factors calculated via bridge sampling (Gronau et al., 2017). Each data set consisted of 300 trials equally split among 6 stimuli, generated with $a = 1.5$ and $ter = 0.3$. Data sets with no systematic variability were generated with $\mu_v = 3$, whereas data sets with systematic variability were generated with $\mu_{v_1} = 4.75$; $\mu_{v_2} = 4$; $\mu_{v_3} = 3.25$; $\mu_{v_4} = 2.5$; $\mu_{v_5} = 1.75$; $\mu_{v_6} = 1$. Data sets with no random variability were generated with $\eta_v^2 = 0$, where data sets with systematic variability were generated with $\eta_v^2 = 1$.

The results of this model recovery simulation can be seen in Table D1 in the form of log-marginal likelihoods, which show several important trends. When the data were generated with systematic variability, then the systematic variability models provided a strong advantage over the models without systematic variability, suggesting that systematic variability is easy to detect in a multi-pass paradigm when it is present. However, this does no appear to be the case for the other situations. When the data were generated without systematic variability, models both with and without systematic variability provided similar log-marginal likelihoods, suggesting a difficulty in detecting an absence of systematic variability. Furthermore, random variability appeared to be difficult to identify in most

situations, and even resulted in the selection of the incorrect model in one instance, where a model with random variability included is selected, despite the data being generated without random variability.

These results suggest that the comparison of discrete models in a multi-pass paradigm may not be able to solve the identifiability issue that we observed within the double-pass paradigm for the η_v^2 and ρ_v parameters. However, it should be noted that our recovery results are dependent on fairly arbitrary decisions in generating parameter values and prior distributions, and therefore, this model recovery analysis should not be used to completely rule out the use of multi-pass paradigms for identifying systematic and random sources of variability in drift rate. However, we also believe that our recovery displays potential issues with the approach, and that a detailed recovery assessment showing the potential success of the approach would be required before the assessment would become tenable.

Table D1: Displays the log-marginal likelihoods for the model recovery in Appendix D. Rows display the models being fit, and columns display the generating model. "Syst" refers to the model with only systematic between-trial variability, and "Rand" refers to the model with only random between-trial variability. The winning model for each generated data set is displayed in bold.

	None	Syst	Rand	Both
None	**198.1**	79.98	121.62	112.39
Syst	196.61	125.86	118.34	148.68
Rand	197.92	82.86	**125.65**	116.07
Both	196.33	**126.63**	124.38	**150.17**