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: 3,748

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

A key assumption of models of human cognition is that there is variability in information processing. 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 between-trial variability in information processing is due to "noise". However, we argue that Ratcliff et al. (2018) failed to distinguish between two different potential sources of between-trial variability: random (i.e., "external noise") and systematic (e.g., item effects). Contrary to the claims of Ratcliff et al. (2018), we show that "external noise" is not required to explain their findings, as the same trends of data can be produced when only item effects are present. 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

A key assumption of models of human cognition is that there is variability in information processing (Shiffrin & Steyvers, 1997; Logan, 1988; Ratcliff, 1978). 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, Smith, Brown, & McKoon, 2016; Evans & Wagenmakers, 2019 for reviews). Most EAMs include two sources of variability in information processing (though see Usher & McClelland, 2001; Brown & Heathcote, 2008), which are both typically implemented as random draws from a normal distribution: within-trial variability and between-trial variability. Within-trial variability represents moment-to-moment fluctuations in our perceptual processing, whereas between-trial variability reflects variability in processing for different items from the same category, fluctuations in processes like attention, or sequential and/or time-based effects. The inclusion of between-trial variability in information processing has become common within the EAMs literature, as it allows these models to account for key qualitative benchmarks observed in decision-making data (Ratcliff, 1978; Brown & Heathcote, 2008).

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 the exact same stimulus was presented on two different trials, assessed whether a diffusion model (Ratcliff, 1978) could explain the level of agreement between these identical trials.

Their findings indicated that only a diffusion model with a non-zero η parameter – the standard deviation of the random between-trial variability in drift rate – could generate the observed trends, which they 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).

There are several key points made by Ratcliff et al. (2018) that we agree with. First and foremost, we agree that Ratcliff et al. (2018) provided evidence that drift rate varies in *some* manner between trials in an experiment, which 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 (e.g., Regenwetter & Robinson, 2017, 2019; Kellen, Klauer, & Singmann, 2012, 2013; Bhatia & Loomes, 2017), which in the case of the diffusion model involves determining whether or not random between-trial variability parameters (e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002) are (1) necessary for explaining empirical trends in choice and response time data, and (2) useful for improving our understanding of decision-making. 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 any of this variability was due to random sources, contradicting the central claim of their article.

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. However, we also wish to explicitly note that our article is *not* an attempt to question previous research that has shown the potential usefulness of between-trial variability parameters in EAMs. Rather, our article aims to provide a critique on the evidence presented in Ratcliff et al. (2018) for the existence of random between trial variability in information processing, and to provide a broader perspective on how we should view random variability parameters in cognitive models.

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. The novel claim of Ratcliff et al. (2018) was that "external noise" was necessary to explain the data from the double-pass experiments. However, this claim would only follow from their findings if evidence for *any* between-trial variability (i.e., a non-zero η parameter) was

also evidence for one specific type of between-trial variability: external noise. However, we argue that this is not the case, as between-trial variability in drift rate can be due to *random* sources (e.g., external noise) or *systematic* sources (e.g., item effects), and that a non-zero η parameter does not indicate the necessity of external noise.

In a theoretical sense, systematic and random sources of variability differ greatly from one another. Systematic variability is caused by factors that are known, such as experimental manipulations, and these factors can be explicitly modelled with different drift rates across the levels of the factor (e.g., difficulty manipulations). In contrast, random variability is caused by factors that are either unknown, or known but not easily modelled (e.g., fluctuations in attention). However, at a practical level these sources of variability are 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. To infer whether there was variability between trials in drift rate, Ratcliff et al. (2018) assessed the "agreement" (i.e., how often participants made the same response to the two identical double-pass trials) and accuracy (i.e., how often participants made the correct response across the entire experiment) observed in the empirical data, and compared these to the predictions of the diffusion model with different levels of between-trial variability (i.e., different values of the η parameter).

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.[1]

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:

[1] 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 $v_{i,j} \sim \text{N}(\mu_{v_i}, \eta_v^2)$, 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.

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

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:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right) \tag{3}$$

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:

$$\begin{bmatrix} v_{i,1} \\ v_{i,2} \end{bmatrix} \sim \text{BN}\left(\begin{bmatrix} \mu_v \\ \mu_v \end{bmatrix}, \eta_v^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \tag{4}$$

as the drift rate on each trial is an independent random draw from the bivariate normal

distribution (i.e., $v_{i,j} \sim \mathrm{N}(\mu_v, \eta_v^2)$).

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.

These formal definitions also clearly show our issues with the analysis method of Ratcliff et al. (2018). Ratcliff et al. (2018) compared the empirical data to predictions from the diffusion model with different potential parameter values for μ_v and η_v^2, finding that (1) $\eta_v^2 = 0$ made a strong prediction about these agreement-accuracy functions, where the agreement would always be lower for a fixed level of accuracy than when $\eta_v^2 > 0$ (see Figure 1, the bottom-right panel), and (2) the strong prediction of $\eta_v^2 = 0$ was not supported by the empirical data, consistent with $\eta_v^2 > 0$. However, these predictions were based on the drift rates being identical for each pass of the same stimulus, which is formally equivalent to $\rho_v = 1$.[2] Therefore, the analysis of Ratcliff et al. (2018) (1) failed to distinguish between the different sources of between-trial variability in drift rate, which is of crucial importance for their main claim, and (2) looked at the predictions of models with only *systematic* between-trial variability in drift rate (i.e., $\rho_v = 1$), meaning that these patterns of data were actually captured *without* the need for random between-trial variability in drift rate, contrary to their claims.

[2] Note that Ratcliff et al. (2018) mention that they performed some additional simulations where the drift rates were not identical for each pass of the same stimulus. However, these simulations only appear to have been used to check whether the strong prediction for $\eta_v^2 = 0$ held when the drift rates on double-pass trials were not identical, meaning that they still failed to distinguish between systematic and random sources of variability.

More generally, our simulations (Figure 1) suggest that the assessment of agreement and accuracy is insufficient to distinguish between systematic and random sources of variability in most cases. Interestingly though, the strong prediction shown by Ratcliff et al. (2018) for $\eta_v^2 = 0$ – which was inconsistent with the empirical data – is also present when $\rho_v = 0$ (top-left panel of Figure 1), making the predictions resemble $\eta_v^2 = 0$ regardless of the actual value of η_v^2. Therefore, only two conclusions logically follow from the data and analysis method of Ratcliff et al. (2018): that (1) between-trial variability in drift-rate is present (i.e., $\eta_v^2 > 0$), and (2) *at least some* of this variability is due to the systematic source of item effects (i.e., $\rho_v > 0$). 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 (Figure 2). Importantly, this trade-off makes it virtually impossible to determine whether the between-trial variability is purely systematic (i.e., $\rho_v = 1$) or a mixture of systematic and random ($0 < \rho_v < 1$), and therefore, whether external noise is actually necessary.

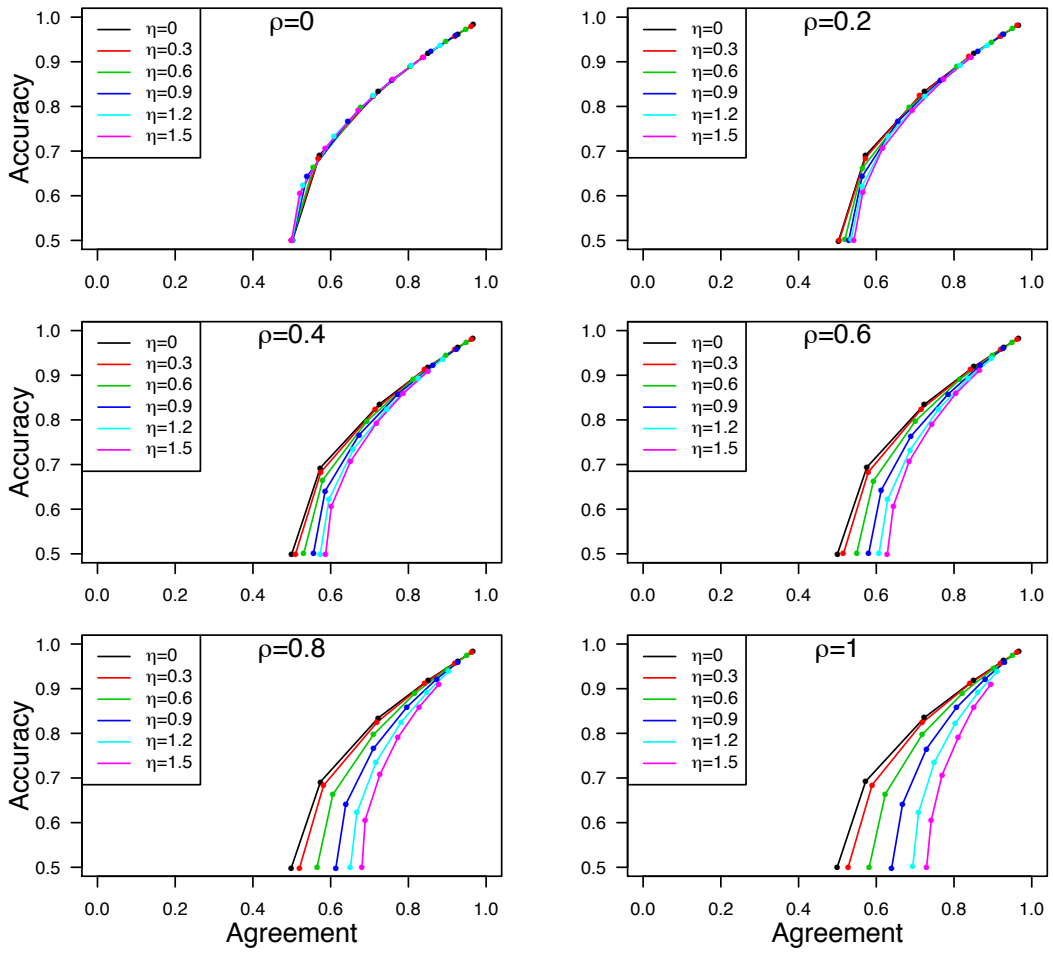

Figure 1. The agreement (x-axis) and accuracy (y-axis) functions for different mean drift rates (points on each line), variability in drift rates (different lines) and correlation in drift rates for identical items (different panels). The bottom-right panel displays the predicted functions when the correlation 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).

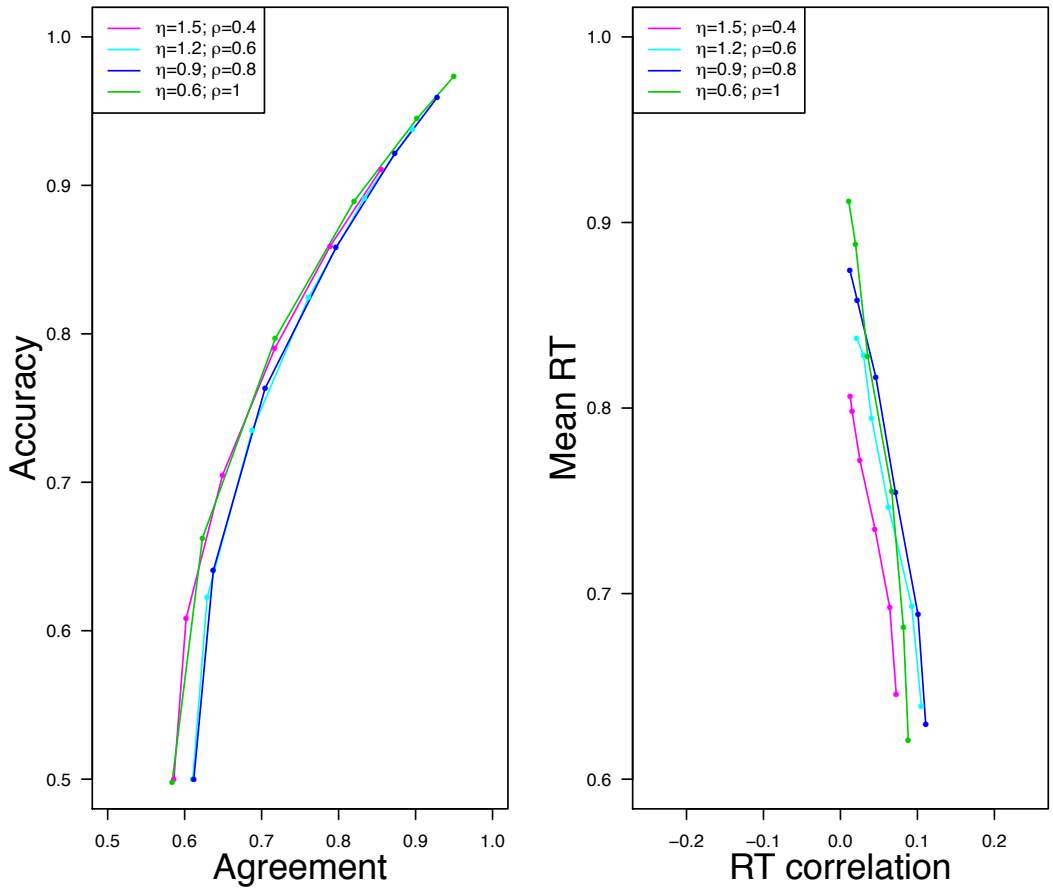

Figure 2. **Left:** The agreement (x-axis) and accuracy (y-axis) functions for different mean drift rates (points on each line), for four selected different combinations of variability and correlation that closely mimic one another, suggesting that the assessment of these summary statistics is of limited 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).

Although this problematic tradeoff could theoretically be solved by first fitting the standard univariate diffusion model to estimate the μ_v and η_v^2 parameters, and then per-

forming 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 obtain reliably (Ratcliff & Tuerlinckx, 2002; Lerche & Voss, 2016) regardless of method (Boehm, Annis, et al., 2018). Note that we are not attempting to question the identifiablility of the η_v^2 parameter in the univariate diffusion model, or suggest that estimates of η_v^2 in previous studies are meaningless. Rather, we are stating that the proposed "two-step" approach requires an unbiased *and* extremely precise estimate of η_v^2, as based on our previous simulations showing the problematic tradeoff between η_v^2 and ρ_v – as well as previous research assessing two-step approaches (Matzke et al., 2017; Boehm, Marsman, Matzke, & Wagenmakers, 2018) – any bias or uncertainty in the estimate of η_v^2 may lead to spurious conclusions. Therefore, based upon previous research (Ratcliff & Tuerlinckx, 2002; Lerche & Voss, 2016; Boehm, Annis, et al., 2018), as well as a small recovery study that we provide in Appendix E, we do not believe that the data from the double pass paradigm of Ratcliff et al. (2018) would be appropriate for the proposed two-step approach, due to the inability to obtain a precise point estimate of the η_v^2 parameter in a data set of this size.

We also considered several other analysis methods and whether they could distinguish between systematic and random sources of between-trial variability in drift rate. Firstly, we considered a similar analysis to the agreement-accuracy functions, though with mean response time and correlation in response time instead of accuracy and agreement. However, similar to agreement, increases in response time correlation can be created by either increasing η_v^2 or increasing ρ_v (Figure 3).

Secondly, we considered whether the joint information from all four variables – accuracy, agreement, mean response time, and correlation in response time – was adequate to separately identify η_v^2 and ρ_v. As this is difficult to determine from assessing separate

functions (e.g., Figure 2), we performed a more rigorous assessment by fitting 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 $\rho_v = 1$ (i.e., only systematic variability) based on it being more parsimonious than ρ_v as a free parameter (i.e., both systematic and random variability; see Appendix B).

Thirdly, 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). However, we were unable to recover the values of ρ_v, which is understandable given the high complexity and dimensionality of the model relative to the data.

Lastly, we considered whether the concept of the double-pass experiment could still be used to distinguish 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), where the majority of stimuli are presented only once and just a small number of stimuli are repeated a large number of times, in order to allow for the adequate estimation of the drift rates for the repeated stimuli. The different potential sources of between-trial variability in drift rate could then be represented by four discrete models – no-variability, random-only variability, systematic-only variability, and systematic-and-random variabil-

ity, and compared using formal model comparison (e.g., Bayes factors; Jeffreys, 1939; Kass & Raftery, 1995; Evans & Brown, 2018. However, our recover 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.

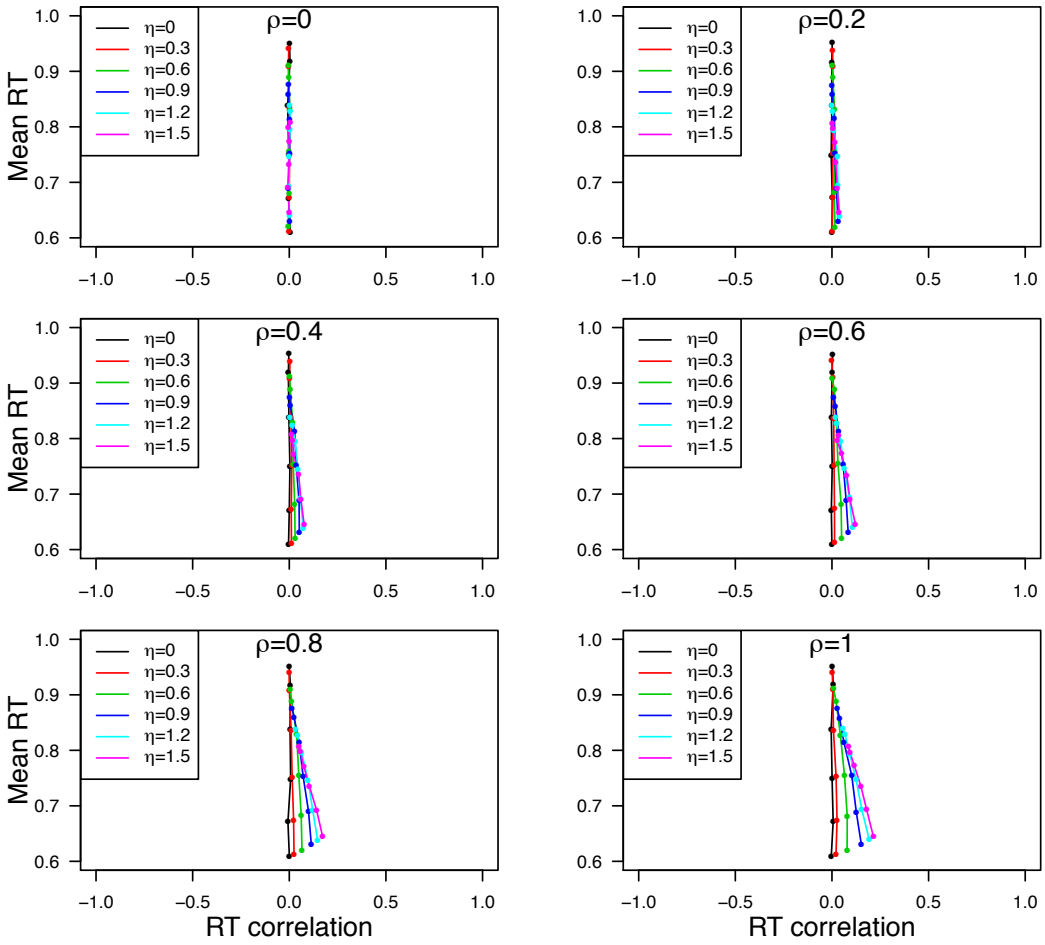

Figure 3. The correlation in (x-axis) and mean (y-axis) response time functions for different mean drift rates (points on each line), variability in drift rates (different lines) and correlation in drift rates for identical items (different panels).

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.

Within cognitive models, we believe that noise simply serves as a convenience parameter that allows the absorption of variability from sources that are difficult to explicitly incorporate into the model, similar to statistical models, where the "noise" (i.e., ϵ) is simply the unexplained variance within the model (De Boeck & Wilson, 2004).

Although noise terms can clearly be useful components of cognitive models, we believe that there are some limitations to focusing too heavily on noise terms. Firstly, based on our definition, noise is highly context dependent – both in terms of the experimental design and the model defined – rather than a general phenomenon that does or does not exist. Therefore, we are uncertain whether general questions about the existence of random variability in specific parameters – such as those posed by Ratcliff et al. (2018) – are actually useful, as they may completely depend on the experimental data set and model defined. Secondly, we believe that adding unnecessary random variability into cognitive models can dilute our understanding of the cognitive processes that they intended to explain. Merely labelling all variability as random does not allow us to gain any additional insight into underlying process causing these variabilities, and including too much random variability within a model can harm its measurement properties (Lerche & Voss, 2016; van Ravenzwaaij, Donkin, & Vandekerckhove, 2017). Therefore, while random between-trial variability parameters within EAMs can serve as a convenient placeholders, which allow the models to account for phenomena that they otherwise could not, researchers should aim to identify and explicitly model the underlying systematic sources of this variability when possible.

There are several methods that researchers can use for identifying different sources of variability, integrating systematic variability into existing models, and/or providing systematic explanations for random variability, which could help to lessen the reliance on random between-trial variability parameters within EAMs. One method involves using ex-

perimental design to tightly constrain the predictions of models, allowing models with different underlying explanations for the variability to be more easily compared (e.g., Ludwig & Davies, 2011; Teodorescu & Usher, 2013; Servant, White, Montagnini, & Burle, 2015; Evans, Dutilh, Wagenmakers, & van der Maas, 2019). Some examples include Ludwig and Davies (2011), who combined the double-pass paradigm with manipulations of stimulus strength and the time of a response signal to discriminate between different observer models for how accumulated evidence changes over time, Evans et al. (2019), who compared a range of different EAM variants that strongly mimic one another by constraining the models to also account for the secondary responses made by participants, and Servant et al. (2015), who provided further constraint on the diffusion model by making it account for electromyography (EMG) recordings.

Another method involves creating a direct mapping between the stimulus features and the parameter values of existing models through *front-end* extensions (e.g., Nosofsky & Palmeri, 1997; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2004; Brown, Marley, Donkin, & Heathcote, 2008), which provide constrained explanations for changes in parameter values. 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. Some examples include models of multi-attribute choice (Roe et al., 2001; Usher & McClelland, 2004; Trueblood, Brown, & Heathcote, 2014), where choices are made between alternatives that each have different values of the same explicit attributes and the front-end extensions transform the attribute values into drift rates for each alternative, and the neural leaky competing accumulator (LCA) extension of Purcell et al. (2010), who constrained the drift rate at each point in time to be a function of the current visual neuron activity that was measured in non-human primates via single-cell recordings.

One final method involves creating new models where the seemingly random variability is a natural by-product of the underlying process, providing a direct and precise explanation for why the variability occurs. These new explanations are often inspired by an attempt to connect knowledge from different fields, such as the concept of *neural plausibility*, where models are designed with processes that are high-level reflections of actual neural processes (Usher & McClelland, 2001, 2004; Verdonck & Tuerlinckx, 2014). Some examples include the LCA Usher and McClelland (2001), where the neurally inspired components allow the model to account for several key qualitative benchmarks in choice and response with only a single (within-trial) source of random variability, and the Ising decision maker (IDM; Verdonck & Tuerlinckx, 2014), where the underlying pools of binary neurons that feed into an accumulation process do not require between-trial variability in drift rate and naturally produce between-trial variability in starting point.

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. Furthermore, we 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, which is contrary to their central claim. 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.

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior research methods*, *50*(4), 1614–1631.

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.

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. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*, 589–603.

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., & Wagenmakers, E.-J. (2019). Evidence accumulation models: Current limitations and future directions. *The Quantitative Methods for Psychology*.

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.

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.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, *90*(430), 773-795.

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.

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.

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, *3*(1).

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*.

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., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

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

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., 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., 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.

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.

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.

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

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, 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. 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 (see Evans & Brown, 2018; Annis, Evans, Miller, & Palmeri, 2019; Evans & Annis, 2019; Evans, Bennett, & Brown, 2018 for examples with cognitive models) 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**

Appendix E

Parameter Recovery of the η_v Parameter in the Double-Pass Paradigm

One potential proposal that we mentioned within the main text for solving the problematic tradeoff between the μ_v, η_v^2, and ρ_v parameters would be to firstly fit the standard univariate diffusion model to the data to estimate the μ_v and η_v^2 parameters, and then perform the agreement-accuracy analyses with the μ_v and η_v^2 values constrained to be the point estimates obtained in the initial fits. However, this approach would be completely dependent on the ability to obtain an accurate and precise estimate of the η_v^2 parameter, which previous research has suggested is difficult to reliably obtain within a reasonable number of trials (Ratcliff & Tuerlinckx, 2002; Lerche & Voss, 2016; Boehm, Annis, et al., 2018). Here, we provide a small parameter recovery assessment for the η_v parameter using the same double-pass design as Ratcliff et al. (2018), with a slightly larger number of trials than their studies (2,000 trials total; 1,000 for each "pass"). Specifically, we generated 216 simulated data sets, which consisted of all potential combinations of the generated values that we used for μ_v (0, 0.5, 1, 1.5, 2, 2.5), η_v (0, 0.3, 0.6, 0.9, 1.2, 1.5), and ρ_v (0, 0.2, 0.4, 0.6, 0.8, 1), with $a = 1.5$ and $ter = 0.3$ in all simulated data sets. We fit a model with 4 free parameters – a, ter, μ_v, and η_v – to each of these data sets using the same methodology as the previous recovery studies, and assessed the posterior estimates of the η_v parameter, which can be seen in Figure E1. As can be seen, in all cases the credible intervals reflect sufficient uncertainty (i.e., a lack of precision) in the estimate of the η_v parameter to make the use of "two-step" analyses questionable, and in many cases the posterior median does not fall along the grey line, indicating the inaccuracy of the central tendency of the estimate.

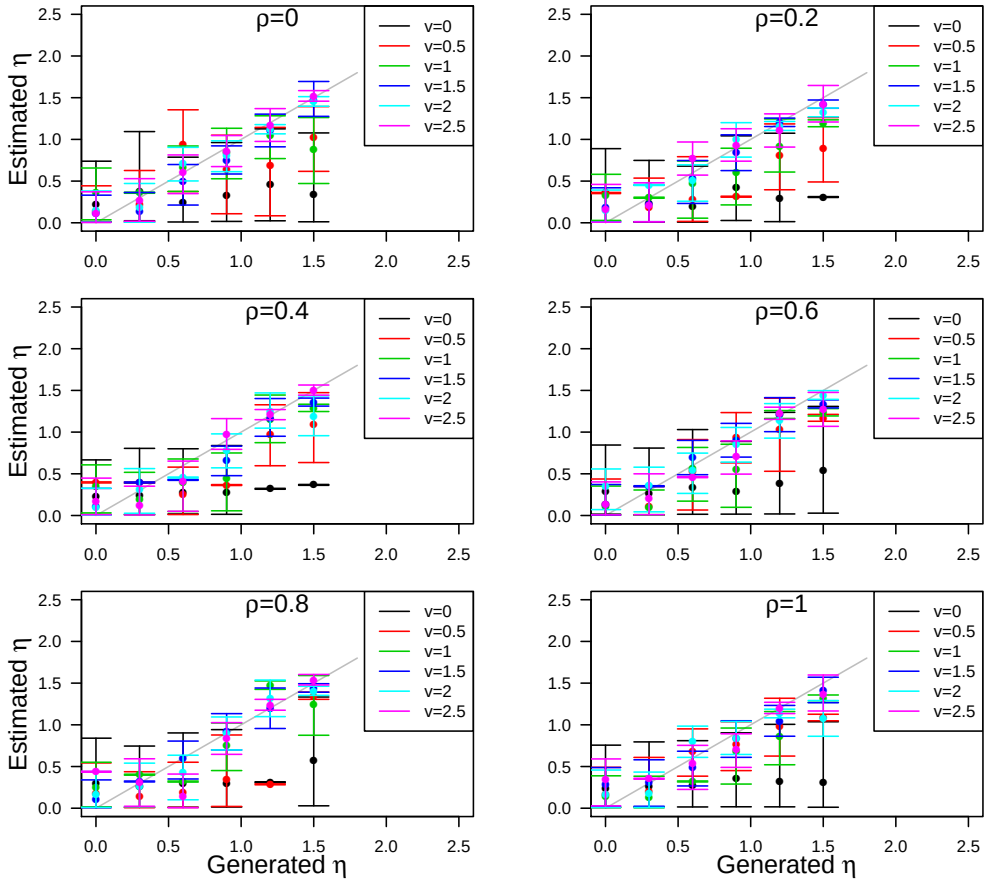

Figure E1. The generated (x-axis) and estimated (y-axis) estimates of η_v for different values of μ_v (different coloured points) and ρ_v (different panels). Each panel displays the posterior median (point) and 95% credible interval (error bars) for the estimate of the η_v parameter, with the grey line displaying correct, precise recovery.