Estimating Across-Trial Variability Parameters of the Diffusion Decision Model: Expert Advice and Recommendations

Udo Boehm[a], Jeffrey Annis[b], Michael J. Frank[c], Guy E. Hawkins[d], Andrew Heathcote[e], David Kellen[f], Angelos-Miltiadis Krypotos[g], Veronika Lerche[h], Gordon D. Logan[b], Thomas J. Palmeri[b], Don van Ravenzwaaij[i], Mathieu Servant[b], Henrik Singmann[j], Jeffrey J. Starns[k], Andreas Voss[h], Thomas V. Wiecki[l], Dora Matzke[m], Eric-Jan Wagenmakers[m]

Author Note

[a]Department of Experimental Psychology, University of Groningen, 9712 TS Groningen, The Netherlands, email Udo Boehm: u.bohm@rug.nl

[b] Department of Psychology, Vanderbilt University, USA, email Jeffrey Annis: jeff.annis@vanderbilt.edu, email Gordon D. Logan: gordon.logan@vanderbilt.edu, email Thomas Palmeri: thomas.j.palmeri@vanderbilt.edu, email Mathieu Servant: servant.mathieu@gmail.com

[c] Department of Cognitive, Linguistic & Psychological Sciences, Brown University, USA, email: Michael_Frank@brown.edu

[d] School of Psychology, University of Newcastle, Australia, email: guy.e.hawkins@gmail.com

[e] School of Medicine, University of Tasmania, Australia, email: andrew.heathcote@utas.edu.au

[f] Department of Psychology, Syracuse University, USA, email: davekellen@gmail.com

[g] Department of Clinical Psychology, Utrecht University, email: amkrypotos@gmail.com

[h] Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Germany, email Veronika Lerche: veronika.lerche@psychologie.uni-heidelberg.de, email Andreas Voss: andreas.voss@psychologie.uni-heidelberg.de

[i] Department of Psychometrics & Statistical Techniques, University of Groningen, The Netherlands, email: d.van.ravenzwaaij@rug.nl

[j] Department of Psychology, University of Zürich, Switzerland, email: singmann@gmail.com

[k] Department of Psychological and Brain Sciences, University of Massachusetts - Amherst, USA, email: jstarns@umass.edu

[l] Cologne, Germany, email: thomas.wiecki@gmail.com

[m] Department of Psychology, University of Amsterdam, The Netherlands, email Dora Matzke: d.matzke@uva.nl, Eric-Jan Wagenmakers: ej.wagenmakers@gmail.com

Authors except UB, EJW and DM are listed in alphabetical order.

Declarations of interest: none.

This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to UB (406-12-125), a European Research Council (ERC) grant to EJW, an NWO Veni grant (451-15-010) to DM, a German Research Foundation grant (VO1288/2-2) to AV and VL, an Australian Research Council Discovery Early Career Researcher Award (DE170100177) to GEH, a Swiss National Science Foundation Grant (100014_165591) to HS and DK, and NSF SBE-1257098, NEI ROI-EY021833, the Temporal Dynamics of Learning Center (NSF SMA-1041755), the Vanderbilt Vision Research Center (NEI P30-EY008126), a Discovery Grant from Vanderbilt University, and a training grant from the NIH (T32-EY007135) to TJP and JA.

Abstract

For many years the Diffusion Decision Model (DDM) has successfully accounted for behavioral data from a wide range of domains. Important contributors to the DDM's success are the across-trial variability parameters, which allow the model to account for the various shapes of response time distributions encountered in practice. However, several researchers have pointed out that estimating the variability parameters can be a challenging task. Moreover, the numerous fitting methods for the DDM each come with their own associated problems and solutions. This often leaves users in a difficult position. In this collaborative project we invited researchers from the DDM community to apply their various fitting methods to simulated data and provide advice and expert guidance on estimating the DDM's across-trial variability parameters using these methods. Our study establishes a comprehensive reference resource and describes methods that can help to overcome the challenges associated with estimating the DDM's across-trial variability parameters.

keywords: Diffusion Decision Model, across-trial variability parameters, parameter estimation

Estimating Across-Trial Variability Parameters of the Diffusion Decision Model: Expert Advice and Recommendations

1 Introduction

The Diffusion Decision Model (DDM) has a long and successful history of accounting for response time (RT) and accuracy data from a wide range of domains, including lexical decision (Yap, Sibley, Balota, Ratcliff, & Rueckl, 2015; Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), memory-retrieval (White, Kapucu, Bruno, Rotello, & Ratcliff, 2014; McKoon & Ratcliff, 1996), perceptual decision-making (Ratcliff, 2002; Smith, Ratcliff, & Wolfgang, 2004; Smith, Ratcliff, & Sewell, 2014), as well as data from neurophysiological studies (Kühn et al., 2011, Philiastides, 2006; for reviews see Forstmann, Ratcliff, & Wagenmakers, 2016, Ratcliff & McKoon, 2008, Ratcliff, Smith, Brown, & McKoon, 2016, Smith & Ratcliff, 2009). The DDM belongs to the class of sequential sampling models for two-choice RT tasks (Ratcliff, 1978; Ratcliff et al., 2004). It conceptualizes RT and accuracy as the result of the accumulation of noisy information over time toward two absorbing boundaries. Figure 1 illustrates the components of the model. The four main parameters are boundary separation a, drift rate v, starting point z, and non-decision time T_{er}. Boundary separation is the distance between the response boundaries and determines the trade-off between response speed and accuracy. Greater boundary separation means that more information needs to be accumulated to trigger a response, which results in longer RTs and higher accuracy. Drift rate v represents the quality of the information that is being accumulated. Higher drift rate means that the mean rate of information accumulation is quicker, which leads to faster and more accurate responses. Starting point z represents an a priori bias towards one of the two response options. A starting point higher than the midpoint between the boundaries, $a/2$, means that less information needs to be accumulated to reach the upper boundary, and the corresponding response option is chosen faster and more frequently. Non-decision time represents processes not related to the decision process, such as stimulus encoding or response execution. In addition to these main parameters, the DDM includes three across-trial variability parameters that we discuss next.

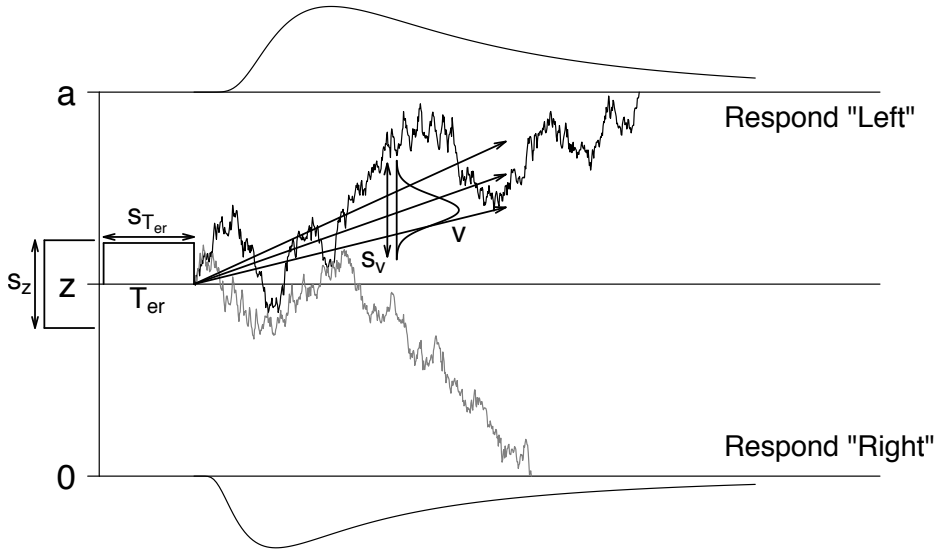

Figure 1: Drift diffusion model (DDM) and its parameters. See section 1 for details.

A key factor in the DDM's success is its ability to account for the different and varied shapes of the RT distributions in a wide range of experimental paradigms. For example, a typical phenomenon in RT experiments is that mean RTs differ between correct and error responses. Such patterns bedevilled early sequential sampling models and several authors suggested adding across-trial variability parameters to account for these phenomena (Laming, 1968; Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002; Smith & Vickers, 1988; Van Zandt & Ratcliff, 1995). Specifically, allowing the starting point of the accumulation process to vary across trials enables models to produce fast errors (Laming, 1968), whereas allowing the drift rate of the accumulation process to vary across trials enables models to produce slow errors (Ratcliff, 1978). These variability parameters allow the DDM to account for the benchmark result that errors tend to be slower than correct responses when accuracy is high, and errors tend to be faster than correct responses when accuracy is low. Moreover, using a combination of both types of variability enables the DDM to also account for crossover patterns where errors are slower than correct responses when accuracy is low, and errors are faster than correct responses when accuracy is high (Ratcliff, McKoon, & van Zandt, 1999; Ratcliff & Rouder, 1998; Wagenmakers et al., 2008). In addition, Ratcliff and Tuerlinckx (2002) have suggested that an across-trial variability component in the non-decision

time parameter might be needed to account for experimental manipulations that affect the leading edge of the RT distribution. The lexical decision data in Ratcliff et al. (2004), for example, required across-trial variability in non-decision time to account for a shift in the 10th percentile of the RT distribution.

Although across-trial variability parameters clearly play an important role in the DDM's ability to fit empirical data, several authors have reported difficulties in estimating the parameter values. For example, Lerche and Voss (2017) assessed the retest reliability of DDM parameter estimates over two separate sessions using a lexical decision task, a recognition memory task, and an associative priming task. In their model fits, Lerche et al. only allowed for across-trial variability in non-decision time but not in drift rate or starting point. Their results for the lexical decision task, for instance, showed that the estimated variability in non-decision time correlated only modestly to weakly between sessions ($r = .20 - .55$). On the other hand, estimates for the four main DDM parameters (i.e., starting point, drift rate, boundary separation, and non-decision time) correlated modestly to strongly between sessions ($r = .30 - .90$). Results for the recognition memory and associative priming tasks were similar. Taken together, the results of Lerche et al.'s study suggest that the DDM's main parameters can be estimated reliably whereas the retest reliability of the variability in non-decision time is notably lower. Results from Lerche, Voss, and Nagler (2017) suggest that this lower retest reliability is due to a lack of true score stability, rather than unreliable estimation; in simulation studies they found a high correlation between true values and estimates of the variability in non-decision time.

In another example, Yap, Balota, Sibley, and Ratcliff (2012) used a large corpus of lexical decision data that had been collected in two sessions (Balota et al., 2007) to evaluate the retest reliability of the DDM parameters. To compute the within-session reliability of the parameter estimates, Yap et al. split the data into halves based on odd and even trials and computed the correlation between parameter estimates from each half of the data. To assess the between-session reliability, Yap et al. computed the correlation between parameter estimates from the first session and parameter estimates from the second session. This analysis showed

that estimates for the main DDM parameters were strongly correlated within ($r = .81 - .93$) as well as between experimental sessions ($r = .65 - .74$). However, although estimates of starting point variability correlated strongly within experimental sessions ($r = .81$), the estimates for drift rate and non-decision time variability correlated less strongly within sessions ($r = .65$ for both parameters), and correlations between parameter estimates from different sessions were relatively weak for all three variability parameters ($r = .39 - .50$). Yap et al. explain the low within-session reliability of the drift rate and non-decision time variabilities with the fact that both model parameters depend on the distribution of error RTs. Because there are typically relatively few observations for error responses, these parameters are not well constrained by the data, which leads to less reliable parameter estimates.

However, Yap et al.'s lexical decision data featured 819 participants with 3374 trials per participant. Together with a mean error rate of 14.4%, this suggests that there was, on average, a total of 486 error RTs for each participant. Consequently, when Yap et al. split their data into two halves to compute the within-session reliability, each half included an average of 243 error responses based on which the across-trial variability parameters could be estimated. If such a sizable data set is insufficient for reliable estimation, this suggests that estimation of the across-trial variability parameters in many other applications of the DDM may also be poor. In functional neuroimaging, one of the fastest growing areas of application of the DDM, there are often practical limitations on the experimental design and the number of trials that can be obtained. This raises the question whether factors beyond the number of trials and conditions can be utilized to improve estimation performance in standard experimental designs.

For example, conventional methods typically fit the DDM on an individual basis and, therefore, require that sufficient data are available for each participant (e.g., Vandekerckhove & Tuerlinckx, 2007, Ratcliff, 2002, Voss & Voss, 2007). Recently developed hierarchical Bayesian methods, on the other hand, use all available data in the group to mutually inform parameter estimates across participants (Vandekerckhove, Tuerlinckx, & Lee, 2011; Wiecki, Sofer, & Frank, 2013). Specifically, hierarchical Bayesian models assume that participants' parameters

are drawn from a common group-level distribution. Because the participant-level and group-level parameters are estimated simultaneously, the parameter estimates for individual participants are informed by the parameter estimates for the rest of the group. This mutual dependence of the parameter estimates reduces the influence of outliers on group-level parameters and yields parameter estimates for individual participants with the smallest estimation error (Efron & Morris, 1977). Hierarchical Bayesian methods might therefore be able to reliably estimate across-trial variability parameters in situations where conventional methods fail.

However, estimating across-trial variablities in hierarchical Bayesian implementations of the DDM comes with its own challenges. For example, the `HDM` package for `JAGS` (Plummer, 2003) and `Stan` (Carpenter et al., 2017) implements a version of the DDM's first-passage time distribution where all across-trial variability parameters are fixed to 0 (Vandekerckhove et al., 2011; Wabersich & Vanderkerckhove, 2014). Nevertheless, trial-to-trial variability in the model parameters can be added using a mixture of first-passage time distributions where the drift rate parameter, for instance, is sampled from a normal distribution for each draw from the first-passage time distribution. Unfortunately, in our experience adding the across-trial variability parameters to the model inevitably leads to erratic behavior of the MCMC chains and a lack of convergence. Specifically, when we generated 5000 trials from the DDM with across-trial variability in drift rate but all other across-trial variabilities fixed to 0, fitting a model with a mixture of first-passage time distributions as described above resulted in MCMC chains that remained stuck at their initial values. The convergence problem might be resolved by using another sampler that is more suitable for the DDM, as for example implemented in the `HDDM` software package (e.g., Wiecki et al., 2013).

However, deciding which sampling algorithm to use requires expert knowledge and experience that is often not available to the naive user. Similar knowledge gaps are likely to also exist for conventional fitting methods, where choosing a suitable numerical optimization algorithm, for example, requires extensive experience. This leaves the practitioner in a precarious situation. On the one hand, across-trial variability parameters can be critical to the DDM's ability

to fit different data patterns. On the other hand, estimating across-trial variability parameters is inherently challenging. Obtaining good parameter estimates might critically depend on expert knowledge that is not available to the average user.

The goal of the present work is, therefore, to conduct a survey of the available methods and to provide a platform for experts from the DDM community to share their knowledge and recommendations for estimating the DDM's across-trial variability parameters. Specifically, we generated three data sets with numbers of trials and experimental conditions as typically used in functional neuroimaging or clinical psychology. We invited experts to apply their preferred fitting methods to the three data sets and give recommendations for estimating across-trial variability parameters in each scenario.

It should be noted that the present work is not a comprehensive parameter recovery study but aims to showcase different fitting methods in a typical application. A comprehensive review on the estimation of the across-trial variability parameters under different experimental designs and generating parameter values with conventional fitting methods can be found in Ratcliff and Tuerlinckx (2002).

2 Structure of the Collaborative Project

We generated three synthetic data sets that differed in complexity and invited researchers from the DDM community to apply their fitting methods to each data set. Collaborators were asked to provide a short summary of their methods and results, including their parameter estimates and a measure of the uncertainty associated with the parameter estimates (e.g., confidence intervals or credible intervals), and to provide advice for other users, including descriptions of problems encountered, workaround solutions, and general recommendations. The invitation letter is available on the project's Open Science Framework (OSF) site: osf.io/fjy8z/.

2.1 Data Sets

We based the structure of the simulated data on a typical setup for a perceptual decision experiment with three conditions that differ only in their level of difficulty (i.e., drift rate). The

data sets were generated from the full DDM using the `rtdists` (Singmann et al., 2016) `R` package (R Core Team, 2015). Each data set was generated with three different drift rates v_{Easy}, v_{Medium}, v_{Hard} for the three experimental conditions, and common values across experimental conditions for boundary separation a, non-decision time T_{er}, relative starting point z (i.e., $z \in [0, 1]$), across-trial variability in drift rate s_v, across-trial variability in non-decision time $s_{T_{er}}$, and across-trial variability in starting point s_z. Here v is mean drift rate and s_v is the standard deviation of the normal distribution from which v is sampled, T_{er} refers to the mean non-decision time and $s_{T_{er}}$ is the range of the uniform distribution from which T_{er} is sampled, and z is the mean relative starting point and s_z is the range of the uniform distribution from which the relative starting point is sampled. Data were generated with the diffusion scale parameter set to $s = 1$.

Table 1 shows the generating parameter values for each data set. The data and detailed descriptions are available at osf.io/fjy8z/. Our generating parameter values were based on Matzke and Wagenmakers' (2009) survey of parameter values estimated in empirical studies.

2.1.1 Level 1. Level 1 of our collaborative project assessed how well the across-trial variability parameters can be estimated for an individual participant independent of the four main DDM parameters. We therefore provided the data-generating values of the main parameters and asked collaborators to estimate the values of the three across-trial variability parameters. The data set consisted of 1000 simulated trials for each experimental condition for a single participant.

The data for Level 1 are shown in the top row of Figure 2. Histograms show the RT distribution for correct (positive x-axis) and incorrect (negative x-axis) responses in the Easy (left column), Medium (middle column), and Hard (right column) condition. As can be seen, RT distributions have the typical right skew. The number of error responses is lowest in the condition with the highest drift rate (i.e., Easy) and increases with decreasing drift rate, thus exhibiting typical patterns produced by the DDM.

2.1.2 Level 2. Level 2 of our collaborative project assessed how well the across-trial variability parameters can be estimated for an individual participant when the values of the main DDM parameters are unknown. We therefore asked collaborators to estimate all DDM

Table 1

Generating parameter values for synthetic data.

		a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
Level 1		1	3.5	2.5	1.5	0.35	0.45	2.2	0.1	0.4
Level 2		0.8	4	3	2	0.43	0.55	1.8	0.1	0.2
Level 3	μ_k	0.8	4	3	2	0.43	0.55	1.6	0.15	0.3
	σ_k	0.3	1	1	1	0.1	0.02	0	0	0
		Individual Participants Level 3								
	PP1	0.54	3.15	1.66	2.37	0.39	0.51			
	PP2	1.52	3.54	3.20	1.29	0.49	0.56			
	PP3	0.32	5.37	2.18	0.03	0.38	0.56			
	PP4	0.58	4.63	3.22	1.27	0.37	0.54			
	PP5	0.49	4.78	4.05	0.92	0.45	0.55			
	PP6	0.86	3.74	3.12	3.18	0.53	0.56			
	PP7	0.73	5.07	2.58	2.67	0.18	0.50			
	PP8	0.53	2.91	2.38	2.41	0.54	0.55			
	PP9	1.27	6.11	1.84	2.07	0.47	0.56			
	PP10	0.53	6.08	3.45	1.95	0.39	0.56			
	PP11	0.39	5.87	5.60	1.55	0.24	0.54			
	PP12	0.48	4.63	5.51	1.17	0.42	0.54			
	PP13	1.37	4.55	3.85	2.27	0.37	0.57			
	PP14	1.32	3.72	5.11	2.98	0.49	0.52			
	PP15	0.71	2.83	1.31	3.10	0.41	0.53			
	PP16	0.87	3.96	2.47	0.83	0.27	0.55			
	PP17	0.70	3.84	3.27	2.42	0.39	0.53			
	PP18	1.11	4.36	3.39	2.76	0.39	0.56			
	PP19	1.20	5.60	4.09	2.42	0.35	0.57			
	PP20	0.90	5.37	2.67	2.19	0.37	0.54			

Note. μ_k is the group-level mean for parameter k, σ_k is the corresponding group-level standard deviation. The diffusion coefficient was $s = 1$ for all data sets. Level 1: data for one participant, main DDM parameters known. Level 2: data for one participant, main DDM parameters unknown. Level 3: data for twenty participants, group-level and individual-level parameters unknown. PPj indicates the generating values for simulated participant j

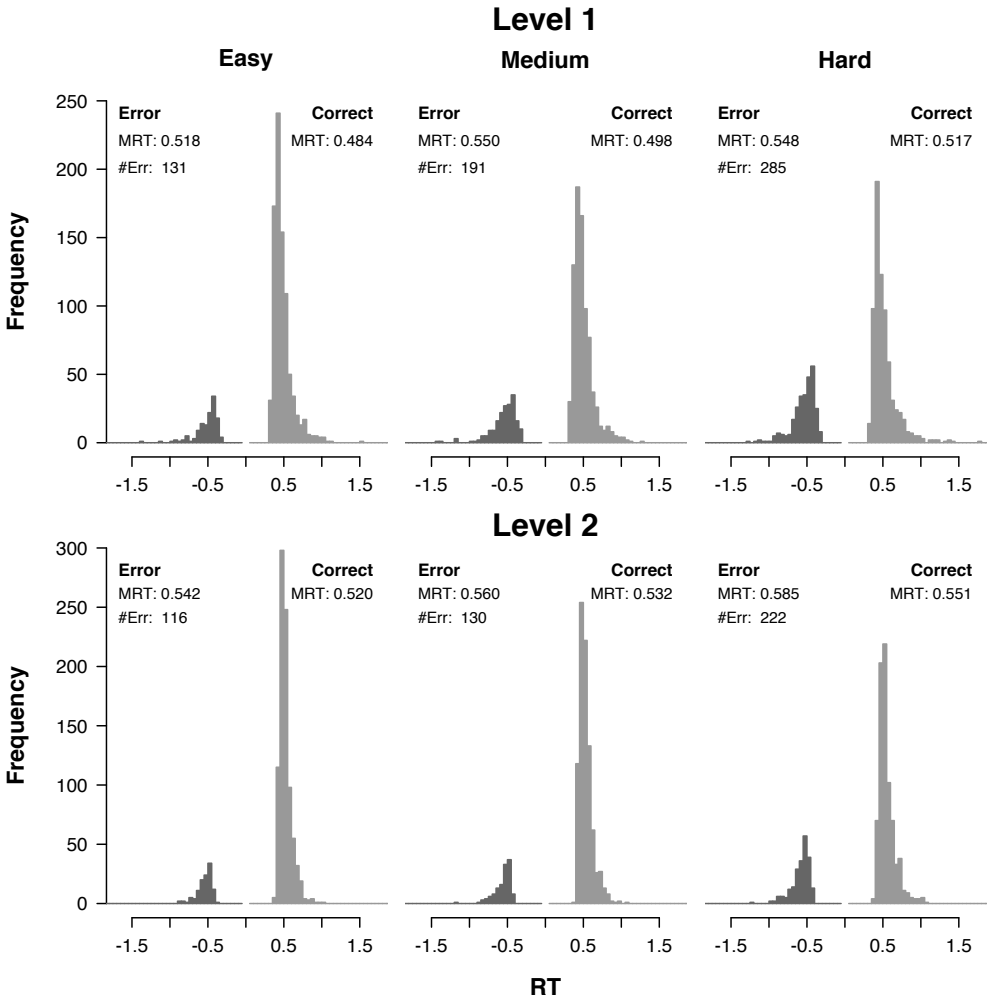

Figure 2: Histograms of simulated RTs for Level 1 and Level 2. Error RTs are shown on the negative x-axis. MRT is the mean response time, #Err is the number of error RTs out of 1000 simulated trials per condition.

parameters from the data. The data set again consisted of data of a single participant with 1000 trials for each of the three experimental conditions. Only drift rate differed between experimental conditions.

The data for Level 2 are shown in the bottom row of Figure 2. RT distributions have a typical right skew. The number of error responses is lower than for Level 1 due to the higher drift rates used to generate the data. Nevertheless, there is a total of 468 error RTs available to characterize the error RT distributions.

2.1.3 Level 3. Level 3 of our project assessed whether pooling data across participants improves estimation of the group-level across-trial variability parameters. We therefore generated a hierarchical data set and asked collaborators to estimate the means and standard deviations of the group-level parameter distributions. The data set consisted of simulated data of 20 participants with 1000 trials for each of the three experimental conditions. The main DDM parameters for each participant had been sampled from a common group-level normal distribution $\mathcal{N}(\mu_k, \sigma_k)$ with mean μ_k and standard deviation σ_k that was truncated to the range of admissible values for each DDM parameter. Across-trial variability parameters were fixed across participants.

The data for Level 3 are shown in Figure 3. Histograms show the average number of trials of 20 simulated participants in each RT bin. The total number of error trials ranged between 150 and 1014, with an average of 498.5.

2.2 Overview of Collaborators and Methods

We received contributions from nine groups of collaborators from the DDM community. Table 2 summarizes the estimation methods and summary statistics used by our collaborators. As the collaborators used three main estimation methods, we will group contributions by method. In what follows we present a brief description of each estimation method followed by a summary of the main results. The full reports by each team of collaborators can be found in the appendix; supplementary materials are available on the project's OSF page (osf.io/fjy8z/).

To foreshadow our main conclusions, all estimation methods used by our collaborators

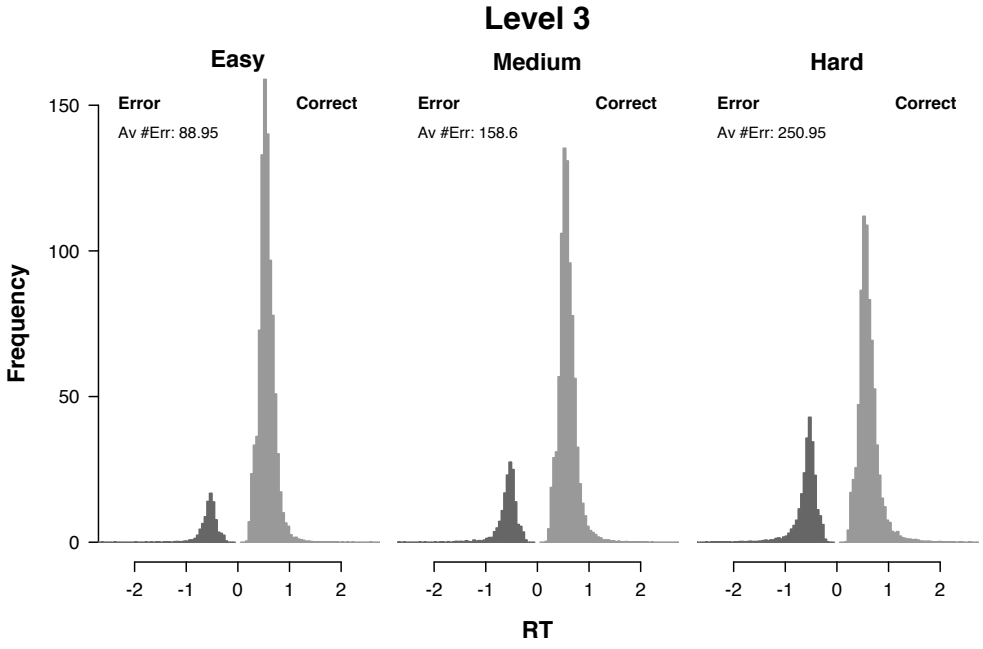

Figure 3: Histograms of simulated RTs for Level 3. Error RTs are shown on the negative x-axis. Histograms show the mean number of observations per RT bin, upper and lower outlines show the 0.9 and 0.1 quantile of the number of observations per RT bin across 20 simulated participants, respectively. Av #Err is the average number of error RTs out of 1000 simulated trials per condition.

Table 2

Estimation methods and measures of uncertainty for parameter estimates used by collaborators.

Collaborator	Fitting Method	Parameter Estimate	Measure of Uncertainty
Annis & Palmeri (Ann)	NHB	PM	95% HDI
Frank, Krypotos, & Wiecki (Fra)	HB	PM	95% HDI
Hawkins (Haw)	HB	PMD	95% HDI
Heathcote (Hea)	HB	PMD	95% HDI
Servant & Logan (Ser)	χ^2	BF	95% BCI
Singmann & Kellen (Sin)	ML	BF	95% BCI
Starns (Sta)	χ^2	BF	L10, 95% CI
Van Ravenzwaaij (Rav)	HB	PMD	95% HDI
Voss & Lerche (Vos)	ML	BF	95% BCI

Note. Abbreviations of contributor names are indicated in brackets.
NHB: non-hierarchical Bayesian, HB: hierarchical Bayesian, χ^2: χ^2-minimization for RT quantiles, ML: maximum-likelihood estimation.
PM: posterior mean, PMD: posterior median, BF: best fitting parameter.
X% HDI: X% highest density interval, X% BCI: X% bootstrap confidence interval, X% CI: X% confidence interval, L10: likelihood-based uncertainty interval.

could accurately recover across-trial variability in non-decision time. Estimates of the across-trial variability in drift rate and starting point, on the other hand, were associated with considerable uncertainty and tended to miss the true parameter value by a wide margin.

3 Estimation Methods

3.1 Bayesian Estimation

Five contributions used Bayesian estimation methods. For Levels 1 and 2, these methods assumed that the DDM parameters were drawn from a parameter-specific prior distribution. Four of the five contributions (Hawkins, van Ravenzwaaij, Frank et al., and Annis & Palmeri) based the parameterization of these prior distributions on Matzke and Wagenmakers's (2009) survey of published parameter estimates. For Level 3, Annis and Palmeri used a two-step analysis for the Level 3 data, in which they first obtained parameter estimates for each participant and subsequently estimated the group-level distributions for these posterior estimates. Heathcote, Hawkins, van Ravenzwaaij, and Frank et al. used a hierarchical modeling approach that assumed that participant-level parameters were drawn from a common group-level distribution. These group-level distributions are characterized by the group-level parameters, which were estimated from the data.

Heathcote, Hawkins, and van Ravenzwaaij assumed all group-level distributions to be normal distributions truncated to the range of plausible values of the particular model parameter (e.g., the distribution of T_{er} was truncated below at 0). The means of these group-level distributions were in turn assigned truncated normal prior distributions; Hawkins and van Ravenzwaaij's parameterization of these prior distribution was again loosely based on Matzke and Wagenmakers's (2009) survey. The standard deviations of the group-level distributions were assigned gamma prior distributions. Frank et al. assumed different group-level distributions for the main DDM parameters that were specific to each parameter (e.g., the a parameter was assigned a gamma distribution). The parameters of these group-level distributions were in turn assigned gamma or truncated normal prior distributions. The across-trial variability parameters, on the other hand, were assigned a single common value for all participants that was sampled

from a half-normal (s_v and $s_{T_{er}}$) or a beta (s_z) prior distribution.

Within the Bayesian framework, point estimates for the parameters are obtained by computing a measure of the central tendency for the marginal posterior distribution of each model parameter. The contributions reported here used the posterior mean or posterior median. Uncertainty about parameter estimates is described by the width of the marginal posterior distribution. All five contributions used the 95% highest density interval (HDI), which, for a unimodal posterior distribution, describes the narrowest interval around the posterior mode that includes 95% of the posterior probability mass.

As the marginal posterior distributions for the DDM are not available in closed-form, numerical methods must be used to approximate the posterior mean or median and the 95% HDI. Heathcote, Hawkins, and van Ravenzwaaij used the Differential-Evolution Markov Chain Monte Carlo (MCMC) algorithm (ter Braak, 2006), and Frank et al. used the Slice-Sampling MCMC algorithm (Neal, 2003). Despite some differences in the implementational details, both algorithms are based on the construction of a number of Markov chains that have the target posterior distribution as their equilibrium distribution. An approximation of the posterior density is obtained by observing the Markov chains after they have converged to their equilibrium distribution, which can then be used to compute relevant summary statistics. Annis and Palmeri used the Laplace approximation of the joint posterior density of all DDM parameters for each participant to estimate the posterior modes and covariance matrix. Based on these estimates, they used numerical integration by Componentwise Adaptive Gauss-Hermite Iterative Quadrature to compute the posterior mean and 95% HDI. For Level 3 they used the same numerical integration method to approximate the posterior means for each participant. These estimates were then combined in a Bayesian model to estimate the group-level mean and standard deviation for each DDM parameter using Hamiltonian MCMC sampling.

3.2 Maximum-Likelihood Estimation

Two contributions used maximum-likelihood estimation. This method uses the DDM's likelihood function to numerically approximate the parameter values that maximize the joint

likelihood of the observed data for each participant. Singmann and Kellen used an algorithm based on Newton's method (Kaufman & Gay, 2003) to find the ML estimators of the DDM parameters; Voss and Lerche computed the ML estimators using a version of the `Simplex` algorithm (Nelder & Mead, 1965).

Both groups used bootstrap confidence intervals (BCI) to quantify the uncertainty associated with the ML estimators. Singmann and Kellen based their BCIs on 1000 bootstrap samples, Voss and Lerche based their BCIs on 200 bootstrap samples and only reported intervals for the across-trial variability parameters.

For Level 3, Voss and Lerche obtained ML estimates of the parameter values for each individual participant and reported the average estimated value across participants. Singmann and Kellen did not fit the Level 3 data.

3.3 χ^2 Minimization

Two contributions used χ^2 minimization. This method estimates the DDM parameters that minimize the deviation between observed and predicted RT quantiles for correct and incorrect responses. Specifically, for the .1, .3, .5, .7, and .9 quantiles, the method minimizes the χ^2 statistic:

$$\chi^2 = \sum_i \frac{N(p_i - \pi_i)^2}{\pi_i}, \tag{1}$$

where N is the total number of observations, p_i and π_i denote the observed and predicted proportions of trials in bin i, respectively, and the summation is over 12 quantiles (6 for correct responses and 6 for error responses).

Servant and Logan excluded errors from the χ^2 computation when their number was below 10, Starns excluded errors from the χ^2 computation when their number was below 5. Both contributions used the `Simplex` algorithm (Nelder & Mead, 1965) to find the parameter values that minimize the χ^2 statistic across experimental conditions. However, whereas Starns estimated separate drift rates for each experimental condition and "left" and "right" stimuli, Servant and Logan estimated a single drift rate for each experimental condition. Moreover, Servant and Logan

imposed a number of constraints on z, s_z, s_v, and $s_{T_{er}}$ to guarantee sensible parameter estimates.

For Levels 1 and 2, Servant and Logan quantified the uncertainty associated with their parameter estimates using parametric BCIs. To this end, they generated 50 bootstrap data sets from the model with the best-fitting parameter values and again fit the DDM to these bootstrap data sets using χ^2 minimization. To quantify the uncertainty associated with his parameter estimates, Starns fixed each DDM parameter in turn to a value above or below the best-fitting value and used χ^2 minimization on the remaining parameters to find the parameter value at which the likelihood of the data was 10 times lower than the likelihood under the best-fitting value.

For Level 3, both contributions used χ^2 minimization to find the best fitting parameter values for each individual participant and reported the average across participants. Starns quantified the uncertainty for his parameter estimates using conventional 95% confidence intervals whereas Servant and Logan did not report measures of uncertainty.

4 Results

Figure 4 presents a summary of the across-trial variability parameter estimates for Level 1 reported by our collaborators and the distribution of parameter values observed in empirical studies reported in Matzke and Wagenmakers (2009) as a reference point. The vertical line indicates the generating value for each parameter, dots indicate point estimates obtained by different estimation methods and error bars show the corresponding measures of uncertainty reported by our collaborators. Results shown in gray are based on fits of the full DDM where the main DDM parameters were not fixed to the true values.

The results for $s_{T_{er}}$ are shown in the left panel. As can be seen, all point estimates for $s_{T_{er}}$ were close to the generating value and the uncertainty intervals were very narrow compared to the range of values typically found in empirical studies, indicating that $s_{T_{er}}$ could be estimated reliably by all estimation methods.

Similarly, most point estimates for s_v, shown in the middle panel, were close to the generating parameter value and uncertainty intervals were relatively narrow compared to the range of values observed in empirical studies. The estimate for s_v reported by Frank et

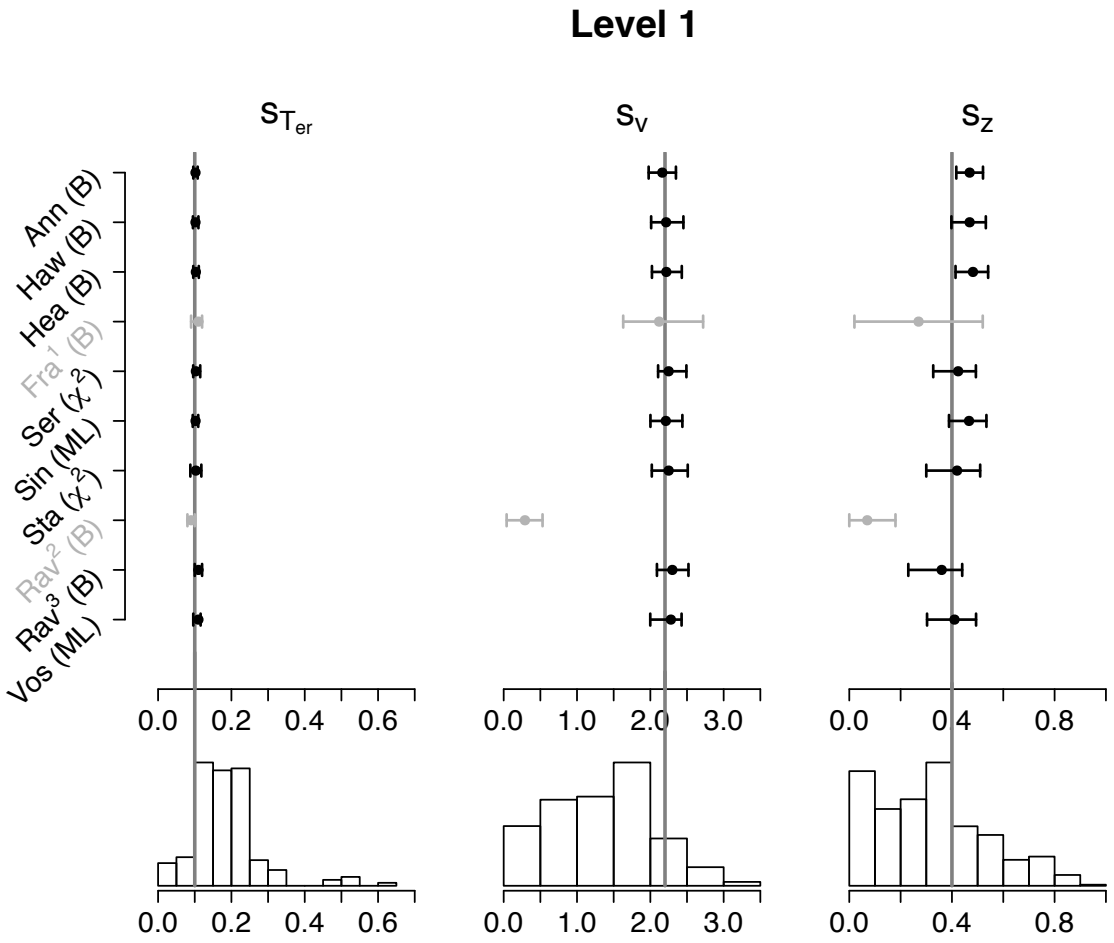

Figure 4: Estimates for across-trial variability parameters for Level 1 obtained with different estimation methods. Histograms at the bottom show the distribution of parameter values observed in empirical studies reported in Matzke and Wagenmakers (2009). The vertical line in each panel shows the generating parameter value. Dots indicate parameter estimates obtained by our collaborators, error bars represent the measures of uncertainty reported by our collaborators (see Table 2). Labels indicate the first author, abbreviations in brackets indicate the fitting methods (B: Bayes, ML: maximum-likelihood estimation, χ^2: χ^2-minimization for RT quantiles). Results shown in gray did not fix the main DDM parameters to their known values. [1]This fit was obtained on request of the organizers after the generating parameter values had been published. [2]This fit was obtained with an incorrectly scaled prior distribution on s_v. [3]This fit was obtained with a corrected prior distribution on s_v after the generating parameter values had been published; see van Ravenzwaaij's contribution in section A.3 for details.

al., shown in gray, was associated with a relatively wide uncertainty interval. As explained in their contribution, Frank et al.'s fitting method does not allow users to fix parameters to a specific value, and thus could not take advantage of the known DDM parameter values for this data set. Similarly, van Ravenzwaaij's initial model fit did not fix the main DDM parameters to their known values. The corresponding point estimate for s_v, shown in gray, missed the generating parameter value by a wide margin. As he explains in his contribution, this was due to a misspecified prior distribution for s_v, which strongly biased the parameter estimate.[1] The second estimate, which fixed the main DDM parameters to their known values and used an appropriate prior distribution, shown in black, was comparable to the estimates obtained with other methods.

Finally, most point estimates of s_z for the Level 1 data, shown in the right panel, missed the generating parameter value. Compared to the range of parameter values observed in empirical studies, the uncertainty intervals associated with these point estimates were relatively narrow. This bias in the estimates of s_z suggests that the parameter might not be sufficiently constrained by the data, even if the value of the z parameter is known exactly. Similar to the results for s_v, Frank et al.'s estimate for s_z was associated with a relatively wide uncertainty interval as their estimation method could not take advantage of the known DDM parameter values. Van Ravenzwaaij's initial point estimate for s_z, shown in gray, also missed the generating parameter value by a wide margin. The second estimate, shown in black, which used an appropriate prior distribution and fixed the known DDM parameters, was comparable to the estimates obtained with other methods.

The results for Level 2 show complementary patterns to the observations above. Figure 5 shows the point estimates and uncertainty intervals for the Level 2 data compared to the distribution of parameter values typically observed in empirical studies. Similar to the results for Level 1, all estimates for $s_{T_{er}}$, shown in the left panel, were close to the generating

[1] Note that van Ravenzwaaij's misspecified prior distribution for s_v also biased the posterior variance for s_v and s_z, resulting in relatively narrow uncertainty intervals.

parameter value and uncertainty intervals were narrow across methods, which again indicates that all estimation methods could reliably recover the value of $s_{T_{er}}$. Moreover, the width of the uncertainty intervals for Level 2 was similar to that for Level 1 for all estimation methods, which further suggests that $s_{T_{er}}$ is sufficiently constrained by the data and is not strongly dependent on the values of the main DDM parameters.

Point estimates of s_v for Level 2, shown in the middle panel, showed relatively small deviations from the generating parameter value compared to the range of values observed in empirical studies. However, across estimation methods there was considerable uncertainty associated with these point estimates, with uncertainty intervals spanning nearly half the range of empirical values. Moreover, compared to Level 1, point estimates for Level 2 showed higher variability around the generating value and the uncertainty associated with these estimates approximately doubled. Interestingly, uncertainty intervals were similar in width across estimation methods and the increase in uncertainty from Level 1 to Level 2 was also comparable across estimation methods. Taken together, these results suggest that s_v is dependent on the values of the main DDM parameters. Indeed, Singmann and Kellen found strong correlations between a, v, z and s_v, and Hawkins found a strong correlation between v and s_v. The initial estimate for s_v reported by van Ravenzwaaij again missed the generating parameter value by a wide margin. However, a second estimate that used an appropriate prior distribution was comparable to the estimates obtained with other methods.

Finally, point estimates of s_z for Level 2, shown in the right panel of Figure 5, deviated considerably from the generating parameter value compared to the range of values observed in empirical studies and uncertainty intervals spanned half the range of empirical values. Moreover, compared to Level 1, point estimates showed increased variability and uncertainty intervals doubled in width for most methods. Similar to s_v, the increase in uncertainty for estimates of s_z from Level 1 to Level 2 was comparable for all estimation methods. However, point estimates obtained from hierarchical Bayesian methods tended to lie closer to the generating parameter value than estimates obtained with other methods, which largely yielded estimates close to

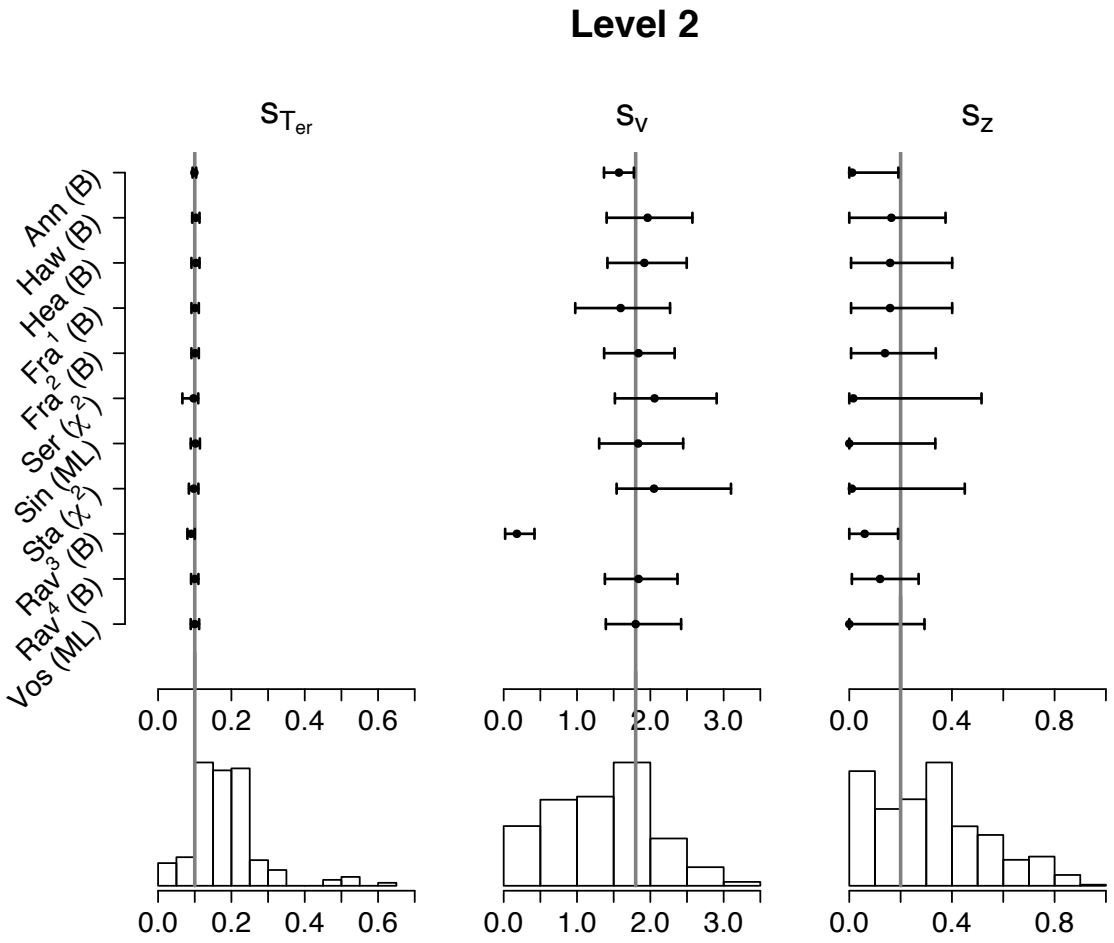

Figure 5: Estimates for across-trial variability parameters for Level 2 obtained with different estimation methods. Histograms at the bottom show the distribution of parameter values observed in empirical studies reported in Matzke and Wagenmakers (2009). The vertical line in each panel shows the generating parameter value. Dots indicate parameter estimates obtained by our collaborators, error bars represent the measures of uncertainty reported by our collaborators (see Table 2). Labels indicate the first author, abbreviations in brackets indicate the fitting methods (B: Bayes, ML: maximum-likelihood estimation, χ^2: χ^2-minimization for RT quantiles). [1]This fit was obtained using accuracy-coding. [2]This fit was obtained using stimulus-coding after the generating parameter values had been published; see Frank et al's contribution in section A.4 for details. [3]This fit was obtained with an incorrectly scaled prior distribution on s_v. [4]This fit was obtained with a corrected prior distribution on s_v after the generating parameter values had been published; see van Ravenzwaaij's contribution in section A.3 for details.

0. This relatively better performance of hierarchical Bayesian methods is likely due to the specification of the prior distribution for s_z, which is mostly based on the empirical distribution of parameter values reported by Matzke and Wagenmakers (2009). Consequently, even if s_z cannot be estimated accurately from the data, the prior distribution will pull point estimates into a region with higher prior probability. These results suggest that s_z, similar to s_v, is not sufficiently constrained by the data and is dependent on the values of the main DDM parameters. This conclusions is again supported by the strong correlations between s_z and T_{er} reported by Hawkins, and Singmann and Kellen.

In contrast to the across-trial variability parameters, the main DDM parameters could be estimated with high precision across estimation methods. The top row in Figure 6 shows the point estimates and uncertainty intervals for the Level 2 data compared to the distribution of parameter values typically observed in empirical studies. As can be seen, point estimates were close to the generating parameter values and uncertainty intervals were narrow for a, T_{er}, and z. Only Starns's, and Voss and Lerche's estimates for T_{er} and Franks et al.'s second estimate for z missed the generating value. The latter result is due to reporting $1 - z$ instead of z and correcting Franks et al.'s estimate for this misreporting yields a value much closer to the generating parameter value. Similarly, point estimates for the drift rates v were close to the generating parameter values across estimation methods; only van Ravenzwaaij's initial estimate missed the generating value. Although uncertainty intervals for v were wider than for the other main DDM parameters, the intervals are relatively narrow compared to the range of parameter values observed in empirical studies. These results suggest that the main DDM parameters can be estimated with relatively high precision at the level of individual participants.

The relationship between the main DDM parameters and the across-trial variability parameters is shown in Figure 7. Gray lines indicate the generating parameter values and black dots show the parameter estimates obtained by our collaborators. The size of each dot indicates how the correlation between the corresponding main DDM parameter and the across-trial variability parameter would change if the data point was removed from the computation of the

correlation, with larger dots being associated with larger changes in the estimated correlation. It is important to note that DDM parameters are generally not independent; Ratcliff and Tuerlinckx (2002), for instance, found correlations between most DDM parameters for individual participants to be at least 0.5. We, therefore, only consider correlations greater than 0.5 to be noteworthy. As can be seen in Figure 7, for $s_{T_{er}}$ (top row) data points for all parameters except a are similar in size, which means that the estimated correlations between $s_{T_{er}}$ and the main DDM parameters are not driven by outliers. For a, removal of the outlier in the bottom right corner of the panel resulted in a correlation of $r = 0.57$, which suggests that estimation of a was strongly dependent on $s_{T_{er}}$. The correlations for the remaining main parameters were small to medium in size, which suggests that the estimation of these parameters was not critically dependent on $s_{T_{er}}$ across estimation methods.

Similarly, for s_v (middle row) data points in the panels for a, T_{er}, and v are similar in size, which suggests that the estimated correlations are not driven by outliers. There are sizable positive correlations between s_v and a, and between s_v and all three drift rates v. This means that estimates of a and v were critically dependent on s_v. The correlation between s_v and z was strongly influenced by a single data point, removal of which increased the correlation to $r = 0.67$. This suggests that estimation z was also critically dependent on s_v.

Finally, for s_z (bottom row) data points in the panels for T_{er} and z are similar in size, which suggests that the estimated correlations are not driven by outliers. The medium-sized negative correlations with T_{er} and z indicate that estimates for these parameters were not critically dependent on s_z. The correlations of s_z with a and v were influenced by a single outlier. However, removal of this outlier did not yield sizable correlations, which suggests that estimates for a and v were not critically influenced by s_z.

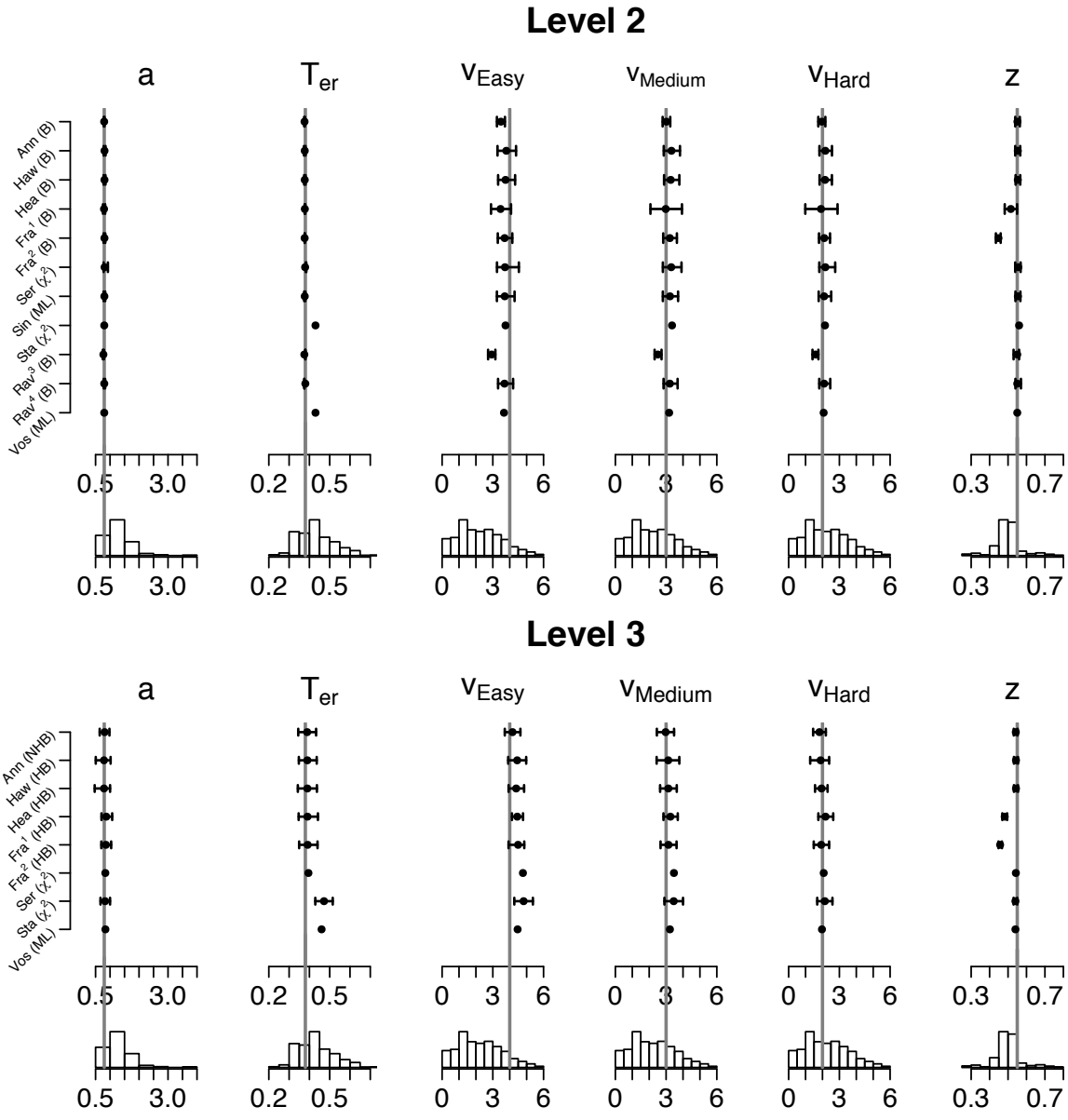
Level 2
a
T_{er}
v_{Easy}
v_{Medium}
v_{Hard}
z
Ann (B)
Haw (B)
Hea (B)
Fra^{1} (B)
Fra^{2} (B)
Ser (χ^2)
Sin (ML)
Sta (χ^2)
Rav^{3} (B)
Rav^{4} (B)
Vos (ML)
0.5 3.0
0.2 0.5
0 3 6
0.3 0.7
Level 3
a
T_{er}
v_{Easy}
v_{Medium}
v_{Hard}
z
Ann (NHB)
Haw (HB)
Hea (HB)
Fra^{1} (HB)
Fra^{2} (HB)
Ser (χ^2)
Sta (χ^2)
Vos (ML)
0.5 3.0
0.2 0.5
0 3 6
0.3 0.7

Figure 6: Estimates for the main DDM parameters for Levels 2 and 3 obtained with different estimation methods. Histograms at the bottom show the distribution of parameter values observed in empirical studies reported in Matzke and Wagenmakers (2009). The vertical line in each panel shows the generating parameter value. Dots indicate parameter estimates obtained by our collaborators, error bars represent the measures of uncertainty reported by our collaborators (see Table 2). Uncertainty intervals for Levels 2 and 3 were not available for some contributions that used χ^2-minimization and maximum-likelihood estimation. Labels indicate the first author, abbreviations in brackets indicate the fitting methods (B: Bayes, HB: hierarchical Bayes, NHB: non-hierarchical Bayes, ML: maximum-likelihood estimation, χ^2: χ^2-minimization for RT quantiles). [1]This fit was obtained using accuracy-coding. [2]This fit was obtained using stimulus-coding after the generating parameter values had been published. The large deviation from the generating value is due to misreporting $1 - z$ instead of z; see Frank et al's contribution in section A.4 for details. [3]This fit was obtained with an incorrectly scaled prior distribution on s_v. [4]This fit was obtained with a corrected prior distribution on s_v after the generating parameter values had been published; see van Ravenzwaaij's contribution in section A.3 for details.

Figure 8 shows the point estimates and measures of uncertainty for the across-trial variability parameters for the Level 3 data reported by our collaborators. The results are similar to those for the participant-level estimates for the Level 2 data. As can be seen, estimates for $\mu_{s_{T_{er}}}$ showed near perfect agreement with the generating parameter value. Moreover, compared to the range of empirical values for $s_{T_{er}}$, uncertainty intervals for the Level 3 data were negligible across estimation methods, which indicates that the parameter $\mu_{s_{T_{er}}}$ could be estimated with high precision. Point estimates for μ_{s_v} showed somewhat higher variability around the generating parameter value. However, this variability was small compared to the range of s_v values observed in empirical studies and uncertainty intervals for the point estimates of μ_{s_v} were relatively narrow. Finally, point estimates for μ_{s_z} deviated considerably from the generating parameter value compared to the range of values in empirical studies and uncertainty intervals were relatively

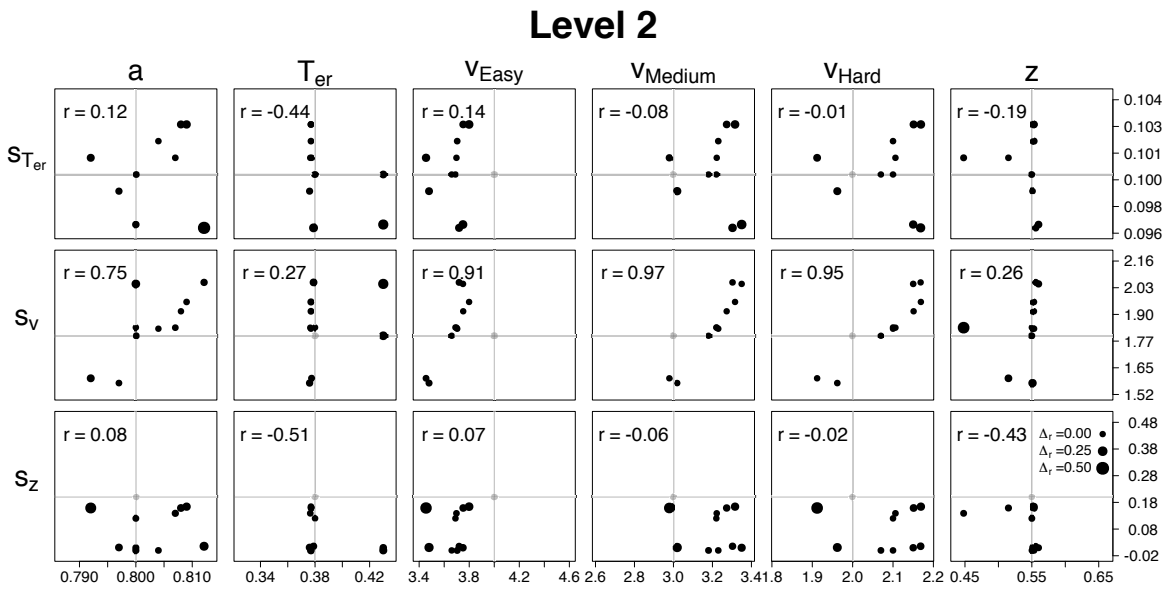

Figure 7: Correlations between the main DDM parameters and the across-trial variability parameters across estimation methods. Thin gray lines in each panel show the generating parameter values. Dots indicate parameter estimates obtained by our collaborators. Dot size represents the change in the estimated correlation if the data point is removed from the computation of the correlation; larger dots correspond to a larger change in correlation, $\Delta_r = |r_{all\,data} - r_{leave\,out\,i}|$. Results from van Ravenzwaaij's initial fit are not included as parameter estimates were considerably biased.

wide for most estimation methods. Similar to the uncertainty intervals for the participant-level estimates for Level 2, uncertainty intervals for μ_{s_z} for Level 3 were relatively wide, which suggests that s_z is insufficiently constrained by the data.

The results for the estimation of the group-level main DDM parameters parallel those for the individual-level parameters. The bottom row in Figure 6 shows the point estimates and uncertainty intervals for the Level 3 data compared to the distribution of parameter values typically observed in empirical studies. As can be seen, point estimates were close to the generating parameter values and uncertainty intervals were narrow for a. Similarly, most contributors' point estimates for T_{er}, v, and z were also close to the generating parameter value and the associated uncertainty intervals were narrow compared to the range of empirical values. As for Level 2, Starns', and Voss and Lerche's estimates for T_{er} were larger than the generating group-level parameter, and Frank et al.'s estimates for z were smaller than the generating group-level parameter. These deviations might, therefore, reflect biases in the estimation of the individual-level parameters. Finally, Servant and Logan's, Starns', and Voss and Lerche's estimates for v overestimated the drift rates in the easy and medium conditions. However, it is hard to assess whether these deviations reflect systematic biases in the estimation methods because there are no uncertainty intervals available for these group-level estimates and there were no comparable deviations visible for the Level 2 data. Taken together, these results show that the group-level main DDM parameters can be estimated with acceptable precision, although some methods might provide biased point estimates for T_{er}, z, and v.

The relationship between group-level estimates of the main DDM parameters and group-level estimates of the across-trial variability parameters is shown in Figure 9. Gray lines indicate the generating parameter values and black dots show the parameter estimates obtained by our collaborators. The size of each dot indicates how the correlation between the corresponding main DDM parameter and the across-trial variability parameter would change if the data point was removed from the computation of the correlation, with larger dots being associated with larger changes in the estimated correlation.

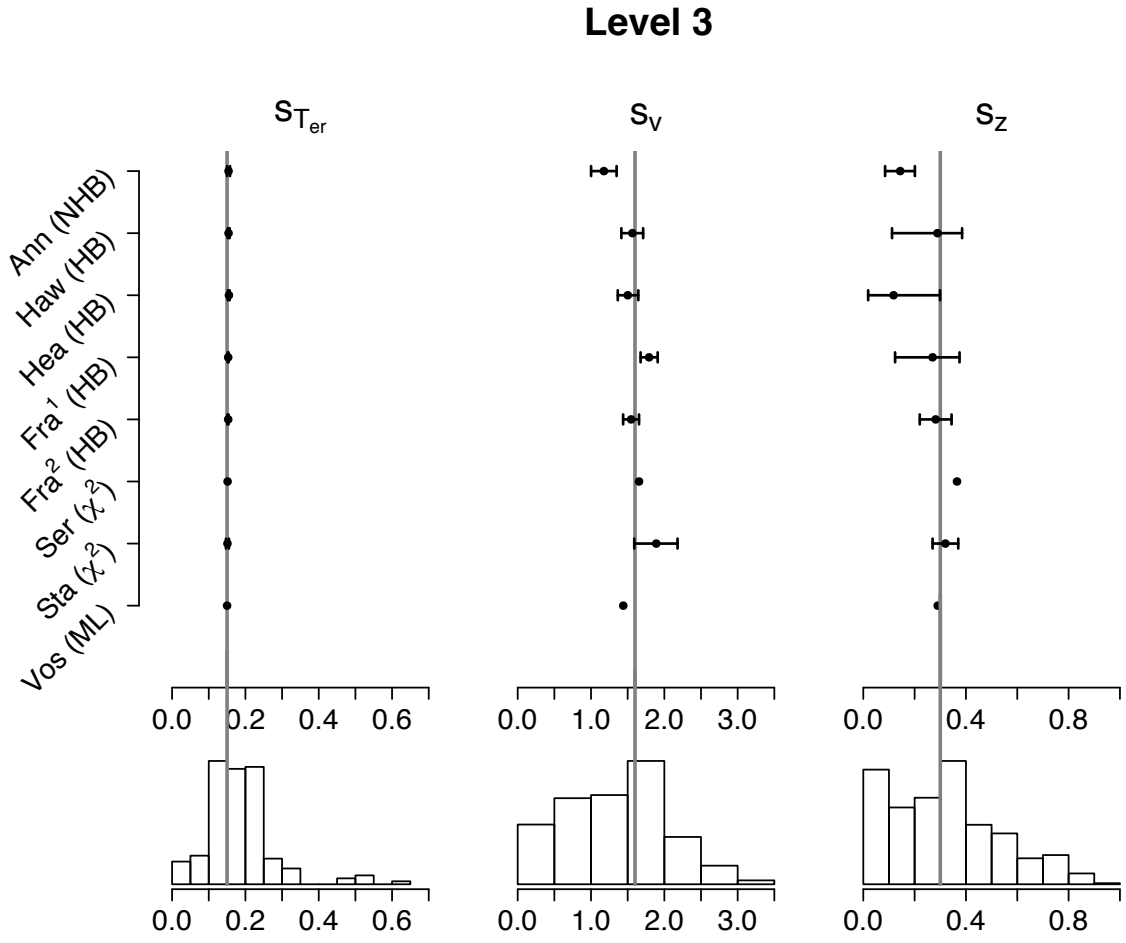

Figure 8: Estimates for the across-trial variability parameters for Level 3 obtained with different estimation methods. Histograms at the bottom show the distribution of parameter values observed in empirical studies reported in Matzke and Wagenmakers (2009). The vertical line in each panel shows the generating parameter value. Dots indicate parameter estimates obtained by our collaborators, error bars represent the measures of uncertainty reported by our collaborators (see Table 2). Labels indicate the first author, abbreviations in brackets indicate the fitting methods (HB: hierarchical Bayes, NHB: non-hierarchical Bayes, ML: maximum-likelihood estimation, χ^2: χ^2-minimization for RT quantiles). [1]This fit was obtained using accuracy-coding. [2]This fit was obtained using stimulus-coding after the generating parameter values had been published; see Frank et al's contribution in section A.4 for details.

As can be seen, for $s_{T_{er}}$ (top row) data points in each panel are similar in size, which suggests that the estimated correlations between $s_{T_{er}}$ and the main DDM parameters are not driven by outliers. Similar to Level 2 after outliers were removed, the correlations for a and z are medium-sized or small, which suggests that estimates of these group-level parameters were not critically dependent on $s_{T_{er}}$ across estimation methods. However, in contrast to Level 2, there are sizable negative correlations between $s_{T_{er}}$ and T_{er}, and between $s_{T_{er}}$ and v_{Easy} and v_{Medium} for Level 3. This suggests that estimation of these group-level parameters was critically influenced by $s_{T_{er}}$.

For s_v (middle row) data points in the panels for a, and v are similar in size, which suggests that the estimated correlations are not driven by outliers. In contrast to Level 2, the correlation between the estimates for s_v and a is only medium-sized, which suggests that estimation of the group-level parameter a was not critically dependent on the estimation of s_v. The estimated correlation between s_v and T_{er} was strongly influenced by two data points. Removal of these data points increased the correlation to $r = 0.61$, which suggests that s_v critically influenced the estimation of the group-level parameter T_{er}. Similar to Level 2, there are sizable positive correlations between s_v and all three drift rates v. This means that, also on the group-level, estimates of v were critically dependent on s_v. Moreover, as for Level 2, there is only a weak correlation between s_v and the group-level parameter z, which suggests that estimates of z were not critically dependent on s_v.

Finally, for s_z (bottom row) data points in all panels are similar in size, which suggests that the estimated correlations are not driven by outliers. Despite some differences in size, similar to Level 2, correlations between s_z and a, s_z and T_{er}, and between s_z and z were not substantial. This means that estimation performance for these group-level parameters was not critically dependent on the estimation of s_z. In contrast to Level 2, the sizable positive correlations between s_z and v_{Easy} and v_{Medium} suggest that estimation of these group-level drift rates was critically influenced by s_z.

Taken together, the results for Level 3 confirm the strong correlations between s_v and

estimates of v observed for Level 2, but suggest additional strong correlations between v and $s_{T_{er}}$, between v and s_z, and between T_{er} and $s_{T_{er}}$. Moreover, the results for Level 3 did not show the strong correlation between a and s_v observed for Level 2. These results might be taken to suggest different dependencies between estimates of DDM group-level parameters than between estimates of participant-level parameters. However, these discrepancies might equally well be a product of chance variation due to the small number of contributions on which the correlations are based.

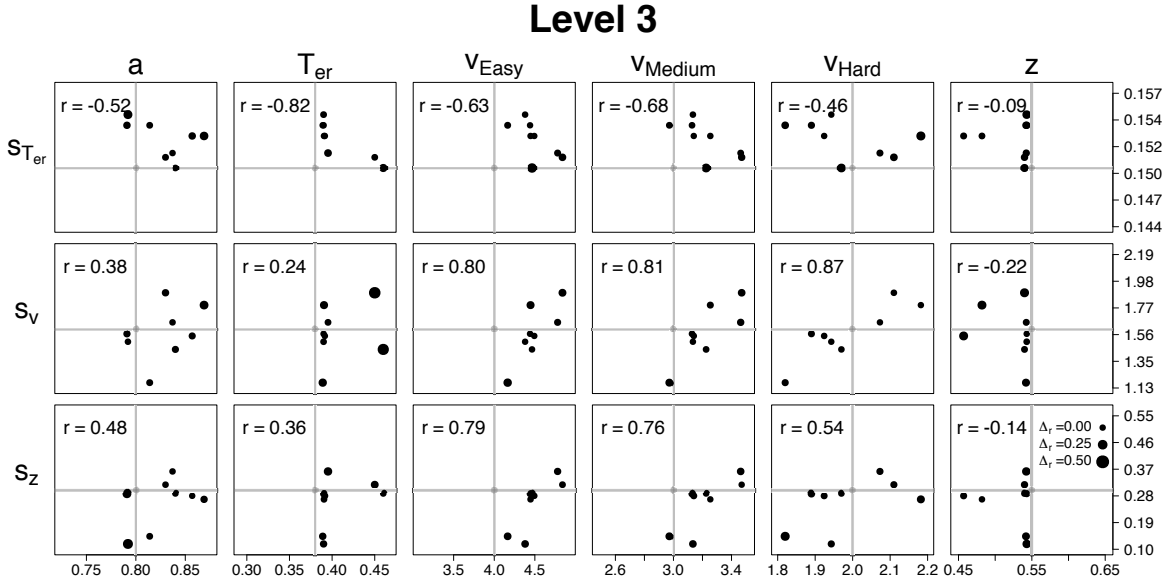

Figure 9: Correlations between group-level means of main DDM parameters and across-trial variability parameters across estimation methods. Thin gray lines in each panel show the generating parameter values. Dots indicate parameter estimates obtained by our collaborators. Dot size represents the change in the estimated correlation if the data point is removed from the computation of the correlation; larger dots correspond to a larger change in correlation, $\Delta_r = |r_{all\,data} - r_{leave\,out\,i}|$.

5 Advice

5.1 Bayesian Estimation

Our collaborators discussed two main problems often encountered with Bayesian methods that rely on MCMC sampling. First, effective approximation of the posterior density requires that MCMC chains have converged to their equilibrium distribution. That is, MCMC samples should reflect genuine samples from the posterior distribution. However, chains might get stuck at a particular value for longer periods of time without having converged, or exhibit a very slow drift towards the equilibrium distribution. In both cases automatic convergence checks might falsely indicate that the chains have converged. Users should, therefore, always visually check that chains have converged and are fluctuating around a common value.

If a sufficient number of chains have been sampled, post-hoc removal of non-converged chains might help address convergence problems without affecting parameter estimates. For the DE-MCMC algorithm, one way to address convergence problems is to use a migration step during burn-in in which samples are exchanged between chains. This allows chains that are far from the other chains to be pulled towards a common value.

Second, the across-trial variability parameters are associated with a relatively flat likelihood function, and hence are not well constrained by the data. In a hierarchical setting in particular, this can result in poor prior updating, where MCMC chains remain stuck in the prior distribution. Such problems can be detected by superimposing the prior distribution and the posterior distributions in a single figure to verify that the estimates reflect the posterior more than the prior. Moreover, repeated sampling with different sensible prior settings should yield similar results for the posterior samples if prior updating occurred.

Users of numerical integration methods might benefit from using estimates of the posterior mode and covariance matrix obtained from a Laplace approximation to initialize the quadrature procedure. The `Simplex` algorithm provides a fast and efficient way to compute Laplace approximations. One limitation of quadrature methods is that their use is limited to models with 10 or fewer parameters, which typically precludes applications to hierarchical

models.

In general, users of Bayesian estimation methods should be aware that these methods are sensitive to serious misspecifications of prior distributions. Users should therefore check that prior specifications are sensible and priors might need to be rescaled for different parameterizations of the DDM (e.g., if the diffusion coefficient s is changed from 0.1 to 1). Lastly, users should be aware that estimating posterior means and HDIs in hierarchical models using MCMC sampling is computationally expensive.

5.2 Maximum-Likelihood Estimation

Parameter estimation using ML methods requires efficient numerical optimization. Within the setup used by Singmann and Kellen, use of the `nlminb` algorithm (Kaufman & Gay, 2003) is recommend as it converges quickly on global optima. In the setup used by Voss and Lerche, the `Simplex` algorithm seems to provide a good compromise between speed of convergence and convergence to global, rather than local, optima.

One drawback of ML estimation methods is their sensitivity to contaminant RTs, which can considerably bias parameter estimates. Whereas in the present study all RTs were known to have been generated by the DDM, in applications to real data more robust estimation methods should be used, such as estimation based on the Kolmogorov-Smirnov statistic.

5.3 χ^2 Minimization

Parameter estimation using χ^2 minimization, similar to ML estimation, requires efficient numerical optimization, in this case of the χ^2 statistic. The optimization method used by our collaborators relies on an iterative procedure using the `Simplex` algorithm. The χ^2 statistic is minimized for a set of starting values, and the resulting parameter estimates are used as starting values for a new iteration of optimization process. This iterative scheme is repeated until the parameter estimates do not change substantially between iterations. Servant and Logan observed that the resulting parameter estimates are dependent on the starting values used in the first iteration, in particular for the parameters v, s_v, and s_z. These instabilities in the parameter

estimates might be due to trade-offs between v and s_v, and a flat likelihood function for s_z, which might be addressed by either fixing or combining parameters that are not well recovered (White, Servant, & Logan, 2017).

Ratcliff and Childers (2015) recently suggested a further refinement of the χ^2 method, where the median RT of errors is used in the computation of the χ^2 statistic, rather than ignoring errors completely if their number is below 10. This refined method might improve parameter estimation for Level 3 where the number of error RTs was small for some data sets.

5.4 General Recommendations

Several of our collaborators reported high correlations and trade-offs between DDM parameters. In particular, s_v, s_z, and v seem to be highly correlated, which complicates their joint estimation. A first way to deal with this problem is to forgo estimation of the across-trial variability parameters altogether and fix their value to 0, based on the motivation that the across-trial variability parameters were introduced into the DDM to account only for fine-grained details of the RT distribution (van Ravenzwaaij, Donkin, & Vandekerckhove, 2017; Ratcliff & Tuerlinckx, 2002, e.g.,). In many practical applications, however, the focus is on the main DDM parameters. In these cases, the across-trial variability parameters increase model complexity without tangible benefits for the estimation of the main DDM parameters; the main DDM parameters can often be estimated precisely even if the data were generated by a DDM with non-zero across-trial variabilities (Lerche & Voss, 2016).

Second, if users decide to estimate the across-trial variability parameters, several steps should be taken to improve the quality and interpretability of parameter estimates. Obtaining a sufficient number of trials is a prerequisite for estimating the across-trial variability parameters. However, simply increasing the length of an experimental session means that participants might lose motivation and focus, which might, in turn, introduce contaminant RTs and thus affect the precision of parameter estimates.

As a general rule, researchers are expected to quantify the error associated with the parameter estimates, for example by obtaining bootstrap confidence intervals. However, in

applications to real data, such confidence intervals are influenced not only by the estimation error but also by potential model misspecification, that is, if data were generated by a different model than the DDM. Therefore, additional methods such as parametric bootstrap should be employed to more appropriately assess estimation error and detect model misspecification.

Finally, due to the high uncertainty associated with the across-trial parameters, comparisons of parameter estimates across participants are notoriously unreliable. In between-subjects designs, all parameters need to be estimated for each participant in each condition, which means that in comparisons across conditions, uncertainty in one parameter can compound uncertainty in another. In within-subjects designs, on the other hand, only the parameters of interest need to be estimated in each condition, all other parameters are assumed to have the same value across conditions. This might allow for meaningful comparisons of across-trial variability parameters in some instances. In memory research, for example, simulations studies indicate that differences in drift rate variability between experimental conditions can be recovered with some reliability (Starns & Ratcliff, 2014) and validation studies were able to detect manipulations of evidence variability in empirical data (Starns, 2014).

6 Discussion

Over the last 40 years, the DDM has become one of the most popular models for explaining RT and accuracy data from a wide range of domains (Forstmann et al., 2016; Ratcliff et al., 2016; Ratcliff & McKoon, 2008). Much of this success is due to the model's ability to fit varied shapes of RT distributions; through the addition of three across-trial variability parameters, the DDM can account for subtle RT patterns that elude most competitor models (Ratcliff, 1978; Ratcliff & Tuerlinckx, 2002; Van Zandt & Ratcliff, 1995). However, several recent studies have reported difficulties estimating these across-trial variability parameters, even in sizable data sets (Lerche & Voss, 2017; Lerche & Voss, 2016; Yap et al., 2012; van Ravenzwaaij & Oberauer, 2009). For example, van Ravenzwaaij and Oberauer (2009) generated data from the full DDM and considered two criteria for fitting the full DDM, one based on a Kolmogorov-Smirnov statistic and one based on a maximum-likelihood type of criterion. They found that both

fitting methods could accurately recover the main DDM parameters as well as the across-trial variability in non-decision time, whereas estimates of the across-trial variability in drift rate and starting point missed the generating parameter values by a wide margin. Ratcliff and Tuerlinckx (2002) found similar results across a wide range of generating parameter values for the main DDM parameters, using a maximum-likelihood and a χ^2-criterion, among others. Moreover, Ratcliff and Tuerlinckx reported sizable correlations between the main DDM parameters and the across-trial variability parameters, which suggests that poor estimation of the across-trial variability parameters might negatively affect estimation of the main DDM parameters.

These findings raise the question whether and how different fitting methods can be optimally used to obtain the best possible estimates of the across-trial parameters. Since van Ravenzwaaij and Oberauer (2009) and Ratcliff and Tuerlinckx's (2002) studies, several new fitting methods and software packages have become available. Using these packages often requires decisions about optimization or sampling algorithms, or adjustments to the implementation, based on expert knowledge of the method. However, many users do not have the required expertise nor the resources to conduct extensive simulation studies to find the best possible approach to fitting their data. Therefore, the current study invited experts from the DDM community to apply their fitting methods to a standard experimental setup and provide recommendations for estimating the DDM's across-trial variability parameters.

The experts contributing to our study used a wide range of fitting methods for the DDM and reported similar difficulties as Lerche and Voss (2017), Lerche and Voss (2016), Yap et al. (2012), and van Ravenzwaaij and Oberauer (2009) when estimating the across-trial variability parameters. Besides practical limitations, such as some methods being unable to fit specific data structures (e.g., the hierarchical structure, or the single-participant structure with some DDM parameters known), the estimation performance of the different methods depended strongly on the specific DDM parameter. Most estimation methods used by our collaborators could accurately recover the main DDM parameters as well as across-trial variability in non-decision time. Estimates of the across-trial variability in drift rate and starting point, on the other hand,

were associated with large uncertainty and tended to miss the generating value by a wide margin. These results are largely in line with those of Ratcliff and Tuerlinckx (2002), who could accurately recover the main DDM parameters on the individual-level but reported large uncertainty for estimates of across-trial variability in drift rate and starting point. Interestingly, uncertainty intervals in our study were similar in width across estimation methods and the increase in uncertainty from a situation where the main DDM parameters were known to a situation where all DDM parameters had to be estimated was comparable for all estimation methods. This indicates that estimation performance was not limited by the estimation methods themselves but rather by the degree to which specific DDM parameters are constrained by the data.

Our results further suggest tradeoffs in the estimation of the main DDM parameters and the across-trial variability parameters. Specifically, we found strong correlations between collaborators' estimates for drift rate variability and drift rate as well as between drift rate variability and boundary separation on the individual-level. Moreover, group-level estimates of all three across-trial variability parameters were strongly correlated with estimates of drift rate, and group-level estimates of variability in non-decision time and drift rate were also correlated with estimates of non-decision time. Although these correlations should be interpreted carefully due to the small number of data points on which the correlations are based, our results generally align with those of Ratcliff and Tuerlinckx (2002). Ratcliff and Tuerlinckx reported strong correlations on the individual-level between drift-rate variability and boundary separation and drift rate, as well as between variability in starting point and boundary separation, non-decision time, and drift rate. Our results suggest that bias in estimates of across-trial variability in drift rate affects estimation performance for the main parameters on all hierarchical levels, and that biased estimates of variability in non-decision time and starting point additionally affect group-level estimates of the main DDM parameters.

6.1 Limitations

There are three aspects of our study that might limit the generalizability of our results. The first aspect concerns the setup of our Level 3 data set, where the across-trial variability parameters were the same for all participants. Hierarchical Bayesian methods often assume that all individual-level parameters are sampled from a non-degenerate group-level distribution with positive variance, and therefore rely on a hyperprior that does not support zero variance (e.g. Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013; Gelman, 2006). However, because we assigned the same value to the across-trial variability parameters for all participants, the true variance of the group-level distribution was zero. This can cause estimation problems in hierarchical Bayesian methods such as `DMC` (Heathcote et al., in press) and, at the same time, gives an unfair advantage to implementations such as `HDDM` (Wiecki et al., 2013), which assumes common across-trial variability parameters across participants. Nevertheless, the results obtained with `DMC` and `HDDM` for the Level 3 were similar; uncertainty intervals for non-decision time and drift rate variability were comparable in width and point estimates were close to the generating value for both implementations. Only uncertainty intervals for starting point variability were wider for `DMC` than for `HDDM` and one `HDDM` point estimate missed the generating value by a sizable margin. Thus, although `DMC` might perform better with a different setup where across-trial variability parameters differ between participants, the present results suggest that `DMC` is relatively robust to such model misspecification.

In addition to these technical considerations, our design choice for the Level 3 data set highlights a more general problem in the specification of cognitive models. Several of our contributors were reluctant to accept certain parameter values as plausible. Heathcote and Hawkins, for example, point out that a group-level variance of 0 for the across-trial variability parameters is implausible. This assumption is embodied in their specification of the prior distribution for the group-level variances, which contains 0 as a boundary value. On the other hand, some modeling practices fix model parameters to a specific value. `HDDM`, for example, assumes a group-level variance of 0. These two modeling approaches, either estimating a

distribution or fixing a parameter to a particular value, essentially represent different a priori choices with regards to model complexity. Although it might be argued that DDM parameters will never be exactly the same across participants, the variance might be so small to be 0 for all intents and purposes. Hence, fixing parameter values might be an appropriate modeling choice in some instances. The question when the simpler model should be preferred is a complicated one and has been discussed extensively in the literature on model selection. We will not pursue this discussion here and refer the interested reader to the relevant literature (e.g., Aho, Derryberry, & Peterson, 2014; Burnham & Anderson, 2002; Jeffreys, 1961; McQuarrie & Tsai, 1998). However, we would like to point out that in applications of hierarchical implementations of the DDM, researchers might need to consider the level of model complexity that is needed to answer their research question. When the focus is on the application of a particular hierarchical model, model complexity is a negligible factor. In contrast, when the goal is statistical inference about or prediction based on model parameters, model complexity plays a pivotal role.

The second limiting aspect concerns our choice of generating parameter values. This issue is best highlighted by a comparison of our results with the results of Ratcliff and Childers's (2015) parameter recovery study. Ratcliff and Childers compared parameter recovery using, among others, χ^2-minimization, `Fast-dm`, and `HDDM` for 48 different combinations of generating parameter values and different numbers of conditions and trials per condition. Similar to our results, all methods in Ratcliff and Childers's study could accurately recover non-decision time. However, whereas all methods in our study could accurately recover boundary separation, in Ratcliff and Childers's study `Fast-dm` and `HDDM` produced biased estimates for some combinations of generating parameter values. Similarly, whereas all methods in our study could accurately recover drift rate, Ratcliff and Childers reported considerable biases in the estimation of drift rate for some generating parameters for `HDDM` and a general tendency to underestimate drift rate for `Fast-dm`. Biases in the estimation of boundary separation and drift rate in Ratcliff and Childers's study were more pronounced for larger generating values.

The interpretation of the discrepancies between Ratcliff and Childers's (2015) and our

results is somewhat hampered by the fact that Ratcliff and Childers did not report results for the across-trial variability parameters. One notable difference between our two studies is the range of generating values. In particular, Ratcliff and Childers used generating values of $s_z = 0.2$, $s_z = 0.6$, and $s_z = 0.8$ whereas we used values of $s_z = 0.2$ and $s_z = 0.3$ for the Level 2 and Level 3 data sets. Moreover, Ratcliff and Childers used generating values of $a = 1$ or $a = 2$, whereas we used a relatively small value of $a = 0.8$ for Level 2 and a small mean $\mu_a = 0.8$ and standard deviation $\sigma_a = 0.3$ for Level 3. At the same time, all other generating parameters in our study fell into the middle range of values reported in the literature.

Hence, the worse recovery performance of some methods for the main DDM parameters in Ratcliff and Childers's study might have been due to trade-offs with the across-trial variability parameters and the large variability in the data resulting from relatively extreme generating values for the main DDM parameters. This suggests that the generally good performance of all methods in our study might not generalize to other settings. In particular, drift rate and to a lesser degree also boundary separation might be estimated with lower precision or estimates might be systematically biased under alternative generating parameter values.

However, we do believe that the good recovery performance for non-decision time variability in our study will likely generalize to other settings despite the low generating values of $s_{T_{er}} = 0.1$ for Level 1 and Level 2. As across-trial variability in non-decision time determines how well-defined the leading edge of the RT distributions is, it might be argued that the small generating value caused minimal smearing of the leading edge, and might hence be responsible for the good recovery results for this parameter. However, despite the larger generating value for the Level 3 data, recovery was also very accurate in this case. Moreover, Ratcliff and Tuerlinckx (2002) could recover a generating value of $s_{T_{er}} = 0.2$ with remarkable accuracy even in the presence of contaminant RTs when using appropriate outlier corrections.

The third limiting aspect concerns the lack of outlier RTs in our simulated data. The results in Ratcliff and Childers (2015) suggest that some of the methods in the present study might show worse recovery performance if outlier RTs are present.

6.2 How and when to estimate across-trial variability parameters

The results of our simulations, in line with previous studies (Lerche & Voss, 2017; Lerche & Voss, 2016; Yap et al., 2012; van Ravenzwaaij & Oberauer, 2009; Ratcliff & Tuerlinckx, 2002; Ratcliff & Childers, 2015), show that the DDM's across-trial variability parameters are notoriously difficult to estimate. At the same time, the high correlations between the main DDM parameters and the across-trial variability parameters (Ratcliff & Tuerlinckx, 2002) imply that misestimation of the latter might bias estimates of the main DDM parameters. This raises the question how such biases can be minimized. A possible solution to the problem of estimating across-trial variability parameters is to place constraints on the admissible range of parameter values. As seen in our study, at least for variability in starting point, hierarchical Bayesian methods tended to yield point estimates that were close to the generating parameter value. This is due to the prior distribution these methods place on the DDM parameters, which pulls parameter estimates towards a priori plausible values if the data provide insufficient information to estimate the parameters. Other fitting methods might similarly benefit from constraining parameters to lie within the range of values observed in previous studies, as was done in the contribution by Servant and Logan. However, as discussed in van Ravenzwaaij's contribution, such constraints need to be carefully adjusted to the specific implementation of the DDM as incorrect prior information can severely bias parameter estimates. A good starting point for constructing constraints on DDM parameters are large-scale surveys of published DDM fits, as provided by Matzke and Wagenmakers (2009).

A further factor that might improve estimation of the across-trial variability parameters is experimental design. The present study aimed to showcase the application of different estimation methods to a standard experimental design as it is often used in functional neuroimaging and clinical psychology. This was motivated mainly by the practical constraints that derive from, for example, neurophysiological recordings. However, in cases where there are weak constraints on the number of conditions, designs that use multiple appropriately spaced difficulty conditions might allow for more precise estimation of the across-trial variability in drift and starting point.

As pointed out in the introduction, the effect of variability in drift rate and starting point is to change the relative speed of correct and error responses. In a quantile probability function these effects are most clearly visible as a change in the left-right symmetry of the highest quantiles (typically the .9 quantile; Ratcliff & McKoon, 2008). Consequently, accurate estimation of these shifts in symmetry requires sufficient information about the tails of the correct and error RT distributions at different accuracy levels that should span a wide range of accuracies. This has two practical implications. First, researchers should include several difficulty conditions (five or six) that are spaced in a way that accuracies span a large part of the range from 0.5 to 1.0. Second, researchers should collect sufficient numbers of trials, especially in low difficulty conditions, to obtain reliable estimates of the highest quantiles of the error RT distribution. It should be noted that, as a consequence of these two recommendations, researchers need to collect sufficient numbers of trials for each difficulty condition. This might not be practically feasible in research areas with strong limitations on the total number of trials, such as clinical psychology or functional neuroimaging.

Another possible approach for improving estimation of the across-trial variability parameters might be to use quantile-averaged data instead of fitting the DDM to individual participants' data. Cohen, Sanborn, and Shiffrin (2008) considered how model recovery is affected if models are selected based either on individual participants' data, or based on averaged data. Their results showed that model recovery based on averaged data could outperform model recovery based on individual participants' data if the number of trials per participant was low. It might therefore be suggested that estimation of the DDM's across-trial variability parameters might also benefit from using averaged data instead of individual data. However, this approach only yields a single group-level estimate for each DDM parameter and provides no information about the variance of the parameter values across participants. This precludes statistical comparisons of parameter estimates between experimental conditions and the computation of correlations with external variables across participants, both of which are often of central interest in experimental studies. A more suitable, hybrid approach might be to use averaged

data to estimate the across-trial variability parameters and subsequently estimate the main DDM parameters for each participant with the across-trial variability parameters fixed to the values obtained from the averaged data. However, Ratcliff and Childers (2015) found that using such a hybrid approach conveyed no improvement in parameter recovery over estimating the across-trial variability parameters at the participant-level.

An interesting difference between the present study and previous studies that tested recovery of across-trial variability parameters is that while we found that all fitting methods could accurately estimate the variability in non-decision time, earlier studies found estimates of non-decision time variability to be unreliable. For example, Lerche and Voss (2017) reported that estimates of variability in non-decision time correlated only weakly between sessions of a lexical decision task, and Yap et al. (2012) found only modest correlations between estimates of variability in non-decision time from the same session of a lexical decision task. This discrepancy in results is most likely due to the use of simulated data from the DDM in the present study whereas Lerche and Voss (2017), and Yap et al. (2012) used experimental data. In experimental data, the true value of the variability in non-decision time might vary over time, which results in a decreased retest reliability (Lerche & Voss, 2017). Moreover, experimental data might contain outlier RTs. Fast outliers in particular affect the location of the leading edge of the RT distribution, which in turn depends on the variability in non-decision time (Ratcliff & Tuerlinckx, 2002), thus leading to biased estimates of non-decision time variability. Although the problem of fast outliers can be addressed to some degree by excluding RTs below a certain cutoff value or by explicitly modeling outliers as being generated by a different process than the DDM, separating genuine responses from outliers is inherently difficult (Ratcliff & Tuerlinckx, 2002). Consequently, estimates of non-decision time variability from experimental data generally need to be interpreted with care. At the same time, this susceptibility to outliers makes non-decision time variability an important DDM parameter. As pointed out by Lerche and Voss (2016), variability in non-decision time can potentially absorb the effects of fast outliers that would otherwise bias estimates of the main DDM parameters. The results of our present study suggest

that variability in non-decision time is only modestly correlated with boundary separation and is not critical to the estimation of the remaining main DDM parameters at the participant-level. Moreover, non-decision time variability is often not of substantial interest to researchers. A pragmatic approach might, therefore, be to view non-decision time variability as a nuisance parameter and forego interpretation of this parameter.

In general, the question when to estimate across-trial variabilities is more difficult to answer. This is also reflected in the diverse recommendations our collaborators provide. Van Ravenzwaaij recommends categorically against estimating across-trial variability parameters. Heathcote suggests that drift rate and non-decision time variability can usually be estimated with reasonable precision while starting point variability should only be estimated if there are clear indications of fast errors. Voss & Lerche prefer simple models that only estimate non-decision time variability and fix the remaining across-trial variabilities to 0. Similarly, Starns recommends always estimating variability in non-decision time and suggests that variability in drift rate and starting point might be fixed to standard values in most applications.

At the core of these diverse recommendations sits the question which DDM parameters can be neglected without negatively affecting the estimation of other parameters. This question has been discussed extensively in the literature. Wagenmakers, Van der Maas, and Grasman (2007) proposed a simplified version of the DDM that fixed the across-trial variability parameters to 0 and assumed that starting point was equidistant to the decision boundaries. In the ensuing debate about the appropriateness of these simplifying assumptions, Ratcliff (2008) pointed out that applying the simplified model to data generated from the full model resulted in biased parameter estimates. However, in applications to real data, the generating model is unknown.

A simple heuristic to decide which across-trial variabilities to include in a model might be to compare the mean correct and error RT and only include the necessary across-trial variability parameters if the means differ. The main drawback of this approach is that mean differences are not necessarily diagnostic. Across-trial variability parameters affect the entire distribution of correct and error RTs, and in particular the tail quantiles. Changes in these quantiles are

notoriously difficult to detect. A more principled approach might be to compare the results of fitting models with and without across-trial variabilities (see also Vandekerckhove & Tuerlinckx, 2007). If both types of models yield the same conclusion, across-trial variabilities can be safely neglected. If the conclusions differ, careful consideration should be given to the possible causes of this discrepancy.

Finally, in recent years there has been increasing interest in the substantive interpretation of the across-trial variability parameters. For example, several authors have argued that variability in drift rate might be related to mind-wandering (McVay & Kane, 2012; Hawkins, Mittner, Forstmann, & Heathcote, 2017). In these cases, where the across-trial variability parameters themselves are of interest, researchers need to ensure that all possible precautions have been taken to optimize estimation of the across-trial variability parameters (i.e., removal or explicit modeling of outlier RTs, sufficient number of difficulty conditions and trials per participant) before proceeding to interpret their results.

To sum up, independent of the particular DDM fitting method used, most of our collaborators agree on two points. First, the DDM's across-trial variability parameters are inherently hard to estimate and there is considerable uncertainty associated with these estimates. A method that can be used to improve this situation is to use parameter estimates from previous studies to inform current estimates. Second, although the across-trial variability parameters afford the DDM a high degree of flexibility, they are often not the focus of inference. Therefore, users should give careful consideration to whether across-trial variability parameters are actually needed in order to fit a particular data set.

Appendix A Individual Contributions - Bayesian Estimation

Unless indicated otherwise, the known parameters for the Level 1 data are set to the true value in each contribution and the DDM parameters are defined as above. In cases where the non-decision time parameter represents the lower bound of the non-decision time distribution, rather than the mean, we will use $T_{er}^{\star}$ instead of T_{er}.

A.1 Heathcote

A.1.1 Methods. Parameter estimates were obtained by Bayesian methods using the Differential-Evolution Markov Chain Monte Carlo (DE-MCMC; ter Braak, 2006) sampler implemented in the `R` language (R Core Team, 2015) in the `Dynamic Models of Choice` (`DMC`; Heathcote, Lin, & Gretton, 2016, Heathcote et al., in press) software.[2] DE-MCMC is a multiple-chain Metropolis sampler with a proposal that automatically adapts to posterior parameter correlations using a "crossover" step, where each chain is updated based on a weighted linear combination of its state and the difference between the states of two other randomly selected chains. During "burn-in" (initial iterations later discarded) we also used "migration" steps to pull in chains stuck in low likelihood areas (see Turner, Sederberg, Brown, & Steyvers, 2013, for a tutorial overview of these methods). The DDM likelihood was calculated using the `rtdists` package (Singmann et al., 2016), with the minimum value for each data point set to 10^{-10} to avoid numerical problems when calculating log-likelihoods.

Sampling used DMC defaults in most cases. For Level 1 and 2 the crossover weight (γ) was set at $2.38/\sqrt{D}$, where D, the number of chains, was set at three times the number of estimated DDM parameters updated in a single block ($D = 3$ for Level 1 and $D = 9$ for Levels 2 and 3). Level 3 estimation was hierarchical, with the same settings when sampling DDM parameters, except group-level parameter crossover weights were sampled from a uniform

[2] DMC is based on code originally written by Brandon Turner and Scott Brown, and comes with a set of tutorials on fitting not only the DDM but also a variety of other models including the LNR (Heathcote & Love, 2012), LBA (Brown & Heathcote, 2008) and the BEESTS model of the stop-signal task with trigger failures (Matzke, Love, & Heathcote, 2017).

Table 3

Specification of prior distributions for DDM parameters in Heathcote's contribution.

	a	v	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
Mean (μ)	1	2.5	0.3	0.5	1	0.5	0.5
Standard Deviation (σ)	1	3	1	1	3	1	1
Lower Bound (L)	0	$-\infty$	0	0	0	0	0
Upper Bound (U)	2	∞	1	1	3	1	1

distribution on $[0.5, 1]$. During burn-in the probability of doing a migration rather than cross-over step was set at 0.05 for both DDM parameters and, where applicable, group-level parameters.

Sampled DDM parameters and corresponding independent truncated normal prior distributions, $\mathcal{N}(\mu, \sigma)[L, U]$, are defined in Table 3. In contrast to the main text, we define non-decision time ($T_{er}^{\star}$) as the lower bound of the uniform non-decision time distribution. The same prior was used for the drift rates in the easy, medium and hard condition, and for fits to Levels 1 and 2. In the hierarchical case individual-level DDM parameters were assumed to come from independent truncated normal group-level distributions (with truncation $[0, \infty]$ for a, v, $T_{er}^{\star}$, s_v and $s_{T_{er}}$, and truncation $[0, 1]$ for z and s_z). The group-level means had the same priors as in Table 3. The group-level standard deviations were all given the same Gamma prior with shape parameter 2 and scale parameter 0.25.

For Levels 1 and 2 burn-in was done in two stages. After obtaining initial starting values by sampling from the prior, the DMC function `run.unstuck.dmc` repeatedly sampled fresh sets of iterations of length *nmc* (here $nmc = 100$) with migration on (each starting from the last value in the previous set). This was repeated until means of each chain's summed posterior log-likelihoods were all less than a criterion absolute difference (by default 10) from the median of the chain means. Subsequently, thinning was set to 10 (i.e., only every 10th set of posterior samples was retained; from here the number of iterations will refer to the number retained), migration was turned off, and the `run.converge.dmc` function used to obtain a set of chains

that are mixed together.

Mixing was quantified by the multivariate potential scale reduction factor $\hat{R}$ (MPSRF; Brooks & Gelman, 1998) calculated by the `CODA` package using transforms to improve normality if appropriate (Plummer et al., 2016). Stationarity was simultaneously checked by splitting the chains in half before calculating $\hat{R}$ (Gelman et al., 2013). The `run.converge.dmc` function first takes a fresh set of iterations (here 100), then sets of *nmc* iterations (here 50) repeatedly sampled and added, with the first *nmc* iterations discarded if that improves $\hat{R}$. The process was run until $\hat{R}$ was close to one (here the default < 1.1 was used). For Level 3 initial fits to each individual data set using the same methods as applied to Level 2 were used to obtain start points for group-level parameters, based on the means and standard deviations of the parameter estimates for individual participants. Hierarchical models were then fit by the `h.run.unstuck.dmc` and `h.run.converge.dmc` functions, which apply the tests to all chains at both levels (i.e., to each participant individually and to the group-level).

The initial instructions for the collaborative project did not specify which boundary corresponded to "left". In all fits it was assumed that "left" corresponded to the lower bound and sampling performed before it was clarified that the opposite was the case. To correct this, we refit Level 1 with z fixed at 0.55, and the complement of the sampled z value (i.e., $1-z$) is reported.[3]

A.1.2 Results. Two and three cycles of `run.unstuck.dmc` were required for Levels 1 and 2, respectively, and for both `run.converge.dmc` completed immediately without the need for any additions. Median posterior estimates and 95% credible intervals for all levels are shown in Table 4. For Level 1, posterior parameter estimates were only weakly correlated (at

[3]This error was actually detected before the clarification was issued as fixing $z = .45$ produced poor fits, and, assuming left corresponded to the lower bound, fitting with $z = .55$ produced good fits and freely estimating all parameters for the Level 1 data produced a median estimate of $z = 0.554$. True values (where known) were within 95% credible intervals for the latter fit ($a = 0.95 - 1.04$, $v_{\text{Easy}} = 2.79 - 3.81$, $v_{\text{Medium}} = 2.1 - 2.91$, $v_{\text{Hard}} = 1.31 - 1.96$, $T^{\star}_{er} = .29 - .30$), with little effect on $s_{T_{er}}$ (0.097-0.117) but much more variability for s_v (1.63-2.68) and s_z (0.03-0.51), consistent with Table 4. As re-doing Level 3 was time consuming and the fix straightforward, refitting was only done for Level 1.

most $r = 0.23$ between s_v and $s_{T_{er}}$). Estimates for Level 2 show much greater uncertainty for s_v and s_z but not for $s_{T_{er}}$. Much stronger correlations were evident between s_v and $T_{er}^{\star}$ (.69), between s_v and a (.59), and between s_v and the drift rates (.76 - .82), as well as among the three drift rates (.64 - .72). For both levels observed and predicted cumulative distribution functions were a close match, indicating a very good fit.

For Level 3 `h.run.unstuck.dmc` and `h.run.converge.dmc` completed immediately (due to the good start points provided by individual fits). There was, however, some visual evidence of a small degree of initial non-stationarity. This was addressed by taking a fresh 100 iterations with results reported in Table 4. There were only weak correlations among group-level parameters. Uncertainty about the group-level mean s_v and $s_{T_{er}}$, was substantially decreased relative to the Level 2 estimates, but this was less so for s_z. The group-level estimate μ_a was surprisingly much wider than the Level 2 estimate for a. Uncertainty in the group-level standard deviation estimates was quite large, reflecting the small sample of 20 participants.

A.1.3 Advice. Although the automatic convergence procedures (i.e., the `run.unstuck.dmc` followed by the `run.converge.dmc` functions described in detail in the methods) worked well in this case they can sometimes fail in a number of ways. Migration can cause false convergence at a local minimum followed by a sometimes long period of apparent stationarity before posterior likelihoods suddenly start increasing, although this is rare if migration probability is low, such as used here. When migration is left on overly long lower likelihood (but still valid) samples are under-represented, especially, causing the initial samples after migration to display a fairly subtle type of non-stationarity that automatic convergence can sometimes fail to detect. It can also fail to pick up other problems, such as slow drifts due to trade-offs between parameters in more complex models. Long time scale waves in chains for highly auto-correlated parameters can be hard for automatic procedures to differentiate from burn-in, although this can be ameliorated by appropriate thinning.[4] Hence, visual inspection

[4]Thinning is not strictly necessary and always throws away some information so is sometimes not recommended. However, as long as it is not excessive it makes handling samples (which can otherwise get very large) more computationally convenient, and it can also make visual inspection easier.

Table 4

Parameter estimates and uncertainty intervals reported by Heathcote.

	a	v_{Easy}	v_{Medium}	v_{Hard}	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.22	0.10	0.48
LB							2.02	0.10	0.41
UB							2.43	0.11	0.54
					Level 2				
Estimate	0.81	3.75	3.27	2.15	0.38	0.55	1.92	0.10	0.16
LB	0.78	3.31	2.88	1.83	0.37	0.54	1.42	0.09	0.01
UB	0.84	4.32	3.78	2.56	0.38	0.57	2.50	0.11	0.40
					Level 3 - μ_k				
Estimate	0.79	4.38	3.13	1.94	0.39	0.54	1.50	0.16	0.19
LB	0.47	3.93	2.65	1.56	0.34	0.53	1.37	0.15	0.02
UB	1.00	4.85	3.63	2.30	0.44	0.55	1.65	0.16	0.30
					Level 3 - σ_k				
Estimate	0.41	1.01	1.10	0.81	0.10	0.02	0.24	0.01	0.18
LB	0.29	0.75	0.84	0.60	0.07	0.02	0.16	0.00	0.10
UB	0.79	1.42	1.50	1.12	0.14	0.03	0.37	0.01	0.32

Note. Estimate: posterior median, LB: lower bound of 95% credible interval, UB: upper bound of 95% credible interval.

of parameters chains, as well as their posterior log-likelihoods, is desirable as a final check[5]. The `plot.dmc` function makes it easy to perform these checks, as illustrated in supplementary materials.

Because across-trial variability parameters have a weak effect it is important to check that priors are not overly influential. This can also be done with `plot.dmc`, which allows priors to be imposed on posterior density estimates. For example, these plots clearly show s_z has the weakest updating among DDM parameters at Level 2, followed by s_v, consistent with the credible intervals in Table 4, with the graphs making this more immediately obvious. Similarly, for Level 3 the weaker updating for a and s_z group-level means is clear, as well as the generally weaker updating of group-level standard deviation parameters. Overall, priors do not seem to have been overly influential for across-trial variability parameters, with group-level standard deviation parameters being the most suspect. In such cases it is advisable to check the sensitivity of estimates to reasonable changes in the prior. However, fits with different priors for these parameters (exponential with scale parameter one) did not affect estimates much.

In real data the minimal individual differences in $s_{T_{er}}$ (median $\sigma = 0.005$) evident in the Level 3 fits would be suspicious, and might indicate hierarchical sampling had fallen into a "zero variance trap" (Lee & Wagenmakers, 2014). The DE-MCMC sampling of DDM between trial variability parameters are prone to this problem if participant and group-level parameter chains are kept in a fixed relationship, but randomly associating chains at the two levels, as was done here, is usually a remedy. Also chain plots did not look characteristic of the zero variance trap (where smallest estimates usually have little variation, whereas here although they were small they were variable), so these results may not be suspicious in the present case.

[5]More robust automatic convergence procedures are under development in DMC and preliminary tests have shown them to perform well in more difficult cases. Such approaches are particularly important in parameter recovery studies when the required large numbers of fits make thorough visual inspection difficult, although even in this context inspection of at least a subset of fits is highly recommended.

Table 5

Specification of prior distributions for DDM parameters in Hawkins' contribution.

	a	v	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
Mean (μ)	2	2	0.5	0.5	0	0	0
Standard Deviation (σ)	2	3	0.5	0.2	1	0.5	0.5
Lower Bound (L)	0	0	0	0	0	0	0
Upper Bound (U)	∞	∞	∞	1	∞	∞	1
Shape (m)	1	1	1	1	1	1	1
Scale (θ)	1	2	1/3	1/3	1	1/3	1/3

A.2 Hawkins

A.2.1 Methods. No pre-processing was performed on any of the data sets. In contrast to the main text, we defined non-decision time ($T_{er}^{\star}$) as the lower bound of the uniform non-decision time distribution. We used the DDM likelihood function as provided in the `rtdists` package for the `R` statistical environment (Singmann et al., 2016).

In the Level 3 analysis we used the hierarchical Bayesian framework described in Heathcote's contribution to simultaneously estimate parameters at the participant and group-levels. The parameterization of the truncated normal prior distributions for the group-level means, $\mathcal{N}(\mu,\sigma)[L,U]$, and the parameterization of the Gamma prior distributions for the group-level standard deviations, $\Gamma(m,\theta)$, are shown in Table 5. The half-normal prior distribution on the three across-trial variability parameters places most density at low values, meaning that non-zero estimates of the across-trial variability parameters were driven by data. The mildly informative prior distributions placed on the group-level parameters were loosely drawn from Matzke and Wagenmakers (cf. Table 3, 2009).

The Level 1 and 2 analyses were not hierarchical (single participant estimation). Therefore, those analyses used the group-level mean ($\mathcal{N}(\mu,\sigma)[L,U]$) prior distributions specified above as participant-level prior distributions.

Parameters were estimated using differential evolution Markov chain Monte Carlo (DE-MCMC; Turner et al., 2013), using the default settings (see Turner et al., 2013). We set the number of MCMC chains to 3 times the number of participant-level parameters (i.e., 9 chains in the Level 1 analysis, 27 chains in the Level 2 and 3 analyses), which is the upper limit recommended by Turner et al. We took 4,000 posterior samples from each chain with a burn-in period of 2,000 samples. Convergence was monitored through visual inspection and the multivariate potential scale reduction factor $\hat{R}$ (Brooks & Gelman, 1998).

To provide point estimates and measures of uncertainty, we summarize the parameter estimates using the posterior median and the 95% highest density interval (HDI; Kruschke, 2011), the smallest interval to contain 95% of the marginal posterior density of a parameter. We summarize individual participant parameter estimates in the Level 1 and 2 analyses, and group-level estimates of the mean and standard deviation parameters in the Level 3 analysis.

A.2.2 Results.

Level 1. Visual inspection and $\hat{R}$ indicated chain convergence ($\hat{R} = 1.01$). The parameter estimates are shown in Table 6. The three across-trial variability parameters appeared to estimate well, with relatively narrow uncertainty intervals.

Level 2. Visual inspection and $\hat{R}$ indicated chain convergence ($\hat{R} = 1.03$). The parameter estimates are shown in Table 6. $s_{T_{er}}$ appeared to estimate well. s_v was strongly correlated with the three drift rate parameters ($r's \geq .78$), which increased the size of its uncertainty interval. The posterior distribution of s_z pushed against the lower boundary (0) so the posterior median may be misleading. s_z was relatively strongly correlated with $T^{\star}_{er}$ ($r = .72$).

Level 3. Visual inspection indicated that the group-level chains of the main model parameters had converged (i.e., v_{Easy}, v_{Medium}, v_{Hard}, a, z, $T^{\star}_{er}$), but the three across-trial variability parameters had not converged. This was at least partially due to the participant-level chains: a few participants had a single chain that had not converged, which predominantly affected one or more of their across-trial variability parameters. Removing 3 (of 27) chains mostly eliminated the problem and led to relatively good convergence for all 20 participants (mean $\hat{R}$ across participants

1.13, range 1.09 -– 1.17). Such post-hoc removal of chains can be justified on the basis that chains are independent, and that removing those chains did not substantially influence the effective sample size. With those chains removed, the group-level $\hat{R}$ was 1.27. This reduced to 1.04 when only considering the main model parameters. The $\hat{R}$s for the group-level across-trial variability parameters were: $\mu_{s_v} = 1.21$, $\sigma_{s_v} = 1.25$, $\mu_{s_z} = 1.19$, $\sigma_{s_z} = 1.12$, $\mu_{s_{T_{er}}} = 1.03$, $\sigma_{s_{T_{er}}} = 1.11$. Some chains for the group-level standard deviation parameters (i.e., σ_{s_v}, σ_{s_z}, $\sigma_{s_{T_{er}}}$) became stuck at low values, which strongly influenced the effective sample size for those parameters (see online appendix). Therefore, estimates of the across-participant variance in the across-trial variability model parameters should be interpreted with caution. The parameter estimates are shown in Table 6.

A.2.3 Advice. The Level 1 and 2 analyses suggest the Bayesian parameter estimation approach outlined here has few difficulties when participants are treated as fixed effects (i.e., each participant's model parameters are estimated independently of all other participants).

In contrast, the Level 3 analysis suggests that hierarchical Bayesian parameter estimation of the DDM can be challenging, at least when attempting to obtain participant-level estimates of the three across-trial variability parameters (i.e., s_v, s_z, $s_{T_{er}}$) when participants are treated as random effects. There was evidence of poor sampling behavior: some chains failed to converge and some group-level chains became stuck at low values. Post-hoc removal of chains that failed to converge at the participant level partially alleviated the problem. Although not principled, this post-processing method is one way to rapidly improve convergence, provided sufficiently many chains were sampled. One alternative would be to run more sampling iterations with the methods outlined above, though we note that we already sampled 4000 iterations so this approach is likely to be very slow. Another alternative is to adopt different sampling rules; for example, incorporating the migration step in the DE-MCMC sampler (see Turner et al., 2013), which can, at times, rapidly improve convergence particularly for participant-level parameter estimates.

However, even implementing these changes might not alleviate the problem where some chains for the group-level scale parameters became stuck at very low values. When a

Table 6

Parameter estimates and uncertainty intervals reported by Hawkins.

	a	v_{Easy}	v_{Medium}	v_{Hard}	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.22	0.10	0.47
LB							2.01	0.10	0.40
UB							2.45	0.11	0.53
					Level 2				
Estimate	0.81	3.80	3.32	2.17	0.38	0.55	1.96	0.10	0.16
LB	0.78	3.28	2.86	1.82	0.37	0.54	1.40	0.09	0.00
UB	0.85	4.37	3.82	2.56	0.38	0.57	2.57	0.11	0.38
					Level 3 - μ_k				
Estimate	0.79	4.44	3.13	1.89	0.39	0.54	1.56	0.15	0.29
LB	0.51	3.91	2.44	1.27	0.35	0.53	1.41	0.15	0.11
UB	1.02	4.97	3.78	2.40	0.44	0.55	1.71	0.16	0.39
					Level 3 - σ_k				
Estimate	0.41	1.10	1.25	0.90	0.10	0.02	0.09	0.00	0.08
LB	0.26	0.76	0.84	0.58	0.07	0.02	0.01	0.00	0.01
UB	0.66	1.55	1.91	1.47	0.13	0.03	0.29	0.01	0.21

Note. Estimate: posterior median, LB: lower bound of 95% highest density interval, UB: upper bound of 95% highest density interval.

parameter exerts only a small influence on the likelihood function of the model - the parameter is not well constrained by data, which is the case for the across-trial variability parameters of the DDM - there is large uncertainty in its corresponding posterior distribution. For example, in the Level 2 analysis the width of the 95% HDI - a measure of uncertainty - for the s_z parameter was over 13 times wider than the 95% HDI for the z parameter. This means there was a much larger range of plausible values for s_z than z; estimates of s_z were less constrained by data. This level of uncertainty can cause problems when hierarchically estimating the across-trial variability parameters of the DDM. This is because participant-level estimates of the across-trial variability parameters are only weakly informed by data, so the hierarchical model shrinks those estimates to very similar values across participants, which produces close-to-zero estimates of the group-level scale parameters. Consequently, caution is warranted when interpreting the group-level estimates of these parameters. It is possible that interpretation of the group-level estimates of the main DDM parameters is largely unaffected.

A.3 Van Ravenzwaaij

A.3.1 Methods.

Inspection of the behavioral data showed that no pre-processing was necessary. All experiments were analyzed using a (hierarchical) Bayesian implementation of the DDM (Ratcliff, 1978, 2002). I used the `rtdists` package in `R` (available from https://cran.r-project.org/web/packages/rtdists/rtdists.pdf) to get densities for the DDM parameters. For optimization, I modified the code for the differential evolution Markov chain Monte Carlo (DE-MCMC) hierarchical Bayesian implementation that was originally developed for the Linear Ballistic Accumulator model (see Brown & Heathcote, 2008 for the model, and Turner et al., 2013 for the DE-MCMC hierarchical Bayesian implementation). Note that the code can be adapted for individual model fits, which is what I did for the first two data sets.

In my original fit for Level 1, I found that fixing the parameters to their known values led to unsatisfactory parameter estimates. As such, I resorted to the procedure I would follow if I had encountered this data set “in the wild”: I left all parameters free to vary (including the known

ones).

Similar to Heathcote's contribution, DDM parameters for Level 1 and Level 2 were sampled from independent truncated normal prior distributions, $\mathcal{N}(\mu,\sigma)[L,U]$, with the parameterization given in Table 7. These mildly informative prior distributions placed on the group-level parameters were loosely drawn from Matzke and Wagenmakers (cf. Table 3, 2009). The first fits led to satisfactory posterior predictives (i.e., the models fit the data well, see section A.3.2 below), but used an unrealistic parameterization for the prior distribution for s_v.[6] The correct prior distribution is $s_v \sim \mathcal{N}(1,1)[0,\infty]$. After consulting with the first and senior author, the decision was made to report here the results of the original fits and of a corrected set of fits that use the correct prior distribution for s_v and fix the known parameters for Level 1 to their true values.

Starting points for the Markov chains were drawn from the following distributions: $v_{\text{Easy}} \sim \mathcal{N}(3.5,0.35)[0,\infty]$, $v_{\text{Medium}} \sim \mathcal{N}(2.5,0.25)[0,\infty]$, $v_{\text{Hard}} \sim \mathcal{N}(1.5,0.15)[0,\infty]$, $a \sim \mathcal{N}(1,0.1)[0,\infty]$, $z \sim \mathcal{N}(0.5,0.05)[0,\infty]$, $T_{er} \sim \mathcal{N}(0.3,0.03)[0,\infty]$, $s_z \sim \mathcal{N}(0.1,0.01)[0,\infty]$, $s_v \sim \mathcal{N}(0.1,0.01)[0,\infty]$, and $s_{T_{er}} \sim \mathcal{N}(0.1,0.01)[0,\infty]$. Note that for the corrected fits, the starting point for $s_v \sim \mathcal{N}(1,0.1)[0,\infty]$.

Similar to Heathcote's contribution, individual-level parameters for the Level 3 data set were sampled from a truncated Gaussian group-level distribution. Thus, for each parameter to be estimated, I estimated a group-level mean parameter and a group-level standard deviation parameter using the parameterization given in Table 7. Priors for all group-level standard deviation parameters were gamma distributions with a shape and a scale parameter of 1, except for parameters σ_z and σ_{s_z} which instead had a shape parameter of 0.1 in order to put more prior mass on low standard deviation values, because the starting point z is naturally bounded between 0 and 1.[7] Starting point distributions for the Markov chains for group-level mean μ

[6] This prior makes sense for a diffusion coefficient $s = 0.1$, but the diffusion coefficient for these data sets is $s = 1$. I detected this error after publication of the generating parameter values.

[7] These prior settings are fairly uninformative. As a result, the specific settings will not have a large influence on the shape of the posterior.

Table 7

Specification of prior distributions for DDM parameters in van Ravenzwaaij's contribution.

	a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
				Levels 1 and 2					
Mean (μ)	1	3.5	2.5	1.5	0.3	0.5	0.1	0.1	0.1
Standard Deviation (σ)	1	3.5	3.5	3.5	0.3	0.2	0.1	0.1	0.1
Lower Bound (L)	0	0	0	0	0	0	0	0	0
Upper Bound (U)	∞	∞	∞	∞	∞	∞	∞	∞	∞
				Level 3					
Mean (μ)	1	3.5	2.5	1.5	0.3	0.5	0.1	0.1	0.1
Standard Deviation (σ)	0.5	1.5	1	0.5	0.1	0.1	0.05	0.05	0.05
Lower Bound (L)	0	0	0	0	0	0	0	0	0
Upper Bound (U)	∞	∞	∞	∞	∞	∞	∞	∞	∞
Shape (m)	1	1	1	1	1	0.1	1	1	0.1
Scale (θ)	1	1	1	1	1	1	1	1	1

were all identical to starting point distributions for the individual parameters, and starting point distributions for group-level σ parameters were derived from starting point distributions for the individual parameters by dividing the mean by 10 and the standard deviation by 2.

For sampling, I used 32 interacting Markov chains for all runs, and ran each for 1,000 burn-in iterations followed by 1,000 iterations after convergence. The interacting chains are an integral component of the DE algorithm and speed up convergence when parameters to be estimated are highly correlated (for details, see ter Braak, 2006). The two tuning parameters of the differential evolution proposal algorithm were set to standard values used in previous work: random permutations drawn uniformly from the interval $[-.001, .001]$ were added to all proposals; and the scale of the difference added for proposal generation was set to $\gamma = 2.38 \times (2K)^{-0.5}$, where K is the number of parameters per participant. No migration step was included. Fitting the data sets for Levels 1 and 2 took about 3 hours each on an Intel Core i3-3220 CPU with 3.30GHz using a single core. Fitting the data set for Level 3 took about 14 hours using four cores.

A.3.2 Results.

Level 1. Convergence of the MCMC chains can be examined in Figure A1, found in the online appendix. Visual inspection of Figure A1 shows that parameter convergence is fine for all parameters except s_z and s_v. For these parameters, the histograms touch the truncation value of zero, and the mixing seems to be relatively poor. The posterior predictive data for the fitted model is compared with the data in Figure A2. The original data are shown by points joined by lines, and distributions of posterior predictive data are shown by box-and-whiskers. Every observation contained in each box-and-whiskers is based on data generated from a sample from the joint posterior. Boxes contain 50% of the observations, and tails extend to 100%. The top-left panel shows correct RTs, the bottom-left panel shows error RTs, and the top-right panel shows proportion correct. Deciles of .1, .5, and .9 are displayed. The figure shows that the model fit the data well, except for an underestimation of error RTs for the slowest quantile.

Convergence of the MCMC chains for the corrected fit can be examined in Figure A3,

found in the online appendix. Visual inspection of Figure A3 shows that parameter convergence is fine for all three parameters. The posterior predictive data for the fitted model is compared with the data in Figure A4. The figure shows that the model fit the data well, except for an overestimation of error RTs for the slowest quantile. This overestimation in the corrected fit, compared to the underestimation in the initial fit, is most likely due to the larger values of the across-trial variability parameters in the corrected fit, which result in higher variability of the predicted error RTs. Interestingly, the initial and corrected model fits both seem to provide a good account of the data and they seem to be qualitatively similar.

Estimated parameters for the two fits can be found in the sections labeled "Level 1" and "Level 1 - Corrected" in Table 8. Interestingly, despite both the initial and corrected fit providing a satisfactory model fit, the estimated parameters are very different. Aside from the expected difference in the s_v parameter, all drift rate v values obtained in the initial fit were lower than the known values, boundary separation a obtained in the initial fit was lower than the known value, and s_z obtained in the initial fit was lower than the value obtained in the corrected fit.

Level 2. Convergence of the MCMC chains for the first fit can be examined in Figure A5, found in the online appendix. Visual inspection of Figure A5 shows that parameter convergence is fine for all parameters except s_z and s_v. For these parameters, the histograms touch the truncation value zero, and the mixing seems to be relatively poor. The posterior predictive data for the fitted model is compared with the data in Figure A6. The figure shows that the model fit the data well, except for an underestimation of error RTs for the slowest quantile.

Convergence of the MCMC chains for the corrected fit can be examined in Figure A7, found in the online appendix. Visual inspection of Figure A7 shows that parameter convergence is fine for all parameters except s_z. For this parameter, the histogram touches the truncation value zero, and the mixing seems to be relatively poor. The posterior predictive data for the fitted model is compared with the data in Figure A8. The figure shows that the model fit the data well, except for an overestimation of error RTs for the slowest quantile. Similar to the results for Level 1, this overestimation in the corrected fit, compared to the underestimation in the initial fit, is

Table 8

Parameter estimates and uncertainty intervals reported by van Ravenzwaaij.

	a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate	0.90	2.22	1.65	1.05	0.43	0.45	0.29	0.09	0.07
LB	0.88	2.05	1.49	0.88	0.33	0.44	0.04	0.08	0.00
UB	0.92	2.40	1.82	0.92	0.34	0.46	0.53	0.10	0.18
					Level 1 - Corrected				
Estimate							2.30	0.11	0.36
LB							2.09	0.10	0.23
UB							2.52	0.12	0.44
					Level 2				
Estimate	0.77	2.93	2.51	1.59	0.42	0.55	0.18	0.09	0.06
LB	0.75	2.72	2.32	1.41	0.42	0.53	0.02	0.08	0.00
UB	0.79	3.15	2.73	1.76	0.43	0.56	0.42	0.10	0.19
					Level 2 - Corrected				
Estimate	0.80	3.69	3.22	2.10	0.43	0.55	1.84	0.10	0.12
LB	0.78	3.31	2.85	1.81	0.42	0.54	1.38	0.09	0.01
UB	0.83	4.20	3.68	2.46	0.43	0.57	2.37	0.11	0.27

Note. Estimate: posterior median, LB: lower bound of 95% credible interval, UB: upper bound of 95% credible interval.

most likely due to the larger values of the across-trial variability parameters in the corrected fit, which result in higher variability of the predicted error RTs. As for the Level 1 data set, the initial and corrected model fits both seem to provide a good account of the data and they seem to be qualitatively similar.

Estimated parameters for the two fits can be found in "Level 2" and "Level 2 - Corrected" in Table 8. Interestingly, despite both the initial and corrected model fit providing a satisfactory model fit, the estimated parameters are very different. Aside from the expected difference in the s_v parameter, all drift rate v values obtained in the initial fit were lower than values obtained in the corrected fit, and s_z obtained in the initial fit was somewhat lower than the value obtained in the corrected fit.

Level 3. Convergence of the MCMC chains for group-level parameters can be examined in Figure A9. The figure shows that not all chains converged. On top of that, based on a visual inspection of the posterior predictives I concluded that the model fit was unsatisfactory, I was unable to get a better fit within the allotted time. As I am not confident about the parameter estimates, I do not report the results of this model fit further here.

A.3.3 Advice.

It is not a secret that I am a proponent of fitting the "simple DDM" without across-trial variability parameters (see e.g., van Ravenzwaaij & Oberauer, 2009, van Ravenzwaaij et al., 2017). The gain of including across-trial variability parameters, being able to capture fast or slow errors in the data as well as the leading edge of RT distributions, is in my opinion outweighed by the cost of a poorer ability to capture individual differences and reduced statistical power to detect experimental effects. It is important to note here that my initial fit with incorrect specification of the prior distribution for s_v led to posterior predictives that were qualitatively similar to those presented for the corrected prior distribution. However, the estimated parameter values were very different, suggesting that the full model with variability parameters may be poorly identified.

Based on the results of fitting the three data sets, it seems that the problem is most

pertinent for parameter s_z. If a researcher does have strong theoretical reasons to fit across-trial variability parameters, they should be aware of the known issues with reliably estimating these parameters (and report those as such in their manuscript). When researchers do wish to fit the full DDM in a Bayesian framework, it is crucial to specify wide priors for the variability parameters. Failing to do so leads to substantial differences in results, as became clear from the initial model fits for Levels 1 and 2.

A.4 Frank, Krypotos, & Wiecki [8]

A.4.1 Methods. We estimated the model parameters for the Level 1 and 2 data sets using Bayesian estimation. Given that in the Level 3 data set responses from multiple participants were available, we used hierarchical Bayesian estimation (Wiecki et al., 2013). A key advantage of this parameter estimation approach is that parameters for each individual participant are estimated while being constrained by the group-level parameter distribution (Wiecki et al., 2013). As a result, the DDM parameters are estimated more accurately than if each participant's data are fit independently.

We quantified the parameter estimates by means of the posterior distributions, which were approximated via slice sampling (Neal, 2003). As we have done in previous studies (e.g., Frank et al., 2015), we estimated the model parameters using the `HDDM` package for `Python` (Wiecki et al., 2013).

In `HDDM`, the top and lower boundary could be defined based on accuracy (i.e., the upper boundary will be coded as correct response and the lower as error response; accuracy-coding) or based on the presented stimulus (i.e., the upper boundary coded as participants pressing the right button, and the lower boundary as participants pressing the left button; stimulus-coding). We initially decided to run an accuracy-coding model, which is most typical. However, in response to a query about this question from the study organizers (and after the results of the initial study had been communicated), we realized that the parameters of this model could not be compared with the parameters of the original study, which were generated using a stimulus-coding model. In

[8] Contributors are listed in alphabetical order.

particular, the z parameter indicates a bias toward left or right responding, which is not possible to capture with accuracy-coding, and it is also possible that this influences the estimates of other parameters. As such, we decided to fit a stimulus-coding model to correspond to the generative model, but using otherwise identical code and procedures. The results of both models were comparable in terms of the across-trial variability performance. Importantly, however, the stimulus-coding model results can be easily interpreted, with their interpretation being in line with that of the initial data set. Here, we present the results of the stimulus-coding models. The results for both, accuracy and stimulus-coding model are available at https://osf.io/fjy8z/.

We used informative (empirical) priors for our model parameters. Specifically, the prior distributions of the group means for each parameter roughly resemble the parameter values reported in the literature, as summarized in Matzke and Wagenmakers (2009). For a visualization of the priors, against the histograms of values summarized in Matzke and Wagenmakers, see Figure 1 of the Supplementary material of Wiecki et al. (2013). Further details on the sampling algorithms used in the model can be found in Wiecki, Sofer, and Frank (2016), and on the `HDDM` website (http://ski.clps.brown.edu/hddm_docs/).

A.4.2 Parameter estimation procedure.

Level 1. We initially did not analyze this data set as the default options of `HDDM` do not allow fixing the main DDM parameters to specific values, but rather require them to be estimated from the data. However, in response to queries from the organizers, we decided to fit the whole model to the data by following the same modeling approach as for the Level 2 data.

Level 2. We ran two Markov Monte Carlo chains, with each chain having 5,000 samples, with 3,000 samples as burn-in. The convergence of each chain was assessed via visual inspection and by computing the potential scale reduction factor $\hat{R}$ (Gelman & Rubin, 1992) for each parameter.

Level 3. We used the same approach as for the Level 2 data set but this time different parameter values were computed for each participant for the main DDM parameters. The across-trial variability parameters were computed only on the group-level. This was done as these

parameters are difficult to estimate for individual participants, and as such it is recommended to be estimated at the group-level (Wiecki et al., 2016). Also, given the long computation times for this model (see section A.4.4 below), we could run only a single chain within the available time. Since we had a single chain, we assessed convergence by the computation of the Geweke statistic (Geweke, 1992), computation of the Monte Carlo error statistic, by visual inspection of the posterior distributions, as well as by visually inspecting the mean and variance across the posterior distributions in windows of 200 samples to ensure they were stable.

A.4.3 Results. We present the results of a single chain for all data sets in Table 9. All parameters reached convergence, although not at the same speed. For example, for the Level 3 data, the s_v, s_z, and $s_{T_{er}}$ parameters could benefit for more samples, despite the visual inspection of the data suggesting convergence.

A.4.4 Advice. There are two issues that deserve attention when using the above approach. The first relates to `HDDM` not allowing setting fixed values for the main parameters. Although in principle the `HDDM` code could be modified to permit this, we did not think that in general one would want to assume fixed parameter values when fitting real data. However, it is remarkable that even without fixing the known parameters, we could recover almost all parameter values by just fitting the full model. The second issue relates to the computing time needed. The user should be aware that estimating the parameters of the full model for a data set with multiple data points (e.g., Level 3 data), including additional participants and/or more trials per participant, will require considerably more time than dealing with a data set with fewer data points or when the between trial parameters are not included. It should also be noted that meaningful estimates of across-trial variability parameters need a lot of MCMC samples to reach convergence (Wiecki et al., 2016). In large parameter recovery experiments we have found the across-trial variability parameters to not be identifiable on the participant-level to any meaningful degree. In addition, convergence of these parameters is very slow, compared to other parameters. Estimating these parameters on the group-level alone overcomes both of these problems. Researchers are advised to estimate these parameters only when they are relevant

Table 9

Parameter estimates and uncertainty intervals reported by Frank, Krypotos & Wiecki.

	a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate	0.99	3.24	2.47	1.61	0.35	0.55[1]	2.12	0.11	0.27
LB	0.95	2.76	2.11	1.30	0.34	0.54[1]	1.63	0.09	0.02
UB	1.04	3.83	2.97	1.98	0.36	0.57[1]	2.72	0.12	0.52
					Level 2				
Estimate	0.81	3.70	3.22	2.11	0.43	0.45[1]	1.84	0.10	0.14
LB	0.78	3.29	2.83	1.79	0.42	0.43[1]	1.37	0.09	0.01
UB	0.84	4.15	3.64	2.45	0.43	0.46[1]	2.33	0.11	0.34
					Level 3 - μ_k				
Estimate	0.86	4.49	3.14	1.92	0.47	0.46[1]	1.55	0.15	0.28
LB	0.70	3.93	2.67	1.48	0.43	0.45[1]	1.44	0.15	0.22
UB	1.04	4.86	3.62	2.39	0.52	0.47[1]	1.66	0.15	0.34

Note. Estimate: posterior mean, LB: lower bound of 95% credible interval, UB: upper bound of 95% credible interval. [1]These parameter estimates were misreported due to a bug; the values reported here are $1-z$ instead of z. Correcting for this misreporting gives estimates much closer to the generating parameter value $z = 0.55$.

to the research question. Alternatively, one could simply continue sampling until chains reach convergence.

A.5 Annis & Palmeri

A.5.1 Methods. Each of the three data sets were fitted within a Bayesian framework. We did not perform any preprocessing. In the results reported below $T_{er}^{\star}$ is the lower bound of the non-decision time distribution.

Level 1 Model. For the Level 1 data set, across-trial variability parameters s_v and $s_{T_{er}}$ were sampled from truncated normal prior distributions, $\mathcal{N}(\mu,\sigma)[L,U]$, and s_z was sampled

Table 10

Specification of prior distributions for DDM parameters in Annis & Palmeri's contribution.

	a	v_{Easy}	v_{Medium}	v_{Hard}	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
				Levels 1 and 2					
Mean (μ)	1	3	2	1	0.3	–	1	0.1	–
Standard Deviation (σ)	1	1	1	1	0.5	–	0.5	0.25	–
Lower Bound (L)	0	0	0	0	0	0	0	0	0
Upper Bound (U)	∞	∞	∞	∞	∞	1	∞	∞	1
				Level 3					
Mean (μ)	1	3	2	1	0.5	0.5	1	0.1	0.1
Standard Deviation (σ)	5	5	5	5	2	1	5	5	2

from a uniform prior $\mathscr{U}(L,U)$. The parameterization of the priors is given in Table 10. These priors were loosely based on those reported in Matzke and Wagenmakers (2009).

Level 2 Model. For the Level 2 data set, we used moderately informative priors based on Matzke and Wagenmakers (2009). All parameters were sampled from truncated normal prior distributions, $\mathscr{N}(\mu,\sigma)[L,U]$, except z and s_z, which were sampled from a uniform prior $\mathscr{U}(L,U)$. The parameterization of the priors is given in Table 10.

Level 3 Model. The Level 3 data set consisted of 20 simulated participants. The Bayesian model described above, applied to the Level 2 data, was also used to estimate each simulated participant's parameters in the Level 3 data; because we were trying out a new Bayesian inference engine (LaplacesDemon) and given the relatively constrained time window required by this collaborative project, we did not have time to develop and fit a hierarchical model and instead took a two-step multilevel approach (e.g., Achen, 2005, Gelman & Hill, 2007, p. 270). After estimating the participant-level posterior means, we treated the participant-level posterior means as observed data in another Bayesian model to estimate group-level means. Participant-level posterior means were assumed to be normally distributed, with the parameterization given in

Table 10. For simplicity and to make the priors less informative we chose not to include bounds on any of the prior distributions. Priors on the standard deviations were weakly informative (Gelman, 2006) half-Cauchy distributions with location parameter 0 and scale parameter 5.

Fitting Methods and Results. Likelihoods for the DDM were obtained from the `rtdists` package (Singmann et al., 2016). Each data set was fit using the `LaplacesDemon` package in `R` Statisticat LLC (2016), which contains a suite of Bayesian tools. We first used Laplace approximation to estimate the modes and covariance matrix for each simulated subject from each data set using initial starting values based on those found in Table 3 of Matzke and Wagenmakers (2009). Next, we obtained estimates of the marginal posterior means and 95% highest density intervals (HDI) via Componentwise Adaptive Gauss-Hermite Iterative Quadrature using the posterior modes and covariance matrix obtained from the Laplace approximation. For the first and second data sets, we report the posterior means and 95% HDI's obtained from the iterative quadrature. For the Level 3 data set, we applied the same model used for the Level 2 data set to each of the simulated participants. This resulted in 20 participant-level posterior means for each parameter. Using these means as data, we then obtained the group-level means and standard deviations. The model was fitted with Stan (Carpenter et al., 2017). We ran 3 chains for 2000 iterations and discarded the first 1000 samples. Chains were visually assessed for convergence and the potential scale reduction factor $\hat{R}$ (Gelman & Rubin, 1992) for all parameters was < 1.1.

A.5.2 Results. Tables 11 shows the posterior means, standard deviations, and 95% HDIs for the DDM parameters for Levels 1 2, and 3. The Level 3 section of the table shows the estimated group-level means and standard deviations of the participant-level parameters. Figures of the fits of the model for each data set can be found in the online appendix. For the Level 1 data, the model provided adequate fits, but less so for incorrect responses for left stimuli, especially in the easy condition. We suspect this is due to the low number of incorrect responses in the easy left stimulus condition. The model also had difficulties fitting the Level 2 data especially for incorrect responses. For Level 3, we obtained reasonable fits with the exception

of some overestimates of error responses times for certain subjects.

A.5.3 Advice. The `LaplacesDemon` package contains numerous methods for fitting Bayesian models. Our advice to users fitting the DDM using this package would be to start with a Laplace approximation to estimate the posterior modes and covariance matrix. We found this step to greatly improve the accuracy of the iterative quadrature and believe it would likely lead to faster convergence of MCMC chains. There are many different optimization algorithms used internally by the Laplace approximation routine. We found that the Nelder-Mead `Simplex` algorithm produced the best fits in the least amount of time. Once the posterior modes and covariance matrix are obtained these can then be input into one of many other algorithms in `LaplacesDemon` such as iterative quadrature, Particle Monte Carlo (PMC), or Markov Chain Monte Carlo (MCMC) to more efficiently obtain posterior estimates. We found that this two-step process led to high quality fits in most cases.

A drawback of the iterative quadrature method we used is that it is only useful for models with 10 or fewer parameters. Therefore, it cannot be used for hierarchical models. If the user is interested in fitting a hierarchical model we recommend first obtaining posterior modes for each subject using Laplace approximation and then using these as starting points in one of the various MCMC or PMC algorithms available in `LaplacesDemon`.

Appendix B Individual Contributions - Maximum-Likelihood Estimation

B.1 Singmann & Kellen

B.1.1 Methods. We estimated the DDM parameters using a trialwise maximum likelihood procedure (Myung, 2003), which was implemented with the statistical software `R` (R Core Team, 2015) and package `rtdists` (Singmann et al., 2016). Because we did not impose any hierarchical structure at the level of the parameters, we confined our analysis to the first two data sets (Level 1 and Level 2). We also did not exclude any trials because all RTs were within normal ranges (fastest RT = 0.307 s, slowest RT = 1.774 s). For each data set we used a wrapper function for the probability density function of the DDM. This wrapper function had separate drift rate value for each condition, with a positive sign for right stimuli and a negative sign for left

Table 11

Parameter estimates and uncertainty intervals reported by Annis & Palmeri.

	a	v_{Easy}	v_{Medium}	v_{Hard}	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.16	0.10	0.47
LB							1.98	0.10	0.42
UB							2.35	0.11	0.52
					Level 2				
Estimate	0.80	3.48	3.02	1.96	0.38	0.55	1.57	0.10	0.01
UB	0.78	3.24	2.79	1.75	0.37	0.54	1.37	0.09	0.00
LB	0.81	3.72	3.25	2.17	0.38	0.56	1.77	0.10	0.19
					Level 3 - μ_k				
Estimate	0.81	4.17	2.97	1.82	0.39	0.54	1.18	0.15	0.14
UB	0.64	3.71	2.45	1.45	0.35	0.53	1.00	0.15	0.09
LB	0.98	4.63	3.47	2.20	0.43	0.55	1.35	0.16	0.20
					Level 3 - σ_k				
Estimate	0.38	1.02	1.13	0.84	0.10	0.02	0.39	0.01	0.13
UB	0.28	0.73	0.81	0.61	0.07	0.02	0.28	0.01	0.10
LB	0.53	1.44	1.60	1.19	0.14	0.03	0.54	0.01	0.19

Note. Estimate: posterior mean, LB: lower bound of 95% highest density interval, UB: upper bound of 95% highest density interval.

stimuli. The data and wrapper function were passed to a non-linear minimization algorithm that searched for the parameters that minimize the negative sum of the log-likelihoods.[9]

Initially, we considered a variety of different non-linear optimization routines (for an overview, see Nash & Varadhan, 2011), but ultimately settled on the `nlminb` algorithm (Kaufman & Gay, 2003), which implements a variation of Newton's method that allows for the use of analytical and approximated (i.e., quasi-Newton) gradients or Hessians (in the present case, we had to rely on the latter). Our preference for this algorithm is in part due to its ability to quickly converge on global optima (i.e., it rarely gets stuck in local optima) but also to our long experience with it when fitting different types of models (see also Singmann & Kellen, 2013).

In order to estimate the uncertainty of our parameter estimates, we implemented a non-parametric bootstrap procedure (Efron & Tibshirani, 1994). For each data set we created 1000 bootstrapped data sets. The bootstrap was performed in a stratified manner: We randomly sampled with replacement from each drift rate *by* stimulus type condition (i.e., the ratio of the different item types remained the same, but the distribution of RTs and responses within each item type was bootstrapped). Note that individual trials (i.e., combination of RT and corresponding response) remained intact throughout this procedure. We avoided local minima by performing five fitting runs with independent initial start values for each (bootstrapped or original) data set, and only considering the results from the best run.

B.1.2 Results. We evaluated model fit by visually comparing the observed RT distributions with the predicted RT distributions. These comparisons suggested a good fit for both data sets (see supplementary materials). In the results reported below, $T_{er}^{\star}$ is the mean of the non-decision time distribution.

Level 1. The across-trial variabilities for the Level 1 data could be estimated with reasonable precision, and the bootstrap parameter-distributions appeared to take on a Gaussian shape (see supplementary materials). The parameter estimates are given in Table 12.

[9]The full `R` scripts for performing the analysis reported here are available in the supplemental materials. See also https://cran.rstudio.com/web/packages/rtdists/vignettes/reanalysis_rr98.html.

Table 12

Parameter estimates and uncertainty intervals reported by Singmann & Kellen.

	a	v_{Easy}	v_{Medium}	v_{Hard}	$T_{er}^{\star}$	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.22	0.10	0.47
LB							2.00	0.09	0.37
UB							2.43	0.11	0.53
					Level 2				
Estimate	0.80	3.63	3.17	2.07	0.38	0.44	1.79	0.10	0.00
UB	0.78	3.21	2.79	1.76	0.37	0.43	1.26	0.09	0.000
LB	0.83	4.28	3.69	2.50	0.38	0.46	2.42	0.11	0.27

Note. Estimate: ML estimate, LB: lower bound of 95% bootstrap confidence interval, UB: upper bound of 95% bootstrap confidence interval.

Level 2. The parameter estimates, the univariate distribution of the bootstrapped parameters, as well as the bivariate scatterplots for the Level 2 data are presented in Figure 10. The histograms clearly show a problem with the s_z parameter as its distribution exhibits a bimodal shape, with one large peak at 0 and a smaller peak around 0.2. None of the other univariate parameter distributions appeared to be pathological. In the case of the bivariate scatterplots, we found considerable correlations among several parameters. These correlations were especially large for s_v (when paired with the other drift rates and $s_{T_{er}}$), between a and z, among the three drift rates, and between s_z and $T_{er}^{\star}$. Moreover, we found the precision of drift rate and drift rate variability parameters to be rather low. The parameter estimates are given in Table 12.

B.1.3 Advice. Our results suggest a differential pattern regarding the utility of estimating the across-trial variability parameters in the DDM. If one has as much data as in the present case, $s_{T_{er}}$, and to a lesser extend s_z, can be estimated with reasonable precision. Regarding s_z, the analysis of the Level 2 data suggests that when the variability is in fact at the lower bound

of zero, any small inaccuracy will lead towards an inflated estimate. Given that it is doubtful whether s_z can be truly zero in any real data set, we do not find this to be a severe problem.

The presence of large variabilities for some parameters, together with very strong correlations among parameters, indicate that the full DDM fails to provide a characterization of the data that is as clear as one would hope. More precisely, even with a large data set, s_v is estimated with little precision (at least if the magnitude of s_v is as large as in the present data). This suggests that when one is interested in the parameter estimates, such as when applying the DDM in a cognitive-psychometric context (Batchelder, 1998), one should have little hopes of ever getting trustworthy estimates.

When considering whether or not to estimate the across-trial variabilities, it is important to consider the role of these parameters. The main motivation behind them is to capture the more fine-grained aspects of the RT distributions (Ratcliff & Tuerlinckx, 2002). Consequently, it should not be surprising that these parameters are difficult to estimate and are particularly vulnerable to the stochastic variability in the data. The presence of strong parameter correlations furthermore suggests that very similar (but not exactly equal) predictions can be obtained when jointly varying some of the parameters. This parameter fungibility suggests that a particularly large value of s_v is more likely to be due to a large value of v than to be a genuine independent effect. Models that contain highly correlated parameters are also known as *sloppy models* (Brown & Sethna, 2003) a class to which we believe the full DDM belongs to.

Our general advice is twofold. When the research goal is along the lines of cognitive psychometrics, simpler models, for example the four-parameter Wiener model (Vandekerckhove, 2014), should probably be preferred. The costs associated with a more complex model do not appear to pay off.

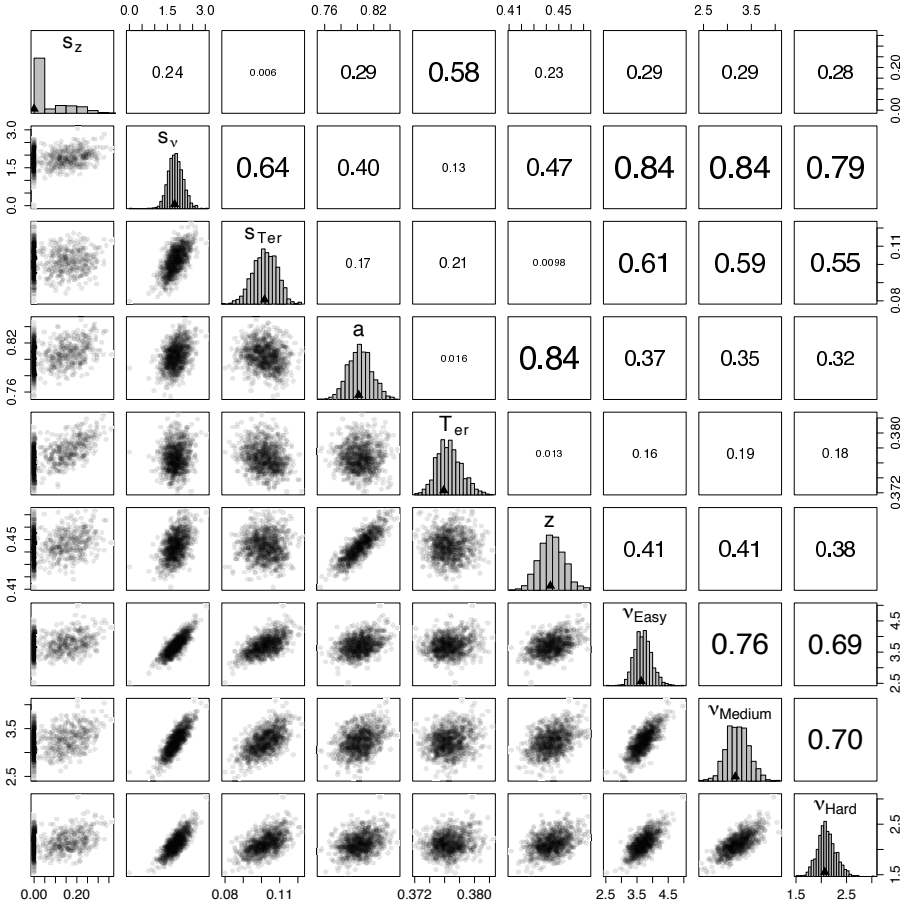

Figure 10: Pairs plot for results of the the Level 2 data reported by Singmann & Kellen. The main diagonal shows the (univariate) histograms of the non-parametric bootstrap based parameter distributions; the maximum likelihood estimate is displayed as a black triangle. The lower triangle shows the bivariate scatterplots of parameter distributions (where each point is plotted with 90% transparency so that larger numbers of overlapping points appear darker). The upper triangle shows the absolute values of the correlations between parameters with larger correlations printed in larger font.

If one nevertheless wants to estimate the across-trial variabilities (e.g., to account for differences in the RT distribution between error responses and correct responses) one should use bootstrap (or similar simulation-based) procedures to estimate the variability of the estimates

obtained. In the case of real data one should complement the present non-parametric bootstrap procedure with a parametric analog (i.e., generate synthetic data from the obtained parameter estimates and use those data to obtain parameter estimates). If the true data-generating process does not conform well to the postulates of the DDM, a comparison of the variability estimates obtained from parametric and non-parametric bootstrap would allow for a fairer assessment of the actual variability and heighten the probability of detecting problems such as the ones associated with s_z in the analysis of Level 2 data.

B.2 Voss & Lerche

B.2.1 Methods.

Overview. Data were analyzed with `fast-dm 30.2` (Voss, Voss, & Lerche, 2015, cf. also Voss & Voss, 2007, 2008). `Fast-dm` is an open source `C` program for parameter estimation in the DDM. Originally, `fast-dm` fitted predicted and observed cumulative RT distributions by minimizing the Kolmogorov-Smirnov statistic (Voss, Rothermund, & Voss, 2004). This method proved to be very robust in the case of contaminated RT distributions (Lerche & Voss, 2016). However, because in the present data there is no evidence for fast outliers or other forms of contamination and because across-trial variabilities are especially difficult to estimate (Lerche & Voss, 2017), a Maximum Likelihood (ML) method recently implemented in `fast-dm` was used for the present project.

Data preparation and model specification. For the analysis, responses "left" and "right" were recoded as 1 and 0, respectively, which are the codes for upper vs. lower thresholds in `fast-dm`, and drift rates were estimated separately for each type of stimuli (i.e., "left" and "right"). Individual data sets (Level 3) were saved into separate files. `Fast-dm` commands for all analyses are presented in Table 13 (see Voss et al., 2015 for further explanations on the handling of `fast-dm`). Note that `fast-dm` currently does not allow setting specific values for parameters that vary between conditions. The commands for the Level 1 analysis (Table 13, left column) result in an estimation of the six drift rates from data. We present results not only from this analysis, but from an additional calculation that fixes also the drift rates to the correct values.

Table 13

`Fast-dm` commands for Voss & Lerche's analysis.

Level 1	Level 2 & 3
`method ml`	`method ml`
`precision 4`	`precision 4`
`set d 0`	`set d 0`
`set p 0`	`set p 0`
`set a 1`	`depends v cond stim`
`set zr 0.45`	`format TIME RESPONSE`
	`cond stim`
`set t0 0.35`	`load *.dat`
`depends v cond stim`	`log level2_3.par`
`format TIME RESPONSE cond stim`	
`load L1.dat`	
`log level1.par`	

However, the latter analysis cannot be performed with the published version of `fast-dm`, but requires changes in the code.

The control commands first set the estimation method to maximum likelihood and set a rather high precision for the calculation of the predicted density functions (default is precision = 3). Then, settings for the standard DDM are given (d=0 indicates that the same non-decision time is used for both thresholds, see Voss, Voss, & Klauer, 2010, and p=0 indicates that the percentage of guessing is set to 0, see Ratcliff, 2002). For the Level 1 analysis the parameters a, z, and T_{er} are set to the true values. The `depends` command allows the drift to vary between conditions. Finally, the names of data columns and of input and output files are specified.

In three further sets of estimation procedures, all analyses were repeated setting one of the three across-trial variability parameters to zero. This allows for testing whether the model fit (i.e.,

the log-likelihood) decreases substantially when the parameter is removed from the model.

B.2.2 Results. `Fast-dm` was run on a PC with an Intel i7 Processor with 2.93 GHz. Mean calculation time was 3,909 seconds (about 1 hour) per data set. Estimated parameters are presented in Table 14.

Estimates of across-trial variability parameters. For the Level 1 data, across-trial variability parameters were estimated first restricting all parameters to the true values (which required an adaptation of the code), and - subsequently -– with the published version of `fast-dm 30.2` (which required the estimation of the six drift rates). Thus, the latter approach has more degrees of freedom, because the six drift rates are estimated. Here we only report the mean drift rate across stimulus types for each experimental condition. As can be seen from Table 14, estimates for the across-trial variabilities are nearly identical for both calculations. The confidence intervals shown in the table were estimated from 200 bootstrap-samples for Level 1 and Level 2 data.

Model fit. Likelihood of model estimation was compared for full models with restricted models, where one of the across-trial variability parameters was fixed to zero. The results, shown in Table 2 in the online appendix, indicated that model fit decreases dramatically for all models when the across-trial variability of non-decision times is set to zero. For the variabilities of starting point and drift rates results are not as clear cut: In many models, the fit is not affected strongly when removing these parameters from the model.

B.2.3 Advice. The present study confirms previous results (e.g., Lerche & Voss, 2016) showing that especially across-trial variabilities of drift and starting point are hard to estimate. Accuracy of estimates for across-trial variability of non-decision time is typically much larger. Whenever these variability parameters are in the focus of interest, researchers need to use all available tools to increase the precision of parameter estimation.

The easiest method to ensure high precision of parameter estimation in general is to use large data sets. Large real data, however, might come along with their own problems, since participants tend to lose motivation and attention while processing large numbers of trials, which

Table 14

Parameter estimates and uncertainty intervals reported by Voss & Lerche.

	a	v_{Easy}	v_{Medium}	v_{Hard}	z	T_{er}	s_v	$s_{T_{er}}$	s_z
Level 1 - Restricted Fit[a]									
Estimate							2.24	0.11	0.42
UB							2.00	0.10	0.30
LB							2.43	0.12	0.49
Level 1 - Full Fit									
Estimate		3.42	2.60	1.71			2.29	0.11	0.41
UB							2.04	0.09	0.30
LB							2.53	0.11	0.51
Level 2									
Estimate	0.80	3.64	3.18	2.07	0.55	0.43	1.80	0.10	0.00
UB							1.39	0.09	0.00
LB							2.42	0.11	0.29
Level 3 - μ_k									
Estimate	0.84	4.47	3.23	1.97	0.46	0.54	1.44	0.15	0.29
Level 3 - σ_k									
Estimate	0.34	1.22	1.24	0.84	0.09	0.03	0.60	0.02	0.18

Note. Estimate: ML estimate for Levels 1 and 2, mean ML estimate across participants for Level 3, LB: lower bound of 95% bootstrap confidence interval, UB: upper bound of 95% bootstrap confidence interval. Drift rate estimates are averaged across left and right stimuli. [a]The code of `fast-dm` was adapted to allow the fixation of drift rates to true values in separate conditions.

in turn could result in an increased number of contaminated trials (contaminated means that the internal response selection mechanisms changes from a diffusion-like mechanism to others, e.g. guessing). Such contamination need not even result in outlier RTs, which makes it hard to detect.

A second recommendation is to use efficient estimation procedures. Here, we decided to use a ML-estimation. However, ML results can be strongly biased when data is contaminated (Lerche & Voss, 2016). So, it might be a safer option to use a more stable procedure (e.g., the Kolmogorov-Smirnov distance) when real data are analyzed to avoid such biases.

Finally, one has to balance advantages and disadvantages of including across-trial variability parameters. On the one hand, only these parameters give the model the full flexibility to account for different patterns observed in real RT data from different tasks. Thus, these across-trial variability parameters seem to be theoretically necessary to make the DDM plausible. On the other hand, the across-trial variabilities are often not the focus of psychological theories, and seem to make the model unnecessarily complex: Recently, Lerche and Voss (2016) demonstrated that the precision of estimates for some model parameters (drift, threshold, and non-decision time) can be increased when across-trial variabilities for drift and starting point were not estimated, even if data were simulated with notable variability (see also van Ravenzwaaij et al., 2017).

Appendix C Individual Contributions - χ^2 Minimization

C.1 Servant & Logan

C.1.1 Methods. The model was simultaneously fit to correct and error RT distributions (.1, .3, .5, .7, .9 quantiles) and to accuracy data using a χ^2 method. Model fits were run in `FORTRAN`. The χ^2 method and the `FORTRAN` code have been fully described by Ratcliff (2002).

The χ^2 statistic has the following form:

$$\chi^2 = \sum_i \frac{N(p_i - \pi_i)^2}{\pi_i} \tag{2}$$

where N is the number of observations grouped into bins bounded by RT quantiles. p_i and π_i are, respectively, the observed and predicted proportions of trials in bin i, and sum to 1 across each

pair of correct and error distributions. The summation over i extends over the 12 bins in each experimental condition (6 bins for correct trials and 6 bins for error trials). Errors were excluded from the χ^2 computation when their number was < 10. The χ^2 statistic was minimized with a `Simplex` routine (Nelder & Mead, 1965). Details regarding the parameterization of `Simplex` are provided in the Advice section.

We added several constraints on the model during the `Simplex` minimization process. First, the (absolute) starting point z was constrained to not exceed 80% of boundary separation a. Secondly, half the width of across-trial variability in starting point $s_z/2$ was constrained to not exceed 90% of the minimal distance between starting point and decision bounds. Thirdly, across-trial variability parameters s_z, s_v, and $s_{T_{er}}$ were constrained to remain ≥ 0. Fourhtly, across-trial variability in drift rate was constrained to remain ≤ 3.29, the maximal value from Matzke and Wagenmakers' (2009) survey of parameter values estimated in empirical studies. Finally, half the width of across-trial variability in non-decision time was constrained to not exceed 90% of mean non-decision time T_{er}.

For Levels 1 and 2, measures of uncertainty for parameter estimates were obtained using a parametric resampling procedure (bootstrap). We generated 50 samples by running 50 simulations from the model using best-fitting parameters. Each sample contained the same number of trials per condition as the original data. The model was then fit to each of the 50 samples. We computed the 95% bootstrap confidence interval (2.5% and 97.5% quantiles) over the 50 bootstrap parameter estimates. For Level 3, we fit the model to each individual data set, and report the mean of parameter estimates over the 20 subjects.

C.1.2 Results. Best-fitting parameters for each Level are presented in Table 15. Plots of observed versus predicted data are provided in the online Appendix (Figure 1). The models with the obtained parameter settings provide a good description of the data sets.

For Levels 1 and 2, the `Simplex` search converged quickly. The uncertainty associated with across-trial variability parameters is much larger for Level 2 than Level 1. In particular, the 95% bootstrap CI for s_z (Level 2) is very large (0-0.5152), which might indicate a sloppy

Table 15

Parameter estimates and uncertainty intervals reported by Servant & Logan.

	a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.25	0.10	0.42
LB							2.11	0.10	0.33
UB							2.49	0.11	0.49
					Level 2				
Estimate	0.81	3.72	3.30	2.17	0.43	0.56	2.06	0.10	0.02
LB	0.78	3.25	2.81	1.81	0.42	0.54	1.52	0.07	0.00
UB	0.93	4.54	3.91	2.75	0.44	0.57	2.90	0.11	0.52
					Level 3 - μ_k				
Estimate	0.84	4.78	3.46	2.07	0.47	0.54	1.66	0.15	0.37
					Level 3 - σ_k				
Estimate	0.36	1.36	1.61	0.90	0.10	0.02	0.90	0.01	0.31

Note. Estimate: best-fitting parameter value for Levels 1 and 2, mean of best-fitting parameter values for individual participants for Level 3. LB: lower bound of 95% bootstrap confidence interval, UB: upper bound of 95% bootstrap confidence interval.

spectrum of sensitivity (i.e., a flat likelihood surface). 95% bootstrap CIs associated with parameter v (Level 2) also appear relatively wide, which might suggest a trade-off between v and s_v.

For Level 3, the `Simplex` search converged generally quickly. Constraints on across-trial variability parameters were critical to keep these parameters in a reasonable range. Without these constraints, s_z and s_v often went negative. In addition, s_v sometimes reached very large values. The best-fitting s_v was equal to the upper bound (3.29) for subjects 3 and 12.

C.1.3 Advice. The `Simplex` search in the `Fortran` code is implemented as follows. One set of starting values is initially entered. We used mean values from Matzke and

Wagenmakers' (2009) survey of parameter values estimated in empirical studies (with $s = 1$, $a = 1.25$, absolute $z = 0.63$, $T_{er} = .435$, $v_{\text{Easy}} = 2.23$,$v_{\text{Medium}} = 2.23$, $v_{\text{Hard}} = 2.23$, $s_{T_{er}} = .183$, $s_v = 1.33$, and absolute $s_z = .37$). `Simplex` is then run several times, using the best-fitting parameters from fit $N-1$ as the starting values for fit N. The process is repeated until the parameters do not change from one iteration to the next by a small amount. We observed, however, that different starting values yielded slightly different best-fitting parameters. For example, we ran additional fits for Level 3 using $v_{\text{Easy}} = 3.5$, $v_{\text{Medium}} = 2.5$, $v_{\text{Hard}} = 1.5$. The best-fitting parameters were $\mu_a = 0.85$, $\mu_{v_{\text{Easy}}} = 5.16$, $\mu_{v_{\text{Medium}}} = 3.71$, $\mu_{v_{\text{Hard}}} = 2.21$, $\mu_{T_{er}} = 0.46$, $\mu_{z\text{ abs}} = 0.47$, $\mu_{s_v} = 0.37$, $\mu_{s_{T_{er}}} = 2.06$, $\mu_{s_{z\text{ abs}}} = 0.15$. Here z_{abs} is the absolute starting point and $s_{z_{\text{abs}}}$ is the corresponding across-trial variability.

Main variations between these additional fits and those reported in Table 15 concern v, s_v and s_z. These variations might be explained by (i) a trade-off between v and s_v and (ii) a relatively flat likelihood surface associated with parameter s_z.[10] To further investigate (i), we computed the correlation between v and s_v across our 50 bootstrap parameter estimates from Level 2 for each difficulty condition. These correlations were very high (easy: $r = .85$; medium: $r = .85$; difficult: $r = .86$), demonstrating a trade-off between v and s_v (larger v is associated with larger s_v; see online Appendix, Figure 2). Parameters that are not well recovered should be fixed or combined (e.g., see our recent parameter recovery work on time-varying DDMs; White et al., 2017).

Ratcliff and Childers (2015) recently introduced some refinements of the χ^2 method. In particular, the median RT of errors is used if the number of errors is lower than the number of quantiles in a given condition. We instead excluded errors from the χ^2 computation when their number was < 10. Using the Ratcliff and Childers refinement in Level 3 (where a few data sets are associated with a small number of errors) might improve parameter recovery.

[10] Alternatively, variations between additional fits and Table 15 may suggest a local minimum problem. However, most of the parameter values are very close, and it seems that `Simplex` ended up in the same region. In addition, χ^2 values associated with parameters in Tables 15 and the additional fits were close (Table 15: mean χ^2 = 59.6; additional fits: mean χ^2 = 61.9).

C.2 Starns

C.2.1 Methods. We performed fits using the χ^2 method described in Ratcliff and Tuerlinckx (2002) using `FORTRAN` programs written by Roger Ratcliff. These programs find the parameter values that minimize χ^2 using the `Simplex` algorithm. We did not remove any trials before fitting. We did not have a method for providing interval measurements on parameters from a single participant's data set, because we have never done this in a paper. If we ever actually have to do single-participant inference for one of our projects, we will develop something more sophisticated like putting together an MCMC chain to estimate posterior distributions for parameter values. However, for the intervals reported for Levels 1 and 2, we used a quick-and-dirty method to get a feel for how tightly the parameters were constrained by data. We fixed the parameter at a value above or below its best-fitting value, reran the fit allowing the other parameters to be optimized, and found the point at which the likelihood with the fixed value was 10 times lower than the likelihood with the optimal value. If the data do not place much constraint on a parameter value, then we should be able to move it over a wide range without substantially affecting the fit (producing a wide interval).

For the Level 3 data, we found the best-fitting parameters for each participant and simply calculated standard 95% confidence intervals using these estimates. This is not an ideal procedure, as it does not acknowledge uncertainty in parameter estimates for individual participants. Nevertheless, this simple technique does a good job in parameter recovery simulations (at least for the main model parameters; Ratcliff & Tuerlinckx, 2002).

C.2.2 Results.

Level 1. The best-fitting parameter estimates are reported in Table 16. With the main model parameters fixed, there was tight constraint on these estimates; that is, the fit deteriorated quickly as we moved the parameters away from their optimal values.

Level 2. For the Level 2 data we estimated separate drift rates for left and right stimuli, in case the drift rates for left and right stimuli were not mirrored (same absolute value with different signs), but it appears that they were based on the fits. The values reported here are the mean

Table 16

Parameter estimates and uncertainty intervals reported by Starns.

	a	v_{Easy}	v_{Medium}	v_{Hard}	T_{er}	z	s_v	$s_{T_{er}}$	s_z
					Level 1				
Estimate							2.25	0.10	0.42
LB							2.02	0.09	0.30
UB							2.51	0.12	0.51
					Level 2				
Estimate	0.80	3.75	3.35	2.15	0.43	0.56	2.05	0.10	0.01
LB							1.54	0.08	0.00
UB							3.10	0.11	0.45
					Level 3 - μ_k				
Estimate	0.83	4.84	3.47	2.11	0.45	0.54	1.89	0.15	0.32
LB	0.67	4.27	2.90	1.69	0.43	0.53	1.59	0.15	0.27
UB	1.00	5.37	4.00	2.59	0.52	0.55	2.18	0.16	0.37

Note. Estimate: best-fitting parameter value for Levels 1 and 2, mean of best-fitting parameter values for individual participants for Level 3. LB: lower parameter value at which the likelihood was 10 times lower than the likelihood under the best fitting value for Levels 1 and 2, lower bound of 95% confidence interval for Level 3, UB: upper parameter value at which the likelihood was 10 times lower than the likelihood under the best fitting value for Levels 1 and 2, upper bound of 95% confidence interval for Level 3. Drift rate estimates are averaged across left and right stimuli.

absolute drift rates across stimuli. In our best-fitting model, χ^2 was 55.6. The corresponding parameter estimates are reported in Table 16. Estimating all of the parameters dramatically reduced the constraint on s_z and s_v. In other words, these parameters had to be moved far from their optimal values to get a substantial change in fit (i.e., a likelihood 10 times lower than the optimal likelihood). The constraint on $s_{T_{er}}$ remained very similar to the Level 1 fit.

Level 3. The same model described for Level 2 was fit to each participant. Across data sets, χ^2 values ranged from 35 to 79.7 with a mean of 56.9. The corresponding parameter estimates are reported in Table 16.

C.2.3 Advice. We would not advise using standard programs if a researcher wants to make conclusions about variability in starting point or non-decision time. No one really seriously endorses the assumed uniform distributions for these parameters; researchers have largely ignored this simplification because it does not seem to affect the types of conclusions they want to make (e.g., detecting effects on speed/accuracy trade-off, information quality, or bias). We do not take estimates of starting point variability from any fit that we run that seriously.

We have, however, been interested in detecting differences in drift rate variability in memory research. Although there is low constraint on this parameter, parameter recovery simulations suggest that the model can detect differences in this parameter between conditions with standard experiment designs (Starns & Ratcliff, 2014). Also, validation studies show that the model can detect manipulations of evidence variability in fits to empirical data (Starns, 2014). So we have more confidence in the s_v estimates, at least in terms of differences across conditions.

In fitting Level 2, we noticed that s_z and s_v strongly covary when the average drift rate is also freely estimated. That is, when we would increase s_v, say, then s_z would also increase and the average drift rate would get farther from zero. This makes sense in hindsight, given that s_z and s_v have opposite effects on the relation between correct and error RTs (increasing s_z promotes fast errors and s_v promotes slow errors; Ratcliff & McKoon, 2008). So a lot of the noise in estimating s_z and s_v comes because they trade off, and the estimates for both would probably get much better if there was a way to place additional constraint on one of them (we are not sure

how this could be achieved). This also makes us more confident in conclusions about changes in s_v across conditions that are constrained to have the same s_z than in conclusions about the absolute value of s_v within a single condition.

References

Achen, C. H. (2005). Two-step hierarchical estimation: Beyond regression analysis. *Political Analysis*, *13*(4), 447–456. doi:http://doi.org/10.1093/pan/mpi033

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. doi:10.1890/13-1452.1

Balota, D. a., Yap, M. J., Cortese, M. J., Hutchison, K. a., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*(4), 331–344. doi:10.1037/1040-3590.10.4.331

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brown, K. S. & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, *68*(2), 021904. doi:10.1103/PhysRevE.68.021904

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. doi:10.1016/j.cogpsych.2007.12.002

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A non-degenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*, *78*(4), 685–709. doi:10.1007/s11336-013-9328-2

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review*, *15*(4), 692–712. doi:10.3758/PBR.15.4.692

Efron, B. & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Efron, B. & Tibshirani, R. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666. doi:10.1146/annurev-psych-122414-033645

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, *35*(2), 485–494. doi:https://doi.org/10.1523/JNEUROSCI.2036-14.2015

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London: Chapman and Hall/ CRC.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511. doi:10.1214/ss/1177011136

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. doi:10.1214/06-BA117A

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Oxford, UK.: Clarendon Press.

Hawkins, G., Mittner, M., Forstmann, B., & Heathcote, A. (2017). On the efficiency of neurally-informed cognitive models to identify latent cognitive states. *Journal of Mathematical Psychology*, *76*, 142–155.

Heathcote, A., Lin, Y., & Gretton, M. B. (2016). DMC: Dynamic Models of Choice [Computer software]. Retrieved from osf.io/5yeh4

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (in press). Dynamic models of choice. *Behavior Research Methods*.

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

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Kaufman, L. & Gay, D. (2003). *The PORT library - optimization*. Murray Hill, NJ: AT&T Bell Laboratories.

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. Burlington, MA: Academic Press.

Kühn, S., Schmiedek, F., Schott, B., Ratcliff, R., Heinze, H.-J., Düzel, E., … Lövden, M. (2011). Brain areas consistently linked to individual differences in perceptual decision-making in younger as well as older adults before and after training. *Journal of Cognitive Neuroscience*, *23*(9), 2147–58. doi:10.1162/jocn.2010.21564

Laming, D. R. J. (1968). *Information theory of choice reaction time*. New York: Wiley.

Lee, M. D. & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lerche, V. & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.01324

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for robust parameter estimation in diffusion modeling? A comparison of different estimation algorithms. *Behavior Research Methods*, *49*(2), 513–537. doi:10.3758/s13428-016-0740-2

Lerche, V. & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*(3), 1–24. doi:10.1007/s00426-016-0770-5

Matzke, D., Love, J., & Heathcote, A. (2017). A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods*, *49*(1), 267–281. doi:10.3758/s13428-015-0695-8

Matzke, D. & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. doi:10.3758/PBR.16.5.798

McKoon, G. & Ratcliff, R. (1996). Separating implicit from explicit retrieval processes in perceptual identification. *Consciousness and Cognition*, *5*(4), 500–511. doi:10.1006/ccog.1996.0029

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and Time Series Model Selection.* Singapore: World Scientific.

McVay, J. C. & Kane, M. J. (2012). Drifting from slow to "d'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 525–549.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi:10.1016/S0022-2496(02)00028-7

Nash, J. C. & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, *43*(9).

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*, 705–767.

Nelder, B. J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. doi:https://doi.org/10.1093/comjnl/7.4.308

Philiastides, M. G. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, *26*(35), 8965–8975. doi:10.1523/JNEUROSCI.1655-06.2006

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., … Magnusson, A. (2016). coda: Output analysis and diagnostics for MCMC [Computer software]. Retrieved from https://cran.r-project.org/package=coda

Plummer, M. (2003). JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international*

workshop on distributed statistical computing (dsc 2003). Vienna, Austria. Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

R Core Team. (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

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

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. doi:10.3758/BF03196283

Ratcliff, R. (2008). The EZ-diffusion method: Too EZ? *Psychonomic Bulletin & Review*, *15*(6), 1218–1228. doi:10.3758/PBR.15.6.1218

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

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182. doi:10.1038/nature13314.A

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., McKoon, G., & van Zandt, T. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300. doi:10.1037/0033-295X.106.2.261

Ratcliff, R. & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. doi:10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. doi:10.1016/j.tics.2016.01.007

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. doi:10.3758/BF03196302

Singmann, H. & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. doi:10.3758/s13428-012-0259-0

Singmann, H., Scott, B., Gretton, M., Heathcote, A., Voss, A., Voss, J., & Terry, A. (2016). rtdists: Response time distributions (R package version 0.6-6) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/rtdists/index.html

Smith, P. L. & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*(2), 283–317. doi:10.1037/a0015156

Smith, P. L., Ratcliff, R., & Sewell, D. K. (2014). Modeling perceptual discrimination in dynamic noise: Time-changed diffusion and release from inhibition. *Journal of Mathematical Psychology*, *59*(1), 95–113. doi:10.1016/j.jmp.2013.05.007

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research*, *44*(12), 1297–1320. doi:10.1016/j.visres.2004.01.002

Smith, P. L. & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*(2), 135–168. doi:10.1016/0022-2496(88)90043-0

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory and Cognition*, *42*(8), 1357–1372. doi:10.3758/s13421-014-0432-z

Starns, J. J. & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language*, *70*(1), 36–52. doi:10.1016/j.jml.2013.09.005

Statisticat LLC. (2016). LaplacesDemon: Complete environment for Bayesian inference. Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing*, *16*, 239–249.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. doi:10.1177/0145721709355835.Continuous

Van Zandt, T. & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, *2*(1), 20–54. doi:10.3758/BF03214411

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*(2), 547–556. doi:10.3758/s13423-016-1081-y

van Ravenzwaaij, D. & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*, 463–473. doi:10.1016/j.jmp.2009.09.004

Vandekerckhove, J. & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–26. doi:10.3758/BF03193087

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. doi:10.1037/a0021765

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. doi:10.1016/j.jmp.2014.06.004

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory and Cognition*, *32*(7), 1206–1220.

Voss, A. & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775.

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

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff 's diffusion model. *British*

Journal of Mathematical and Statistical Psychology, *63*, 539–555. doi:10 . 1348 / 000711009X477581

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with Diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg. 2015.00336

Wabersich, D. & Vanderkerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, *46*, 15–28. doi:10.1016/j.cognition.2008.05.007

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. doi:10.3758/ BF03194023

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140–159. doi:10.1016/j.jml.2007.04.006

White, C. N., Kapucu, A., Bruno, D., Rotello, C. M., & Ratcliff, R. (2014). Memory bias for negative emotional words in recognition memory is driven by effects of category membership. *Cognition & Emotion*, *28*(5), 867–80. doi:10.1080/02699931.2013.858028

White, C. N., Servant, M., & Logan, G. D. (2017). Practical considerations for using conflict-based diffusion models to interpret choice RT data. *Psychonomic Bulletin & Review*.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2016). HDDM 0.6.0 documentation. Retrieved from http: //ski.clps.brown.edu/hddm%7B%5C_%7Ddocs/index.html

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. doi:10.3389/fninf.2013. 00014

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English lexicon project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. doi:10.1037/a0024177

Yap, M. J., Sibley, D. E., Balota, D. a., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597–613. doi:10 . 1037/xlm0000064