Measuring the completeness of race models for perceptual decision-making

Anjali Sifar (sanjali@iitk.ac.in)
Dept of Cognitive Science, IIT Kanpur, India

Hariharan Purohit (hariharan22@iitk.ac.in)
Dept of Cognitive Science, IIT Kanpur, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Depts of Cognitive Science and Computer Science, IIT Kanpur, India

Abstract

Computational models of perceptual decision-making depend heavily on empirical goodness-of-fit measures for model selection. However, it is not possible to improve models' fit to data indefinitely, particularly when the data in question are variable across multiple elicitations. The completeness of a model or a theory assesses the extent to which it can predict observations in comparison with an ideal model. We measure the completeness of contemporary race models on a paradigmatic perceptual decision-making task - random dot motion discrimination - and show that the simple drift diffusion model is already close to complete in describing random dot motion discrimination data, with more complex models being in fact overfit to datasets. Thus, in this paper, we quantitatively demonstrate limits to the ability of conventional choice fraction and response time data to disambiguate complex models of perceptual decision-making.

Keywords: random dot motion discrimination, retest reliability, completeness, race models, drift diffusion

Introduction

Random dot motion discrimination (RDM) task is a workhorse experimental paradigm in the study of perceptual decision-making (Gold, Shadlen, et al., 2007; Ratcliff & McKoon, 2008). The attractiveness of this task is amplified by the illustrious history of connections between psychology, modelling and neuroscience made possible by the use of RDMs in primate studies to understand the 'evidence-integrative' behavior of LIP neurons (Britten et al., 1992; Gold, Shadlen, et al., 2007).

A variety of computational race models have been applied in the study of the perceptual decisions inherent in tasks like random dot motion discrimination (Brown & Heathcote, 2008; Ditterich, 2006; Ratcliff & McKoon, 2008; Usher & McClelland, 2001). While these models share some basic assumptions about the nature of the choice process, viz. that it comprises of temporal integration of evidence in favor of various alternatives up to some evidence threshold (Brown & Heathcote, 2008; Ratcliff & McKoon, 2008), they also differ along important dimensions, e.g. whether the evidence threshold is fixed, or dynamic (Cisek et al., 2009; Thura et al., 2012). While some researchers have sought to differentiate these models using empirical evaluation, (Hawkins et al., 2015; Thura et al., 2012), this is not straightforward to accomplish. For example, Donkin et al. (2011) show that two models may fit a dataset equivalently well, but show different trends in the direction of a particular shared model parameter, suggesting that the mapping of model parameters to latent cognitive variables may not be accurate, which may limit the interpretability of model-based analyses.

An additional challenge to the interpretability of model-based analyses, and even predictions, for RDM tasks lies in the currently inadequate characterization of sources of noise for the task. While race models have several sources of variance built into their equations, these values are empirically known to be unreliable (Lerche & Voss, 2017), and have weak theoretical bases (Brown & Heathcote, 2008). Recently, Ratcliff et al. (2018) used a double pass paradigm to characterize how well drift diffusion models can accommodate the effect of exogenous noise while fitting RDM data. However, such efforts do not confront the fundamental problem that variance parameters in these models are not systematically grounded and serve almost entirely as mathematical receptacles of otherwise inexplicable variance in data fitting exercises (Brown & Heathcote, 2008).

Both these lacunae in current computational models of perceptual decision-making, in our view, stem from an overly statistically-driven focus towards model evaluation and selection (Yarkoni, 2022). If fitting the data as precisely as possible, modulo statistical regularization, is the primary criteria for model goodness, then parameter interpretability becomes a secondary concern, and the ultimate provenance of noise in models a tertiary one.

When can we say that models are good 'enough' in terms of empirical goodness-of-fit for us to consider other criteria for model selection? The concept of 'completeness', recently introduced by Fudenberg et al. (2019), offers an elegant answer to this question. The completeness of a model, in simple terms, is defined as the ratio of the improvement in predictive performance a model produces vis-a-vis a naive or random baseline to the improvement in predictive performance produced by the *best possible* model vis-a-vis the same naive baseline. The best possible model is taken simply to be the unparameterized table of mappings from the independent variable to the dependent variable.

When this mapping is not bijective, the best possible model will not be perfect. For example, if a person responds to the presentation of a stimulus X with the response y_1 at one time, and y_2 at another, then a predictor will necessarily incur some loss when trying to assign a value to at least one of the two observations in this dataset (Fudenberg et al., 2019). This is particularly true for behavioral datasets, since behav-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

ioral responses frequently have low retest reliability. It has recently been observed that complex models fit to individual elicitations of behavior, in fact, run the risk of becoming overprecise, in the sense that their predictions explain the dataset rather than the phenomenon (Sifar & Srivastava, 2022). Thus, in such settings, knowing the completeness of a class of models can inform scientists about the extent to which further improvements in empirical goodness-of-fit can convey scientific insight.

In this paper, we characterize the completeness of well-known race models applied to random dot motion discrimination (RDM) data. The primary unknown in this exercise is the performance of the ideal model for this task, which we measure using a test-retest reliability experiment, showing subjects the exact same RDM stimuli one week apart. We then use this estimate of the best possible model's performance to assess the completeness of several race models commonly used to study RDM and other related tasks. As a corollary to our main result, we also identify the test retest reliability of the parameters of these models for RDM.

Methods

Design

Participants performed a 2-alternative forced choice (2AFC) random dot motion (RDM) discrimination task which required them to indicate the direction of a set of apparently moving dots, see Figure 1a. We followed the task specifications from the RDM experiment in Ratcliff et al. (2018). Five white dots, each of radius 1×1 pixel, move apparently in space and time by virtue of quick successive presentation at every frame in an invisible circular aperture (of radius 100px) on a black background.

Every trial is described by *coherence* level and *direction* of movement. Coherence determines the probability of each dot being a signal dot per frame, refer Figure 1b for more details. The signal dot moves by 4 pixels in the *direction* of motion at the next frame while the noise dot is placed randomly. The dots are regenerated every 3 (dot life) frames such that the motion is described globally (Pilly & Seitz, 2009). A total of 24 frames were presented for 400 ms with the screen refresh rate of 60 Hz. The participant was required to respond within 1500 ms of the stimulus onset with the designated key press (Left arrow key for left direction, Right arrow key for right direction). The trial structure was self paced such that the user had to press a key to continue to the next trial.

The task was designed using the Pygame module (v2.1.2) of Python. The monitor used for display was ASUS VG248QE, 24-inch FHD with a screen resolution of 1920x1080 (width $\times$ height) at 40% brightness level (to reduce eye strain). The task was designed for retest reliability estimation ensuring that every participant viewed the exact same stimuli, in the same sequence, in both sessions. The three levels of coherence and 2 possible directions (left and right) were randomly interleaved, and then we pre-generated the dot positions for the 24 frames of every trial in every block. This means that the dot position at every frame was exactly the same between the two sessions for a given trial.

Procedure

We recruited a total of 69 (25F) participants for the experiment. We provided a compensation of USD 6 for all four sessions, and USD 2.5 in case the participant did not achieve the required accuracy in the staircase procedure (explained later). The overall design is illustrated in Figure 1c.

Each participant was seated on a height adjustable chair in front of the monitor display in a dark room. The viewing distance (60 cm) was manually measured for each session from the center of the screen to the nose of the individual.

Following a practice session for task familiarity, we conducted a 4 up-1 down staircase to estimate the psychophysical threshold with correct response probability of 84% (Lu & Dosher, 2013). The staircase continued for a total of 30 reversals. For a given participant, we estimated the subjective threshold of coherence level as the mean of the last six reversals. We used this threshold (JND) $\pm$ 0.05 as the three levels of subjective coherence per participant. We had to drop 24 participants from the experiment at this stage because they could not achieve the accuracy level needed within 30 reversals.

Finally, we continued to the main experiment session with a total of 45 (17F) participants. The two main sessions were exactly the same with respect to task conditions and stimulus presentation. With the trial structure depicted in Figure 1a, each participant performed 90 trials per block. Error feedback was provided for 300 ms.

Response Variables

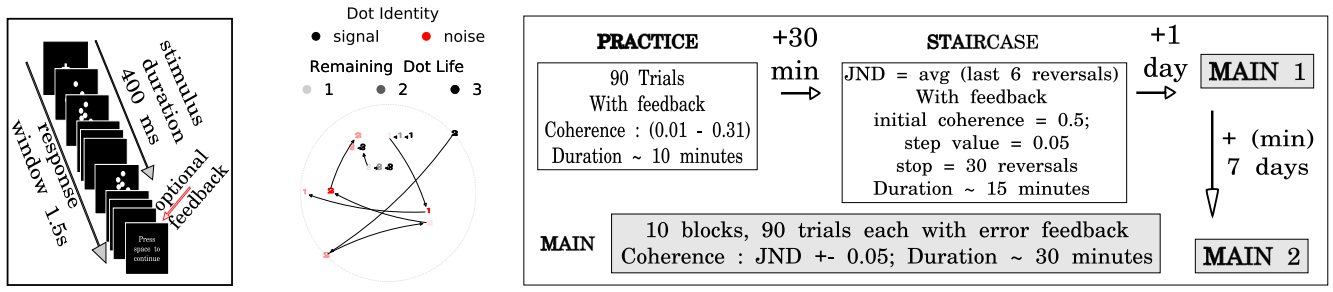
For every trial, we recorded the reaction time (RT) and response correctness (1/0).

Data cleaning

An average of 300 trials per coherence level were presented to each participant. Trials across sessions for each participant with missing responses in either session were removed from the analysis, yielding an average of 298 trials per coherence level per participant. Three participants' data had to be removed because of poor accuracy and more than fifteen sub-200ms RT trials. With this last exclusion, we were left with a total of 42 (17F, mean age 23.7) participants for the final analysis.

Analysis

We measured retest reliability of random dot motion task separated by at least one week at three subjective coherence levels. Assuming the entire source of variability is accounted for by the external stimulus, we expect the participant performance to approach its *true* value with a large number of trials per coherence level assuming the underlying perceptual decision-making process has high reliability.



(a) (b) (c)

Figure 1: **Task design a.** Trial structure. Error feedback only in main sessions. Dot positions at every frame for every trial in the two main sessions was exactly the same for the retest paradigm. **b.** RDM implementation details using five consecutive frames for three dots marked by numerical digits. Signal dot movement is restricted to horizontal axis while noise dot movement is random. Dot life controls the number of frames a dot maintains its identity (signal/noise) to prevent local movement inference. **c.** Experiment design. The four sessions conducted per participant with specific details. The results presented in this study belong to the two main retest sessions conducted at least a week apart.

Models

We use three variants of accumulator models in our analysis. The simplest model is the drift diffusion model (DDM), which also acts as a baseline for other variants of choice models (Stone, 1960). According to this model, noisy evidence

$$dx = \nu dt + c\sqrt{dt}$$

accumulates until a fixed decision boundary (threshold) a is reached. ν is the drift (slope) towards the threshold starting from z which is diffused by wiener noise represented by $c\sqrt{dt}$. An additive term of Non decision time (Ter) on top of accumulated evidence is used which is the time required for encoding and motor responses. We assume z=0, fixed values for scaling parameters c and timestep dt across all models. Assuming one level of drift rate per coherence (difficulty) level, a total of five parameters were free for model fitting (ν_1,ν_2,ν_3, a, Ter).

The second model is a variant of DDM - the collapsing bounds model - which additionally assumes that the decision threshold collapses as a function of time, given by the following equation (Hawkins et al., 2015) :

$$u(t) = a - (1 - exp(-(\frac{t}{\lambda})^k))(0.5a - a')$$

where λ is the shaping parameter, while a' is the asymptotic threshold.

The third model is an urgency gating model which assumes that instead of a drop in threshold, there is an urgency gain in the evidence accumulation process during the course of a trial given by the following equation (Hawkins et al., 2015) :

$$\gamma(t) = b_0 + \frac{s_y exp(s_x(t-d))}{1+exp(s_x(t-d))} + \frac{1+(1-s_y)exp(-s_x d)}{1+exp(-s_x d)}$$

where s_x, s_y are the shape parameters, d is the delay and b_0 is the intercept.

We also test more complex variants of each of these base models, which add variability parameters for the drift $\sim \mathcal{N}(0,\eta)$, starting point ($sz \sim \mathcal{U}(z_{min},z_{max})$) and non decision time ($st \sim Ter \pm st_0$).

Analytic tools

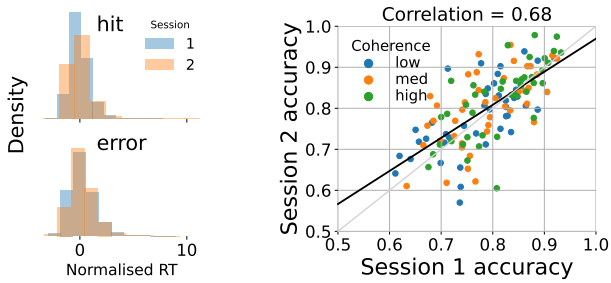
For continuous variables like mean RT, choice accuracy and individual trial level RT, we used Pearson's correlation r to measure the reliability across the retest sessions. Additionally, we used κ to measure binary individual level choice agreement.

We used the model fitting routine provided by CHaRT-R package for parameter estimation (Chandrasekaran & Hawkins, 2019). Model parameters were estimated using the QMPE statistic for each participant independently. For each coherence level, error and correct RT distributions were split into 10 quantile bins. Fits were obtained by maximising $ln(L(\theta|T)) \propto \sum_{j=1}^{m} N_j ln \int_{q_{j-1}}^{j} f(t,\theta)dt$ for m quantiles, N_j = number of observations in each quantile of a given coherence level and $f(t,\theta)$ is the corresponding likelihood value. This gave us an approximation of the log likelihood value (LL) for the model fit (Heathcote et al., 2002).

Following Sifar and Srivastava (2022), we define the ideal model simply as the *predictions* made by a participant when (s)he is presented with the exact same stimulus in a different session. In other words, session i is treated as the ideal model for session j data, $i \neq j$. The log likelihood (as measured by QMPE) was used to measure the completeness of the models. Following Fudenberg et al. (2019), we measure completeness as,

$$\frac{LL_{Random} - LL_{Model}}{LL_{Random} - LL_{Ideal}}.$$

We also report BIC as a measure of model goodness-of-fit. For the ideal model, the number of free parameters is not straightforward to establish. We set it at 10 for the calculations we report here, but values between 1-20 reproduced similar results.



(a) (b)

Figure 2: **Observed retest reliability a.** For the histogram, we first normalised observed RTs per participant to ensure that within and between participant variability is not mixed. We plot normalised RT histograms for session 1 and 2, segregated by hit($\sim$ 30,000 trials per session in upper panel) and error($\sim$ 7,500 trials per session in lower panel). **b.** Each dot in the scatter plot depicts the accuracy for a given coherence level and participant.

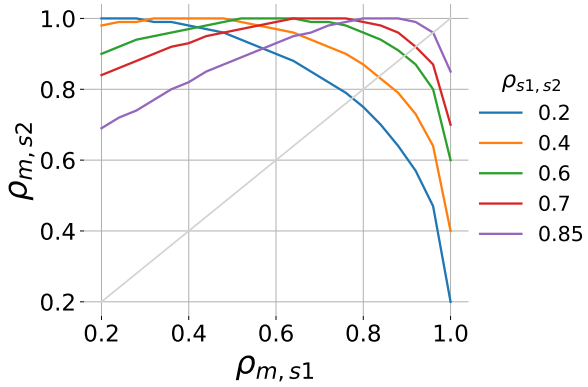


Figure 3: Limits on model-data correlations imposed by test-retest reliability of the underlying construct. All points below the $x = y$ line on each of the curves represent instances of models overfit to session 1's data; correlation of such a model with session 2's data is guaranteed to be lower.

Results

We present three sets of results. First, we present our estimates of the test-retest reliability of the random dot motion discrimination task. Second, we present our estimates of the completeness of different choice models with respect to the task. Third, we demonstrate the reliability of different parameters in these models to data fits across sessions.

Reliability estimates for RDM

A two-way repeated measures ANOVA was conducted with coherence and session as the independent factors and mean RT as the dependent variable. We observed main effects for session [F(1,41) = 17.66, $p < 0.001$] as well as coherence [F(2,82) = 34.86, $p < 0.001$] with no interaction effect. Treating session as a treatment, the difference in mean RT for subjects, across conditions, yields an effect size of Cohen's $d = 0.3$.

We found retest reliability of the mean RT across coherence levels and subjects to be r = 0.84. The corresponding value for accuracy across all coherence levels is $r = 0.68$ (also see Figure 2b).

Trial level consistency for accuracy measured using Cohen's κ was none to slight (κ = 0.2). Similarly, the average retest reliability for individual trial RT was negligible (r = 0.15), across all coherence levels and participants.

It is unsurprising to see low reliability for individual trials, given that evidence integration is unlikely to recruit the same neural pathways across multiple task elicitations (Beck et al., 2012). However, the moderate correlations seen for accuracy across sessions is surprising, given that RDM is meant to be a perceptual decision-making task with little cognitive variability. The reliability for accurate choices we find in our experiment is similar to values seen for much more abstract risky economic decisions (Sifar & Srivastava, 2022), suggesting more cognitive involvement in this task than has been appreciated heretofore.

Another underappreciated aspect of retest reliability is that it imposes strong limits on the degree of correlation fits to one instrument may achieve on another (Vul et al., 2009). In particular, the consistency of a model m with data observed in two sessions s_1 and s_2 is limited by a statistical identity

$$\rho_{m,s2} \leq \rho_{s1,s2}\rho_{m,s1} + \sqrt{(1-\rho^2_{s1,s2})(1-\rho^2_{m,s1})}.$$

This relationship is graphically illustrated in Figure 3, showing that for low retest reliability, extremely high correlations between the model and one session's data is guaranteed to produce much lower correlations for that model for the other session's data, even if both sessions use the same target stimuli and protocol.

This observation is of immediate relevance for scientists seeking to fit models to behavioral data as inputs to downstream neural modelling. For example, Pisauro et al. (2017) predict value based decisions using a dynamic sequential sampling model with very high precision (accuracy : r = 0.96; RT : r = 0.91) in order to subsequently identify clusters of EEG activity monitored while doing the task. Given the substantially lower reliability of accuracy we have seen for RDM and Sifar and Srivastava (2022) have documented for value-based choice, it is unclear to what extent the bijective mapping from behavior to brain regions posited in such a study can be assuredly inferred.

Completeness of race models on RDM

Figure 4 (left panel) plots our estimates of completeness for six race models - simple DDM, collapsing bounds, urgency gating, and variants thereof with additional parameters meant to represent intra- and inter-trial variability, for both sessions individually. For the random predictor, we used the observed RT and generated random labels for accuracy.

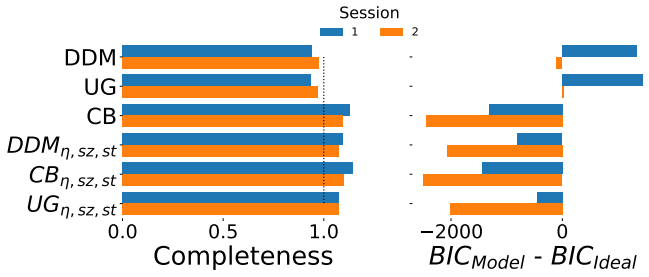


Figure 4: Completeness measures for each session. **Left panel** Completeness of race models for RDM. Black line indicates 100% completeness. **Right panel** Difference of BIC for each model from ideal model BIC.

We note that both the simple DDM and UG models demonstrate more than 90% completeness for the random dot motion discrimination task. The other models used in our comparison are actually over-complete, in the sense that they show lower residuals with respect to the first session's data from participants, than those participants' own responses to the same stimuli do from a second session.

The difference between BIC of the race models from the ideal model's BIC (taking the average of BIC calculated using either session as the ideal model), plotted in Figure 4 (right panel), yields an identical conclusion. Negative values of this quantity indicate that the BIC for a model of participants' data is a closer empirical fit than participants' test data is to the same participants' retest data.

This pattern of classic models already being close to complete appears to be a recurrent pattern across studies. Fudenberg et al. (2019) found this to be true for predicting certainty equivalent assignments for monetary gambles by human respondents, and in predicting first move patterns in game-theoretic games. Sifar and Srivastava (2022) found this to be true for predicting choice proportions in risky economic decisions. In continuation of this trend, we find this to be true for predicting accuracy and RT patterns in an RDM task.

Reliability of race model parameters for RDM

Finally, we measured the reliability of recovered model parameters across sessions. The upper panel in Figure 5 shows the reliability of each model parameter for the three basic models alongside the simplified, but widely used, EZ-DDM model (Wagenmakers et al., 2007).

The EZ-DDM model's parameter estimation process is simply an analytical transformation of the choice proportion, mean and variance of the RT distributions to three prominent DDM model parameters under some simplifying assumptions (Wagenmakers et al., 2007). We find that, despite (or perhaps because of) its simplicity, EZ-DDM's parameter estimates were the most reliable across the two sessions, with the five parameter simple DDM coming in second in terms of reliability. We also observed the highest retest reliability for non-decision time across all models, followed by drift rate. All other parameters exhibited very poor reliability, even becoming negatively correlated in some cases.

The lower panel in Figure 5 shows the reliability of parameters for the more complicated variants of the base models. In addition to poor reliability of the added parameters, we also observed reduced consistency for drift rate for all models in this category, compared to the simpler models in the upper panel.

General Discussion

In this paper, we described our efforts at measuring the completeness of race models for RDM. Three results emerged from our analysis. Our first result is the demonstration of moderate retest reliability for accuracy and high retest reliability for response time distributions for the RDM task in aggregate, but extremely low trial-level reliability for both variables. The fact that cohort-level reliability of choice accuracy in a perceptual decision-making task is less reliable than choice proportions in abstract economic decisions is surprising, and warrants further investigation.

Our second result is the observation that simple race models are already more than 90% complete, in the sense that the *empirically best* model for this task would not reduce mean squared loss by more than 10% from what simple DDMs alone can accomplish. We note that this finding is not entirely unexpected. A recent parameter recovery study across multiple researchers' DDM model fitting pipelines revealed that fits produced by the simplest models tended to provide the best explanations for the test data (Dutilh et al., 2019).

While these earlier results point to simpler race models being better in a relative sense than complex ones, our results lay down an objective marker for just how well they work, and how little explainable variance is genuinely left in the residuals of their predictions on RDM data. Put simply, DDM is a 'good enough' model of RDM and related tasks; more complex models can fit individual choice proportion and response time datasets better as a curve-fitting exercise, without generating reliable insights. Thus, we suggest that better insights into perceptual decision-making require richer data, not more complex models (Sifar & Srivastava, 2022).

Our third result is the observation that all parameters in the EZ-DDM model, but only drift rates and non-decision time parameters for other models, are reliable across session fits. Lerche and Voss (2017) make similar observations in reliability testing of race models across a battery of cognitive tasks.

Overall, we show that more precise models of the RDM task may not necessarily be better, since their precision arises from over-fitting to single observations of intrinsically variable behavior. As we demonstrate, maximizing model-data correlations for one session's observations of low reliability data guarantees that the fit models will fail to predictively generalise even to the same participants' responses to the same stimulus. Thus, our results support a recent theoretical claim that the presence of irreducible noise can easily defeat empirical claims proved using Fisherian hypothesis testing,

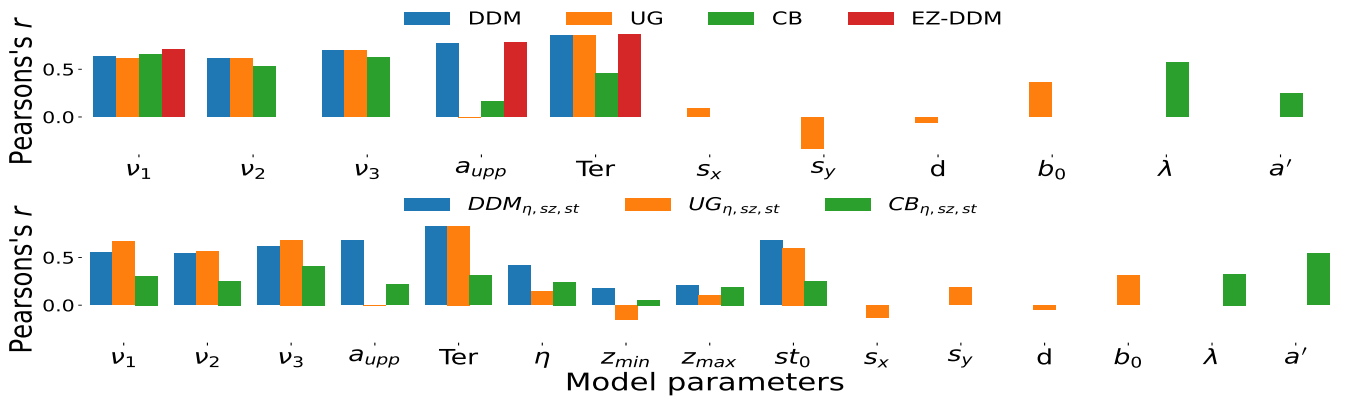


Figure 5: **Model parameter reliability** Upper and lower panel show the reliability between each model's parameters across two session fits for simple and complicated models respectively. A single bar for a subset of parameters indicate that parameter is specific to the model represented by the color of the bar. Parameter details in text.

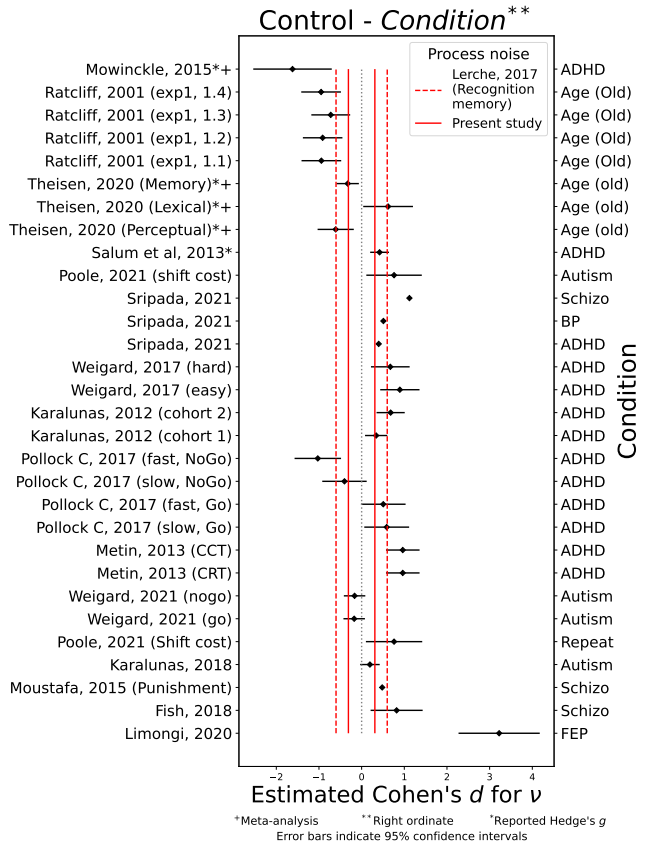


Figure 6: Estimated empirical effects for difference in drift rate between two groups of population or task conditions under a variety of conditions. Left ordinate axis specifies the study and task condition for the result represented by black markers, while right ordinate axis briefly specifies the population under investigation. Error bars represent 95% CI.

since this inference engine is not designed to reason about non-stationary observations (Yarkoni, 2022).

A practical implication of accepting this argument is that small to moderate empirical effects documented in the literature on the basis of statistically significant group differences in DDM parameters must be interpreted with caution. As a visual demonstration of the breadth of studies affected by this cautionary principle, we show a subset of such results, focusing only on drift rate-based results for succinctness, in Figure 6 (Fish et al., 2018; Huang-Pollock et al., 2017; Karalunas et al., 2014; Karalunas et al., 2018; Karalunas et al., 2012; Limongi et al., 2020; Metin et al., 2013; Moustafa et al., 2015; Poole et al., 2021; Ratcliff et al., 2001; Sripada & Weigard, 2021; Theisen et al., 2021; A. Weigard & Huang-Pollock, 2017; A. S. Weigard et al., 2021). For each study, we used the reported mean and standard deviations to measure Cohen's d between groups with statistically significant differences. Figure 6 shows the claimed results depicted by black markers. We show the effect size of drift rate measurements obtained by treating session as a condition in our analysis by solid red vertical lines. Dashed lines are used to report a similar effect size in drift rate for a recognition memory task in DDM parameters retest reliability study conducted by Lerche and Voss (2017). In view of the high unreliability of these parameters in both RDM and memory tests, only effect sizes substantially larger than the session-based sizes, can be interpreted unambiguously.

Our results may seem counter-intuitive in seeming to favor older, simpler models over newer, complex ones. Progress towards better explanations inevitably involves the use of more complex models. It is equally necessary to appreciate, as we show in this paper, that unreliability of data used to estimate model parameters fundamentally limits the granularity at which models can be discriminated based on empirical goodness-of-fit. As Almaatouq et al. (2022) have recently argued, progress in understanding requires the complexity of models and richness of data to expand together.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, 1–55.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

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

Chandrasekaran, C., & Hawkins, G. E. (2019). Chartr: An r toolbox for modeling choices and response times in decision-making tasks. *Journal of neuroscience methods*, *328*, 108432.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience*, *29*(37), 11560–11571.

Ditterich, J. (2006). Evidence for time-variant decision making. *European Journal of Neuroscience*, *24*(12), 3628–3641.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic bulletin & review*, *18*(1), 61–69.

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., et al. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic bulletin & review*, *26*(4), 1051–1069.

Fish, S., Toumaian, M., Pappa, E., Davies, T. J., Tanti, R., Saville, C. W., Theleritis, C., Economou, M., Klein, C., & Smyrnis, N. (2018). Modelling reaction time distribution of fast decision tasks in schizophrenia: Evidence for novel candidate endophenotypes. *Psychiatry research*, *269*, 212–220.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2019). Measuring the completeness of theories. *arXiv preprint arXiv:1910.07022*.

Gold, J. I., Shadlen, M. N. et al. (2007). The neural basis of decision making. *Annual review of neuroscience*, *30*(1), 535–574.

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*(6), 2476–2484.

Heathcote, A., Brown, S., & Mewhort, D. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic bulletin & review*, *9*(2), 394–401.

Huang-Pollock, C., Ratcliff, R., McKoon, G., Shapiro, Z., Weigard, A., & Galloway-Long, H. (2017). Using the diffusion model to explain cognitive deficits in attention deficit hyperactivity disorder. *Journal of abnormal child psychology*, *45*(1), 57–68.

Karalunas, S. L., Geurts, H. M., Konrad, K., Bender, S., & Nigg, J. T. (2014). Annual research review: Reaction time variability in adhd and autism spectrum disorders: Measurement and mechanisms of a proposed trans-diagnostic phenotype. *Journal of Child Psychology and Psychiatry*, *55*(6), 685–710.

Karalunas, S. L., Hawkey, E., Gustafsson, H., Miller, M., Langhorst, M., Cordova, M., Fair, D., & Nigg, J. T. (2018). Overlapping and distinct cognitive impairments in attention-deficit/hyperactivity and autism spectrum disorder without intellectual disability. *Journal of abnormal child psychology*, *46*(8), 1705–1716.

Karalunas, S. L., Huang-Pollock, C. L., & Nigg, J. T. (2012). Decomposing attention-deficit/hyperactivity disorder (adhd)-related effects in response speed and variability. *Neuropsychology*, *26*(6), 684.

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the ratcliff diffusion model. *Psychological research*, *81*(3), 629–652.

Limongi, R., Jeon, P., Mackinley, M., Das, T., Dempster, K., Théberge, J., Bartha, R., Wong, D., & Palaniyappan, L. (2020). Glutamate and dysconnection in the salience network: Neurochemical, effective connectivity, and computational evidence in schizophrenia. *Biological psychiatry*, *88*(3), 273–281.

Lu, Z.-L., & Dosher, B. (2013). *Visual psychophysics: From laboratory to theory*. MIT Press.

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). Adhd performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology*, *27*(2), 193.

Moustafa, A. A., Kéri, S., Somlai, Z., Balsdon, T., Frydecka, D., Misiak, B., & White, C. (2015). Drift diffusion model of reward and punishment learning in schizophrenia: Modeling and experimental data. *Behavioural Brain Research*, *291*, 147–154.

Pilly, P. K., & Seitz, A. R. (2009). What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision research*, *49*(13), 1599–1612.

Pisauro, M. A., Fouragnan, E., Retzler, C., & Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous eeg-fmri. *Nature communications*, *8*(1), 15808.

Poole, D., Miles, E., Gowen, E., & Poliakoff, E. (2021). Shifting attention between modalities: Revisiting the modality-shift effect in autism. *Attention, Perception, & Psychophysics*, *83*(6), 2498–2509.

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

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

Sifar, A., & Srivastava, N. (2022). Over-precise predictions cannot identify good choice models. *Computational Brain & Behavior*, 1–19.

Sripada, C., & Weigard, A. (2021). Impaired evidence accumulation as a transdiagnostic vulnerability factor in psychopathology. *Frontiers in psychiatry*, *12*, 627179.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Theisen, M., Lerche, V., von Krause, M., & Voss, A. (2021). Age differences in diffusion model parameters: A meta-analysis. *Psychological Research*, *85*(5), 2012–2021.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of neurophysiology*, *108*(11), 2912–2930.

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

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Voodoo correlations in social neuroscience. *Perspectives on psychological Science*, *4*(3), 274–290.

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An ez-diffusion model for response time and accuracy. *Psychonomic bulletin & review*, *14*(1), 3–22.

Weigard, A., & Huang-Pollock, C. (2017). The role of speed in adhd-related working memory deficits: A time-based resource-sharing and diffusion model account. *Clinical Psychological Science*, *5*(2), 195–211.

Weigard, A. S., Brislin, S. J., Cope, L. M., Hardee, J. E., Martz, M. E., Ly, A., Zucker, R. A., Sripada, C., & Heitzeg, M. M. (2021). Evidence accumulation and associated error-related brain activity as computationally-informed prospective predictors of substance use in emerging adulthood. *Psychopharmacology*, *238*(9), 2629–2644.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.