1.5em 0pt
v5.3Version 1, July 2024

Do modeling choices matter for the reliability of individual difference measures in conflict tasks?

Michelle C. Donzallaz[1], Udo Boehm[2], Andrew Heathcote[1,3], Chris Donkin[4], Dora Matzke[1], & Julia M. Haaf[5]

[1] Psychological Methods, University of Amsterdam
[2] Centrum Wiskunde & Informatica, Amsterdam
[3] University of Newcastle, Australia
[4] LMU Munich
[5] Department of Psychology, University of Potsdam

Abstract

There is a growing realization that experimental tasks that produce reliable effects in group comparisons can simultaneously provide unreliable assessments of individual differences. Proposed solutions to this "reliability paradox" range from collecting more test trials to modifying the tasks and/or the way in which effects are measured from these tasks. Here we systematically compare two proposed modeling solutions in a cognitive conflict task. Using the ratio of individual variability of the conflict effect (i.e., signal) and the trial-by-trial variation in the data (i.e., noise) obtained from Bayesian hierarchical modeling, we examine whether improving statistical modeling may improve the reliability of individual differences assessment in four Stroop datasets. The proposed improvements are 1) increasing the descriptive adequacy of the statistical models from which conflict effects are derived, and 2) using psychologically-motivated measures from cognitive models. Our results show that modeling choices do not have a consistent effect on the signal-to-noise ratio: the proposed solutions improved reliability in only one of the four datasets. We provide analytical and simulation-based approaches to compute the signal-to-noise ratio for a range of models of varying sophistication and discuss their potential to aid in developing and comparing new measurement solutions to the reliability paradox.

Keywords: reliability, cognitive conflict, Bayesian hierarchical modeling

Cognitive conflict is commonly assessed using experimental paradigms by comparing conditions requiring higher versus lower levels of cognitive control. One such paradigm is the classic color-Stroop task (Stroop, 1935), in which participants indicate the color in which a color-word is printed. The Stroop effect is conventionally measured by the response time (RT) difference between trials in which the printed color and the meaning of the word are different (i.e., incongruent; e.g., "red" written in blue) and trials in which they match (i.e., congruent; e.g., "red" written in red). Similar cognitive conflict paradigms with a congruency manipulation are the Flanker (Eriksen & Eriksen, 1974) and the Simon task (Simon & Rudell, 1967). In all of them, the incongruent condition requires more control to inhibit automatic associations between the wrong response and irrelevant information (e.g., color). Positive conflict effects, that is the RT difference between the incongruent and the congruent conditions, provide an index of cognitive control necessary to manage the conflict (Hommel, 2011; MacLeod, 1991; Ridderinkhof, Wylie, van den Wildenberg, & et al., 2021).

Conflict effects are highly robust at the population-level – in fact, the Stroop effect is even considered universal (Haaf & Rouder, 2017; MacLeod, 1991). At the individual-level, however, the picture is more complicated. While almost everyone seems to show a conflict effect, true individual differences are assessed with a lot of uncertainty and are masked by measurement error. This situation has been labeled the "reliability paradox" (Hedge, Powell, & Sumner, 2018): experimental control, which has the desirable effect of increasing validity, also tends to reduce individual differences, and hence decreases reliability (Keye, Wilhelm, Oberauer, & et al., 2009; Paap & Greenberg, 2013; Pettigrew & Martin, 2014; Rey-Mermet, Gade, & Oberauer, 2018).

To better understand reliability in experimental tasks, consider typical Stroop task data from I participants with two conditions (i.e., congruent and incongruent trials), and K trials per condition. For illustrative purposes, we use the simplest model possible, the normal-normal model. Individual i's mean RT differences between the conditions follow a normal population-level distribution,

$$\overline{Y}_{i\text{inc}} - \overline{Y}_{i\text{con}} \sim \text{Normal}(\mu_\theta, \sigma_\theta^2 + \frac{2\sigma^2}{K}),$$

with mean μ_θ and variance $\sigma_\theta^2 + \frac{2\sigma^2}{K}$, where μ_θ is the population mean of the conflict effect, σ_θ the true individual variation and σ the measurement error[1]. By "true individual variation", we mean the extent to which people differ quantitatively in their conflict effect if we had an infinite number of observations. By measurement error, we mean the confounding variability that exists because participants' RTs vary across trials, for example due to motor and perceptual processes. Using hierarchical modeling, one can estimate both σ_θ and σ as well as the population mean μ_θ.

Reliability is commonly defined as the ratio between true between-subjects variance σ_θ^2 and total variance, where the latter is the sum of true between-subjects variance in the effect of interest and an error term:

$$r = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_E^2}.$$

[1] In this paper, we use the terms "variability" and "variation" interchangeably to refer to both variance and standard deviation (i.e., the square root of the variance)

MCD, AH, and DM are supported by a Vidi grant to DM (VI.Vidi.191.091) from the Dutch Research Council (NWO). UB is supported by a Veni grant (VI.Veni.201G.045) from NWO. JMH is supported by a Veni grant (VI.Veni.201G.019) from NWO. AH is supported by the Australia-US Multidisciplinary University Research Initiative (AUSMURIV000003). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5776.

The authors have no relevant financial or non-financial interests to disclose.

We would like to thank Suzanne Hoogeveen, Michael Nunez, Jeff Rouder, and Niek Stevenson for valuable discussions.

Correspondence concerning this article: m.c.donzallaz@uva.nl.

In experimental tasks with two main conditions of interest, such as the Stroop task, the error term is twice the squared standard error (see Rouder & Haaf, 2019, for more details),

$$\sigma_E^2 = \frac{2\sigma^2}{K},$$

where K is the number of trials performed per condition. Therefore, the reliability term becomes

$$r = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \frac{2\sigma^2}{K}}.$$

Rouder and Haaf (2019) pointed out that to understand the reliability paradox it is important to acknowledge that reliability is not ”portable”: even for otherwise identical tasks, reliability differs with K. This is due the fact that as K grows, the standard error decreases and the reliability increases. Rouder, Kumar, and Haaf (2019) proposed that the suitability of a conflict task for individual-differences research instead be evaluated in terms of the signal-to-noise ratio (SNR)

$$\gamma = \frac{\sigma_\theta}{\sigma},$$

where the signal is the individual variability of the conflict effect, σ_θ, and the noise is the trial-to-trial variability in RT, σ. Using hierarchical modeling Rouder et al. (2019) found that the ratio, γ, was around 1/7 across a large set of conflict tasks, leading them to conclude that for conventional conflict tasks, $K > 500$ (i.e., over 1000 trials per participant) is required to obtain sufficiently reliable measurement.

Low reliability, and the impracticality of collecting that many trials, has spurred research to develop more reliable ways of measuring cognitive control in general and cognitive conflict in particular. One direction has been to abandon conflict measures based on RT differences, as taking a difference doubles the measurement error (Draheim, Mashburn, Martin, & Engle, 2019). Instead, it has been suggested that cognitive control should be assessed using accuracy-based, adaptive tasks or the number of correct (in)congruent trials performed within a fixed time frame, with confounding from overall speed partialed out. However, some of these approaches sacrifice the validity of RT difference scores, which is evidenced by a large body of experimental literature, and risk confounding from individual differences in processing speed (Kucina et al., 2023; Rouder et al., 2019).

Kucina et al. (2023) used an alternative approach to improve reliability. They used the SNR γ to compare, combine, and refine traditional conflict tasks. They also refined the RT difference measure itself and took into account that RT distributions are well described by a shifted log-normal distribution (Heathcote & Love, 2012), improving the descriptive adequacy of the model used. They reported SNRs for conflict effects in the range of 1/3 to 1/2, corresponding to classic reliability values of 0.8 or higher, with only 100 trials. In related work, Haines et al. (2020) found that modeling RT as lognormally distributed improved test-retest reliability in several conflict datasets (Hedge, Powell, & Sumner, 2018), a delay-discounting task (Gawronski, Morrison, Phills, & Galdi, 2017), and an Implicit Association Test (Ahn et al., 2020). More generally, Haines et al. (2020) argued that simple atheoretical summaries, such as conflict effects based on mean RT, are unable to validly characterize the underlying psychological processes and recommended a “generative modeling approach“ instead to overcome a “theory-description gap”. Consistent with this recommendation, Hedge, Powell, Bompas, Vivian-Griffiths, and Sumner (2018) demonstrated that it can be important to simultaneously account for accuracy and RT because inconsistent relationships between accuracy-based and RT-based effects are widespread. Individual differences in this speed-accuracy trade-off may potentially contribute to the unreliability of RT-based conflict effects. Evidence-accumulation models can disentangle these confounding processes (Hedge, Powell, Bompas, et al., 2018).

In summary, the reliability paradox has triggered a wide range of responses in the community. The proposed solutions range from refining and reinventing traditional tasks and measures, to advocating for theoretically-motivated modeling approaches that provide psychologically meaningful measures of cognitive conflict and a more complete and descriptively more accurate account of choice RT performance. Here we investigate the degree to which different modeling choices influence the SNR.

Comparing the Reliability of Model-Based Conflict Effects

Does improving the descriptive adequacy of the model from which conflict effects are derived improve the SNR? Can a cognitive model that provides a psychologically meaningful characterization of performance detect relatively more signal in a given dataset than a purely statistical model? And to what extent does this depend on the dataset? We will use the SNR of conflict effects as measured by variants of the Stroop task to answer these questions.

Rouder and Mehrvarz (2024) suggested that the SNR can be used as an indicator for reliability that is independent of the trial size. To see the relationship between reliability and SNR, note that reliability can be re-expressed in terms of the ratio γ,

$$r = \frac{\gamma^2}{\gamma^2 + \frac{2}{K}}.$$

In theory, the independence of the number of trials makes γ a portable indicator of reliability (Rouder & Haaf, 2019). In practice, however, the SNR does not always prove to be portable. Kucina et al. (2023) modeled the possibility that individual differences in conflict effect and trial-to-trial variability changes with K and computed the SNR as a function of the number of trials. After de-trending to account for decreased mean RT with practice (Evans, Brown, Mewhort, & Heathcote, 2018; Heathcote, Brown, & Mewhort, 2000), their results indicated that the between-subjects variability σ_θ decreased as more trials were performed (i.e., individual differences decreased with practice). Although this result raises doubts about the portability of γ, they showed that the SNR can be useful for comparing reliability among tasks for a fixed K. Whether the SNR is portable or not, however, does not affect its usefulness for our current purpose: comparing reliability among models of the same task for a fixed K.

We investigate the effects of closing the "theory-description gap" (Haines et al., 2020) on the SNR by progressively improving the descriptive adequacy and theoretical underpinning of the models used to quantify the conflict effect. We examine how reliably various models can detect individual differences in four Stroop datasets: we move from purely statistical models, such as the normal or (shifted) lognormal distribution that provide a descriptive characterization of RT distributions, to cognitive models, such as the lognormal race and the racing diffusion models that simultaneously account for choice accuracy and RTs and that provide a psychologically meaningful account of performance. We show how the SNR can be computed analytically for these five models and outline a simulation-based approach that is more broadly applicable to any model[2].

The main argument for more complex statistical models of the conflict effect is that descriptive adequacy may increase the SNR (Haines et al., 2020). A better fit means more accurate measurement of individual differences and the noise term, the components of the SNR. Additionally, cognitive models may further increase the SNR. By adding psychologically meaningful parameters, some variability that is attributed to noise in statistical models may be explained by cognitive

[2] Any model with unbiased (e.g., maximum likelihood) estimators

models. Therefore, cognitive modeling may increase the SNR by decreasing its denominator, the noise term, which is critical for reliable assessment of individual differences in Stroop effects.

Method

From the Normal Model to the Racing Diffusion Model

The five models employed are illustrated in Figure 1. The first three are purely RT-based Bayesian hierarchical models that were also used by Haines et al. (2020). They are shown in increasing order of descriptive adequacy: a normal model, which makes the unrealistic assumption that RTs are normally distributed and assigns probability to negative RTs, followed by two models that account for the characteristic right skew of empirical RT distribution by assuming a lognormal parametrization, one with and one without an estimated shift parameter. The shift parameter ensures that the lower bound of the RT distribution is shifted away from zero, excluding unrealistically fast RTs.

We then use two psychologically-grounded evidence-accumulation models, the lognormal race model (LNR; Heathcote & Love, 2012) and the racing diffusion model (RDM; Tillman, Van Zandt, & Logan, 2020), both providing a comprehensive account of performance by simultaneously accounting for response choices and the corresponding RTs (see also Matzke, Logan, & Heathcote, 2020). Racing evidence-accumulation models have a long history in psychology because they provide a principled yet flexible approach to describe and explain performance in a broad range of tasks (Heathcote & Matzke, 2022). Both the LNR and the RDM assume that decisions are made by a process of accumulating evidence until a threshold amount is reached, which then triggers a response. In particular, the models assume an independent race between a set of evidence accumulators, corresponding to the different response options. In terms of the color-Stroop task, this means separate racers for each response option (e.g., left button for "green" and right button for "red") that race against each other. The racer that wins determines the response. Individuals may start accumulating evidence at different start points and/or set different thresholds to vary the trade-off between the speed of their responses and the accuracy of their choices (Donkin & Brown, 2018). For example, a smaller distance from the starting point of evidence accumulation to the response threshold leads to faster but less accurate responses. Conversely, larger distances lead to slower but more accurate responses. Similar to the shifted lognormal model, the cognitive models also feature a ”non-decision time” parameter that accounts for time spent to encode evidence from the choice stimulus and to produce a response. This parameter shifts the finishing time distributions and ensures that the lower bound of the RT distributions is greater than zero.

The LNR (Heathcote & Love, 2012) was chosen because it forms a natural extension of the RT-only models by assuming a lognormal parameterization but also accounting for choice accuracy. The model assumes that the rate of evidence accumulation and the distance between the start point and threshold follow independent lognormal distributions. As a result, the time for a single accumulator to reach its threshold (i.e., the finishing time) also follows a lognormal distribution. When accuracy is perfect, the observed RT distribution follows a shifted lognormal distribution, consistent with the use of this distribution by Haines et al. (2020). However, when performance is not perfect, as is typically the case for Stroop data (see Table 1 for average error rates in the four data sets we examine), the only RT-based shifted lognormal model is mis-specified. Here we use the LNR model to assess whether the SNR improves when accuracy is also taken into account.

A limitation of the LNR model is that the rate of evidence accumulation and the distance from starting point to threshold are not separately identified because they combine linearly. As a result, the conflict effect of the LNR confounds individual differences in response cuation and the effects of congruency on the rate of evidence accumulation. The RDM (Tillman et al., 2020) addresses this challenge to validity by allowing us to separately estimate evidence-accumulation rates and response thresholds.

The RDM (Tillman et al., 2020) assumes that each of the accumulators is a Wiener diffusion process with an evidence-accumulation rate v, a starting point of 0, and a response threshold β. Trial-to-trial variability in behavior is caused by stochastic accumulation, that is, during accumulation, the evidence-accumulation rate changes from moment to moment due to the addition of fractional Gaussian noise. The finishing time distribution of each accumulator is an inverse Gaussian distribution (i.e., Wald distribution; Wald, 1947). Here we use the RDM to investigate whether using psychologically-grounded measures of cognitive conflict as captured in the evidence-accumulation rate parameter improves the SNR.

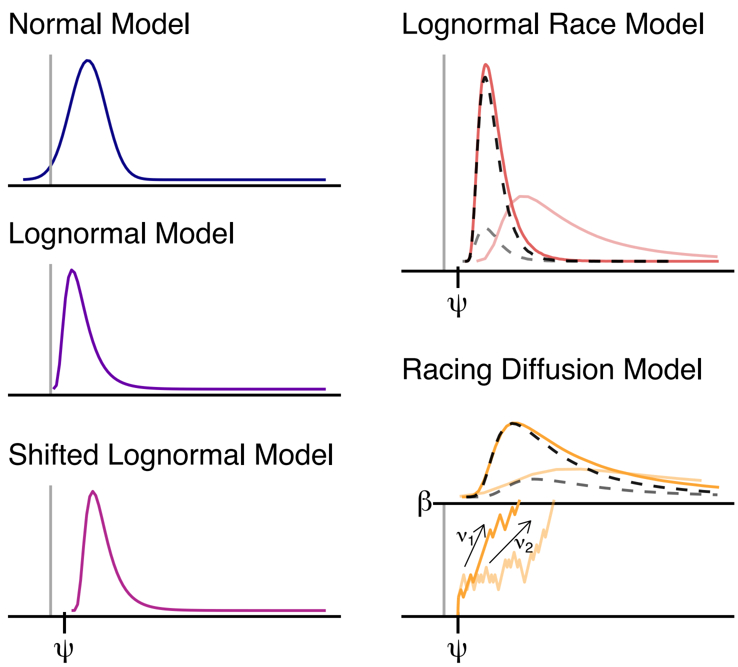

Figure 1. Model overview. The plots show exemplary shapes of the corresponding response and finishing time distributions as assumed by the respective models. The lightgrey vertical line indicates response time (RT) = 0 on the x-axis. The **normal model** crosses this line, demonstrating that the normal model also allows for negative RTs. The **lognormal model** accounts for the fact that RTs can only be positive and that their distribution is typically skewed. The **shifted lognormal model** additionally accounts for the fact that RT distributions are shifted away from zero, whereby ψ indicates the shift parameter. The **lognormal race model** (LNR) additionally accounts for choice errors. The model conceptualizes decision-making as an independent race between evidence accumulators, one for each response option, until a threshold is reached which then triggers a response. Depicted are the finishing time distributions of the matching and mismatching accumulators (i.e., stimulus and response (mis)match). The dashed lines show observed distributions of the respective racers (i.e., those that won the race), scaled by their winning proportions. ψ indicates the non-decision time parameter. The **racing diffusion model** (RDM) separately estimates the evidence-accumulation rate and the threshold parameters. The yellow paths depict an exemplary race between the matching and mismatching accumulators. The matching accumulator (darker yellow) reaches the threshold β first, resulting in a correct response on that trial. ν_1 and ν_2 refer to the matching and matching evidence-accumulation rate parameters.

We deliberately chose to explore the performance of relatively standard evidence-accumulation models as opposed to more complex models specifically developed for conflict tasks. Although conflict models such as the spotlight diffusion model (White, Ratcliff, & Starns, 2011) align more closely with the cognitive processes assumed to underlie performance in these tasks, they are limited in terms of tractability: the models grapple with pronounced parameter trade-offs and hierarchical estimation is not straightforward. As a result, despite their theoretical merits, they are less suitable as measurement models to examine individual differences in conflict tasks.

The full Bayesian hierarchical model specification can be found in the supplementary materials. The three purely RT-based models, the normal, the lognormal, and the shifted lognormal models, are parameterized such that the (log) means are decomposed into an intercept α_i and a conflict or congruency effect θ_i for individual i. We allow the (log) standard deviation to vary across individuals, σ_i, but not across conditions (i.e., congruent vs. incongruent). For the shifted lognormal model, we also estimate individual shift parameters ψ_i which remain the same across conditions. The LNR is parameterized such that the log mean of the accumulator that *matches* the stimulus is decomposed into an intercept α_{1i} and a conflict effect θ_i. The corresponding log standard deviation, σ_{1i}, is allowed to vary across individuals but again not across conditions. Due to high accuracy rates in our analyzed datasets (see Table 1), we fix the log mean and log standard deviation of the accumulator that mismatches the stimulus across individuals and estimate an intercept only and no conflict effect (i.e., we estimate α_2 and σ_2). Just like in the shifted lognormal model, we also estimate a non-decision time parameter for each individual, ψ_i. Lastly, for the RDM, we decompose the matching evidence-accumulation rate into an intercept α_{1i} and a conflict effect θ_i. Note that we only model a conflict effect on the rate because we expect that the cognitive conflict will impact the evidence accumulation rate at which participants respond to the (in)congruent stimuli. In contrast, the individual response thresholds are set at the beginning of a trial and we do not expect stimulus encoding or motor control processed to differ across conditions. Similar to the LNR specification, we estimate an intercept only and do not allow for individual differences in the mismatching evidence-accumulation rate, α_2 (see Lüken, Heathcote, Haaf, & Matzke, 2023). Furthermore, we estimate individual response thresholds, β_i, and non-decision time parameters, ψ_i, which were fixed across congruency conditions and accumulators.

Estimating Signal-to-Noise Ratios

All models have a designated parameter reflecting the numerator of the SNR, σ_θ (i.e., individual differences in the conflict effect). However, this is not the case for its denominator. To determine the trial-by-trial variation, we estimate the standard error of the conflict effect. Note that in the case of the normal model, the standard error is simply $\frac{2\sigma^2}{K}$ and that σ can be directly estimated using hierarchical modeling and interpreted on the real line (see Rouder & Haaf, 2019; Rouder et al., 2019).

For the normal model, we could use the formula provided by Rouder and Mehrvarz (2024) to obtain a point estimate of γ from RT data. However, we prefer to use Bayesian hierarchical modeling (Lee, 2011) for all models for several reasons. First, hierarchical models partition the variability in data into between and within-participant components. Second, Bayesian estimation naturally provides measures of uncertainty in γ estimates, providing a basis for inference about the SNR. Third, Bayesian hierarchical models are well suited for estimating the parameters of evidence-accumulation models such as the LNR and RDM, as well as the shift parameter of the lognormal distribution when it is estimated (e.g., Heathcote et al., 2019; Stevenson et al., 2024).

Hierarchical modeling typically directly provides an estimate of the between-subjects variability of interest, σ_θ for all models. The models just need to have parameters reflecting the difference between the congruent and incongruent conditions (i.e., θ_i). For all but the RDM, the individual noise parameters are also directly provided (i.e., σ_i for the normal and the three lognormal models): σ_i, the shape parameter of the lognormal distribution and standard deviation of the log-RTs, directly reflects the individual trial-by-trial variation. Moreover, within models, σ_θ and σ_i are on the same scale (i.e., on the real line in the normal model and on the log scale in the three lognormal models). Therefore, the units cancel out, making the ratios comparable to each other. Since we assume individual noise parameters, we need to average estimates of σ_i across individuals, $i = 1, ..., I$:

$$\gamma = \frac{\hat{\sigma}_\theta}{\sqrt{\frac{1}{I}\sum_{i=1}^{I}\hat{\sigma}_i^2}} \tag{1}$$

Finding an expression for the noise term σ_i in the RDM is more complicated. Contrary to the other models, trial-by-trial variation is not explicitly represented as a model parameter. Besides the hierarchical structure that we place on top of the model parameters, our specification of the RDM has one inherent source of variability: within-trial variation of the drift rate, which is described by a Wiener diffusion process (Tillman et al., 2020). However, this parameter reflects variation *within* trials and is conventionally set to 1 to make the model identifiable. To get an estimate of between-trial variation or measurement error, the standard error of the conflict effect needs to be determined. Let $\hat{\nu}_{i_{con}}$ and $\hat{\nu}_{i_{inc}}$ be individual i's matching evidence-accumulation rate estimates in the congruent and incongruent condition such that $\Delta_{\hat{\nu}_i} = \hat{\nu}_{i_{con}} - \hat{\nu}_{i_{inc}}$ represents the conflict effect. To get an expression for the noise, we need to compute the variance $V(\Delta_{\hat{\nu}_i})$. An expression for this variance can be derived as follows:

$$\begin{aligned}
\mathrm{V}(\hat{\nu}_{i_{con}} - \hat{\nu}_{i_{inc}}) &= \mathrm{V}(\hat{\nu}_{i_{con}}) + \mathrm{V}(\hat{\nu}_{i_{inc}}) - 2\mathrm{Cov}(\hat{\nu}_{i_{con}}, \hat{\nu}_{i_{inc}}) \\
\mathrm{V}(\hat{v}) &= (\mathrm{E}(\hat{v}))^2 - \mathrm{E}(\hat{v}^2) \\
\mathrm{E}(\hat{v}) &= \sqrt{K}\frac{\Gamma(K-3/2)}{\Gamma(K-1)}\left(\frac{1}{K\beta} + v\right) \\
\mathrm{E}(\hat{v}^2) &= \frac{K}{K-2}\left(v^2 + \frac{3}{K^2\beta^2} + \frac{3v}{K\beta}\right) \\
\mathrm{E}(\hat{v}_1\hat{v}_2) &= \frac{K}{K-2}\left(\frac{1}{K\beta} + \nu_1\right)\left(\frac{1}{K\beta} + \nu_2\right) \\
\mathrm{Cov}(\hat{v}_1, \hat{v}_2) &= \mathrm{E}(\hat{v}_1\hat{v}_2) - \mathrm{E}(\hat{v}_1)\mathrm{E}(\hat{v}_2)
\end{aligned}$$

Note that Γ is the gamma function. The proof is provided in the supplementary materials. This effect variance reflects the standard error $\frac{\sqrt{2}\sigma_i}{\sqrt{K_i}}$, which can be transformed to get an estimate of the individual noise terms σ_i: $\sqrt{\frac{V(\Delta_{\hat{\nu}_i})K_i}{2}}$ where K_i is the number of trials per condition that individual i performed. Averaging across individuals, the SNR for the RDM is:

$$\gamma = \frac{\hat{\sigma}_\theta}{\sqrt{\frac{1}{I}\sum_{i=1}^{I}\frac{V(\Delta_{\hat{\nu}_i})K_i}{2}}}. \tag{2}$$

In order to capture uncertainty in γ estimates, we calculate the ratio for all available posterior samples of the individual RDM parameters.

The standard error of the effect of interest within a model can also be computed using a simulation-based approach. The details are provided in the supplementary materials. The advantage of the simulation-based approach is that it is more broadly applicable than an analytical expression which is specific to the particular model. However, the simulation-based approach

Table 1
Descriptive statistics (mean RT and accuracy), number of participants (I), number of trials per congruency condition (K), and task type.

	RT (SD)	*Accuracy*	*I*	*K*	*Task type*
Enkavi et al. (2019)	.73 (.21)	.96	522	48	Color Stroop
Pratte et al. (2010)	.77 (.37)	.96	38	164	Color Stroop
Rey-Mermet et al. (2018)	.59 (.15)	.96	128	95	Number Stroop
Von Bastian et al. (2016)	.75 (.25)	.97	121	48	Number Stroop

requires an unbiased maximum-likelihood estimator for the quantity of interest, which may not always be available. For the RDM, this quantity is the evidence-accumulation rate, which has such a maximum-likelihood estimator (i.e., the natural estimator; see supplementary materials).[3]

Results

We applied the five models to four datasets: number-Stroop tasks by (Rey-Mermet et al., 2018, younger age group) and by Von Bastian, Souza, and Gade (2016), color-Stroop tasks by Pratte, Rouder, Morey, and Feng (2010) and by Enkavi et al. (2019). We only analyzed the (in)congruent conditions and retrieved all datasets from the preprocessed database made available by Haaf, Hoffstadt, and Lesche (2024).

Before estimation, we removed the first three trials of every block, neutral trials, and RTs <= 0.25 seconds. Two studies (Enkavi et al., 2019; Rey-Mermet et al., 2018), used response windows to ensure fast responding. This approach lead to somewhat bimodal RT distributions and convergence problems for some models. We therefore removed very slow responses, specifically responses that were at the upper limit of the response window (Rey-Mermet et al., 2018: RT >= 1.9 s;, Enkavi et al., 2019: RT > 1.5 s). Pratte et al. (2010) did not have a response window, but the very long RTs caused some convergence problems, so we excluded them (RT >= 3 s). We also excluded participants with fewer than 50% remaining trials. Overall, only few trials and participants were excluded (von Bastian et al., 2016: 0.9%, Rey-Mermet et al., 2018: 5% of the observations including all observations from one participant due to having less than 50% trials after data cleaning; Pratte et al., 2010: 5%; Enkavi et al., 2019: 4% of the observations including all observations from one participant due to having less than 50% trials after data cleaning). For the fitting of the (shifted log)normal models, we also removed all incorrect trials.

We estimated all models using stan(Stan Development Team, 2023)[4] using weakly informative priors based on prior predictive checks and known plausible ranges of RTs in seconds[5]. The descriptive adequacy of the normal model was consistently worse than for the other four models

[3]There exists also a straightforward method-of-moments estimator for the evidence-accumulation rate, but this is not a maximum-likelihood estimator and is biased (see the supplementary materials).

[4]We used the following R packages: R (Version 4.2.1; R Core Team, 2022) and the R-packages *DBI* (Version 1.1.3; R Special Interest Group on Databases (R-SIG-DB), Wickham, & Müller, 2022), *dplyr* (Version 1.1.3; Wickham, François, Henry, Müller, & Vaughan, 2023), *extraDistr* (Version 1.9.1; Wolodzko, 2020), *ggplot2* (Version 3.4.3; Wickham, 2016), *papaja* (Version 0.1.1; Aust & Barth, 2022), *patchwork* (Version 1.1.3; Pedersen, 2023), *RSQLite* (Version 2.3.1; Müller, Wickham, James, & Falcon, 2023), *rstan* (Version 2.21.5; Stan Development Team, 2022), *scales* (Version 1.2.1; Wickham & Seidel, 2022), *StanHeaders* (Version 2.21.0.7; Stan Development Team, 2020), *stringr* (Version 1.5.0; Wickham, 2022), *tidybayes* (Version 3.0.2; Kay, 2022), *tidyr* (Version 1.3.0; Wickham, Vaughan, & Girlich, 2023), and *tinylabels* (Version 0.2.3; Barth, 2022).

[5]See the supplementary materials for the prior specification, estimation results, and the detailed posterior predictive assessment of the absolute descriptive adequacy of the models (i.e., the match between the observed response time distributions and the posterior predictions).

because the normal model cannot account for the characteristic skewness of RT data (see Figure 2). The shifted lognormal, the LNR, and the RDM showed good descriptive adequacy, however the RDM and the LNR both had some difficulties capturing slow errors in the incongruent condition, particularly for the dataset by (Rey-Mermet et al., 2018, see the Appendix and supplementary materials). To quantify descriptive adequacy, we computed the root mean squared error (RMSE) for the correct RTs and in the case of the LNR and RDM, additionally for the proportion of correct responses[6]. We simulated 500 datasets from the posterior predictive distributions and computed the RMSE for each of the simulated datasets and several quantiles (see Table 2 and 3 in the Appendix). For most datasets, RMSE steadily decreased for the first three models and the RMSE for the LNR tended to be smaller than for the RDM.

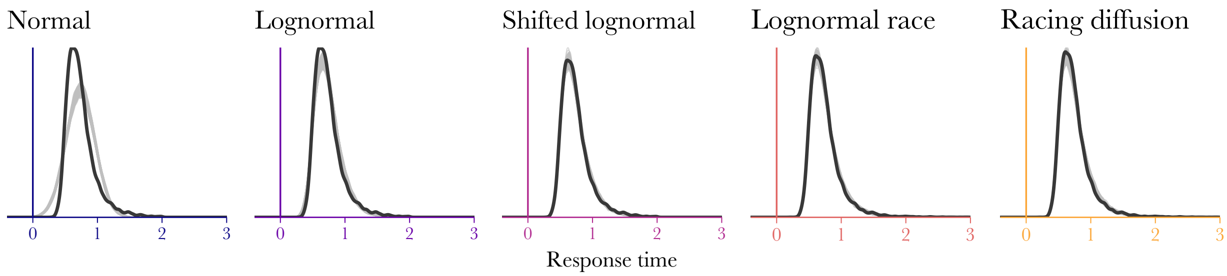

Figure 2. Observed (black) and predicted (grey) correct response time distributions collapsed across participants and conditions for the dataset by Von Bastian et al. (2016). The grey lines represent 500 samples from the posterior predictive distribution. The lognormal race and the racing diffusion model distributions also contain the incorrect responses.

We then computed the SNRs by plugging the parameter posterior samples into Equations (1) and (2) (see Figure 3). Note that the simulation-based approach leads to approximately the same result[7]. Within datasets, the credible intervals overlap substantially, suggesting that the choice of the model does not substantially influence the ratio in the analyzed datasets. The only exception is the Von Bastian et al. (2016) dataset, where the posterior median increases from below $\frac{1}{10}$ for the normal model to around $\frac{1}{5}$ for the RDM. Notably, the SNRs of the dataset by Enkavi et al. (2019) are considerably higher than those of the other datasets.

[6]We did not use model selection criteria such as the deviance information criterion (DIC) due to the fact that the models are not comparable as they were fitted to different data (i.e., the (shifted log)normal models were fitted to correct responses only).

[7]see the supplementary materials for the comparisons

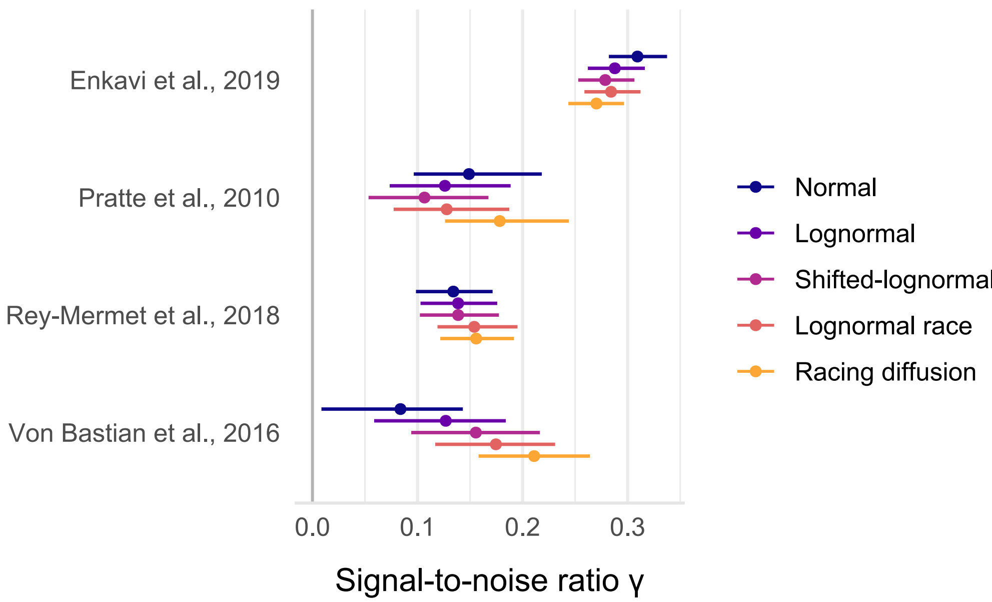

Figure 3. Posterior medians and 95% credible interval of the signal-to-noise ratios γ as computed by the analytical method.

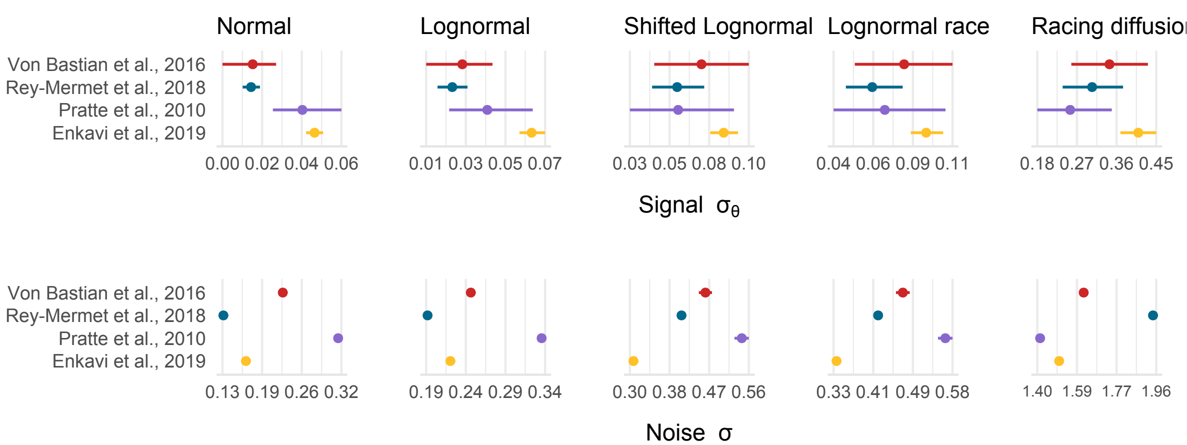

Figure 4. Posterior medians and 95% credible intervals of the numerator (i.e., signal) and denominator (i.e., noise) of the signal-to-noise ratios.

Looking at the components of the SNRs separately (Figure 4), the ratio differences across datasets and within models seem to be driven by both the extent of individual differences in the Stroop effect (i.e., signal) and by trial-by-trial variation (i.e., noise). Note that the credible interval (CrI) of the noise term are narrower than those of the signal term because the former is estimated from all observations. In contrast, the CrI of σ_θ differ in width across datasets as more participants leads to more precise estimation (see Table 1). Across models, the order of the terms stays largely the same – with two exceptions. (1) Moving from the normal model to the RDM for the data by Von Bastian et al. (2016), we gain relatively more signal, though note that the scales of the signal and the noise terms are not directly comparable. Within the components of the other datasets, there does not appear to be a dominating signal or noise term across models, which is not surprising given that the CrI of the SNRs are overlapping.

Discussion

Experimental tasks that produce robust group-level differences can simultaneously produce unreliable individual differences. Proposed solutions to this "reliability paradox" (Hedge, Powell, & Sumner, 2018) include collecting more trials per participant, modifying the cognitive tasks, using more descriptively adequate models, and using psychologically-motivated measures derived from cognitive models to quantify individual differences in the construct of interest. Here we focused on the last two solutions, and built on the work by Rouder and Mehrvarz (2024) and Rouder et al. (2019) by using signal-to-noise ratios (SNRs) as a tool for comparing models of varying sophistication. The SNR is a useful measure of reliability quantifying how well individual differences can be detected relative to measurement error. We provided analytical solutions to compute the ratio for five models from the output of Bayesian hierarchical modeling: the normal, lognormal, shifted lognormal, lognormal race (LNR), and racing diffusion models (RDM). Moreover, we provided a general algorithm that - whenever an unbiased estimator of the quantity of interest is available - can be used to compute the SNR for any model. We then applied the five models to four Stroop datasets.

Our analysis showed that modeling choices do not have a consistent effect on the SNR: the proposed solutions improved reliability in only one of the four datasets, and we found no consistent ordering or pattern as to which model yielded the highest SNR across datasets. Notably, improved descriptive accuracy did not correspond to a higher SNR. The normal model is clearly unable to accommodate the slow tail of typical RT distributions, yet the corresponding SNRs are not consistently worse than those of the RDM or lognormal models that can account for the skewness in observed RTs. These differences did not appear to be explained by an effect of trial number or sample size. Perhaps a more systematic assessment of more datasets might help identify dataset or task characteristics that are predictive of higher SNRs for one model over another. As of now, the preferred model can only be established after fitting all models using hierarchical modeling and computing the SNRs.

In our analyses, we only looked at a subset of potential models. For example, other relevant models are the diffusion decision model (Ratcliff & McKoon, 2008; Ratcliff & Smith, 2004) or the linear ballistic accumulator model (Heathcote & Love, 2012). However, our approach can be extended to these models. In both cases the trial-by-trial variation is not directly reflected in a model parameter. Yet, there are two ways of computing the measurement error term of the SNR: (1) one can try to derive it analytically by computing the variance of the effect of interest or standard error, as we have done for the RDM. (2) one can use a simulation-based approach if unbiased maximum likelihood estimators are available.

The focus of this paper is the reliability of measures. However, the validity of a measure (i.e., its ability to measure the intended construct) is also important. Cognitive model parameters have a clear psychological interpretation and so intrinsically support valid inferences as long as the model accurately represents the data and the process that generates it. So, even if a particular model parameter might not lead to a higher SNR, it might still be a better measure. For instance, suppose we analyze a dataset in which speed-accuracy trade-offs are present. Because the LNR or RDM can account for such trade-offs by combining information about accuracy and speed they are likely to provide more valid measures. However, estimating a model providing a valid psychological account may also require particular qualities in the data and the design from which it comes. For example, Lüken et al. (2023) showed that low error rates compromised the quality of estimates of the parameters of the diffusion decision model and the linear ballistic accumulator. It is possible that the low error rates in the datasets we analyzed here may have have had the same effect on the LNR and RDM, and that is the reason why they did not consistently perform better than the rest.

Clearly, both reliability and validity need to be taken into account whenever researchers attempt to answer substantive questions about the nature of individual differences in cognitive control. Descriptive adequacy is also important, as a model that clearly misfits the data is unlikely to be valid. However, good fit alone does not ensure validity, the model must also provide a sensible account of psychological processes that could plausibly generate the data, preferably one that is backed up by converging evidence from prior literature (i.e., the same type of model has provided accurate and coherent accounts of data from related tasks and manipulations).

In sum, the SNR can be used as a tool to identify the statistical or cognitive model that is best suited to examine individual differences in conflict data. We have provided analytical and simulation-based approaches to compute the SNR for a range of models of varying sophistication and showed that models that provide a better, and potentially more valid, description do not necessarily consistently improve the reliability with which individual differences in cognitive conflict are measured. Hence, we recommend that our methodology be deployed on a case-by-case basis to assess the effects of model choices on reliability.

Declarations

Funding: MCD, AH, and DM are supported by a Vidi grant to DM (VI.Vidi.191.091) from the Dutch Research Council (NWO). UB is supported by a Veni grant (VI.Veni.201G.045) from NWO. JMH is supported by a Veni grant (VI.Veni.201G.019) from NWO. AH is supported by the Australia-US Multidisciplinary University Research Initiative (AUSMURIV000003). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5776.

Conflicts of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The archival data used in this article is available at `https://osf.io/fq8ep`

Code availability (software application or custom code): The code is available at `https://osf.io/fq8ep`

Authors' contributions: MCD: conceptualization, methodology, formal analysis, writing and editing, visualization; UB: methodology, writing and editing, visualization; AH: conceptualization, methodology, writing and editing, supervision CD: conceptualization, editing, supervision; DM: conceptualization, methodology, writing and editing, supervision, funding acquisition; JMH: conceptualization, methodology, writing and editing, supervision.

References

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., ... Pitt, M. A. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports*, *10*(1), 12091. doi: 10.1038/s41598-020-68587-x

Donkin, C., & Brown, S. D. (2018). Response times and decision-making. In E.-J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience: Vol. 5. methodology* (4th ed., pp. 349–382). John Wiley & Sons.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives. *Psychological Bulletin*, *145*(5), 508–535. doi: 10.1037/bul0000192

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. doi: 10.1073/pnas.1818430116

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149. doi: 10.3758/BF03203267

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the Law of Practice. *Psychological Review*, *125*(4), 592–605. doi: 10.1037/rev0000105

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. doi: 10.1177/0146167216684131

Haaf, J. M., Hoffstadt, M., & Lesche, S. (2024). *Attentional control data collection: A resource for efficient data reuse.* `https://doi.org/10.31234/osf.io/4evy6`.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. doi: 10.1037/met0000156

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., ... Turner, B. (2020). Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences.
doi: 10.31234/osf.io/xr7y3

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185 – 207. doi: 10.3758/BF03212979

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior Research Methods*, *51*, 961–985. doi: 10.3758/s13428-018-1067-y

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in psychology*, *3*. doi: 10.3389/fpsyg.2012.00292/abstract

Heathcote, A., & Matzke, D. (2022). Winner takes all! what are race models, and why and how should psychologists use them? *Current Directions in Psychological Science*, *31*(5), 383–394. doi: 10.1177/09637214221095852

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and Variable Correlation Between Reaction Time Costs and Accuracy Costs Explained by Accumulation Models: Meta-Analysis and Simulations. *Psychological Bulletin*, *144*(11), 1200–1227. doi: 10.1037/bul0000164

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. doi: 10.3758/s13428-017-0935-1

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189–202. doi: 10.1016/j.actpsy.2010.04.011

Keye, D., Wilhelm, O., Oberauer, K., & et al. (2009). Individual differences in conflict-monitoring: testing means and covariance hypothesis about the simon and the eriksen flanker task. *Psychological Research*, *73*, 762–776. doi: 10.1007/s00426-008-0188-9

Kucina, T., Wells, L., Lewis, I., Salas, K. d., Kohl, A., Palmer, M. A., ... Heathcote, A. (2023). Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nature Communications*, *14*(1), 2234. doi: 10.1038/s41467-023-37777-2

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. doi: 10.1016/j.jmp.2010.08.013

Lüken, M., Heathcote, A., Haaf, J. M., & Matzke, D. (2023, October 18). *Parameter identifiability in evidence-accumulation models: The effect of error rates on the diffusion decision model and the linear ballistic accumulator.* Retrieved from `https://doi.org/10.31234/osf.io/wsgnt`

MacLeod, C. M. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological bulletin*, *109*(2), 163 – 203. doi: 10.1037/0033-2909.109.2.163

Matzke, D., Logan, G. D., & Heathcote, A. (2020). A cautionary note on evidence-accumulation models of response inhibition in the stop-signal paradigm. *Computational Brain Behavior*, *3*, 269–288. doi: 10.1007/s42113-020-00075-x

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258. doi: 10.1016/j.cogpsych.2012.12.002

Pettigrew, C., & Martin, R. C. (2014). Psychology and aging. *Psychology and Aging*, *29*(2), 187–204. doi: 10.1037/a0036085

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between stroop and simon effects using delta plots. *Attention, Perception, & Psychophysics*, *72*, 2013–2025. doi: 10.3758/APP.72.7.2013

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi: 10.1162/neco.2008.12-06-420

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. doi: 10.1037/0033-295X.111.2.333

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501–526. doi: 10.1037/xlm0000450

Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P. M., & et al. (2021). The arrow of time: Advancing insights into action control from the arrow version of the eriksen flanker task. *Attention, Perception, & Psychophysics*, *83*, 700–721. doi: 10.3758/s13414-020-02167-z

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. doi: 10.3758/s13423-018-1558-y

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail.
doi: 10.31234/osf.io/3cjr5

Rouder, J. N., & Mehrvarz, M. (2024). Hierarchical-model insights for planning and interpreting individual-difference studies of cognitive abilities. *Current Directions in Psychological Science*, *33*(2), 128–135. doi: 10.1177/09637214231220923

Simon, J. R., & Rudell, A. P. (1967). Auditory s-r compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. doi: 10.1037/h0020586

Stan Development Team. (2023). Stan modeling language users guide and reference manual [Computer software manual]. Retrieved from `https://mc-stan.org`

Stevenson, N., Donzallaz, M. C., Innes, R. J., Forstmann, B., Matzke, D., & Heathcote, A. (2024, January 30). Emc2: An r package for cognitive models of choice. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/2e4dq` doi: 10.31234/osf.io/2e4dq

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643. doi: 10

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, *27*, 911–936. doi: 10.3758/s13423-020-01719-6

Von Bastian, C. C., Souza, A. S., & Gade, M. (2016). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General*, *145*(2), 246. doi: 10.1037/xge0000120

Wald, A. (1947). *Sequential analysis.* John Wiley & Sons.

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210-238. doi: 10.1016/j.cogpsych.2011.08.001

Appendix

Below, we present plots of the model fits (observed and predicted cumulative distribution functions) and tables displaying the RMSE for all models and datasets. Additional figures assessing the fits can be found in the supplementary materials.

Posterior predictive checks

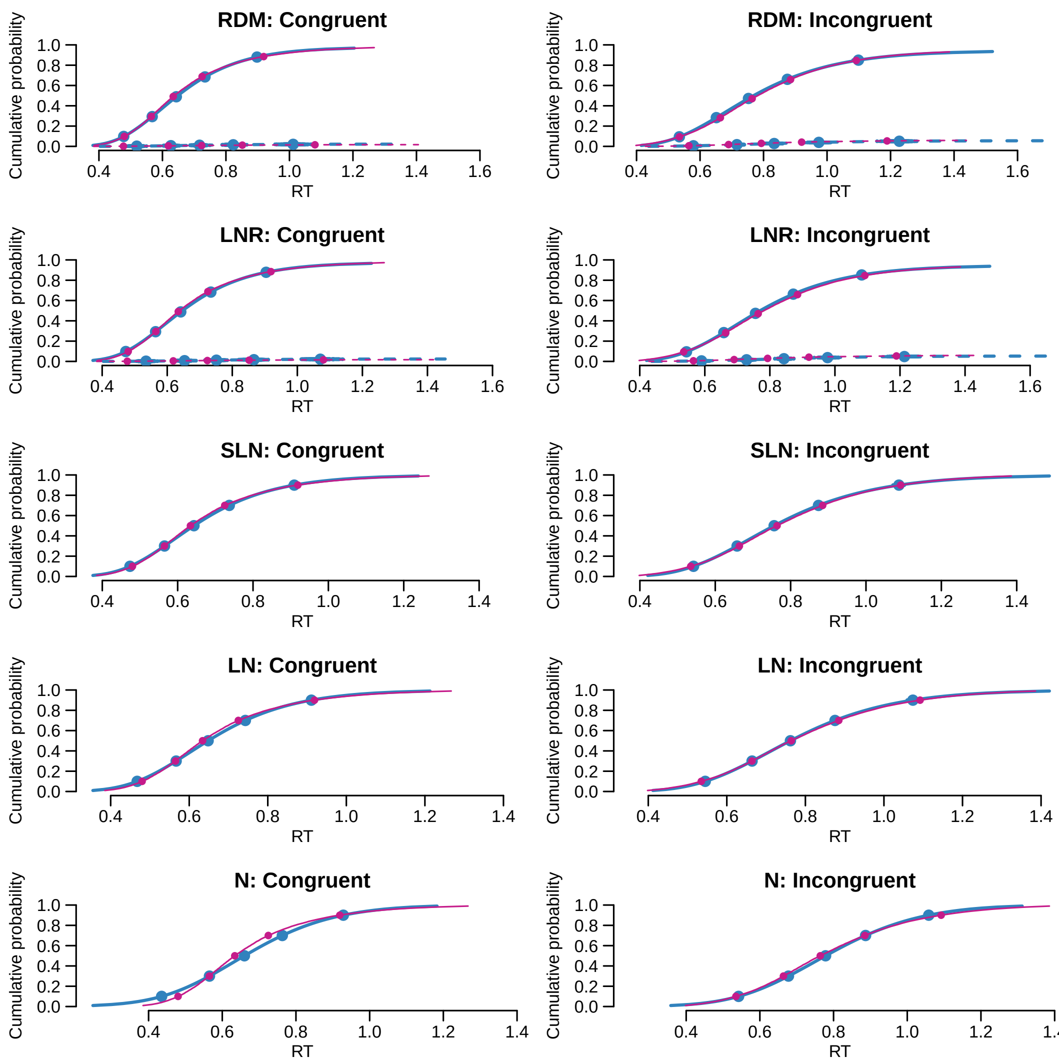

Figure 5. Observed and predicted cumulative distribution functions (CDFs) for Enkavi et al. (2019). Observed = pink, blue = predicted, RDM = racing diffusion model, LNR = lognormal race model, SLN = shifted lognormal model, LN = lognormal model, N = normal model. The predictions are based on 500 sampled datasets from the posterior predictive distribution and plotted is the median including 95% credible interval. The points show the 10%, 30%, 50%, 70%, and 90% quantiles averaged across participants, separately for the two congruency conditions. Note that for the RDM and the LNR, both the (defective) CDFs of correct and incorrect responses are shown, whereas for the other models, only the CDFs of correct responses are depicted.

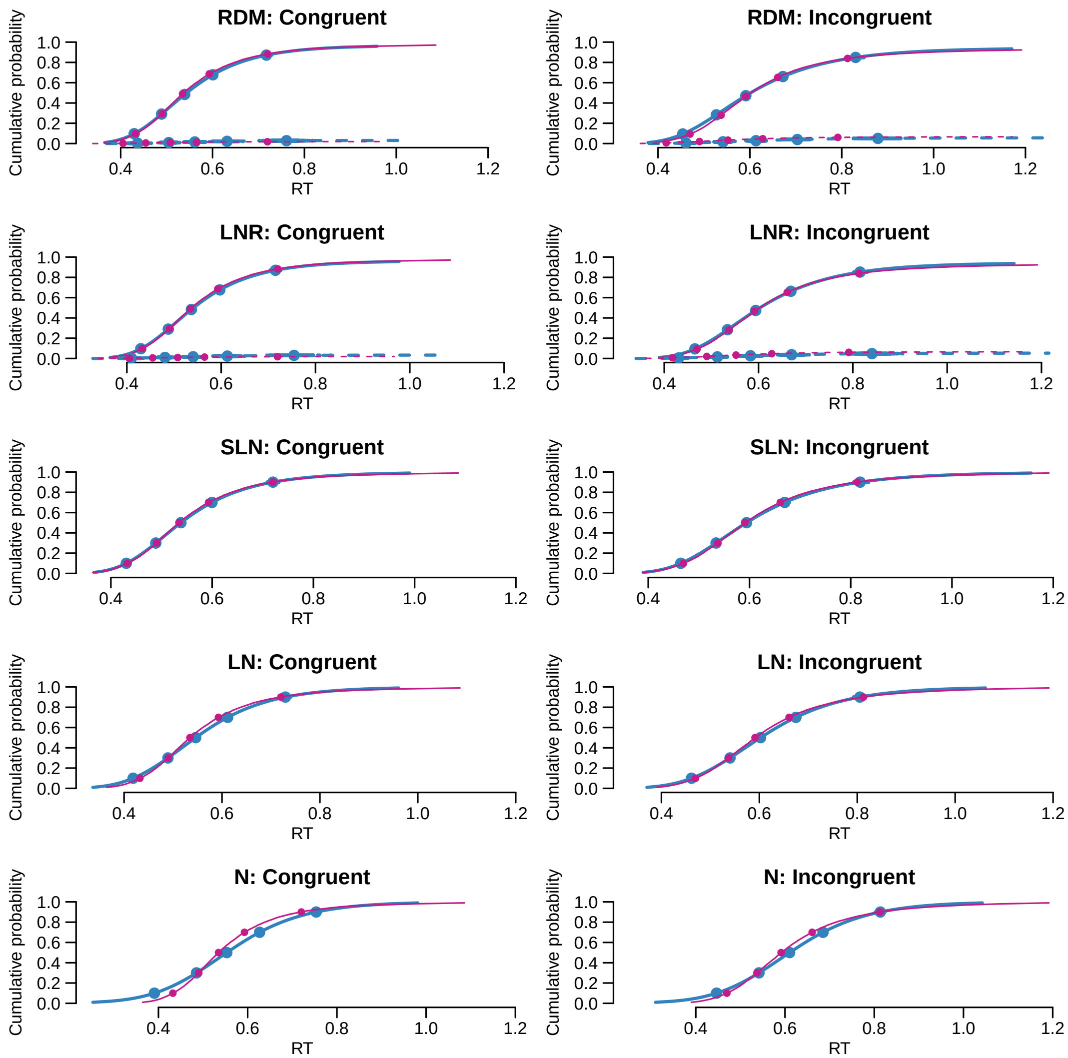

Figure 6. Observed and predicted cumulative distribution functions (CDFs) for Rey-Mermet et al. (2018). Observed = pink, blue = predicted, RDM = racing diffusion model, LNR = lognormal race model, SLN = shifted lognormal model, LN = lognormal model, N = normal model. The predictions are based on 500 sampled datasets from the posterior predictive distribution and plotted is the median including 95% credible interval. The points show the 10%, 30%, 50%, 70%, and 90% quantiles averaged across participants, separately for the two congruency conditions. Note that for the RDM and the LNR, both the (defective) CDFs of correct and incorrect responses are shown, whereas for the other models, only the CDFs of correct responses are depicted.

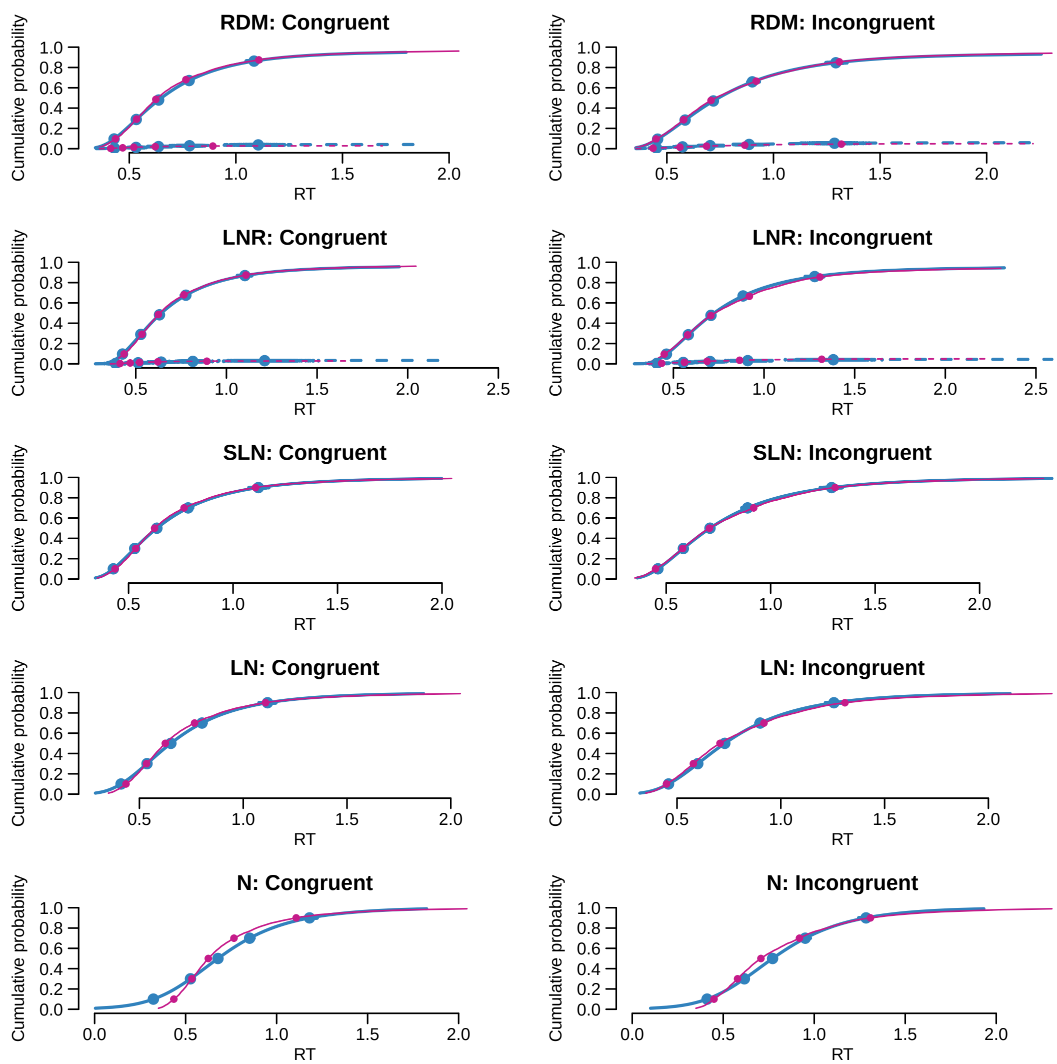

Figure 7. Observed and predicted cumulative distribution functions (CDFs) for Pratte et al. (2010). Observed = pink, blue = predicted, RDM = racing diffusion model, LNR = lognormal race model, SLN = shifted lognormal model, LN = lognormal model, N = normal model. The predictions are based on 500 sampled datasets from the posterior predictive distribution and plotted is the median including 95% credible interval. The points show the 10%, 30%, 50%, 70%, and 90% quantiles averaged across participants, separately for the two congruency conditions. Note that for the RDM and the LNR, both the (defective) CDFs of correct and incorrect responses are shown, whereas for the other models, only the CDFs of correct responses are depicted.

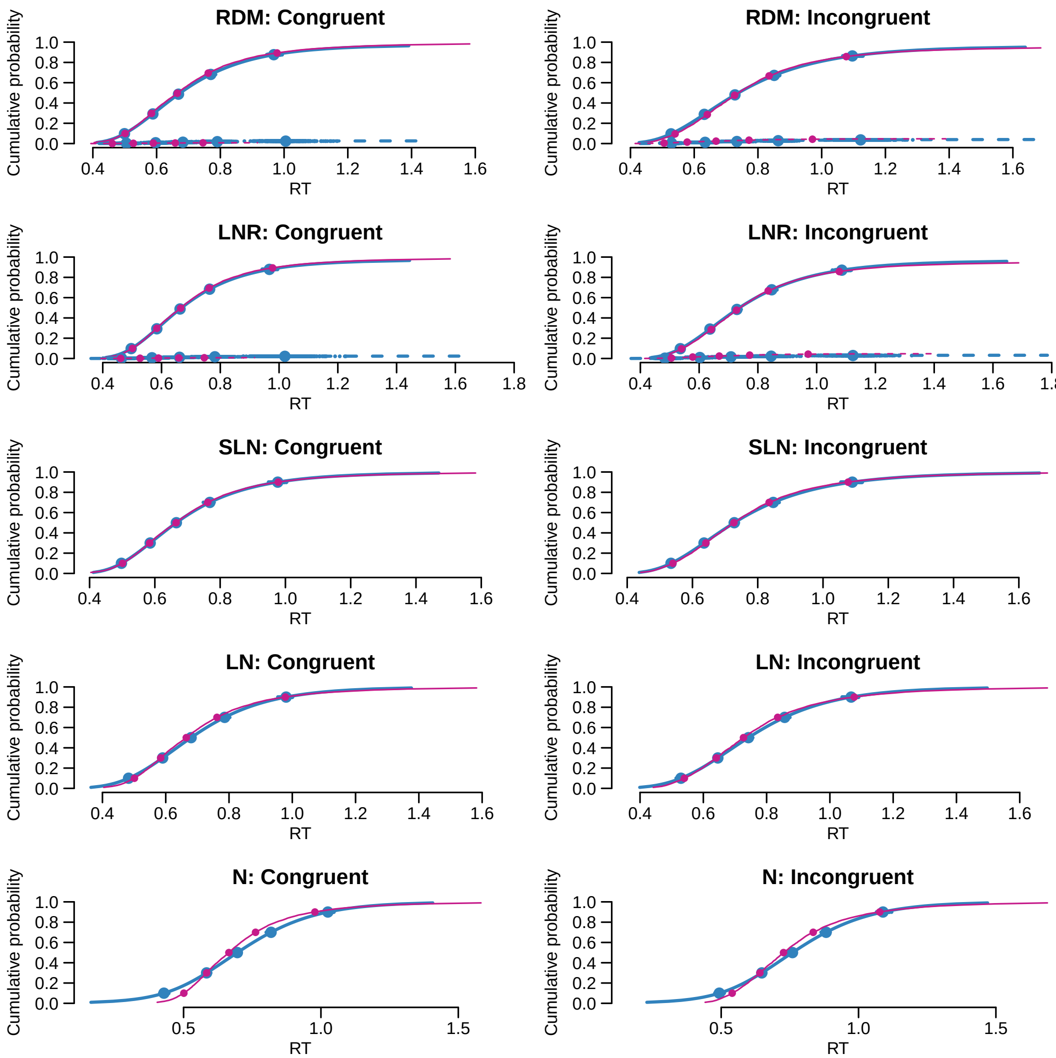

Figure 8. Observed and predicted cumulative distribution functions (CDFs) for Von Bastian et al. (2016). Observed = pink, blue = predicted, RDM = racing diffusion model, LNR = lognormal race model, SLN = shifted lognormal model, LN = lognormal model, N = normal model. The predictions are based on 500 sampled datasets from the posterior predictive distribution and plotted is the median including 95% credible interval. The points show the 10%, 30%, 50%, 70%, and 90% quantiles averaged across participants, separately for the two congruency conditions. Note that for the RDM and the LNR, both the (defective) CDFs of correct and incorrect responses are shown, whereas for the other models, only the CDFs of correct responses are depicted.

Goodness of fit (RMSE)

Table 2

Median and 95 % credible interval of the root mean squared error based on the 1%, 10%, 30%, 50%, 70%, 90%, and 99% quantiles of the correct response times and 500 samples from the posterior predictive distribution.

	Enkavi et al. (2019)	Pratte et al. (2010)	Rey-Mermet et al. (2018)	Von Bastian et al. (2016)
N	0.054 [0.051, 0.057]	0.168 [0.152, 0.184]	0.069 [0.064, 0.073]	0.126 [0.116, 0.136]
LN	0.011 [0.009, 0.013]	0.073 [0.045, 0.100]	0.054 [0.047, 0.059]	0.086 [0.071, 0.100]
SLN	0.019 [0.013, 0.024]	0.016 [0.006, 0.049]	0.027 [0.018, 0.035]	0.033 [0.013, 0.051]
LNR	0.016 [0.011, 0.021]	0.018 [0.007, 0.048]	0.031 [0.023, 0.038]	0.031 [0.023, 0.038]
RDM	0.025 [0.018, 0.031]	0.049 [0.018, 0.077]	0.027 [0.019, 0.034]	0.027 [0.019, 0.034]

Table 3

Median and 95 % credible interval of the root mean squared error computed on the proportion of correct responses using 500 samples from the posterior predictive distribution.

	Enkavi et al. (2019)	Pratte et al. (2010)	Rey-Mermet et al. (2018)	Von Bastian et al. (2016)
LNR	0.001 [0.000, 0.003]	0.002 [0.000, 0.005]	0.001 [0.000, 0.004]	0.002 [0.000, 0.005]
RDM	0.001 [0.000, 0.003]	0.011 [0.006, 0.016]	0.001 [0.000, 0.004]	0.006 [0.002, 0.010]