Inclusion Bayes Factors for Mixed Hierarchical Diffusion Decision Models

Udo Boehm[1,*], Nathan J. Evans[2,*], Quentin F. Gronau[1], Dora Matzke[1], Eric-Jan Wagenmakers[1], and Andrew J. Heathcote[1,3]

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

[2]University of Queensland, Australia

[3]University of Newcastle, Australia

[*]These authors contributed equally to the manuscript.

Author Note

The authors declare no competing financial interests. This research was supported by an European Research Commission Advanced Grant (743086 UNIFY) to EJW, an NWO Vidi grant to (VI.Vidi.191.091) DM, an NWO Veni grant to UB (VI.Veni.201G.045), an Australian Research Council grant (DP200100655) to AH, and an Australian Research Council Discovery Early Career Researcher Award (DE200101130) to NJE.

A preprint of the manuscript has been made available on PsyArXiv (https://psyarxiv.com/45t2w/, DOI: 10.31234/osf.io/45t2w). Part of the ideas presented here have been presented at the conference "Sequential Sampling Models of Decision Making", 2016, Emetten, Switzerland (Boehm et al., 2016). R code for our models and simulations is available on OSF: https://osf.io/v6u8e/

Correspondence concerning this article should be addressed to Udo Boehm, Department of Psychology, University of Amsterdam, Nieuwe Prinsengracht 129B, 1018 WS Amsterdam, The Netherlands, Email: u.bohm@uva.nl

Abstract

Cognitive models provide a substantively meaningful quantitative description of latent cognitive processes. The quantitative formulation of these models supports cumulative theory building and enables strong empirical tests. However, the non-linearity of these models and pervasive correlations among model parameters pose special challenges when applying cognitive models to data. Firstly, estimating cognitive models typically requires large hierarchical data sets that need to be accommodated by an appropriate statistical structure within the model. Secondly, statistical inference needs to appropriately account for model uncertainty to avoid overconfidence and biased parameter estimates. In the present work we show how these challenges can be addressed through a combination of Bayesian hierarchical modelling and Bayesian model averaging. To illustrate these techniques, we apply the popular diffusion decision model to data from a collaborative selective influence study.

keywords: Diffusion model, Bayes factors, response time data

Inclusion Bayes Factors for Mixed Hierarchical Diffusion Decision Models

Introduction

Cognitive models provide many advantages over a-theoretical statistical and psychometric measurement models of psychological data. Moving beyond the merely descriptive, their parameter estimates support a theoretically motivated account of latent psychological processes that leverages the cumulative results of previous research (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2014). The way in which these estimates change as a function of experimental manipulations, and of group and individual differences, is, therefore, better able to support psychologically meaningful explanations (Tuerlinckx & Boeck, 2005; van der Maas et al., 2011) as well as to link behavior to other types of data, such as from the neurosciences (Forstmann & Wagenmakers, 2015). They are intrinsically more parsimonious than purely descriptive statistical models because their parameters can be restricted to take on ranges, or to change, in ways that are consistent with their process interpretations (Heathcote, 2019). Cognitive models can also avoid pitfalls associated with the many untested assumptions that usually have to be made in purely statistical tests of psychological explanations. Statistical tests need to operationalize assumptions in terms of simple observed differences and/or relationships. Cognitive models, on the other hand, provide a comprehensive account of both manifest and latent processes, and ground the necessary assumptions either in cumulative research findings or make these assumptions more easily testable. Finally, because they provide principled explanations, the predictions of cognitive models are more easily generalized to new situations than those of their purely statistical counterparts.

One of the leading classes of cognitive models is built on the idea of evidence accumulation. The popularity of evidence accumulation models derives from their ability to account simultaneously for response time (RT) and choice data, which are frequently used in psychology and the cognitive sciences. Reflecting the advantages of the cognitive-modeling approach, the most widely adopted model of this class, the diffusion

decision model (Ratcliff, 1978; Ratcliff & McKoon, 2008, DDM:) has been applied to numerous areas of research, ranging from lexical (Ratcliff et al., 2004; Wagenmakers et al., 2008; Yap et al., 2015) and perceptual (Ratcliff, 2002; Smith, Ratcliff, & Sewell, 2014; Smith et al., 2004) decision making to recognition memory (McKoon & Ratcliff, 1996; White et al., 2014) and cognitive aging (Ratcliff et al., 2006, 2007). The aim of the present paper is to develop a coherent approach to estimation and inference for cognitive models such as the DDM that addresses the particular statistical challenges they present.

Statistical Challenges for Cognitive Modeling

Estimation and inference for cognitive models is difficult because they are usually highly non-linear, and their parameters are typically highly correlated. The latter "sloppiness" (a term introduced in mathematical biology, see Apgar et al., 2010; Gutenkunst et al., 2007) is shared with many models of biological systems that characterize observed input and outputs in terms of intervening processes that cannot be directly observed. When "sloppiness" is combined with non-linearity several challenges arise that can detract from the potential advantages offered by the cognitive modeling approach.

Challenge I: Hierarchical Data Structures. As a first challenge, cognitive models need to appropriately account for hierarchical data structures. "Sloppiness" and the pervasive non-linearity mean that successfully fitting cognitive models like the DDM to individual participants' data often requires each participant to perform a large number of trials (Lerche et al., 2016). Because large numbers of trials reduce measurement noise, they can help remedy problems of identifiability caused by parameter correlations (Kolossa & Kopp, 2018; Smith & Little, 2018). Unfortunately, simple techniques that average small amounts of data from each member of a large group of participants to compensate for the small number of trials per participant can produce misleading results due to non-linearity (Brown & Heathcote, 2003; Heathcote et al., 2000). This limits the effectiveness of cognitive modeling in settings such as clinical psychology and neuroscience where it is often

not practical to obtain many trials from each participant. A further issue is that many DDM applications rely on a two-step analysis procedure. In a first step, the model is fit to the data of individual participants and experimental conditions (e.g., Voss & Voss, 2007; Wagenmakers et al., 2007). In a second step, statistical hypothesis tests are applied to the individual participants' parameter estimates using techniques such as analysis of variance or regression (e.g., Karayanidis et al., 2009; Schmitz & Voss, 2012; Voss et al., 2004; Voss et al., 2013). This two-step procedure has several major problems.

First, in order to obtain stable parameter estimates, not only must each participant perform a considerable number of trials overall, but they must also do so in each condition (Wagenmakers, 2009). Fitting all within-subject conditions simultaneously for each participant can ameliorate this problem to some degree, at least in parsimonious models that share parameters across conditions (e.g., Ratcliff, 1978, 2002; Ratcliff et al., 2004; Ratcliff & McKoon, 2008; Ratcliff et al., 2006, 2007; Smith, Ratcliff, & Sewell, 2014; Smith et al., 2004; Wagenmakers et al., 2008; Yap et al., 2015). However, major problems remain with separate fits to each participant. As has been argued repeatedly in the mixed-modelling literature, failing to treat participants as a random effect can bias parameter estimates (e.g., Agrestia et al., 2004; Barr et al., 2013; Freeman et al., 2010; Grilli & Rampichini, 2015). Uncertainty in individual participants' point estimates is also ignored in second-step parameter tests (Boehm, Marsman, et al., 2018).

Mixed or hierarchical models, which provide simultaneous estimates for a group of participants, provide a potential solution to these challenges. They avoid the problems associated with simple averaging while improving estimation efficiency by shrinking individual participant estimates toward the central tendency of the group (Evans et al., 2018; Rouder & Lu, 2005; Rouder et al., 2003; Shiffrin et al., 2008). Bayesian methods make hierarchical approaches possible for non-linear cognitive models where the evaluation of high-dimensional integrals required at the group level is otherwise impractical. This has enabled applications of hierarchical Bayesian DDMs in both clinical psychology

(Huang-Pollock et al., 2017) and neuroscience (Forstmann et al., 2016; Kühn et al., 2011; Philiastides, 2006). More generally it has been argued that the hierarchical Bayesian approach provides the best statistical methodology for cognitive modeling (Lee, 2016). For the DDM in particular Vandekerckhove et al. (2011) argue that it provides advantages to both of Cronbach's (1957) two disciplines of psychology (see also Cronbach, 1975), underpinning a coherent account of both fixed and random effects for experimentalists while grounding psychometric measurement models in a substantive theory (see also Borsboom, 2006). Bayesian approaches are also advantageous for model selection, providing a comprehensive account of model flexibility that goes beyond simple parameter counts to address the way in which interactions among parameters are constrained by the form of the model's likelihood. Accounting for such "functional-form complexity" (Pitt & Myung, 2002) is particularly important for cognitive models because strong correlations between parameters can substantially impact on their flexibility.

Challenge II: Inference Under Model Uncertainty. As a second challenge, inference for cognitive models needs to appropriately account for model uncertainty. Many applications of cognitive models aim to identify relationships between cognitive processes that are represented by model parameters and a manifest variable, such as an experimental manipulation or individual differences in some observable property. To this end, researchers specify a set of candidate models, each of which allows a subset of the model parameters to covary with the manifest variable while constraining all other model parameters to be equal across levels of the manifest variable. Inference can then proceed by selecting the model that best accounts for the data. Bayes factors are a classical method for model selection that appropriately penalizes for model complexity (Heck et al., 2022; Kass & Raftery, 1995). However, it may be undesirable to base inference on a single model due to model uncertainty. Because only finite amounts of data are available, the correct model can never be known with certainty. For example, chance variation in sampling participants and stimulus materials together with the "sloppiness" of cognitive models can introduce

spurious covariation between model parameters and the manifest variable. The resulting uncertainty about the most appropriate model can lead to incorrect conclusions if inference is only based on a single model. Fortunately, inference can instead be based on a weighted average of the complete set of candidate models that takes both model complexity and model uncertainty into account (Hinne et al., 2020; Hoeting et al., 1999).

As well as aiding parameter estimation and interpretation, basing inference on model combinations has the potential to enhance other virtues of cognitive models. In simple statistical models, model averaging has been found to improve prediction (e.g., Quinn et al., 2017). Model averaging can also have important theoretical implications by establishing the selective influence of manipulations on cognitive processes (Ashby & Townsend, 1980; Sternberg, 1969). In many cases, identifying invariances—that is, determining which processes do not contribute to an observed phenomenon—provides important theoretical constraints on cognitive models (Rouder et al., 2009). Hence, methods such as complexity-penalized model averaging that appropriately accommodate model uncertainty when quantifying the evidence for the absence as well as the presence of an influence are highly desirable. Consequently, model-averaging techniques are key enablers for realizing the advantages of cognitive modeling.

These considerations have led to the increasing availability of general-purpose hierarchical Bayesian estimation and model-selection techniques for the DDM and other evidence accumulation models (Evans & Annis, 2019; Gronau, Heathcote, et al., 2019; Gunawan et al., 2020; Heathcote et al., 2019; Vandekerckhove et al., 2011; Wiecki et al., 2013). In the next section we discuss advantages and disadvantages of these implementations, with a focus on the DDM. We then propose an alternative set of methods based on Rouder et al.'s (2012) approach to Bayesian linear mixed models and Gronau, Heathcote, et al.'s (2019) approach to Bayesian inference. In the remainder of this paper we implement the framework for the DDM and test its application to Dutilh et al.'s (2019) data from a blinded collaborative study that challenged analysts to identify selective

influences of a range of experimental manipulations on evidence-accumulation processes. We assess the performance of our estimation and inference methods with these data and with synthetic data generated from parameters estimated from the empirical data. We conclude by discussing how this approach, and potential extensions, constitutes a promising framework for providing coherent estimation and inference for cognitive models.

Novelty

Our solution to the two statistical challenges in cognitive modeling combines different methods from Bayesian statistics. Although the use of some of these methods in psychology, and cognitive modeling in particular, has been advocated before, our work uniquely combines and extends these earlier approaches. Moreover, we provide a concrete example of the practical feasibility of a coherent, fully Bayesian approach to inference and estimation for a complex and widely used cognitive model.

Our solution to the first challenge embeds the DDM in a Bayesian hierarchical mixed modeling framework. Although hierarchical Bayesian methods have been advocated in cognitive modeling in general (e.g., Rouder & Lu, 2005; Rouder et al., 2005; Rouder et al., 2007), and for the DDM in particular (e.g., Vandekerckhove et al., 2011; Vandekerckhove, 2014; Wagenmakers, 2009; Wiecki et al., 2013), our approach differs in two important respects. First, most existing hierarchical implementations of the DDM ignore the fixed effects structure of experimental conditions (e.g., Hawkins & Heathcote, 2021; Turner et al., 2015; Wiecki et al., 2013), or implement a factor analysis model at the population level (e.g., Turner et al., 2017). In contrast, we adopt a hierarchical mixed modeling framework that affords a high level of flexibility in accommodating statistical structures at the population level, such as mixed effects ANOVAs, ANCOVAs, and regression models (see section "Fully Hierarchical Implementation of the Full DDM" for a detailed discussion). Second, we adopt an effect-size parameterization from Bayesian linear regression for condition and subject effects. This allows us to impose default priors on the

effect sizes and priors that are strongly informed by the literature on intercept terms. Thus, in contrast to earlier work (e.g., Vandekerckhove et al., 2011; Vandekerckhove, 2014), our approach eliminates the need for researchers to develop priors ad hoc for every new application of the DDM.

Our solution to the second challenge uses Bayesian model averaging (BMA) to account for model uncertainty. BMA has been advocated for inference and prediction in various contexts (e.g., Hoeting et al., 1999), including the assessment of replication studies (Iverson et al., 2010), weather forecasting (Raftery et al., 2005), and hydrology (Höge et al., 2019). Fragoso et al. (2018) in their review find that the number of publications on BMA has grown considerably over the past 20 years. However, application in the social sciences are dominated by the field of economics. The advocacy of BMA in psychology, in particular, has focused on (linear) statistical models such as structural equation models (Kaplan, 2021; Kaplan & Lee, 2016; Kaplan & Yavuz, 2020), network analysis, and ANOVA-type linear models (Hinne et al., 2020). On the other hand, applications to highly non-linear cognitive models are missing, which is presumably due to the absence of efficient sampling algorithms and easy-to-use software for hierarchical Bayesian cognitive models in the past.

In summary, despite the well-known theoretical advantages of hierarchical modeling and BMA, only the recent advent of efficient sampling algorithms has made it possible to realize these theoretical advantages for complex, highly non-linear cognitive models. In the present work we demonstrate how Bayesian hierarchical mixed modeling and BMA can be combined to obtain a coherent approach for estimation and inference for cognitive models. Our hope is that this Bayesian framework will in the future enjoy a similar level of applicability and popularity in cognitive modeling as it has achieved in statistical applications.

Models

Diffusion Decision Model (DDM)

The full 7-parameter DDM is illustrated in Figure 1. As currently used in most applications the DDM is based on three elaborations of the basic Wiener diffusion model for binary choice first proposed by Stone (1960).

Basic Diffusion Model

The basic diffusion model is the continuous time limit of a random walk process in which an evidence total starts at a point x_0 between two boundaries. Each boundary corresponds to one of the two response options and we assume without loss of generality that one boundary is located at 0 and the other boundary is located at $a > 0$. Evidence fluctuates from moment to moment as it accrues at a mean rate v due to the effect of zero-mean Gaussian noise with standard deviation s. Boundaries do not change with time, and when a boundary is reached the corresponding response is initiated. Stimuli associated with the upper boundary response have a positive v and stimuli associated with the lower boundary a negative v. Errors occur when the wrong boundary is reached first due to the effects of accumulation noise. The observed RT is the sum of a decision component and a non-decision component. The decision component is the time required for the process to move from the start point x_0 to the boundary, and the non-decision component, t_0, is the total time required to initially encode evidence from the choice stimulus and to produce a response once a boundary is reached. An increased distance a between the boundaries represents an increase in response caution. An increase in response caution increases both decision time—because the evidence total has to travel farther—and accuracy—because there is a longer time to average out the effects of accumulation noise. Wabersich and Vandekerckhove (2014) provide an R package to compute the Wiener decision model distribution functions.

Formally, the basic diffusion model is the solution to the first-passage problem with

boundaries located at 0 and a for a stochastic differential equation driven by white noise (Ratcliff, 1978; Ratcliff & McKoon, 2008),

$$X(t) = x_0 + v(t - t_0) + sB(t - t_0), \quad t \in [t_0, \infty). \tag{1}$$

Here, $X(t)$ is the solution process at time t. The non-decision component of the decision process is incorporated via the time shift $t - t_0$, so that the process has starting point x_0 at time $t = t_0$. The term $v(t - t_0)$ is a deterministic offset from the starting point with slope given by the drift rate v. The term $sB(t - t_0)$ is a time-shifted Brownian motion scaled by the diffusion coefficient s, which represents a normally distributed stochastic offset with mean 0 and variance $s^2(t - t_0)$. The first-passage distribution of the model remains unchanged if the parameters a, v and x_0 are divided by s.[1] Therefore, it is commonly assumed that $s = 1$.

This Wiener diffusion process produces uni-modal and positively skewed decision time distribution shapes that are characteristic of empirical RT distributions (see Figure 1). Different series expressions are available for the density of the first-passage time distribution (see, e.g., Foster & Singmann, 2021; Ratcliff, 1978). A small-time representation that converges quickly for small values of t is particularly attractive for applications in psychology (Gondan et al., 2014; Van Zandt et al., 2000). The first-passage density at the lower bound is:

$$f_-(t; a, v, x_0, t_0) = \frac{1}{a^2\sqrt{2\pi(t - t_0)^3}} e^{-vx_0 - \frac{1}{2}v^2(t-t_0)} \sum_{n=-\infty}^{\infty} \frac{x_0 + 2an}{a} e^{-\frac{(x_0+2an)^2}{2a^2(t-t_0)}}, \tag{2}$$

[1] The invariance of the first-passage distribution under affine transformations can be directly seen from the full expression for the first-passage density given in Ratcliff (1978), where all terms involving a, v and x_0 appear as ratios of s. For an intuitive argument, dividing the term $sB(t - t_0)$ in Equation (1) by s changes the size of the stochastic moment-to-moment fluctuations by a factor $1/s$. Therefore, rescaling the deterministic moment-to-moment fluctuations (i.e., v) as well as the spatial domain of the process (i.e., the starting point x_0 and the boundary separation a) by the same factor yields a stochastic process with the same first-passage distribution.

and the density for the upper bound is obtained by reversing the sign of the drift rate v and replacing x_0 by $a - x_0$, that is, $f_+(t; a, v, x_0, t_0) = f_-(t; a, -v, a - x_0, t_0)$.

Full Diffusion Decision Model

Although the basic diffusion model produces the correct shape for a decision time distribution, it is not by itself a plausible model of binary choice because when responding is unbiased (i.e., $x_0 = a/2$) correct and error decision-times have identical distributions. In contrast, participants responding cautiously generally have slower correct than error responses, whereas when they prioritize speed over accuracy errors can be slower than correct responses. Building on the work of Laming (1968) and Ratcliff (1978), Ratcliff and Rouder (1998) pointed out that this pattern could be accommodated by Gaussian across-trial variability in the rate of accumulation, with mean v and standard deviation s_v, and uniformly distributed across-trial variability in the starting point, with mean x_0 and range s_{x_0}. Rate variability causes slow "stimulus-quality" errors that cannot be entirely eliminated by increasing a, and starting-point variability causes fast "response-caution" errors that can be eliminated by increasing a (see Damaso et al., 2020, for further discussion of these two types of errors in evidence-accumulation models).

The third elaboration, which completes the specification of the full DDM that we focus on here, is uniform across-trial variability in non-decision time with mean t_0 and range s_t (Ratcliff & Tuerlinckx, 2002). Although these three elaborations make the DDM a realistic model of choice RT, the across-trial variability parameters can be hard to estimate (Boehm, Annis, et al., 2018; Evans et al., 2020; Lerche & Voss, 2016; Lerche et al., 2017; van Ravenzwaaij & Oberauer, 2009).

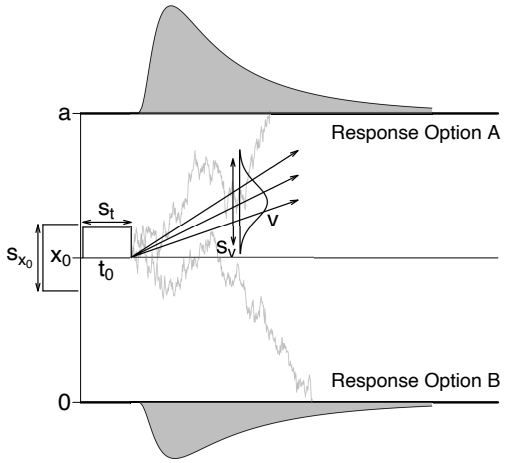

Figure 1

Diffusion Decision Model. The model describes decision making as the accumulation of noisy evidence to one of two thresholds located at 0 *and a that each correspond to one response option. The evidence signal on a single trial is described as a Wiener process with diffusion constant* 1 *and a drift coefficient that is sampled from a normal distribution with mean* v *and standard deviation* s_v*. The starting point of the Wiener process on each trial is sampled from a uniform distribution with mean* x_0 *and range* s_{x_0}*. An additive non-decision term that is sampled from a uniform distribution with mean* t_0 *and range* s_{t_0} *on each trial accounts for the time required for sensory encoding and response execution. The core model parameters* a, v, t_0, x_0 *together with the stochastic within-trial variability of the Wiener process account for the general right-skewed shape of empirical response time distributions with a long right tail. The across-trial variability parameters* s_v, s_{t_0}, s_{x_0} *account for fine-grained details in the leading edge and the relationship between the means of the observed correct and error RT distributions.*

Taken together, the seven parameters of the full DDM $a, v, t_0, x_0, s_v, s_{x_0}, s_{t_0}$ are the boundary separation, drift rate, non-decision time, starting point, and across-trial variability in drift rate, starting point, and non-decision time, respectively. In practical applications it is often convenient to parameterize the model in terms of the relative

starting point $z = x_0/a$ and variability in the relative starting point $s_z = s_{x_0}/a$, which is the parameterization that we will use in the sequel.

The first-passage density of the full DDM at the lower bound is:

$$f_-(t; a, v, x_0, t_0, s_v, s_{x_0}, s_{t_0}) = \int_{x_0-\frac{s_{x_0}}{2}}^{x_0+\frac{s_{x_0}}{2}} \int_{t_0-\frac{s_{t_0}}{2}}^{t_0+\frac{s_{t_0}}{2}} \frac{1}{a^2 s_{x_0} s_{t_0} \sqrt{2\pi(t-\tau)^3(1+s_v^2(t-\tau))}} \times e^{\frac{s_v^2\zeta^2 - 2v\zeta - v^2(t-\tau)}{2(1+s_v^2(t-\tau))}} \sum_{n=-\infty}^{\infty} \left(\frac{\zeta}{a} + 2n\right) e^{-\frac{(\zeta+2an)^2}{2a^2(t-\tau)}} \, \mathrm{d}\tau \, \mathrm{d}\zeta. \tag{3}$$

As can be seen from the form of the density, whereas across-trial drift variability leads to an analytic expression, the density with across-trial variability in non-decision time and start-point requires numerical integration over the uniform non-decision time and start-point distributions. An efficient implementation of the density and distribution function is available in an R package (Singmann et al., 2020) based on the C-code of Voss and Voss (2007) (see also Voss et al., 2015).

Bayesian Implementation of Diffusion Models

Pioneering work on developing a Bayesian implementation of evidence accumulation models based on the basic diffusion model was conducted by Vandekerckhove et al. (2011). In their first application fitting data from single participants Vandekerckhove et al. introduced all three types of across-trial variability as random effects. Their second application assumed unbiased responding (i.e., $z = 1/2$) and used a multi-level structure, with random trial effects on drift rates and non-decision times at the first level. At the second level there were random participant effects for response caution, a, and for the mean and across-trial variability parameters for drift rates and non-decision times.

Vandekerckhove et al. (2011) note that their way of introducing across-trial variability, which they call a hierarchical diffusion model (HDM), is "akin" to the DDM. However, it is important to acknowledge a key difference: as the number of trials increases the effect of the assumed form of the across-trial distributions washes out. The upshot is that the across-trial distribution functions can take on whatever arbitrary form that best

fits the data, at least in the limit of a large number of trials. Subsequently, Jones and Dzhafarov (2014) showed that a diffusion process with arbitrary across-trial distributions, which they call the general diffusion model (gDM), is unfalsifiable because it can fit any pattern of data. Hence, the HDM becomes equivalent to the gDM as the number of trials grows. However, leading proponents of the DDM see the fixed forms of the across-trial variability distributions as key assumptions (Smith, Ratcliff, & McKoon, 2014). Therefore, a Bayesian implementation of across-trial variability as a random effect does not appear to be desirable.

A more recent hierarchical Bayesian implementation of the full DDM takes account of the exact form of across-trial variability but assumes only a group level. In their hierarchical drift-diffusion model (HDDM)[2] Wiecki et al. (2013) chose to implement the three across-trial variability parameters as fixed effects. They did so because "the influence of these parameters onto the likelihood is often so small that very large amounts of data would be required to make meaningful inference at the individual level" (p. 3). Although this greatly helps Markov chain Monte-Carlo (MCMC) estimation of the model, it seems plausible that levels of across-trial variability might differ between individuals. It is also unclear whether high correlations among parameters might not cause biases in estimates of other parameters when this is the case.

Fully Hierarchical Implementation of the Full DDM

Each of the Bayesian implementations of the diffusion model discussed so far ignores some important aspects of the (hierarchical) structure of real data. Vandekerckhove et al. (2011) only implement the basic diffusion model likelihood (Equation (2)). Although they argue that the three across-trial variability parameters can be introduced as random

[2] The terms "drift-diffusion" is commonly used in neuroscience, sometimes referring to the simple Wiener diffusion model and sometimes to versions with across-trial variability. We adopt Ratcliff and McKoon's (2008) "diffusion-decision model" terminology for the version with all three types of within-participant variability.

effects, doing so causes considerable problems for MCMC algorithms that are used to obtain posterior samples (Boehm, Annis, et al., 2018). Moreover, the resulting model is ill-specified as the distribution of the across-trial parameters is data-dependent and will thus fit any sufficiently large data set (i.e., the model cannot be falsified). Wiecki et al. (2013) implement the full DDM likelihood (Equation (3)) but only allow for fixed effects on the three across-trial variability parameters, which embodies the assumption that these parameters are the same across individuals. Our hierarchical implementation of the DDM addresses these shortcomings of earlier implementations by using the full DDM likelihood together with a complete mixed modeling framework that can flexibly accommodate fixed and random effects of participants and experimental conditions, as well as being amenable to extension to general linear models (e.g., ANOVA, ANCOVA, and regression designs). In addition, we adopt the convention of parameterizing condition and participant effects in terms of standardized, dimensionless effect sizes. This approach, which was developed in statistical modeling but has not hitherto been applied to diffusion modeling, allows us to impose default priors on the effect sizes and supports comparability of results across studies.

Heathcote et al.'s (2019) Dynamic Models of Choice (DMC) software package provides R functions for hierarchical Bayesian estimation of a range of evidence-accumulation models, including the full DDM and simplified versions with one or more sources of across-trial variability removed. MCMC estimation is achieved with the Differential Evolution algorithm (ter Braak, 2006; Turner et al., 2013) that is well suited to models with highly correlated parameters. Hence, across-trial variability parameters can be estimated as random effects although, as noted by Wiecki et al. (2013), this does require larger trial numbers, particularly for the starting-point range. HDDM only allows for fixed truncated normal non-informative priors or informative priors based on Matzke and Wagenmakers's (2009) literature review (see Tran et al., 2021, for a more recent update with similar findings), whereas DMC provides a wide range of prior distribution forms with

parameters set by the user. DMC also provides a range of model-selection criteria, including the *deviance information criterion* (DIC; Spiegelhalter et al., 2002), which is a hierarchical generalization of the Aikaike information criterion (AIC; Aikaike, 1974), the Bayesian predictive information criterion (BPIC, Ando, 2007), which is an approximation of the posterior mean of the expected log-likelihood of the predictive distribution, the widely applicable information criterion (WAIC; Vehtari et al., 2016), which is an asymptotic approximation to Bayesian cross-validation, and Bayes factors (Gronau, Heathcote, et al., 2019), which are the ratio of the marginal likelihood of two candidate models. The marginal likelihood in DMC is estimated through Warp-III bridge sampling (Gronau, Wagenmakers, et al., 2019; Meng & Schilling, 2002), which iteratively updates an estimator of the marginal likelihood based on a proposal distribution whose first three moments are matched to the posterior distribution.

Hierarchical mixed modeling in DMC

The approach we develop here is built on DMC. We extend DMC's hierarchical DDM implementation by a Bayesian hierarchical mixed modeling framework that supports model selection as well as model-averaging. Our approach to inference is based on inclusion Bayes factors. For ease of exposition we will discuss this approach for an experiment where P participants are tested in two experimental conditions. However, we emphasize that the same principles can be applied to arbitrary study designs that can be accommodated through linear functions of the cognitive model parameters and covariates. Moreover, we assume that all model parameters can be transformed to have a normal distribution. While several DDM parameters are supported only on a bounded interval or the positive real line, the sampling algorithm in DMC assumes that parameters are supported on the entire real line. We accommodate this constraint by applying transformations that map the support of the parameter to the entire real line. For convenience we assume that the transformed parameter is normally distributed.

A first consideration in the development of our hierarchical implementation is that it needs to take random participant effects into account. These describe individual differences in parameter values, and it seems reasonable that such effects do exist. For example, if a participant's evidence accumulation is more efficient than average (i.e., they have a higher v than other participants) in one condition it seems likely, all other things being equal, that they would be more efficient in another condition. Similarly, if a participant was more cautious than average (i.e., they had a higher a) then that would likely apply across all conditions. However, all hierarchical DDM analyses discussed previously neglect these individual differences. Let θ be any one DDM parameter. Earlier analyses assumed that the model parameter $\theta_{p,c}$ for participant $p = 1, \ldots, P$ is sampled independently in each experimental condition $c = 1, 2$ from, say a normal distribution with mean μ_c and variance σ^2

$$\theta_{p,c} \sim \mathcal{N}(\mu_c, \sigma). \tag{4}$$

Note that in this model specification the effect of the experimental condition is treated as a fixed effect.

In our hierarchical implementation, we account for individual differences in terms of additive random participant effects, which induce a positive correlation in parameters over conditions. Specifically, we assume that the model parameter $\theta_{p,c}$ for participant $p = 1, \ldots, P$ is sampled from a normal distribution

$$\theta_{p,c} \sim \mathcal{N}(\mu_p + \mu_c, \sigma), \tag{5}$$

where μ_p is the additive participant effect. It has been recommended in the linear mixed-modeling literature (Barr et al., 2013) to use maximal random effects structures (i.e., including all interactions between fixed and random factors). However, using maximal random effects can negatively affect statistical power (Matuschek et al., 2017) and will potentially aggravate existing sampling and sample size issues already posed by the DDM. We leave the exploration of more elaborate random effects models to future work.

A second consideration in the development of our hierarchical implementation is the choice of prior distributions for the random participant effects and fixed condition effects. A possible approach to this problem might be to base the prior distributions on parameter estimates obtained in earlier studies. Surveys conducted by Matzke and Wagenmakers (2009) and Tran et al. (2021) examined published parameter values that were, in the main, averaged over participants and experimental conditions. Hence, such surveys can provide priors for the mean DDM parameter values, but do not provide information about differences between experimental conditions (i.e., fixed condition effects) nor about average differences between participants (i.e., random participant effects). Moreover, the size of fixed and random effects will often be context-dependent. In clinical samples, for instance, DDM parameters will tend to vary more widely between participants than in healthy controls (e.g., Dillon et al., 2015; Huang-Pollock et al., 2017; Weigard et al., 2018), which would need to be reflected in wider priors for random participant effects in clinical samples.

Instead of specifying priors for fixed and random effects on an absolute scale, we adopt Jeffreys's (1961) recommendation to use standardized effect sizes. Standardized effect sizes (e.g., the difference between parameters in two conditions divided by a measure of variability in parameters over participants) are dimensionless quantities for which it is easier to specify "default" priors (Liang et al., 2008). In order to take advantage of the available information about mean parameter values, we introduce an intercept term that is modeled on an absolute scale, and whose prior distribution we base on Matzke and Wagenmakers's (2009) survey. That is, we assume that the model parameter $\theta_{p,c}$ for participant $p = 1, \ldots, P$ is sampled from a distribution

$$\theta_{p,c} \sim \mathcal{N}(\mu_\theta + \underbrace{\sigma_{res}\alpha_p}_{=\mu_p} + \underbrace{\sigma_{res}\beta_c}_{=\mu_c}, \sigma_{res}). \tag{6}$$

Here, μ_θ is the intercept for the DDM parameter θ, which we assign an informative prior based on surveys of published parameter estimates, independent of the priors for the remaining terms $\sigma_{res}\alpha_p + \sigma_{res}\beta_c$. The random person effect on the absolute scale, μ_p, is

now expressed as the product $\sigma_{res}\alpha_p$, where σ_{res} is the residual standard deviation[3] of the DDM parameter θ, and α_p is the standardized person effect. Taking the product $\sigma_{res}\alpha_p$ scales the dimensionless standardized effect size α_p to the absolute scale of the DDM parameter θ. Similarly, the fixed condition effect on the absolute scale, μ_c, is expressed as the product $\sigma_{res}\beta_c$, where β_c is the standardized condition effect. Up to some technical details that we will deal with later, the standardized effects α_p and β_c can be thought of as the deviation from μ_θ in units of residual standard deviations.

The parameterization of the DDM in terms of dimensionless standarized effect sizes and an intercept on an absolute scale has two main advantages. First, the estimated effect sizes can be more readily compared across studies than effects on the absolute scale. Second, default priors guarantee the consistency of inference across studies. The choice of prior distributions affects the value of the Bayes factors described in the next section (Kass & Raftery, 1995), which means that different choices of priors can lead to different conclusions based on the same data. Assigning default priors to the standardized effect sizes and basing the prior for the intercept term on surveys of previous studies eliminates the need to develop priors ad hoc, and thus supports consistent and comparable inference across studies.

Bayesian Model Averaging for Inference Under Model Uncertainty

Bayesian methods offer a natural way of addressing the second challenge for cognitive modeling, the need for statistical inference and estimation to appropriately account for model uncertainty. We begin with a brief summary of Bayesian estimation theory, which provides the main motivation for the development of our approach to inference under model uncertainty.

---

[3] Since the mixed effects framework we use here was developed for regression models, we refer to σ_{res} as the "residual standard deviation". However, as we explain in the section "Model specification", the DDM parameters θ only make contact with the data via the DDM likelihood.

Bayesian estimation theory

Bayesian estimation theory considers the problem of obtaining a guess for the value of an unknown quantity of interest θ, based on observed data D, which is optimal in a sense that will be made precise shortly. Uncertainty about the true value of θ is modelled by assigning a prior distribution $\pi(\theta)$ that describes the a priori plausibility of different values of θ before any data are observed. An estimator is a map $\hat{\theta}(D)$ that provides a guess of the value θ based on the observed data D. The cost of guessing θ incorrectly is expressed by a loss function $L(\theta, \hat{\theta})$ that depends on the unknown true value of θ and the the value of the estimator $\hat{\theta}$. A common choice for the loss function is the squared error $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. An estimator is called a *Bayes estimator* if it minimizes the expected loss under the prior distribution $\mathbb{E}_\pi[L(\theta, \hat{\theta})]$.

Bayesian Model Averaging

Our approach to inference here is based on Bayesian model averaging (BMA), which results in optimal predictions under squared error loss (and other convex loss functions; Hoeting et al., 1999). More generally, BMA results in better predictions than any individual model (Beck et al., 2008; Höge et al., 2019; Iverson et al., 2010; Raftery et al., 2005; Vehtari & Ojanen, 2012). In particular, our main inferential tool will be *inclusion Bayes factors*, which offer a principled solution to the problem of making inferences about model parameters in the face of model uncertainty (Gronau, Wagenmakers, et al., 2019; Hinne et al., 2020; Hoeting et al., 1999; Jeffreys, 1961; Jevons, 1874). Inclusion Bayes factors can be used to combine the information about a parameter of interest across models, weighing the contribution of each model by its marginal likelihood. The marginal likelihood includes an implicit penalty for model complexity (Jefferys & Berger, 1992; Jeffreys, 1939; MacKay, 2003; Myung & Pitt, 1997), which means that inferences based on inclusion Bayes factors take both model uncertainty and model complexity into account. As Hinne et al. (2020) point out, basing inferences on the complete set of candidate models

has several advantages over single-model inference. Weighing the contribution of each candidate model by its a posteriori plausibility avoids placing too much confidence in any single model. This also acts as a safeguard against model misspecification. Whereas in single-model-inference conclusions stand and fall with the winning model, basing inferences on a set of models means that models that are approximately correct can compensate for the misspecification of other models. The BMA framework also offers advantages for parameter estimation. Collecting additional data may change which model has the highest a posteriori plausibility. If parameter estimates are based on a single winning model, switching between models can result in sudden changes in parameter estimates. BMA, on the other hand, bases parameter estimates on a weighted average across candidate models, which means that parameter estimates will change gradually as more data are collected.

Returning to our setup where a group of P participants is tested in two experimental conditions, the researcher wants to test the hypothesis that a DDM parameter θ differs between the two experimental conditions. In most applications only the four main DDM parameters are of interest: $\theta \in \{a, v, z, t_0\}$. Assume that the parameter of interest is response caution a. The question now arises what assumptions the researcher should make about the remaining main parameters, v, z, t_0. In the BMA approach we advocate here, instead of committing to a single set of assumptions about the remaining main parameters, the researcher bases inference on the set of all hierarchical models that can be specified by letting a subset of the main DDM parameters differ between conditions. That is, each of the four main DDM parameters can either be free to differ between experimental conditions, or be fixed to be equal in both conditions. This yields a total of 2 (a differs between conditions vs. a does not differ between conditions) $\times 2$ (v differs between conditions vs. v does not differ between conditions) $\times 2$ (z differs between conditions vs. z does not differ between conditions) $\times 2$ (t_0 differs between conditions vs. t_0 does not differ between conditions) $= 16$ candidate models $\mathcal{H} = \{H_1, \ldots, H_{16}\}$, as illustrated Table 1. In each model the fixed condition effect for some parameters is zero

(i.e., $\mu_1 = \mu_2$), which is indicated by a space, whereas the fixed condition effect for the remaining parameters is allowed to be non-zero (i.e., $\mu_1 \neq \mu_2$), which is indicated by a "+". We denote by $\mathcal{H}_0 = \{H_1, \ldots, H_8\}$ the subset of models in which a is equal across conditions (i.e., the models shown in the first 8 columns of Table 1 with a space in the row for a), which in psychological terms means that participants' response caution is the same in both conditions while any possible combination (represented by the different models) of their rate of evidence accumulation v, their start point z, and their non-decision time t_0 may differ between conditions. Similarly, we denote by $\mathcal{H}_1 = \{H_9, \ldots, H_{16}\}$ the subset of models that allow a to differ between conditions (i.e., the models shown in the last 8 columns of Table 1 with a "+" in the row for a), which in psychological terms means that participants' response caution may differ between conditions, in addition to any combination of the remaining three main parameters.

Table 1

Full set of candidate models. Each shows the specification of one candidate model, where a '+' indicates that the fixed condition effect for the DDM parameter is allowed to be non-zero.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}	H_{13}	H_{14}	H_{15}	H_{16}
a									+	+	+	+	+	+	+	+
v					+	+	+	+					+	+	+	+
z			+	+			+	+			+	+			+	+
t_0		+		+		+		+		+		+		+		+

Next, the researcher fits all 16 candidate models to the data D and computes the marginal likelihood of each. Each model H_k has a likelihood function $\ell(\cdot \mid H_k, \theta)$ that depends on a set of parameters with parameter space Θ_k. The researcher assigns a joint prior distribution π_k to these parameters and computes the *marginal likelihood* of model

H_k by integrating over the prior:

$$p(D \mid H_k) = \int_{\Theta_k} \ell(D \mid H_k, \theta)\pi_k(\theta)\,\mathrm{d}\theta. \tag{7}$$

Intuitively, the marginal likelihood describes the predictive adequacy for the data of model H_k, averaged over all possible values of the model's parameters.

The marginal likelihoods can now be used to compute the inclusion Bayes factor, which describes the change in the plausibility of a condition effect on a after observing the data. The researcher assigns a prior probability $p(H_k)$ to each of the candidate models such that $\sum_{k=1}^{16} p(H_k) = 1$. The prior probabilities describe the plausibility of each model before observing any data. In the simplest case, the researcher might deem all candidate models equally plausible a priori and assign prior probability $p(H_k) = \frac{1}{16}$ to all models. Consequently, the *prior inclusion odds* are given by the prior plausibility of the eight models in which the fixed condition effect for a is allowed to be non-zero relative to the prior plausibility of the eight models in which the fixed condition effect for a is zero.

Uniform prior probabilities are appropriate when no prior information is available (Jeffreys, 1939). However, in the presence of strong prior knowledge one may instead use non-uniform prior probabilities. For instance, if data from an earlier study with a comparable setup are available, the posterior model probabilities from that study might be used as prior probabilities for the present analysis. Moreover, simple models with few parameters are typically preferable to complex models with many parameters (Wrinch & Jefferys, 1921). This preference can be accommodated by basing each candidate model's prior probability on the number of parameters in that model (Consonni et al., 2018; Jeffreys, 1939; Wilson et al., 2010).

The plausibility of each model after observing data D is computed via Bayes' rule. This yields the *posterior probability*:

$$p(H_k \mid D) = \frac{p(H_k)p(D \mid H_k)}{\sum_{k=1}^{16} p(H_k)p(D \mid H_k)}. \tag{8}$$

The posterior probability only provides information about the plausibility of a single model

after the data have been observed. The *inclusion Bayes factor* describes how the plausibility of the presence of a condition effect on a, which is expressed by the prior inclusion odds, changes after observing the data. The inclusion Bayes factor is computed as:

$$\text{BF}_{\text{inc } a} = \underbrace{\frac{\sum_{H_k \in \mathcal{H}_1} p(H_k \mid D)}{\sum_{H_l \in \mathcal{H}_0} p(H_l \mid D)}}_{\substack{\text{Posterior inclusion} \\ \text{odds}}} \Bigg/ \underbrace{\frac{\sum_{H_k \in \mathcal{H}_1} p(H_k)}{\sum_{H_l \in \mathcal{H}_0} p(H_l)}}_{\substack{\text{Prior inclusion} \\ \text{odds}}}. \tag{9}$$

BMA is not only applicable to the posterior probability of the candidate models, it can also be applied to the (marginal) posterior distribution of the model parameter of interest. To obtain the model-averaged posterior distribution of a parameter, the researcher first computes under each candidate model the marginal posterior distribution for the parameter of interest. We denote the parameter of interest by $\tilde{\theta}$ with domain $\tilde{\Theta}_k$, the marginal posterior distribution is then given by:

$$p(\tilde{\theta}_k \mid D, H_k) = \frac{\int_{\Theta_k \setminus \tilde{\Theta}_k} \ell(D \mid H_k, \theta) \pi_k(\theta) \, \mathrm{d}\theta}{\int_{\Theta_k} \ell(D \mid H_k, \theta) \pi_k(\theta) \, \mathrm{d}\theta}, \tag{10}$$

where $\Theta_k \setminus \tilde{\Theta}_k$ is the parameter space of the remaining model parameters. The model-averaged posterior distribution is the weighted average of the posterior distribution for the parameter of interest under each model, where the weights are given by the posterior model probabilities:

$$\bar{p}(\tilde{\theta}_k | D) = \sum_{k=1}^{16} p(H_k | D) p(\tilde{\theta}_k | D, H_k). \tag{11}$$

In particular, the model-averaged posterior distribution can be used to obtain point estimates for a parameter of interest. If the parameter of interest is $\tilde{\theta} = a$, for instance, a model-averaged posterior estimate $\hat{a}$ can be obtained by computing the weighted average of the posterior mean $\hat{a}_k$ under each model of the 16 models, where the weights are again the models' posterior probability:

$$\hat{a} = \sum_{k=1}^{16} \hat{a}_k p(H_k | D). \tag{12}$$

A credible interval can be computed based on the quantiles of the model-averaged posterior distribution.

In practice, the marginal likelihoods of the models as well as the posterior distributions of the parameters under each model are not available in closed form. Instead, posterior samples obtained through MCMC methods are used to approximate the marginal model likelihoods and the posterior parameter distributions under each candidate model. In the remainder of this work we will illustrate how the hierarchical mixed modelling and BMA methods discussed above can be implemented in practice. To this end we will reanalyse the data from Dutilh et al.'s (2019) study.

Real Data Analysis

We carried out our analysis of Dutilh et al.'s (2019) data in two steps. In a first step we applied Bayesian model averaging in combination with mixed-effect diffusion modeling to identify which cognitive processes had been manipulated in each experimental condition of Dutilh et al.'s study. In the second step of our analysis we conducted a simulation study to test whether our Bayesian model averaging approach could correctly identify the putative selective-influence manipulations. We begin by describing our re-analysis of Dutilh et al.'s data.

Methods

Data

Dutilh et al. (2019) conducted a collaborative blinded selective influence study. They collected data from twenty students performing a binary random-dot-motion-direction classification task in which speed-accuracy emphasis instructions (instruction), frequency of left and right motion directions (bias), and motion coherence (difficulty) were varied between blocks. The speed-accuracy instructions aimed to selectively manipulate participants' response caution. The frequency manipulation aimed to selectively manipulate participants' response bias, and the varying motion coherence aimed to selectively manipulate participants' rate of information processing.

Data from different experimental blocks were subsequently combined to create fourteen different data sets and collaborators were asked to use evidence-accumulation models to infer which combination of putative selective-influence had been manipulated between the experimental blocks that comprised the fourteen data sets. Our analysis was based on this data after removing trials with RTs less than 0.25s (0.6% of responses, see Smith and Lilburn (2020) for a similar fast-guess exclusion criterion).

Dutilh et al. (2019) report extensive statistical analyses on their behavioral data, which we summarize briefly here. A Bayesian ANOVA on the arcsine-transformed accuracy data showed that participants responded more accurately under accuracy vs. speed instructions, and in the easy vs. hard condition, while the bias manipulation did not affect response accuracy. A Bayesian ANOVA on the mean RTs showed that participants responded faster under speed vs. accuracy instructions, and in the easy vs. hard condition, while the bias manipulation did not affect mean RTs. Moreover, Dutilh et al. computed the mean response criterion from signal detection analysis based on the proportions of hits and false alarms for each data set. A Bayesian ANOVA showed that the response criterion was affected by the bias manipulation under speed instructions but not under accuracy instructions. Taken together, their analyses indicated that the experimental manipulations had the desired effects on participants' behavior and that the effects of the manipulations were of sufficient size to be detectable by the DDM.

Table 2 shows the experimental manipulations and descriptive statics for the fourteen data sets used in Dutilh et al.'s study. Response accuracy was relatively high but well below ceiling in nearly all data sets, which means that sufficient incorrect responses were available for the estimation of the DDM parameters. Moreover, the RT quantiles indicate that RT distributions had the typical right-skewed shape with a long right tail (i.e., the 10% quantile was much closer to the 50% quantile than the 90% quantile), which makes the data amenable to modeling with the DDM (Ratcliff & McKoon, 2008).

Data format for DMC. We used the DMC software package for R (Heathcote et al., 2019) to fit the mixed-effects DDMs to the data. The DMC functions require all of the data to be in a long data frame of 7 columns, with each of these columns requiring a specific aspect of the data, a specific column name, and a specific naming convention for the data itself. Each row should be the data from a single trial, meaning the total number of rows in the data frame should be $subjects \times trials$ (assuming that all subject completed the same number of trials).

The columns of the data frame should have the following format. The first column should be labeled "s", and be an R data type factor containing the subject numbers for each participant, reflected as integers labeled from 1 to the total number of participants. The second column should be labeled "S", and be a factor containing the stimulus identity for each trial, reflected by the labels of "s1" for one stimulus and "s2" for the other stimulus. The third column should be labeled "F", and be a factor containing the condition identity for each trial, reflected by the labels of "f1" for one condition and "f2" for the other condition. The fourth column should be labeled "R", and be a factor containing the response identity for each trial, reflected by the labels of "r1" for one response and "r2" for the other response, where r1/r2 would lead to a correct response for stimulus s1/s2, respectively. The fifth column should be labeled "RT", and be numeric values

Table 2

Descriptive statistics for the 14 data sets from Dutilh et al. (2019).

Data Set	Condition	Manipulation			Mean Accuracy	Mean RT Quantile (s)		
		Difficulty	Instruction	Bias		0.1	0.5	0.9
1	A	hard	speed	no bias	0.75	0.37	0.49	0.7
	B	hard	speed	no bias	0.74	0.38	0.49	0.69
2	A	hard	speed	no bias	0.96	0.48	0.6	0.83
	B	easy	speed	no bias	0.74	0.37	0.6	0.69
3	A	hard	speed	no bias	0.86	0.37	0.48	0.63
	B	hard	accuracy	no bias	0.86	0.5	0.48	1.15
4	A	hard	speed	no bias	0.74	0.37	0.49	0.69
	B	hard	speed	bias	0.85	0.5	0.49	1.13
5	A	hard	speed	no bias	0.95	0.48	0.6	0.84
	B	easy	accuracy	no bias	0.74	0.38	0.6	0.69
6	A	hard	speed	no bias	0.84	0.36	0.47	0.62
	B	easy	speed	bias	0.86	0.49	0.47	1.05
7	A	hard	speed	no bias	0.74	0.38	0.51	0.69
	B	hard	accuracy	bias	0.86	0.5	0.51	1.12
8	A	easy	speed	no bias	0.86	0.5	0.68	1.13
	B	hard	accuracy	no bias	0.84	0.36	0.68	0.62
9	A	easy	speed	no bias	0.74	0.37	0.49	0.69
	B	hard	speed	bias	0.86	0.37	0.49	0.64
10	A	hard	accuracy	no bias	0.75	0.37	0.49	0.7
	B	hard	speed	bias	0.74	0.37	0.49	0.69
11	A	easy	speed	no bias	0.73	0.37	0.49	0.68
	B	hard	accuracy	bias	0.86	0.49	0.49	1.04
12	A	hard	accuracy	no bias	0.74	0.38	0.5	0.69
	B	easy	speed	bias	0.84	0.36	0.5	0.62
13	A	hard	speed	bias	0.75	0.36	0.49	0.69
	B	easy	accuracy	no bias	0.84	0.36	0.49	0.62
14	A	hard	speed	no bias	0.95	0.47	0.59	0.82
	B	easy	accuracy	bias	0.75	0.37	0.59	0.69

corresponding to the response time in seconds for each trial. The sixth and seventh columns, which should be labeled "Ff1" and "Ff2" respectively, are dummy-coded versions of the third "F" column used to define the effects. Specifically, the sixth column, "Ff1", should have a numeric value of 1 in all cases where the trial is from condition "f1", and a numeric value of 0 in all cases where the trial is from the condition "f2". The seventh column, "Ff2", should have a numeric value of 0 in all cases where the trial is from condition "f1", and a numeric value of 1 in all cases where the trial is from the condition "f2".

Modeling

We applied Bayesian model averaging in combination with mixed-effect diffusion modeling to identify which cognitive processes had been manipulated in each experimental condition of Dutilh et al.'s study. Each experimental condition of the study targeted a subset of the four cognitive processes that are represented by the four core DDM parameters. We therefore implemented sixteen different hierarchical DDMs (listed in Table 1), one for each possible combination of targeted cognitive processes (including no effect of any manipulation). These models allowed the DDM parameters that correspond to the targeted cognitive process to vary between experimental conditions, while all other parameters were held constant across experimental conditions. We fit these sixteen models to each of the fourteen data sets, estimated the marginal likelihood of each model, and computed the posterior inclusion probability for each of the four DDM parameters.

Model specification

We specified our mixed-effect DDMs at two levels. Let C be the number of experimental conditions, P the number of participants and T the number of trials in each experimental condition. We denote the joint DDM likelihood of response times and decisions by $DDM(a, v, t_0, z, s_v, s_z, s_{t_0})$ (i.e., we avoid having to differentiate between the densities $f_{\pm}(a, v, t_0, x_0, s_v, s_z, s_{t_0})$ for correct and incorrect responses and use the

parameterization in terms of the relative starting point z), where the model parameters $a, v, t_0, z, s_v, s_z, s_{t_0}$ are mean boundary separation, drift rate, non-decision time, relative starting point, and across-trial variability in drift rate, relative starting point, and non-decision time, respectively.[4] At the trial level, we assumed that the response time $RT_{c,p,t}$ and decision $d_{c,p,t}$ for trial t of participant p performing experimental condition c is distributed as:

$$(RT_{c,p,t}, d_{c,p,t}) \sim DDM(a_{c,p}, v_{c,p}, t_{0_{c,p}}, z_{c,p}, s_{v_p}, s_{z_p}, s_{t_0}). \tag{13}$$

As can be seen, we assumed that the across-trial parameters s_v and s_z were constant across experimental conditions as large amounts of data are required to estimate these parameters (Boehm, Annis, et al., 2018). Moreover, we set $s_{t_0} = 0$ as obtaining convergence of MCMC chains for this parameter is notoriously slow.

To specify the participant level of our model, we used a probit-transformed version—$\tilde{z}$ and $\tilde{s_z}$—of the z and s_z parameters, respectively, and log-transformed versions—$\tilde{a}$, $\tilde{t}_{t_0}$ and $\tilde{s}_v$—of the a, t_0, and s_v parameters, respectively, which have support on the entire real line. We modeled the seven DDM parameters as being independent normally distributed and imposed random participant effects. In addition, we assumed fixed condition effects on the four main DDM parameters, a, v, t_0, z, whereas the across-trial parameters were only assigned a fixed intercept. Although our implementation supports estimation of non-decision time variability, we set $s_{t_0} = 0$ in all models to reduce the already considerable computing times for our model fits.

Fixed condition and random participant effects. We followed Rouder et al. (2012) for the specification of the random and fixed effects parts of the model. If X_C is the $(P \cdot T) \times C$ design matrix for the effect of experimental condition, we computed the

[4] The relative starting point parameters are restricted to the unit interval. The mean relative starting point is $z = x_0/a$, where x_0 is the absolute starting point. To ensure that the uniform distribution from which the absolute starting point is sampled on each trial does not exceed the interval $[0, a]$, its range is defined as $s_z \times 2 \times \min\{x_0, a - x_0\}$.

$(P \cdot T) \times (C-1)$ design matrix for the corresponding fixed effect of experimental condition, X_C^*, using the matrix of orthonormal contrasts Q_C. That is, we computed $X_C^* = X_C Q_C$, where Q_C is the matrix of eigenvectors of $\Sigma_C = I_C - \frac{1}{C} J_C$, I_C is the $C \times C$ identity matrix, and J_C is the $C \times C$ matrix of with all entries equal to 1 (see Equation (13) in Rouder et al. (2012)). Using the matrix X_C^* instead of the full design matrix X_C ensures that the fixed condition effects sum to zero. For specificity, the experimental design we analyze below was fully balanced, meaning that every one of the P persons completed T trials in each of the $C = 2$ experimental conditions. Then

$$\Sigma_2 = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \text{and} \quad Q_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}. \tag{14}$$

is the single eigenvector corresponding to the non-zero eigenvalue 1. The full design matrix is

$$X_C = \left.\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \end{pmatrix}\right\} \text{block of } 2T \text{ rows is repeated } P \text{ times}, \tag{15}$$

and the corresponding design matrix for the fixed effects is

$$X_C^* = X_C Q_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ \vdots \\ -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ \vdots \end{pmatrix}. \tag{16}$$

The random person effects were modeled using the full design matrix because no constraint is imposed that the random effects should sum to zero.

The linear model for the four main DDM parameters was:

$$\theta_{c,p} \sim \mathcal{N}\left(\mu_\theta + \sigma_{res\,\theta}((X_P)_{\cdot,c,p} \cdot \boldsymbol{\alpha}_\theta + (X_C^*)_{\cdot,c,p} \cdot \boldsymbol{\beta}_\theta), \sigma_{res\,\theta}\right), \tag{17}$$

where $\theta \in \{\tilde{a}, v, \tilde{t}_0, \tilde{z}\}$, $(X_\cdot)_{\cdot,\,c\,p}$ denotes the row of the design matrix corresponding to person p in condition c, and $\boldsymbol{\alpha}_\theta$ and $\boldsymbol{\beta}_\theta$ denote the vectors of P and $C-1$ standardized effect sizes, respectively. Moreover, μ_θ denotes the intercept term for the model parameter vector θ, and $\sigma_{res\,\theta}$ denotes its residual variance. For the two across-trial variability parameters s_z and s_v, the participant-level model was:

$$\theta_p \sim \mathcal{N}\left(\mu_\theta, \sigma_{res\,\theta}\right). \tag{18}$$

Finally, we assumed the priors for the vectors of standardized condition effects had multivariate Cauchy distributions. That is, the prior distribution for the fixed condition effects was specified as:

$$\boldsymbol{\beta}_\theta \sim \text{Cauchy}_C\left(\mathbf{0}, r_{C,\theta} I\right), \tag{19}$$

where Cauchy_C denotes the C-dimensional Cauchy distribution, $\mathbf{0}$ is a location vector with C entries, I is the $C \times C$ identity scale matrix for the multivariate Cauchy distribution and $r_{C,\,\theta}$ is the scaling factor for the fixed condition effects. For the standardized effect sizes for the random participant effects we assumed a multivariate normal distribution:

$$\boldsymbol{\alpha}_\theta \sim \mathcal{N}_P\left(\mathbf{0}, r_{P,\theta} I\right), \tag{20}$$

where $\mathcal{N}_P$ denotes the P-dimensional normal distribution, and $r_{P,\theta}$ is the scaling factor for the random participant effects. The latter choice of prior distribution, which deviates from Rouder et al.'s (2012) approach, was made for two reasons. First, because the random subject effects were not the target of our statistical inference, we considered normal priors to be sufficiently non-informative to not affect our inferences about the fixed condition effects. Second, the heavy tails of the Cauchy distribution can slow convergence of MCMC posterior sampling considerably. We assigned weakly informative truncated normal priors

to the residual standard deviations,

$$\sigma_{res\,\theta} \sim \mathcal{N}\left(\mu_{\sigma\,\theta}, \sigma_{\sigma\,\theta}\right)[0, \infty), \tag{21}$$

where $[0, \infty)$ indicates truncation to the positive real line. Our use of informative priors for the residual standard deviations, instead of the improper priors suggested by Rouder et al. (2012), was due to the need to approximate the marginal model likelihoods by sampling. Rouder et al. consider single-model comparisons of linear models against a null model, in which case the improper prior drops out of the computation, and the Bayes factor is obtained by low-dimensional numerical integration. In contrast, the marginal likelihoods required for the comparison of several highly nonlinear models we consider here makes it necessary to approximate high-dimensional integrals by MCMC-sampling, which renders the use of improper priors impractical.

Construction of informative priors for intercepts. As discussed earlier, the cognitive-process interpretation of the parameters of a cognitive model means that informative priors should be used for the intercept terms in our model. Here, we assigned informative normal priors to the intercepts for the DDM parameters:

$$\mu_\theta \sim \mathcal{N}\left(\mu_{\mu\,\theta}, \sigma_{\mu\,\theta}\right). \tag{22}$$

The values of the hyperparameters for $\mu_{\sigma\,\theta}$ and $\sigma_{\sigma\,\theta}$ were chosen to be weakly informative whereas the values of the hyperparameters for $\mu_{\mu\,\theta}$ and $\sigma_{\sigma\,\theta}$ were informed by Matzke and Wagenmakers's (2009) survey of DDM parameter estimates. We computed the sample mean $\bar{M}_\theta$ and sample standard deviation $\bar{S}_\theta$ of the parameter estimates in Matzke and Wagenmakers's survey (using the relative values of z and s_z) and applied appropriate transformations to specify the prior distribution with support on the entire real line. For v no transformation was required. For $\theta \in \{a, t_0, s_v\}$ we used $\mu_\theta = \log(\bar{M}_\theta/(\bar{S}_\theta^2/\bar{M}_\theta^2 + 1)^{1/2})$ and $\sigma_\theta = \left(\log(\bar{S}_\theta^2/\bar{M}_\theta^2 + 1)\right)^{1/2}$. For $\theta = z$, we approximated the probit transformation by sampling and truncation so that the prior distributions would yield values in the range $[0, 1]$. To this end, we generated 10,000 samples from a normal distribution with mean $\bar{M}_z$

and standard deviation $\bar{S}_z$, removed all samples that fell outside $[0, 1]$, applied the probit transformation to the truncated samples and used the mean and standard deviation of these truncated samples to parameterize the prior distribution. For $\theta = s_z$ we followed the same procedure to determine the prior mean but used 1.1 times the standard deviation of the truncated samples for z to compensate for the reduction in the variance of the samples introduced by the truncation. The values of all prior means and standard deviations were rounded to two decimals. The values of the hyperparameters $r_{C,\theta}, r_{P,\theta}, \mu_{\sigma\,\theta}, \sigma_{\sigma\,\theta}, \mu_{\mu\,\theta}, \sigma_{\mu\,\theta}$ are given in Table 3.

Table 3

Hyperparameter values for the hierarchical DDM parameters, transformed to the real line.

Parameter	$r_{C,\theta}$	$r_{P,\theta}$	$\mu_{\sigma\,\theta}$	$\sigma_{\sigma\,\theta}$	$\mu_{\mu\,\theta}$	$\sigma_{\mu\,\theta}$
$\tilde{a}$	1	0.5	0.36	0.15	0.16	1
v	1	1	1.27	0.61	2.23	1
$\tilde{t}_0$	1	0.5	0.23	0.13	-0.87	1
$\tilde{z}$	1	0.25	0.59	0.05	0	1
$\tilde{s}_v$	–	0.5	0.48	1.06	0.17	1
$\tilde{s}_z$	–	0.25	0.65	1.15	-0.52	1

Posterior sampling

We implemented our hierarchical DDMs through a modification of the DMC R software (Heathcote et al., 2019). For each model we ran the automatic convergence algorithm implemented within the DMC software. We obtained 500 posterior samples from $3k$ chains (where k is the number of participant-level free parameters in the model) to assess convergence after the automatic convergence algorithm, followed by 25,000 total samples across all chains to estimate the marginal likelihood. The Gelman-Rubin statistic $\hat{R}$ (Gelman & Rubin, 1992) was smaller than 1.1 for all chains, which indicates acceptable

convergence.

Estimation of marginal likelihoods

We used the bridge sampling estimator (Bennett, 1976; Meng & Wong, 1996) implemented in DMC (Gronau, Heathcote, et al., 2019) to compute the marginal likelihood of each of our sixteen candidate models for a given data set. Our estimates of the marginal likelihood were based on 25,000 posterior samples. We set the maximum number of iterations for the iterative scheme to 500 and used a multivariate normal proposal distribution that matched the first three moments of the posterior samples.

Results

For our reanalysis of Dutilh et al.'s (2019) data we computed the posterior probability for each of the 16 candidate models and inclusion Bayes factor for each of the four core DDM parameters for each data set. For the computation of the posterior probabilities we assigned a uniform prior distribution to the models, that is, each model had prior probability $p(H_k) = \frac{1}{16}$.

As a first step, we verified the numerical stability of the estimated log-marginal likelihoods on which the computation of the posterior model probabilities and the inclusion Bayes factors are based. The variability of the bridge-sampling estimator decreases as the number of posterior samples increases. To assess whether the number of posterior samples used in our analyses yielded a sufficiently stable bridge-sampling estimator, we compared the estimated log-marginal likelihood at different numbers of posterior samples. Due to the considerable computational costs, we limited this analysis to a single data set.

Figure 2 shows the change in the estimated log-marginal likelihood as the number of posterior samples increases for data set 8 from Dutilh et al. (2019). Numerical instabilities of a given size have a larger impact on the inclusion Bayes factor when the estimated log-marginal likelihood is small compared to when it is large. Therefore, the figure shows the percentage change in estimated log-marginal likelihood relative to the estimated

log-marginal likelihood at the largest posterior samples size. Specifically, the figure shows the deviation of the log-marginal likelihood estimated at 75,000 and 50,000 samples from the estimate at 25,000 samples, divided by the value at 75,000 samples. Each line shows the percentage change for one of the 16 models. As can be seen, for fifteen of the models the log-marginal likelihood did not change by more than 0.1% of its initial value. The largest change was 0.7%, which was observed for the model that allowed all core DDM parameters to differ between conditions except for a while there was overwhelming evidence for a condition effect on a in this data set (set 8, see below). Hence, this larger change in the log-marginal likelihood was most likely due to the maximal misspecification. Taken together, these results indicate that 25,000 posterior samples are sufficient to obtain stable estimates of the log-marginal likelihood.

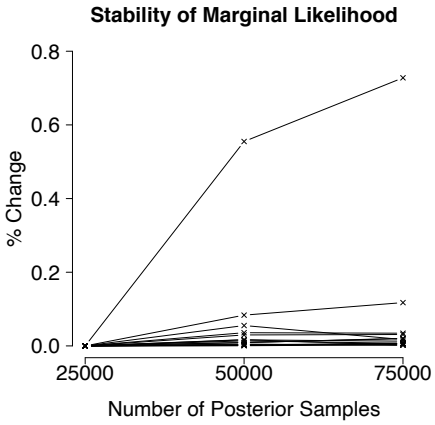

Figure 2

Stability of the bridge sampling estimate of the marginal likelihood. Lines show the percentage change in the estimated marginal likelihood of the 16 candidate models for data from Dutilh et al. (2019) (data set 8) at different numbers of posterior samples.

Figure 3 shows the posterior probability for each of the 16 candidate models (bars) for each of Dutilh et al.'s (2019) 14 data sets (the exact numerical values are presented in Table A1). Models are ordered by decreasing posterior probability. The vertical string below each bar indicates which DDM parameters were free to vary between conditions in

the corresponding model. A '+' indicates that the corresponding parameter was free to vary between conditions, a space indicates that the parameter was fixed across conditions. The blue bar in each panel highlights the model that corresponds Dutilh et al.'s assumed selective influence manipulation for the data set, which we call the target set. The outline bar highlights the model that corresponds to the model predicted by the results of Voss et al. (2004) and Rae et al. (2014), which we refer to as the refined target set.

As can be seen, the target model never had the highest posterior probability, whereas the refined target model had the highest posterior probability of 9/14 cases. The five exceptions all involved an additional t_0 term, two an extra a term and one an extra z term. These results suggest that the misspecification of the DDM may have had a wider influence than suggested by Smith and Lilburn's (2020) results, especially in the case of data set 10 where there was no manipulation of any kind, yet both t_0 and z effects were supported. A final possibility, that we investigate further in the simulation study, is that our methods are not capable of reliably identifying the data-generating model.

Figure 4 shows the log-inclusion Bayes factors for the four core DDM parameters for each of Dutilh et al.'s 14 data sets (the exact numerical values are presented in Table A2). Blue bars indicate model parameters that were the target of the selective influence manipulations. Importantly, for all data sets the inclusion Bayes factors indicate evidence for all model parameters that were targeted by the selective influence manipulations. However, support for z in data data sets 11 and 14 is considerably weaker than for other parameters ($BF_{\text{inc } z} = 10.22$ and $BF_{\text{inc } z} = 1.08$, respectively) and in general it is always relatively weak. Manipulations of response caution and the rate of information processing, on the other hand, resulted in strong evidence for a difference in the a and v parameter,

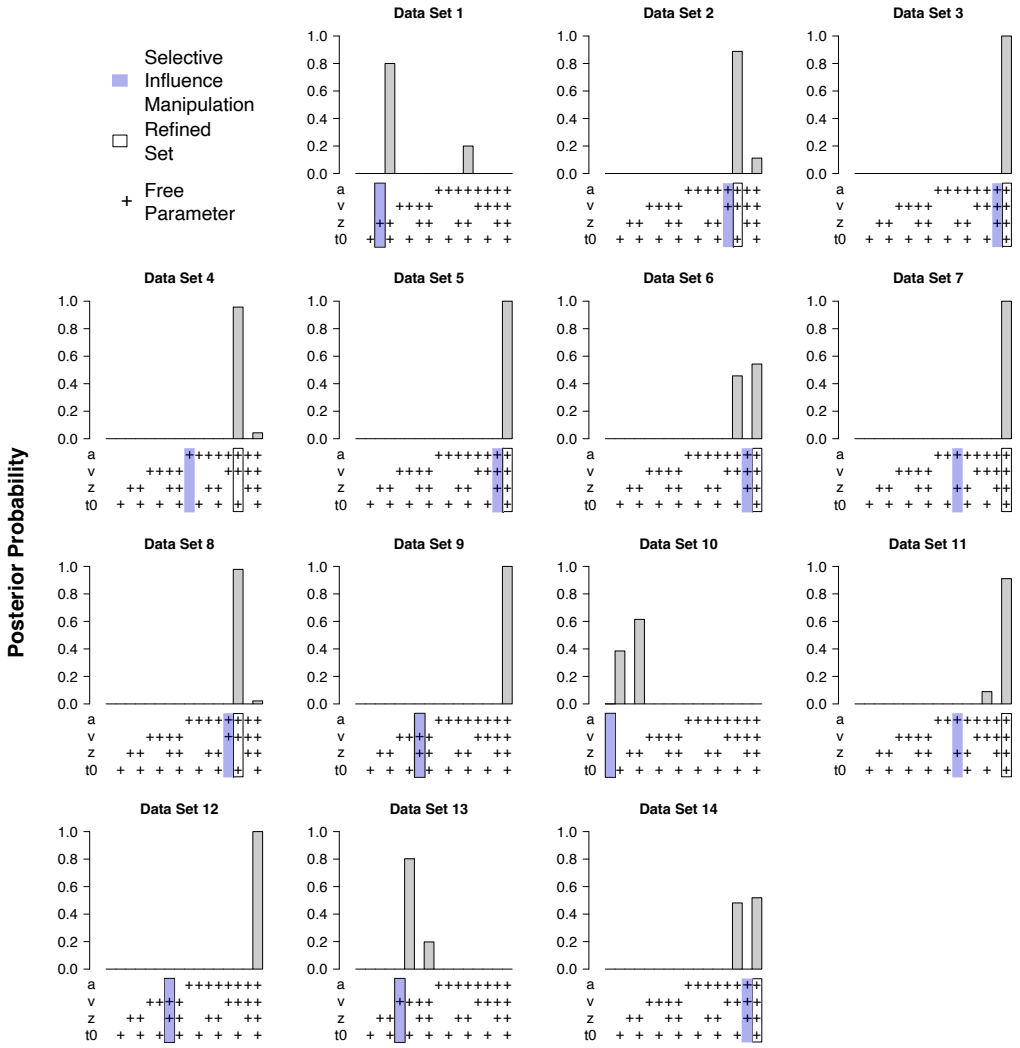

Figure 3

Estimated posterior model probabilities for 14 data sets from Dutilh et al. (2019). Each bar shows the log-marginal likelihood for one of the 16 candidate models. The symbol '+' below each bar indicates whether the model allowed the corresponding core DDM parameter to vary between experimental conditions. Blue bars indicate the model that corresponds to the selective influence manipulation. Outlines indicate models that were included in the refined set.

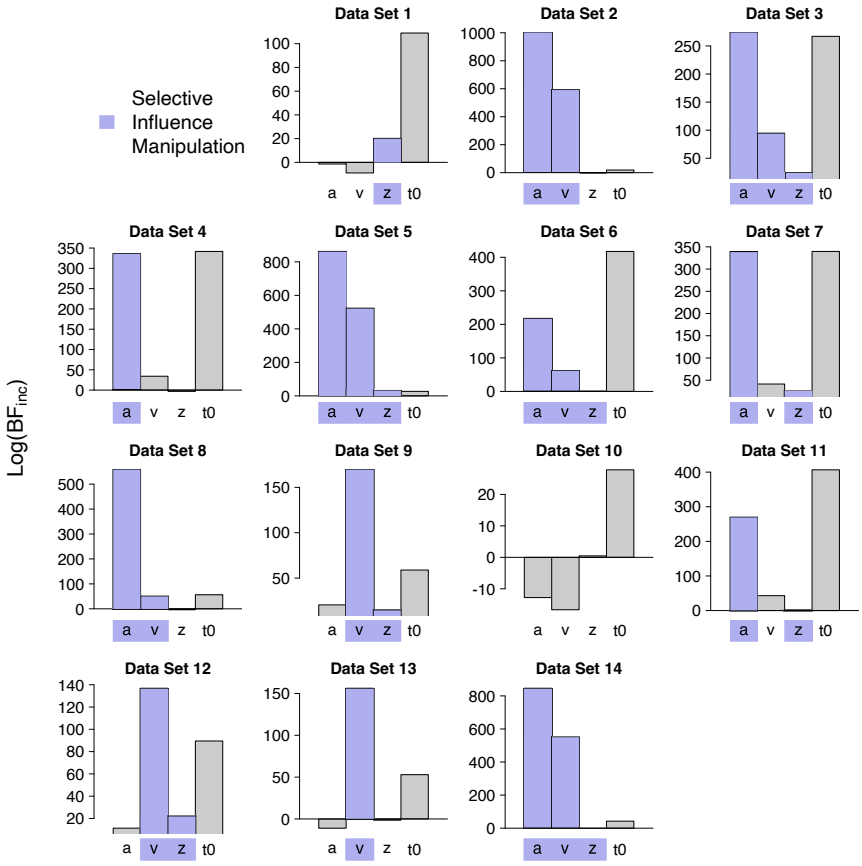

Figure 4

Log-inclusion Bayes factors for 14 data sets from Dutilh et al. (2019). Blue bars indicate parameters that were target of the selective influence manipulation.

respectively. In all data sets except data sets 2 and 5, the inclusion Bayes factors also indicated strong evidence for an effect on t_0 while this parameter was not targeted by any of the experimental manipulations. The additional effects in the refined set (i.e., t_0 effects in data sets 2-8, 11 and 14 and v effects in data sets 4, 7 and 11) were supported in every case.

As pointed out earlier, the random dot task may cause violations of the DDM's selective influence assumptions, which can result in biased parameter estimates. Therefore, we do not consider the application of BMA to the DDM parameter estimates for Dutilh et al.'s (2019) data, but return to this issue in the simulation study that we report next.

Simulations

In the second step of our analysis we conducted a simulation study to test whether our Bayesian model averaging approach could correctly identify the putative selective-influence manipulations. To this end we generated sixteen data sets from the mixed-effect DDM, one for each possible combination of cognitive processes that might be manipulated. As in our real data analysis, we implemented sixteen different hierarchical DDMs (listed in Table 1), one for each possible combination of targeted cognitive processes. We subsequently fit these sixteen models to each of the fourteen data sets, estimated the marginal likelihood of each model, and computed the posterior inclusion probability for each of the four DDM parameters.

Methods

Simulated Data

The simulated data matched the structure of the data in Dutilh et al.'s study. Each data set consisted of 20 simulated participants performing 200 trials for each of two experimental conditions. The generating population-level parameters for the intercepts ($\mu_a = 1.37$, $\mu_v = 1.62$, $\mu_z = 0.5$, $\mu_{t_0} = 0.27$, $\mu_{s_v} = 0.24$, $\mu_{s_z} = 0.09$) and residual standard deviations on the transformed (i.e., unbounded) scale ($\sigma_a = .15$, $\sigma_v = 0.61$, $\sigma_z = .05$,

$\sigma_{t_0} = 0.13$, $\sigma_{s_z} = 1.15$, $\sigma_{s_v} = 1.06$) were based on the estimated population-level values for Dutilh et al.'s data. We also used the standard deviations of the posterior estimates (i.e., posterior means) of the participant effects α_θ (again on the unbounded scale) to generate participant random effects ($a = 0.16$, $v = 0.31$, $z = 0.08$, $s_z = 0.08$, $s_v = 0.15$, $t_0 = 0.18$). For the core DDM parameters a, v, z, t_0 the effect sizes were based on the average effect size estimated from Dutilh et al.'s data in manipulations where the parameter was assumed to vary between conditions, where our assumptions differed from Dutilh et al.'s based on the following considerations.

Voss et al. (2004) found that speed vs. accuracy emphasis can affect t_0, and Rae et al. (2014) found that it can affect both t_0 and v in perceptual, lexical, and recognition memory choices. Further, Smith and Lilburn (2020) suggested that some perceptual properties of the random dot task used in Dutilh et al.'s (2019) study do not comply with the DDM's process assumptions. The DDM assumes that stimulus encoding is completed before evidence accumulation begins. This assumption is appropriate for experimental tasks where the encoding time of the stimulus is short compared to the time period over which information is integrated to reach a decision. However, psychophysical studies suggest that the encoding time in the random dot task is in the range of 400ms, which is relatively long compared to the duration of the decision process, and that evidence accumulation begins before encoding is complete, so the DDM is driven by a drift rate that changes during the decision. Smith and Lilburn suggest further that this model misspecification might lead to violations of the selective influence assumptions for non-decision time, non-decision time variability, and drift rate. As discussed further below, this was particularly borne out by Dutilh et al.'s (2019) results, which in most cases identified a non-decision time effect where none was present. We note, however, that Dutilh et al.'s expectation that the DDM is appropriate is very well justified on the basis of past practice, as it has been used extensively for modeling choices in the random dot-motion task (e.g., Boehm et al., 2014; Mulder et al., 2007; Palmer et al., 2005; Ratcliff & McKoon, 2008; van Vugt et al., 2012).

Based on these findings we assumed that speed vs. accuracy emphasis instructions affected the v and t_0 parameters as well as the a parameter when we constructed our simulation. That is, the effect size we used for (1) $v = 0.98$ was based on the average of the estimated effect sizes under manipulations of coherence and/or emphasis, (2) $a = 1.62$ and $t_0 = 0.72$ were both based on the average estimated effect sizes for the emphasis manipulations, and (3) $z = 0.82$ was based on the average estimated effect sizes for the frequency manipulations (as originally assumed by Dutilh et al., 2019). We note, however, that Smith and Lilburn's (2020) findings indicate that it is possible that there are even more widespread effects of the misspecification.

Model specification

We used the same model specification as in the real data analysis. That is, we specified our mixed-effect DDMs at two levels. The fixed condition and random participant effects were modeled as before, using the sum-to-zero constraint for fixed effects suggested by Rouder et al. (2012). We assigned the vectors of standardized condition effects multivariate Cauchy distributions. Moreover, we assigned informative normal priors to the intercept terms on the transformed scale, with the values of the hyperparameters given in Table 3.

Posterior sampling

We performed posterior sampling in the same way as in our real data analysis, using the DMC R software (Heathcote et al., 2019). For each model we ran the automatic convergence algorithm implemented within the DMC software. We obtained 500 posterior samples from $3k$ chains (where k is the number of participant-level free parameters in the model) to assess convergence after the automatic convergence algorithm, followed by 25,000 total samples across all chains to estimate the marginal likelihood. The Gelman-Rubin statistic $\hat{R}$ (Gelman & Rubin, 1992) was smaller than 1.1 for all chains, which indicates acceptable convergence.

Estimation of marginal likelihoods

We computed the marginal likelihoods in the same way as in our real data analysis, using the bridge sampling estimator implemented in DMC (Gronau, Heathcote, et al., 2019). Our estimates of the marginal likelihood were based on 25,000 posterior samples. We set the maximum number of iterations for the iterative scheme to 500 and used a multivariate normal proposal distribution that matched the first three moments of the posterior samples.

Results

Figure 5 shows the posterior probabilities for the 16 candidate models for each of the 16 simulated data sets (the exact numerical values are presented in Table A3). As can be seen, the estimated posterior probability was highest for the generating model in all data sets except data set G. In data set G the model with the highest posterior probability did not include an effect on z, but the generating model did have the second-highest posterior probability. These results make it clear that the differences between the models in the refined set and those selected by our methods are likely due to misspecification of the DDM for Dutilh et al.'s (2019) data. In particular, there was no problem determining when data were generated without a t_0 effect, and no problems with additional a or z effects being identified as was the case for Dutilh et al.'s data over and above those in the refined set.

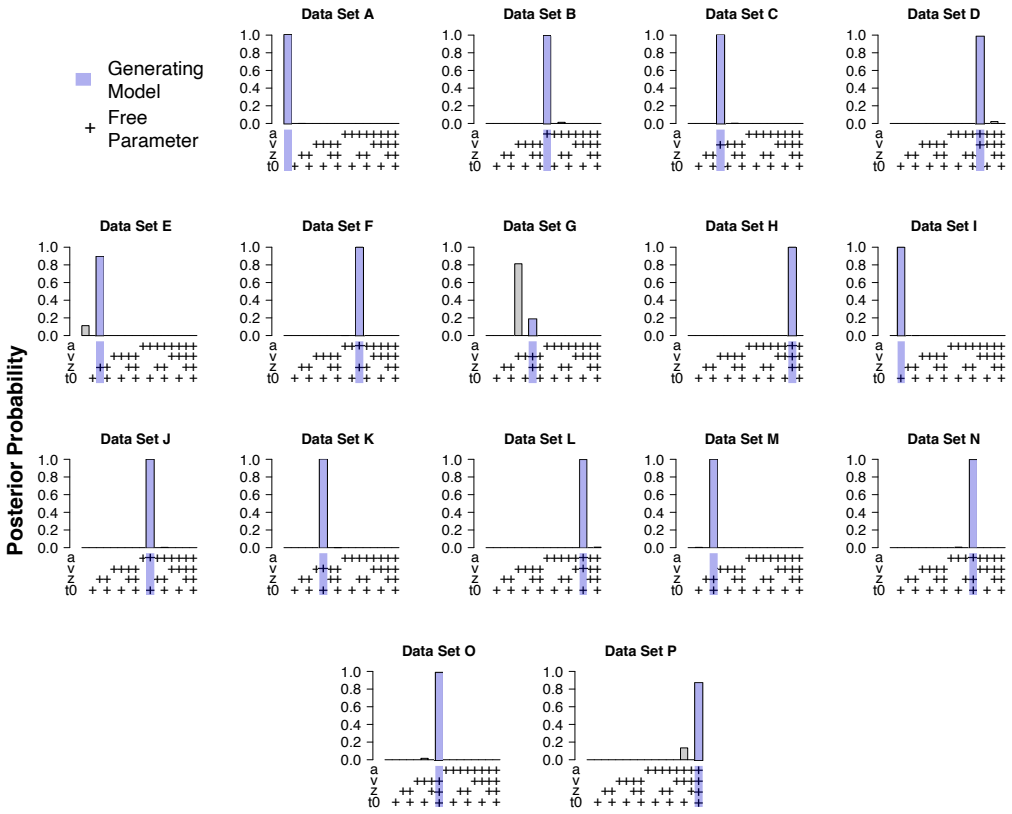

Figure 5

Estimated posterior probabilities for 16 simulated data sets in which different sets of core DDM parameters differed between conditions. Each bar shows the log-marginal likelihood for one of the 16 candidate models. The symbol + below each bar indicates whether the model allowed the corresponding core DDM parameter to vary between experimental conditions. Blue bars indicate the data generating model.

Figure 6 shows the log-inclusion Bayes factors for the four core DDM parameters for the 16 simulated data sets (the exact numerical values are presented in Table A4). They support the inclusion of all of the parameters that differed between conditions in the generating model. However, in the data sets where the z parameter was manipulated (data sets E - H and M - P), the inclusion Bayes factors only provide weak support. The evidence against the inclusion of parameters that were not manipulated is generally much weaker than the evidence for the inclusion of parameters that were manipulated, which is a general property of Bayes factors where the point of test falls inside the prior distribution (see Bahadur & Bickel, 2009; Jeffreys, 1939; Johnson, 2010). However, with the exception of z, the evidence was usually still overwhelming (i.e., < -20), which can be best seen with data set A because of the smaller range of values displayed.

Finally, we illustrate the utility of BMA parameter estimates. Figure 7 shows effect estimates (β) for the the four main DDM parameters produced by the 16 models fit to data set O, where true effects were present for all but a. Gray bars show the individual model effect estimates, the shaded bar shows the model-averaged estimate, and the orange bar shows the generating value (for β_a the latter two are zero). As can be seen, parameter estimates varied considerably between the individual models. For instance, models that allowed non-decision time but not boundary separation to vary between conditions (indicated by the leftmost brace) produced non-decision time effect estimates close to the true effect. In contrast, models that allowed both non-decision time and boundary separation to vary between conditions (indicated by the rightmost brace), severely underestimated the non-decision time condition effect. Hence, whereas model-averaged parameter estimates (shaded bars) are close to the true value for all four parameters, if researchers base parameter estimation on a single model, selecting the wrong model can considerably bias parameter estimates.

In most of our simulations, model uncertainty was relatively low, which is reflected by the generating model nearly always having the highest posterior probability. Therefore,

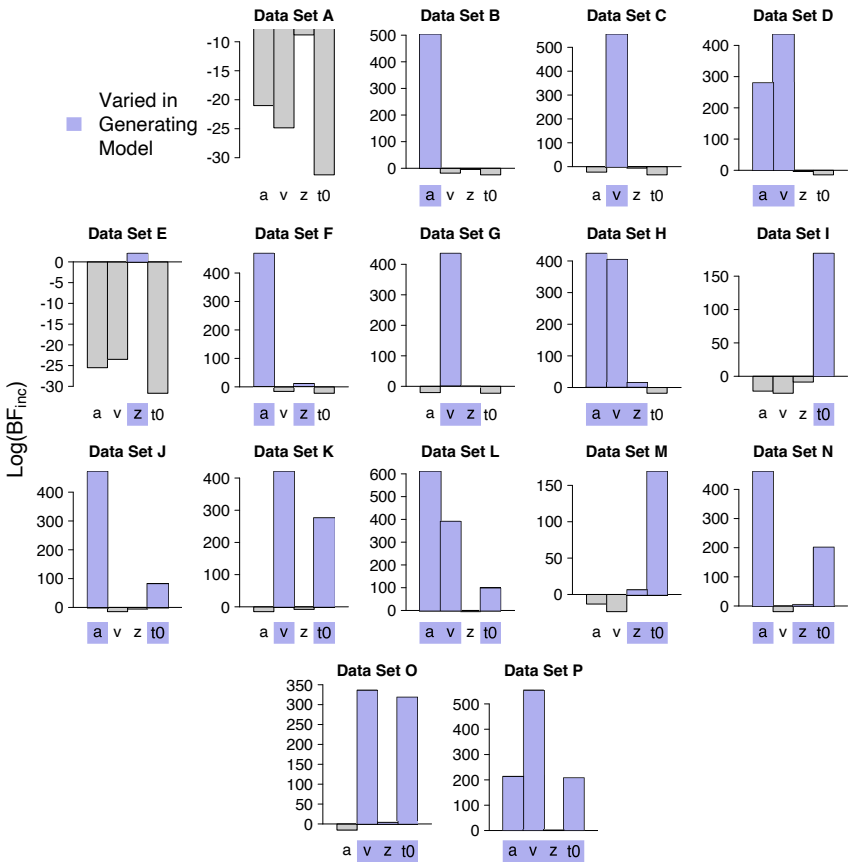

Figure 6

Log-inclusion Bayes factors for 16 simulated data sets. Blue bars indicate parameters that were varied between conditions in the data generating model.

the model-averaged parameter estimates were similar to those obtained from the generating model. However, the results for data set G illustrate the dangers of basing parameter estimates on a single model that has the highest posterior probability but is not the generating model. The estimated effect sizes for this data set for the 16 models are shown in Figure 8. The generating model (indicated by the blue bar at the bottom) included non-zero effects for v and z but the model with the highest posterior probability (indicated by the green bar at the bottom) only included an effect in v. A researcher who bases parameter estimates only on the model with the highest posterior probability would conclude that the effect on the z parameter is zero. In contrast, the model-averaged

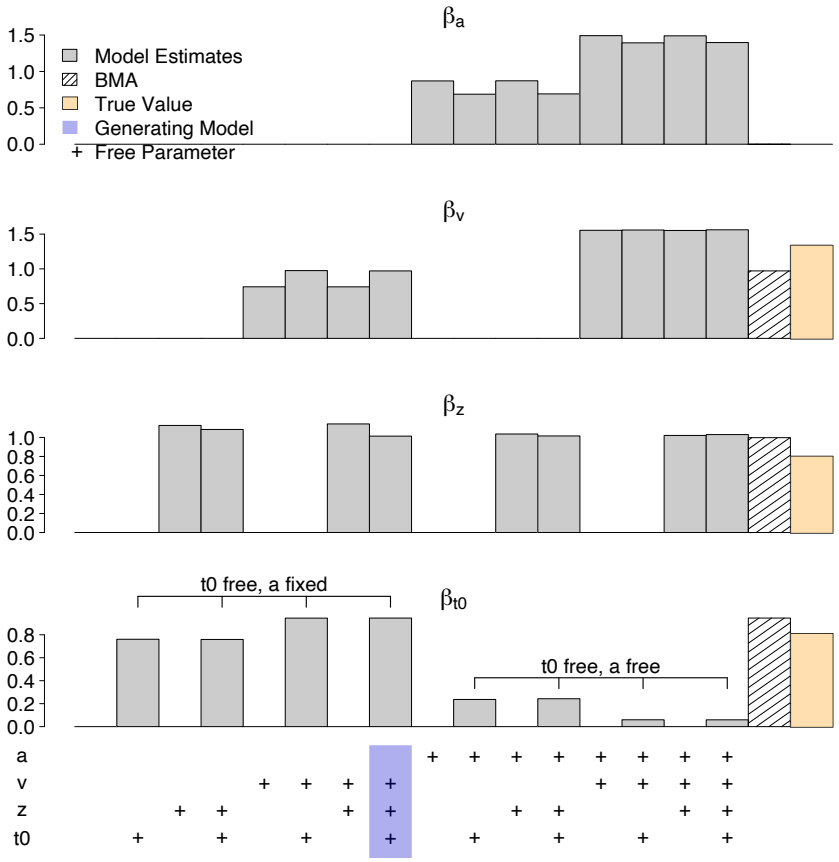

Figure 7

Estimates of the condition effects (β_θ) on the core DDM parameters for simulated data set O. The blue bar at the bottom indicates the generating model. The BMA estimate for β_a is numerically indistinguishable from 0 and is therefore not visible.

parameter estimate combines the estimates from all models, and, in particular, includes the non-zero estimate for β_z obtained from the generating model (see shaded bar at the right). Hence, the non-zero model-averaged estimate is closer to the generating value.

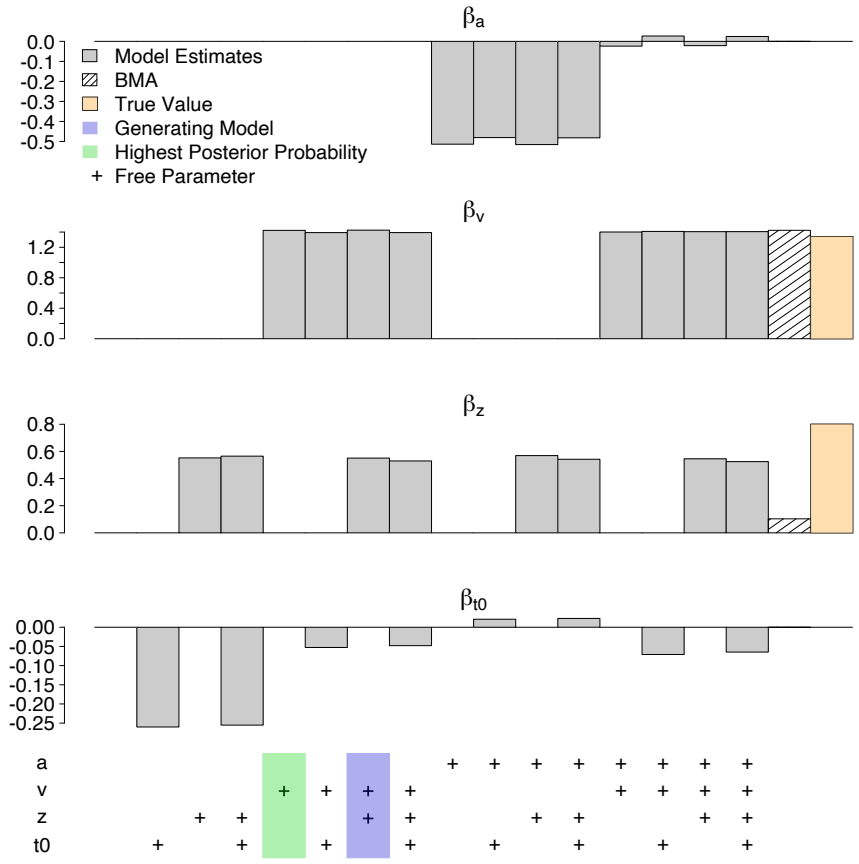

Figure 8

Estimates of the condition effects (β_θ) on the core DDM parameters for simulated data set G. The blue bar at the bottom indicates the generating model, the green bar indicates the model with the highest posterior probability. The BMA estimate for β_a is numerically indistinguishable from 0 and is therefore not visible.

Discussion

In the present study we illustrated how Bayesian hierarchical mixed modeling and Bayesian model averaging can be combined to obtain a coherent approach for estimation and inference for cognitive models. Many popular cognitive models are non-linear with highly correlated parameters. As a consequence, fitting cognitive models often requires large amounts of data per participant and the "sloppiness" of these models introduces a high level of model uncertainty. Our combined approach of hierarchical mixed modeling and Bayesian model averaging addresses both of these problems in a coherent manner. By treating individual differences in DDM parameters as random effects and differences between experimental conditions as fixed effects, all available data can be accommodated in a single hierarchical model. This increases the efficiency of our models compared to existing hierarchical DDM implementations that ignore within-person correlations of DDM parameters. Moreover, the use of an effect-size parameterization and default priors in our models simplifies the interpretation of differences in DDM parameters between experimental conditions and supports the integration of results across studies. Finally, basing inference on inclusion Bayes factors provides a coherent approach to inference under model uncertainty while penalizing model complexity in a principled way (Myung & Pitt, 1997). Inclusion Bayes factors weigh all possible fixed-effects configurations by their a posterior plausibility, and thus avoid giving too much weight to individual models that are subject to sampling variation.

As an illustrative example, we re-analyzed the data from Dutilh et al.'s (2019) blinded collaborative study, which sought to determine how well researchers could identify experimental manipulations that putatively selectively influenced only one DDM parameter. In line with the results reported by the collaborators in Dutilh et al.'s study, our inclusion Bayes factors indicated strong evidence for an effect on boundary separation and drift rate when these were the target of the selective influence manipulations. For experimental manipulations that were aimed at starting point the inclusion Bayes factors

indicated considerably weaker but generally positive evidence for a starting point effect. Finally, in 12 of the 14 experiments the inclusion Bayes factors also indicated moderate to strong evidence for the inclusion of a non-decision time effect while this parameter was not targeted by any of the experimental manipulations. In 9 of these cases these results are likely due to known effects of a speed vs. accuracy emphasis manipulation on non-decision time (Rae et al., 2014; Voss et al., 2004), as were three cases where drift rates were also affected by this manipulation. The remaining cases of non-decision effects, and a few cases of apparent non-selective influence related to boundary separation and starting point effects, may be due to the dot-motion stimuli used in Dutilh et al.'s choice task. Although this task is widely modeled by the DDM, Smith and Lilburn (2020) presented convincing evidence that the DDM is misspecified for such stimuli.

An alternative possibility is that our method cannot reliably recover the true model in Dutilh et al.'s (2019) design, perhaps because insufficient trials were performed by each participant. To check this possibility we conducted a simulation study using the same design. We generated data from the DDM for all possible combinations of manipulations of the core model parameters. The inclusion Bayes factors indicated strong evidence for an effect on boundary separation, drift rate and non-decision time when these varied between conditions in the generating model. Similar to the empirical data, the evidence was considerably weaker for a starting point effect but generally pointed towards the presence of an effect. These results show that our method is able to accurately detect which core DDM parameters are affected by experimental manipulations. The weaker evidence for manipulations of starting point might be due to the probit transformation we applied to the starting point, which might have shifted some of the prior mass of the heavy-tailed Cauchy distribution away from 0, and thus given undue a priori plausibility to large effects.

An interesting result was that log-Bayes factors against an effect on a DDM parameter were generally smaller in absolute value than log-Bayes factors for the presence of an effect. The reason for this phenomenon is that a model in which the DDM parameter

is fixed to be equal in both experimental conditions, the null model, is nested under the corresponding model in which the DDM parameter may vary between conditions, the alternative model. Specifically, the two models only differ in the prior distribution for the effect size for the DDM parameter in question. In the null model the prior is a point-mass at 0 while in the alternative model the prior is a Cauchy distribution centered at 0. This means that data that are generated by the null model are also plausible under the alternative model but not vice versa. As a consequence, the Bayes factor against an effect on the DDM parameter grows more slowly as the sample size increases than the Bayes factor for an effect (Bahadur & Bickel, 2009; Jeffreys, 1939; Johnson, 2010).

The mixed effects modeling approach taken in the present work can be further generalized. The fixed and random effects structures in our hierarchical DDMs were based on Dutilh et al.'s (2019) experimental setup with two factor levels. However, the Bayesian mixed modeling framework on which our hierarchical DDMs were based can accommodate more general multi-way ANOVA designs with more than two factor levels (Rouder et al., 2012). Moreover, the mixed modeling framework can be extended to include continuous covariates for ANCOVA and regression designs (Liang et al., 2008; Rouder & Morey, 2012). Both of these extensions are easily implemented as additive terms in Equation (17), although at present our software allows only a single factor with multiple levels.

The implementation of mixed effects modeling we propose here relies on default priors for the mixed effects but requires the specification of informed priors for the intercept terms so as to retain the cognitive process interpretation of the DDM parameters. The informed priors used in the present work were based on a large survey of published parameter estimates for the DDM. However, such surveys might not be readily available for other cognitive models. In that case researchers will need to either conduct such a survey themselves, or might even need to obtain a sufficiently large set of parameter estimates first. At least for evidence-accumulation models, there are large numbers of published data sets with sufficient numbers of trials per participant available so that reliable parameter

estimates can be obtained by fitting the target model to each participant's data individually. Informed priors can then be obtained as described in the present work, by fitting a suitable continuous distribution to the first and second moments of the empirical distribution of parameter estimates. The resulting priors can subsequently be updated as more parameter estimates from further studies become available.

A practical limitation of our method is its high computational costs. In order to compute inclusion Bayes factors, we had to fit all sixteen possible configurations of fixed effects structures to the experimental data and compute the marginal likelihood for each model and data set. Fitting the 16 models to Dutilh et al.'s (2019) data, for instance, took between 19 and 27 hours per model. A first way to decrease the associated computational costs is through the selection of more suitable priors for the mean DDM parameters. Our current implementation specifies an informative prior distribution separately for each DDM parameter and thus ignores the covariance structure among the parameters. As a consequence, the posterior sampling algorithm might spend considerable time exploring parameter combinations with a low likelihood. Including the covariance structure in the prior distribution can increase sampling efficiency and the quality of parameter estimates (Gunawan et al., 2020).

A second way to decrease computational costs is through the choice of a more efficient sampling algorithm. Our current implementation first generates a large number of posterior samples using differential evolution MCMC sampling (ter Braak, 2006; Turner et al., 2013), and subsequently uses bridge sampling to compute the marginal likelihood (Bennett, 1976; Meng & Wong, 1996). Recently developed sampling algorithms for the Linear Ballistic Accumulator (Brown & Heathcote, 2008) model provide a more efficient, integrated approach for parameter estimation and model selection that is amenable to a high level of parallelization (Gunawan et al., 2020; Tran et al., 2020). Hence, adapting these algorithms to the DDM might help further reduce the computational costs for the computation of inclusion Bayes factors.

A potential alternative to BMA is the use of regularization methods such as Lasso (Gelman et al., 2013). Applying this approach to our mixed modeling framework, a single hierarchical model with a maximal mixed effects structure is fit to the data. The priors in the model are chosen in such a way that many model parameters (i.e., fixed and random effects) are shrunk to near-zero values. Kang et al. (2022), for instance, suggest this approach for their hierarchical DDM implementation in which a structural equation model (instead of a mixed effects structure) is used to relate large neural data sets (e.g., the BOLD response of several thousand voxels in fMRI measurements) to DDM parameters. If sufficient data are available for all participants and trials, the regularization approach yields similar results to BMA in the sense that the model-averaged parameter estimates for parameters whose value is shrunk to a near-zero value by the Lasso would also be dominated by a single model in which the corresponding parameter value is fixed to zero. While the regularization approach is computationally more efficient than BMA, it suffers from the same shortcoming as single-model inference and estimation; whereas in model-averaging parameter estimates change continuously as more data become available, regularization methods lead to abrupt changes in the estimated parameter values.

In summary, in the present work we showed how a mixed effects modeling framework for the DDM can address several shortcomings of earlier hierarchical implementations of the DDM. Our implementation allows researchers to appropriately account for within-subjects correlations by means of additive random subjects effects. The use of default priors in combination with an effect size parameterization simplifies the comparison and integration of results across studies. Moreover, the implementation of our models in the DMC software package allows researchers to readily compute inclusion Bayes factors, and thus facilitates inference under model uncertainty. We believe that this approach represents an advance over previous methods of simultaneously analyzing choice and RT data (Heathcote et al., 2019; Van der Linden, 2007; Vandekerckhove et al., 2011; Wiecki et al., 2013) and advances the development of a cognitively grounded psychometrics

(Batchelder, 1998; Riefer et al., 2002).

Appendix A: Exact analysis results

The tables in this section show the exact numerical values for the posterior probabilities (Tables A1 and A3) and inclusion Bayes factors (Tables A2 and A4) for the analysis of Dutilh et al.'s (2019) and our simulated data. Note that the posterior probability for some models (last row in Table A1 and next-to-last row in Table A3) is shown as 1 for some data sets due to rounding. For data set 3 in Table A1, for instance, the posterior probability for the maximal model differs from 1 only in the tenth decimal, so its exact decimal value (to 10 digits) 0.9999999998 is rounded to 1.

Table A1

Estimated posterior model probabilities for 14 data sets from Dutilh et al. (2019).

Model				Data Set													
a	v	z	t_0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
				$9.53 \cdot 10^{-55}$	$4.78 \cdot 10^{-581}$	$3.60 \cdot 10^{-604}$	$5.97 \cdot 10^{-263}$	$3.16 \cdot 10^{-225}$	$3.37 \cdot 10^{-295}$	$3.19 \cdot 10^{-245}$	$1.20 \cdot 10^{-194}$	$1.86 \cdot 10^{-4}$	$6.76 \cdot 10^{-13}$	$1.16 \cdot 10^{-228}$	$1.19 \cdot 10^{-11}$	$5.35 \cdot 10^{-89}$	$1.47 \cdot 10^{-226}$
			+	$1.40 \cdot 10^{-9}$	$1.55 \cdot 10^{-492}$	$6.80 \cdot 10^{-233}$	$2.78 \cdot 10^{-154}$	$1.61 \cdot 10^{-451}$	$2.17 \cdot 10^{-158}$	$6.15 \cdot 10^{-175}$	$4.51 \cdot 10^{-364}$	$4.56 \cdot 10^{-80}$	$3.85 \cdot 10^{-1}$	$6.83 \cdot 10^{-125}$	$2.49 \cdot 10^{-79}$	$1.25 \cdot 10^{-73}$	$1.78 \cdot 10^{-425}$
		+		$4.65 \cdot 10^{-48}$	$4.62 \cdot 10^{-580}$	$6.07 \cdot 10^{-590}$	$2.38 \cdot 10^{-561}$	$1.29 \cdot 10^{-500}$	$1.30 \cdot 10^{-594}$	$1.50 \cdot 10^{-517}$	$2.23 \cdot 10^{-493}$	$4.57 \cdot 10^{-100}$	$1.61 \cdot 10^{-13}$	$3.28 \cdot 10^{-526}$	$5.33 \cdot 10^{-103}$	$8.57 \cdot 10^{-90}$	$7.24 \cdot 10^{-526}$
		+	+	$8.00 \cdot 10^{-1}$	$6.65 \cdot 10^{-493}$	$6.57 \cdot 10^{-222}$	$1.57 \cdot 10^{-152}$	$1.50 \cdot 10^{-428}$	$3.21 \cdot 10^{-156}$	$2.67 \cdot 10^{-156}$	$4.02 \cdot 10^{-367}$	$3.29 \cdot 10^{-75}$	$6.15 \cdot 10^{-1}$	$2.11 \cdot 10^{-119}$	$1.44 \cdot 10^{-70}$	$3.81 \cdot 10^{-74}$	$1.74 \cdot 10^{-425}$
	+			$6.58 \cdot 10^{-60}$	$2.31 \cdot 10^{-545}$	$4.31 \cdot 10^{-387}$	$1.64 \cdot 10^{-542}$	$1.06 \cdot 10^{-488}$	$1.27 \cdot 10^{-419}$	$3.08 \cdot 10^{-527}$	$1.49 \cdot 10^{-309}$	$6.94 \cdot 10^{-49}$	$5.43 \cdot 10^{-20}$	$1.21 \cdot 10^{-522}$	$6.77 \cdot 10^{-61}$	$1.25 \cdot 10^{-32}$	$2.78 \cdot 10^{-496}$
	+		+	$9.17 \cdot 10^{-13}$	$2.86 \cdot 10^{-434}$	$1.72 \cdot 10^{-131}$	$2.78 \cdot 10^{-148}$	$4.61 \cdot 10^{-397}$	$5.84 \cdot 10^{-98}$	$2.16 \cdot 10^{-166}$	$8.86 \cdot 10^{-243}$	$4.00 \cdot 10^{-15}$	$3.07 \cdot 10^{-8}$	$2.14 \cdot 10^{-123}$	$4.54 \cdot 10^{-15}$	$8.02 \cdot 10^{-1}$	$2.16 \cdot 10^{-368}$
	+	+		$4.13 \cdot 10^{-53}$	$1.42 \cdot 10^{-544}$	$7.27 \cdot 10^{-373}$	$4.56 \cdot 10^{-541}$	$3.38 \cdot 10^{-464}$	$1.28 \cdot 10^{-418}$	$5.37 \cdot 10^{-499}$	$5.86 \cdot 10^{-308}$	$1.75 \cdot 10^{-44}$	$7.64 \cdot 10^{-21}$	$1.72 \cdot 10^{-520}$	$1.85 \cdot 10^{-51}$	$3.08 \cdot 10^{-33}$	$9.18 \cdot 10^{-496}$
	+	+	+	$1.27 \cdot 10^{-4}$	$8.74 \cdot 10^{-436}$	$2.62 \cdot 10^{-119}$	$1.51 \cdot 10^{-146}$	$3.01 \cdot 10^{-375}$	$2.32 \cdot 10^{-95}$	$5.82 \cdot 10^{-148}$	$8.40 \cdot 10^{-246}$	$1.24 \cdot 10^{-9}$	$2.79 \cdot 10^{-8}$	$5.89 \cdot 10^{-118}$	$1.25 \cdot 10^{-5}$	$1.98 \cdot 10^{-1}$	$5.06 \cdot 10^{-369}$
+				$1.39 \cdot 10^{-57}$	$6.10 \cdot 10^{-272}$	$1.86 \cdot 10^{-164}$	$1.81 \cdot 10^{-164}$	$8.48 \cdot 10^{-261}$	$2.09 \cdot 10^{-207}$	$4.36 \cdot 10^{-181}$	$4.06 \cdot 10^{-47}$	$1.34 \cdot 10^{-105}$	$3.20 \cdot 10^{-21}$	$1.48 \cdot 10^{-196}$	$3.84 \cdot 10^{-110}$	$4.14 \cdot 10^{-91}$	$5.63 \cdot 10^{-265}$
+			+	$2.57 \cdot 10^{-10}$	$1.24 \cdot 10^{-257}$	$1.75 \cdot 10^{-51}$	$1.36 \cdot 10^{-15}$	$2.63 \cdot 10^{-240}$	$1.50 \cdot 10^{-27}$	$3.36 \cdot 10^{-30}$	$3.83 \cdot 10^{-23}$	$9.47 \cdot 10^{-80}$	$1.76 \cdot 10^{-6}$	$1.53 \cdot 10^{-20}$	$1.08 \cdot 10^{-69}$	$1.14 \cdot 10^{-68}$	$1.59 \cdot 10^{-239}$
+		+		$3.40 \cdot 10^{-50}$	$4.30 \cdot 10^{-273}$	$6.89 \cdot 10^{-155}$	$2.45 \cdot 10^{-165}$	$5.80 \cdot 10^{-245}$	$7.09 \cdot 10^{-209}$	$6.47 \cdot 10^{-168}$	$1.49 \cdot 10^{-48}$	$1.05 \cdot 10^{-100}$	$1.58 \cdot 10^{-21}$	$8.47 \cdot 10^{-198}$	$7.42 \cdot 10^{-101}$	$8.29 \cdot 10^{-92}$	$9.50 \cdot 10^{-266}$
+		+	+	$2.00 \cdot 10^{-1}$	$4.44 \cdot 10^{-257}$	$7.43 \cdot 10^{-42}$	$5.67 \cdot 10^{-17}$	$2.25 \cdot 10^{-226}$	$6.75 \cdot 10^{-28}$	$1.03 \cdot 10^{-18}$	$1.14 \cdot 10^{-24}$	$2.60 \cdot 10^{-74}$	$1.06 \cdot 10^{-6}$	$1.87 \cdot 10^{-19}$	$3.55 \cdot 10^{-60}$	$1.97 \cdot 10^{-69}$	$1.96 \cdot 10^{-240}$
+	+			$1.68 \cdot 10^{-61}$	$4.82 \cdot 10^{-9}$	$1.73 \cdot 10^{-127}$	$3.61 \cdot 10^{-149}$	$6.37 \cdot 10^{-29}$	$4.34 \cdot 10^{-182}$	$1.34 \cdot 10^{-160}$	$3.35 \cdot 10^{-25}$	$1.29 \cdot 10^{-32}$	$4.16 \cdot 10^{-27}$	$1.68 \cdot 10^{-177}$	$4.90 \cdot 10^{-49}$	$7.98 \cdot 10^{-24}$	$1.55 \cdot 10^{-19}$
+	+		+	$3.25 \cdot 10^{-14}$	$8.88 \cdot 10^{-1}$	$4.71 \cdot 10^{-11}$	$9.57 \cdot 10^{-1}$	$2.19 \cdot 10^{-14}$	$4.57 \cdot 10^{-1}$	$1.50 \cdot 10^{-11}$	$9.78 \cdot 10^{-1}$	$2.71 \cdot 10^{-7}$	$1.46 \cdot 10^{-12}$	$8.91 \cdot 10^{-2}$	$2.50 \cdot 10^{-10}$	$1.29 \cdot 10^{-5}$	$4.81 \cdot 10^{-1}$
+	+	+		$3.63 \cdot 10^{-54}$	$5.77 \cdot 10^{-10}$	$9.43 \cdot 10^{-117}$	$4.51 \cdot 10^{-150}$	$1.87 \cdot 10^{-12}$	$1.70 \cdot 10^{-183}$	$3.02 \cdot 10^{-148}$	$1.17 \cdot 10^{-26}$	$2.62 \cdot 10^{-26}$	$1.11 \cdot 10^{-27}$	$7.38 \cdot 10^{-178}$	$1.31 \cdot 10^{-39}$	$2.37 \cdot 10^{-24}$	$1.69 \cdot 10^{-19}$
+	+	+	+	$5.88 \cdot 10^{-6}$	$1.12 \cdot 10^{-1}$	$1 \cdot 10^{0}$	$4.28 \cdot 10^{-2}$	$1 \cdot 10^{0}$	$5.43 \cdot 10^{-1}$	$1 \cdot 10^{0}$	$2.17 \cdot 10^{-2}$	$1 \cdot 10^{0}$	$5.66 \cdot 10^{-13}$	$9.11 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$3.65 \cdot 10^{-6}$	$5.19 \cdot 10^{-1}$

Table A2

Log-inclusion Bayes factors for 14 data sets from Dutilh et al. (2019).

	Data Set													
Parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a	−1.39	998.24	273.04	335.74	862.37	217.9	339.02	557.35	20.51	−12.78	269.9	11.29	−11.01	846.37
v	−8.93	590.03	94.7	34.19	519.57	61.39	41.41	51.59	169.32	−16.65	43.05	136.89	156.29	549.74
z	20.22	−2.07	23.78	−3.11	31.45	0.17	24.92	−3.81	15.12	0.47	2.32	22.11	−1.4	0.08
t_0	108.98	19.04	267.16	341.68	27.01	417.56	339.68	56.32	58.91	27.81	406.68	89.53	52.92	42.57

Table A3

Estimated posterior model probabilities for 16 simulated data sets.

a	v	z	t_0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
				$1.00 \cdot 10^{-1}$	$6.27 \cdot 10^{-330}$	$5.08 \cdot 10^{-266}$	$6.94 \cdot 10^{-335}$	$1.12 \cdot 10^{-1}$	$6.47 \cdot 10^{-325}$	$1.43 \cdot 10^{-221}$	$4.47 \cdot 10^{-440}$	$3.89 \cdot 10^{-136}$	$6.33 \cdot 10^{-612}$	$5.90 \cdot 10^{-400}$	$1.30 \cdot 10^{-661}$	$1.41 \cdot 10^{-136}$	$4.95 \cdot 10^{-580}$	$7.71 \cdot 10^{-372}$	$3.03 \cdot 10^{-660}$
			+	$4.86 \cdot 10^{-15}$	$7.86 \cdot 10^{-224}$	$1.12 \cdot 10^{-272}$	$2.62 \cdot 10^{-264}$	$1.55 \cdot 10^{-15}$	$5.10 \cdot 10^{-212}$	$1.27 \cdot 10^{-223}$	$1.38 \cdot 10^{-343}$	$1.00 \cdot 10^{-1}$	$2.78 \cdot 10^{-237}$	$1.03 \cdot 10^{-200}$	$4.69 \cdot 10^{-367}$	$3.30 \cdot 10^{-3}$	$3.73 \cdot 10^{-213}$	$4.41 \cdot 10^{-164}$	$6.45 \cdot 10^{-315}$
		+		$1.48 \cdot 10^{-4}$	$8.23 \cdot 10^{-327}$	$1.88 \cdot 10^{-268}$	$7.06 \cdot 10^{-337}$	$8.88 \cdot 10^{-1}$	$5.67 \cdot 10^{-318}$	$2.07 \cdot 10^{-222}$	$2.53 \cdot 10^{-434}$	$2.10 \cdot 10^{-139}$	$1.58 \cdot 10^{-607}$	$1.38 \cdot 10^{-402}$	$2.59 \cdot 10^{-653}$	$1.67 \cdot 10^{-135}$	$6.99 \cdot 10^{-575}$	$1.09 \cdot 10^{-370}$	$1.50 \cdot 10^{-658}$
		+	+	$8.88 \cdot 10^{-19}$	$5.60 \cdot 10^{-225}$	$2.98 \cdot 10^{-275}$	$4.46 \cdot 10^{-266}$	$1.61 \cdot 10^{-14}$	$4.07 \cdot 10^{-207}$	$2.18 \cdot 10^{-224}$	$7.54 \cdot 10^{-337}$	$1.70 \cdot 10^{-4}$	$2.42 \cdot 10^{-239}$	$3.35 \cdot 10^{-204}$	$1.81 \cdot 10^{-366}$	$9.97 \cdot 10^{-1}$	$2.56 \cdot 10^{-211}$	$1.85 \cdot 10^{-162}$	$1.79 \cdot 10^{-314}$
	+			$1.62 \cdot 10^{-11}$	$3.36 \cdot 10^{-304}$	$9.98 \cdot 10^{-1}$	$2.17 \cdot 10^{-209}$	$6.81 \cdot 10^{-12}$	$9.27 \cdot 10^{-304}$	$8.13 \cdot 10^{-1}$	$5.72 \cdot 10^{-292}$	$2.36 \cdot 10^{-134}$	$1.82 \cdot 10^{-493}$	$1.52 \cdot 10^{-210}$	$6.75 \cdot 10^{-583}$	$3.23 \cdot 10^{-134}$	$1.36 \cdot 10^{-503}$	$3.78 \cdot 10^{-231}$	$2.35 \cdot 10^{-485}$
	+		+	$2.06 \cdot 10^{-25}$	$2.54 \cdot 10^{-218}$	$1.21 \cdot 10^{-15}$	$4.58 \cdot 10^{-122}$	$9.89 \cdot 10^{-26}$	$3.89 \cdot 10^{-209}$	$1.04 \cdot 10^{-10}$	$2.40 \cdot 10^{-191}$	$8.91 \cdot 10^{-12}$	$2.94 \cdot 10^{-205}$	$1.00 \cdot 10^{-1}$	$1.33 \cdot 10^{-265}$	$1.61 \cdot 10^{-13}$	$1.32 \cdot 10^{-202}$	$1.61 \cdot 10^{-2}$	$3.93 \cdot 10^{-95}$
	+	+		$1.77 \cdot 10^{-15}$	$2.09 \cdot 10^{-301}$	$2.01 \cdot 10^{-3}$	$1.84 \cdot 10^{-210}$	$5.58 \cdot 10^{-11}$	$1.44 \cdot 10^{-296}$	$1.87 \cdot 10^{-1}$	$1.93 \cdot 10^{-285}$	$1.38 \cdot 10^{-137}$	$2.45 \cdot 10^{-492}$	$6.52 \cdot 10^{-213}$	$6.58 \cdot 10^{-574}$	$3.35 \cdot 10^{-133}$	$3.91 \cdot 10^{-502}$	$3.51 \cdot 10^{-230}$	$8.02 \cdot 10^{-484}$
	+	+	+	$2.85 \cdot 10^{-29}$	$7.68 \cdot 10^{-220}$	$2.68 \cdot 10^{-18}$	$9.12 \cdot 10^{-124}$	$9.72 \cdot 10^{-25}$	$1.64 \cdot 10^{-204}$	$2.98 \cdot 10^{-11}$	$5.78 \cdot 10^{-185}$	$1.03 \cdot 10^{-15}$	$6.35 \cdot 10^{-207}$	$3.83 \cdot 10^{-4}$	$8.08 \cdot 10^{-266}$	$5.46 \cdot 10^{-11}$	$7.98 \cdot 10^{-201}$	$9.84 \cdot 10^{-1}$	$9.74 \cdot 10^{-95}$
+				$7.51 \cdot 10^{-10}$	$9.87 \cdot 10^{-1}$	$1.90 \cdot 10^{-240}$	$2.93 \cdot 10^{-189}$	$7.90 \cdot 10^{-13}$	$3.25 \cdot 10^{-5}$	$3.29 \cdot 10^{-190}$	$3.40 \cdot 10^{-183}$	$9.66 \cdot 10^{-81}$	$7.22 \cdot 10^{-36}$	$4.58 \cdot 10^{-320}$	$1.67 \cdot 10^{-232}$	$2.33 \cdot 10^{-75}$	$3.66 \cdot 10^{-90}$	$5.43 \cdot 10^{-293}$	$2.50 \cdot 10^{-383}$
+			+	$2.69 \cdot 10^{-19}$	$1.86 \cdot 10^{-11}$	$3.88 \cdot 10^{-254}$	$1.91 \cdot 10^{-189}$	$3.60 \cdot 10^{-23}$	$1.10 \cdot 10^{-14}$	$7.03 \cdot 10^{-198}$	$1.15 \cdot 10^{-186}$	$2.12 \cdot 10^{-10}$	$9.97 \cdot 10^{-1}$	$5.65 \cdot 10^{-183}$	$8.27 \cdot 10^{-171}$	$4.70 \cdot 10^{-9}$	$5.55 \cdot 10^{-3}$	$4.39 \cdot 10^{-149}$	$4.86 \cdot 10^{-242}$
+		+		$8.01 \cdot 10^{-14}$	$1.28 \cdot 10^{-2}$	$1.02 \cdot 10^{-242}$	$4.38 \cdot 10^{-191}$	$7.56 \cdot 10^{-12}$	$1.00 \cdot 10^{-1}$	$9.14 \cdot 10^{-191}$	$1.49 \cdot 10^{-176}$	$2.51 \cdot 10^{-84}$	$2.23 \cdot 10^{-38}$	$5.23 \cdot 10^{-323}$	$3.64 \cdot 10^{-234}$	$4.85 \cdot 10^{-74}$	$4.60 \cdot 10^{-88}$	$1.02 \cdot 10^{-291}$	$2.46 \cdot 10^{-383}$
+		+	+	$5.00 \cdot 10^{-23}$	$2.02 \cdot 10^{-13}$	$8.55 \cdot 10^{-257}$	$4.13 \cdot 10^{-191}$	$4.96 \cdot 10^{-22}$	$1.74 \cdot 10^{-10}$	$2.11 \cdot 10^{-198}$	$6.16 \cdot 10^{-180}$	$2.06 \cdot 10^{-14}$	$3.38 \cdot 10^{-3}$	$6.91 \cdot 10^{-187}$	$3.37 \cdot 10^{-172}$	$1.88 \cdot 10^{-6}$	$9.94 \cdot 10^{-1}$	$7.06 \cdot 10^{-147}$	$1.13 \cdot 10^{-241}$
+	+			$2.10 \cdot 10^{-20}$	$1.11 \cdot 10^{-8}$	$1.09 \cdot 10^{-10}$	$9.79 \cdot 10^{-1}$	$4.18 \cdot 10^{-22}$	$2.74 \cdot 10^{-12}$	$9.96 \cdot 10^{-10}$	$2.88 \cdot 10^{-7}$	$1.83 \cdot 10^{-90}$	$8.72 \cdot 10^{-42}$	$4.44 \cdot 10^{-121}$	$2.73 \cdot 10^{-43}$	$1.13 \cdot 10^{-84}$	$2.99 \cdot 10^{-98}$	$1.51 \cdot 10^{-140}$	$1.20 \cdot 10^{-92}$
+	+		+	$1.04 \cdot 10^{-29}$	$1.46 \cdot 10^{-19}$	$9.05 \cdot 10^{-24}$	$4.48 \cdot 10^{-7}$	$2.92 \cdot 10^{-32}$	$8.50 \cdot 10^{-22}$	$7.80 \cdot 10^{-20}$	$2.92 \cdot 10^{-15}$	$1.54 \cdot 10^{-20}$	$4.29 \cdot 10^{-7}$	$3.02 \cdot 10^{-7}$	$9.95 \cdot 10^{-1}$	$1.90 \cdot 10^{-18}$	$2.68 \cdot 10^{-11}$	$3.70 \cdot 10^{-9}$	$1.34 \cdot 10^{-1}$
+	+	+		$2.77 \cdot 10^{-24}$	$1.31 \cdot 10^{-10}$	$2.67 \cdot 10^{-13}$	$2.14 \cdot 10^{-2}$	$3.49 \cdot 10^{-21}$	$1.19 \cdot 10^{-7}$	$1.82 \cdot 10^{-10}$	$1.00 \cdot 10^{0}$	$3.10 \cdot 10^{-94}$	$2.43 \cdot 10^{-44}$	$3.77 \cdot 10^{-124}$	$7.43 \cdot 10^{-45}$	$4.19 \cdot 10^{-83}$	$2.75 \cdot 10^{-96}$	$2.14 \cdot 10^{-139}$	$1.08 \cdot 10^{-91}$
+	+	+	+	$2.19 \cdot 10^{-33}$	$3.33 \cdot 10^{-21}$	$1.47 \cdot 10^{-26}$	$5.87 \cdot 10^{-9}$	$3.64 \cdot 10^{-31}$	$1.25 \cdot 10^{-17}$	$1.71 \cdot 10^{-20}$	$1.30 \cdot 10^{-8}$	$1.20 \cdot 10^{-24}$	$2.35 \cdot 10^{-9}$	$8.04 \cdot 10^{-11}$	$5.49 \cdot 10^{-3}$	$5.10 \cdot 10^{-16}$	$3.56 \cdot 10^{-9}$	$2.09 \cdot 10^{-7}$	$8.66 \cdot 10^{-1}$

Table A4

Log-inclusion Bayes factors for 16 simulated data sets.

	Data Set															
Parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	−21.01	501	−22.94	279.37	−25.51	469.23	−20.56	424.22	−22.27	470.93	−15.01	609.43	−13.18	460.73	−15.36	216.13
v	−24.85	−18.3	551.97	433.59	−23.49	−15.95	436.05	404.86	−25.44	−14.66	419.64	391.59	−23.63	−19.45	336.52	554.44
z	−8.82	−4.34	−6.21	−3.82	2.07	10.33	−1.47	15.06	−8.68	−5.69	−7.87	−5.2	5.71	5.19	4.12	1.86
t_0	−32.96	−24.7	−34.34	−14.61	−31.67	−22.47	−22.73	−18.16	184.24	80.91	277.12	97.98	168.76	201.09	319.23	209.35

Appendix B: Implementation of the DDM in the DMC toolbox

Here, we provide a brief tutorial on how to implement the code associated with our proposed methodology. While the methodology may seem complex, the underlying DMC functions (code provided on the OSF: https://osf.io/v6u8e/) makes our proposed methodology extremely simple to implement, with the user only needing to parse the data into the correct format, and adjust a few lines of code at the beginning of the fitting script. However, while this section will cover how to run and implement the code, users unfamiliar with how to interpret the output of the code should see Heathcote et al.'s (2019) and Gronau et al.'s (2020) for existing in-depth tutorials on interpreting the DMC sampling output and the bridge sampling output, respectively.

The script "implementationFit.R" provides code that can be easily adapted to allow users to apply the proposed methodology to their data sets. Specifically, users only need to change at most 6 lines of code in the script: lines 3-8. Lines 3-6 are where users need to specify what parameters are allowed to differ between conditions in the model; line 3 for threshold (a), line 4 for drift rate (v), line 5 for starting point (z), and line 6 for non-decision time (t_0). In each case, setting the "index" variable (e.g., "thresIndex" for threshold on line 3) to 1 will constrain the parameter to have the same value in both conditions (i.e., no effect of the manipulation on this parameter), and setting it to 2 will allow the parameter to have different values in each condition (i.e., an effect of the manipulation on this parameter). Line 7 is where the user needs to specify the ".RData" file containing the data (e.g., "myData.RData"), which needs to already be parsed into the format described above. Finally, line 8 specifies the number of CPU cores to parallelize the fit over, which is currently set to 1 to try and ensure users do not accidentally overload their computers when initially testing out the code on computers with a small number of CPU cores.

Provided that the data is in the correct format, the correct data file is loaded in, and proper indexes (i.e., 1 or 2) are specified for all of the parameters, the code should

automatically perform the sampling and save all of the relevant sampling and diagnostic information. Specifically, the final saved ".RData" file from a fit will begin with "BF_FIT_", and then contain a 4 character string consisting of Ns and Vs, corresponding to whether each variable was (V) or was not (N) allowed to vary across conditions, with parameters in the same order as specified by the user at the start of the code. This file will contain two variables: "newHsamples", which contains all of the posterior samples needed for posterior estimation and inference methods (see Heathcote et al.'s (2019) for an in-depth tutorial on how to interpret this output), and "r", which contains the output for the bridgeSampling package used to calculate the marginal likelihood (see Gronau et al.'s (2020) for an in-depth tutorial on how to interpret this output).

R Script "implementationFit.R"

```
rm(list=ls())

# The following indexes are set to values to specify what
model we are fitting to the data, in terms of what
parameters are allowed to change over the conditions of the
experiment. Setting an index to 1 means that the parameter
stays constant over conditions, and setting the index to 2
means that the parameter is allowed to vary across the
conditions. In this case, we have all indexes set to 1,
making this the null model

thresIndex=1     #Set to 1 to constrain threshold over
conditions, or to 2 to let it vary
driftIndex=1     #Set to 1 to constrain drift rate over
conditions, or to 2 to let it vary
```

```
startIndex=1     #Set to 1 to constrain starting point over conditions, or to 2 to let it vary
nonDecIndex=1     #Set to 1 to constrain non-decision time over conditions, or to 2 to let it vary
dataFileName="myData.RData"
cores=1 # Set the number of cores; important if multi-core parallelization is possible!

# Make dmc functions available in the current R environment
source ("dmc/dmc_ES.R")


# Effect size DDM model
load_model ("DDM-ES","ddmES.R")



# Transform the previous indexes into strings that are used throughout the script, to create some automation. For example, the previous threshold index that we set to 1 is then used to make the "doesThresVary" variable (used for naming save files etc.) equal to "N" (for "not varying"), and the mapThres variable (used to create the dmc model and specify whether the parameter is constant/varying)
doesThresVary=c("N","V") [thresIndex]; mapThres=c("1","F") [thresIndex];
doesDriftVary=c("N","V") [driftIndex]; mapDrift=c("1","F") [driftIndex];
doesStartVary=c("N","V") [startIndex]; mapStart=c("1","F") [startIndex];
doesNonDecVary=c("N","V") [nonDecIndex];
mapNonDec=c("1","F") [nonDecIndex];
```

```
24
25 stringIndexes=paste(doesThresVary, doesDriftVary,
   doesStartVary, doesNonDecVary, sep="")
26
27
28 set.seed(666)
29
30 # Specify the priors for the random subject effects on all
   estimated parameters
31 subject.sd <- c(a=0.5,v=1,z=0.25,sz=0.25,sv=0.5,t0=0.5)
32
33 # Create the parameter mapping for DMC based on the previous
   indexes, in order to match to the internal specifications of
   DMC. In all cases, a parameter should have a value of "1" if
   it is constant across conditions, and a value of "F" if it
   differs across experimental conditions (specifically, factor
   "F", which is the standard naming convention in DMC)
34 p.map <- list(a=mapThres, v=mapDrift ,z=mapStart, d="1",
   sz="1", sv="1", t0=mapNonDec, st0="1")
35
36 # Unit effects for hyper-prior; essentially, this creates a
   list called effects, which stores which parameters will vary
   over conditions (needed for the internal specifications of
   DMC  for the random effects models)
37
38 effects = list()
39 if (doesThresVary=="V") effects[["a"]]=1
40 if (doesDriftVary=="V") effects[["v"]]=1
41 if (doesStartVary=="V") effects[["z"]]=1
```

```
42 if (doesNonDecVary=="V") effects[["t0"]]=1
43
44 sampFileName <- paste("FIT_",stringIndexes,".RData",sep="")
45
46 #### Model specification ----------------------------------
47
48 # Here we specify information about the data needed for the
   internal specifications of DMC, as well as the constants in
   the model (i.e., parameters that are fixed to specific
   values, are aren't estimated at all). S refers to the
   stimuli, F refers to the factors (i.e., conditions), and R
   refers to the responses
49
50 factors=list(S=c("s1","s2"),F=c("f1","f2")) # Specify the
   conditions (stimuli and factors)
51 responses=c("r1","r2") # Specify the responses
52 match.map=list(M=list(s1="r1",s2="r2")) # Specify which
   response is correct for each stimulus
53 const <- c(st0=log(1e-6),d=0) # Set constants in the model;
   note that you can't set st0 to exactly zero as undefined on
   log scale
54 type="rd"
55
56 # Set up the DMC model based on all of the previous
   specifications; specifically, the "model.dmc" function is
   used to bind the information together, which is needed for
   the internal specifications of DMC
57 model <- model.dmc(
58   p.map=p.map,
```

```
59    constants=const,
60    match.map=match.map,
61    factors=factors,
62    responses=responses,
63    type=type)
64
65  #### Make Gaussian data level prior including DMs, mu_indx,
    alpha_indx; specifically, this sets up the priors for the
    internal specifications of DMC
66  p.vector <- attr(model,"p.vector") # Doesnt have to have
    content
67  p.prior <- prior.p.dmc(
68    p1=p.vector,
69    p2=rep(1,length(p.vector)),
70    model=model
71  )
72
73  #### Population distribution ------------------------------
74
75  # Population parameter intercepts and scales; specifically,
    this line is loading the estimated intercepts from Matzke &
    Wagenmakers (2009), which are used to create the informed
    priors
76  tmp=load("MatzkeIntercepts.RData")
77
78  # Load the data and connect it with the DMC model
79  load(dataFileName)
80  dm <- data.model.dmc(data = datas, model = model)
```

```
81
82
83
84 #### Sampling ----------------------------------------------
85
86 # hyper-prior for fitting; specifically, this code sets up
   the priors in the format needed for the internal
   specifications of DMC
87 pp.prior <- prior.pp.dmc(
88   p1.mu=intercept.mean,           # mean of normal with SD=1
89   p1.sigma=intercept.sd,          # mean of default normal
90   p2.ce=effects,p2.se=subject.sd) # default cauchy
91
92
93 # Get some (amount specified by nmc) initial MCMC samples
   from the model, which will be used as the starting point of
   the auto-convergence algorithm in the next step
94 hsamples <- h.samples.dmc(nmc = 120,
95                           p.prior = p.prior,
96                           pp.prior = pp.prior,
97                           data = dm)
98 save(datas,hsamples,file=sampFileName) # Save these initial
   samples
99
100
101 # Run fits ------------------------------------------------
102
103 tmp=load(sampFileName) # Load in initial samples from above
```

```
104
105 # Run the auto-convergence algorithm in DMC, the standard
    version used in other implementations of DMC. In the initial
    run sampling mixes the standard cross-over step with
    migration steps on a randomly chosen 5\%  of iterations at
    both participant and population levels.
106
107 hsamples <- h.run.unstuck.dmc(hsamples, cores=cores,
    p.migrate=0.05, h.p.migrate=0.05, max.try=10)
108 save(hsamples,file = sampFileName)
109
110 # This continues until no chains differ from the median of
    the others in likelihood by more than 10 units. Next
    migration is turned off, 120 samples taken, then on each
    cycle 40 more added and either retained or the first 40
    removed, based on whichever results in better mixing as
    measured by the ‘‘gelman.diag’’ statistic provided by the
    CODA package function of that name. This processes continues
    until gelman.diag < 1.1. At that point previous samples are
    discarded and a further 300 iterations performed.
111
112 hsamples <-
    h.run.converge.dmc(h.samples.dmc(samples=hsamples, nmc=120,
    thin=10), cores=cores, report=10, max.try=10, finalrun=TRUE,
    finalI=300)
113 save(hsamples,file = sampFileName) # Once the sampling
    appears to have converged, store these samples
114
```

```
115 # Occasionally visual inspection of the output from the
    previous step reveals that some chains have diverged. In
    this case these can be removed and then a further set of
    samples obtained as follows, usually fixing the problem. As
    doing so does not hurt in cases where there are no problems
    this was done in all cases.
116
117 hsamples <- h.run.dmc(h.samples.dmc(samples=hsamples,
    nmc=50, replace.bad.chains=TRUE), cores=cores) # Start by
    getting some new initial samples
118
119 # A final set of 500 iterations is obtained and saved for
    later analysis.
120
121 hsamples <-
    h.run.dmc(h.samples.dmc(samples=hsamples,nmc=500),cores=cores)
122 save(hsamples,file = sampFileName) # Save the posterior
    samples
123
124
125 # Summarize fit ------------------------------------------
126
127 Smry <- summary.dmc(hsamples,hyper=FALSE) # Obtain summaries
    of the individual-level parameter estimates
128 h.Smry <- summary.dmc(hsamples,hyper=TRUE) # Obtain
    summaries of the group-level parameter estimates
129 print(Smry) # Display the individual-level parameters summary
130 print(h.Smry) Display the group-level parameters summary
131 save(Smry,h.Smry,hsamples,file = sampFileName) # Save the
    summaries with the samples
132
```

```
133 # Get the R.hat values for the individual and group level
    parameters to ensure that they are still reasonable (i.e.,
    that chains didn't drift away from the stationary
    distribution after the auto-convergence algorithm finished)
134 h.gelman.diag.dmc(hsamples)
135 gelman.diag.dmc(hsamples,hyper=TRUE)
136 effectiveSize.dmc(hsamples,hyper=TRUE)
137
138 # Make plots of the chains of the posterior distributions to
    diagnose any potential issues
139 pdf(paste(use.file,".pdf",sep=""),height=6,width = 8)
140 plot.dmc(hsamples,hyper=TRUE,pll.chain=TRUE)
141 plot.dmc(hsamples,hyper=TRUE,layout=c(2,5)) # mu
142 plot.dmc(hsamples,hyper=TRUE,layout=c(2,3),hyper.par="sigma")
143 plot.dmc(hsamples,hyper=TRUE,layout=c(3,5),hyper.par="alpha")
144 par(mfrow=c(3,5))
145 for (i in 1:length(hsamples))
146   plot.dmc(hsamples[[i]], pll.chain=TRUE,
    main.pll=paste("Subject", i))
147 dev.off()
148
149
150 # Calculate the marginal likelihood for the model (required
    to calculate Bayes factors and for model averaging) based on
    25000 new samples (assuming that is enough) ------------
151
```

```
totalIter=round(25000/hsamples$`1`$n.chains,0) # Get the number of iterations of the sampler that are needed to obtain 25000 new samples
newHsamples <- h.run.dmc(h.samples.dmc(samples=hsamples, nmc=totalIter), cores=cores) # Run the sampler to obtain the 25000 new samples
save(newHsamples, file = paste("BF_", sampFileName,sep="")) # Save the new samples
r <- h.bridge.sampler.es.dmc(samples = newHsamples, cores = cores) # Run bridge sampling on the new samples to obtain an estimate of the marginal likelihood
save(newHsamples, r, file = paste("BF_", sampFileName, sep="")) # Save the new samples and the estimated marginal likelihood
```

References

Agrestia, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, *47*(4), 639–653. https://doi.org/https://doi.org/10.1016/j.csda.2003.12.009

Aikaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, *94*(2), 443–458.

Apgar, J. F., Witmer, D. K., White, F. M., & Tidor, B. (2010). Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, *6*(10), 1890.

Ashby, F. G., & Townsend, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology*, *21*(2), 93–123.

Bahadur, R. R., & Bickel, P. J. (2009). An optimality property of Bayes' test statistics. In J. Rojo (Ed.), *Optimality: The third Erich L. Lehmann symposium* (pp. 18–30). https://doi.org/10.1214/09-LNMS5704

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

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

Beck, J., Ma, W., Kiani, R., Hanks, T., Churchland, A., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, *60*, 1142–1152. https://doi.org/https://doi.org/10.1016/j.neuron.2008.09.021

Bennett, C. H. (1976). Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, *22*, 245–268. https://doi.org/https://doi.org/10.1016/0021-9991(76)90078-4

Boehm, U., Gronau, Q. F., Matzke, D., Singmann, H., Sarafoglou, A., Vandekerckhove, J., Ly, A., Steingroever, H., Marsman, M., Leslie, D., Forster, J., & Wagenmakers, E.-J. (2016). Bayes factors for the diffusion model [invited presentation]. *Sequential sampling models of decision making*. Emmetten, Switzerland.

Boehm, U., van Maanen, L., Forstmann, B., & van Rijn, H. (2014). Trial-by-trial fluctuations in cnv amplitude reflect anticipatory adjustment of response caution. *NeuroImage*, *96*, 95–105. https://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2014.03.063

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*, 1614–1631. https://doi.org/10.3758/s13428-018-1054-3

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 11–21.

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

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective bayesian analysis. *Bayesian Analysis*, *13*, 627–679.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cronbach, L. J. (1975). Beyond the Two Disciplines of Scientific Psychology. *American Psychologist*, *2*, 116–127.

Damaso, K., Williams, P., & Heathcote, A. (2020). Evidence for different types of errors being associated with different types of post-error changes, 1–6.

Dillon, D. G., Wiecki, T., Pechtel, P., Webb, C., Goer, F., Murray, L., Trivedi, M., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Kurian, B., Adams, P., Carmody, T., Weyandt, S., Shores-Wilson, K., Toups, M., McInnis, M., Oquendo, M. A., ... Pizzagalli, D. A. (2015). A computational analysis of flanker interference in depression. *Psychological Medicine*, *45*, 2333–2344. https://doi.org/10.1017/S0033291715000276

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., ... Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*, 1051–1069.

Evans, N. J., & Annis, J. (2019). Thermodynamic integration via differential evolution: A method for estimating marginal likelihoods. *Behavior Research Methods*, *51*, 930–947.

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592–605.

Evans, N. J., Tillman, G., & Wagenmakers, E.-J. (2020). Systematic and random sources of variability in perceptual decision-making: Comment on ratcliff, voskuilen, and mckoon. *Psychological Review*, *127*(5), 932–944.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

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. https://doi.org/10.1146/annurev-psych-122414-033645

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An Introduction to Model-Based Cognitive Neuroscience*. Springer.

Foster, K., & Singmann, H. (2021). Another approximation of the first-passage time densities for the ratcliff diffusion decision model. *arXiv e-prints*, Article arXiv:2104.01902.

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, *86*(1), 1–28.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, *62*(1), 1–18. https://doi.org/https://doi.org/10.1016/j.jml.2009.09.004

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

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

Gondan, M., Blurton, S. P., & Kesselmeier, M. (2014). Even faster and even more accurate first-passage time densitites and distributions for the Wiener diffusion model. *Journal of Mathematical Psychology*, *60*, 20–22.

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Quality & Quantity*, *49*, 967–976. https://doi.org/https://doi.org/10.1007/s11135-014-0060-5

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: Bridge sampling for for marginal likelihoods and Bayes factors (R package version 1.0-0) [Computer software]. https://cran.r-project.org/web/packages/bridgesampling/index.html

Gronau, Q. F., Heathcote, A., & Matzke, D. (2019). Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior Research Methods*, *52*(2), 918–937.

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2019). A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using Warp-III bridge sampling. *Psychometrika*, *84*, 261–284.

Gunawan, D., Hawkins, G. E., Tran, M.-N., Kohn, R., & Brown, S. D. (2020). New estimation approaches for the hierarchical Linear Ballistic Accumulator model. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2020.102368

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Computational Biology*, *3*(10), e189.

Hawkins, G. E., & Heathcote, A. (2021). Racing Against the Clock: Evidence-Based Versus Time-Based Decisions. *Psychological Review*, *128*(2), 222–263. https://doi.org/10.1037/rev0000259

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior Research Methods*, *51*(2), 961–985.

Heathcote, A. (2019). What Do the Rules for the Wrong Game Tell us About How to Play the Right Game? *Computational Brain & Behavior*, *2*(3), 187–189.

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

Heck, D. W., Boehm, U., Florian, B.-M., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., ... Hoijtink, H. (2022). *Psychological Methods*. https://doi.org/10.1037/met0000454

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). Linear deterministic accumulator models of simple choice. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417.

Höge, M., Guthke, A., & Nowak, W. (2019). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107.

Huang-Pollock, C., Ratcliff, R., McKoon, G., Shapiro, Z., Weigard, A., & Galloway-Long, H. (2017). Using the diffusion model to explain cognitive deficits in Attention Deficit Hyperactivity Disorder. *Journal of Abnormal Child Psychology*, *45*(1), 57–68. https://doi.org/10.1007/s10802-016-0151-y

Iverson, G., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p-rep. *Psychological Methods*, *15*, 172–181. https://doi.org/10.1037/a0017182

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

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

Jevons, W. S. (1874). *The principles of science: A treatise on logic and scientific method*. Macmillan.

Johnson, V. E. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B (Methodological)*, *72*, 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

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

Kang, I., Yi, W., & Turner, B. M. (2022). A regularization method for linking brain and behavior. *Psychological Methods*, *27*(3), 400–425. https://doi.org/10.1037/met0000387

Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika*, *86*(1), 215–238. https://doi.org/https://doi.org/10.1007/s11336-021-09754-5

Kaplan, D., & Lee, C. (2016). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 343–353. https://doi.org/https://doi.org/10.1080/10705511.2015.1092088

Kaplan, D., & Yavuz, S. (2020). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*, *55*(4), 553–567. https://doi.org/https://doi.org/10.1080/00273171.2019.1657790

Karayanidis, F., Mansfield, E. L., Galloway, K. L., Smith, J. L., Provost, A., & Heathcote, A. (2009). Anticipatory reconfiguration elicited by fully and partially informative cues that validly predict a switch in task. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(2), 202–215.

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

Kolossa, A., & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *NeuroImage*, *172*, 775–785.

Kühn, S., Schmiedek, F., Schott, B., Ratcliff, R., Heinze, H.-J., Düzel, E., Lindenberger, U., & 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. https://doi.org/10.1162/jocn.2010.21564

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* Academic Press SN -.

Lee, M. D. (2016). Bayesian methods in cognitive modeling. In J. T. Wixted & E.-J. Wagenmakers (Eds.), *The stevens handbook of experimental psychology and cognitive neuroscience* (pp. 37–84).

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, *7*. https://doi.org/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. https://doi.org/10.3758/s13428-016-0740-2

Lerche, V., Voss, A., & Nagler, M. (2016). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, 1–25.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixture of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge University Press.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

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. https://doi.org/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. https://doi.org/10.1006/ccog.1996.0029

Meng, X.-L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, *11*, 552–586.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical explanation. *Statistica Sinica*, *6*, 831–860.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2007). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience*, *32*(7), 2335–2343.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*, 376–404.

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. https://doi.org/10.1523/JNEUROSCI.1655-06.2006

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425.

Quinn, R. K., James, M. H., Hawkins, G. E., Brown, A. L., Heathcote, A., Smith, D. W., Cairns, M. J., & Dayas, C. V. (2017). Temporally specific miRNA expression

patterns in the dorsal and ventral striatum of addiction-prone rats. *Addiction Biology*, *23*(2), 631–642.

Rae, B., Heathcote, A., Donkin, C., Avarell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226–1243. https://doi.org/http://dx.doi.org/10.1037/a0036801

Raftery, A., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.

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. https://doi.org/10.3758/BF03196283

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182. https://doi.org/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. https://doi.org/10.1162/neco.2008.12-06-420

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

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*(4), 626–635.

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75–90 years old. *Psychology and Aging*, *22*(1), 56–66.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaching to dealing with contaminant reaction and parameter variability.

Psychonomic Bulletin & Review, *9*(3), 438–481. https://doi.org/10.3758/BF03196302

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical model with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection model with random participant and item effects. *Psychometrika*, *72*(4), 621–642.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Rouder, J. N., Sun, D., Speckman, P. L., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606.

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 222–250.

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1248–1284.

Singmann, H., Brown, S., Gretton, M., & Heathcote, A. (2020). *Rtdists: Response time distributions* [R package version 0.11-2]. https://CRAN.R-project.org/package=rtdists

Smith, P. L., & Lilburn, S. D. (2020). Vision for the blind: visual psychophysics and blinded inference for decisionmodels. *Psychonomic Bulletin & Review*, *27*, 882–910. https://doi.org/https://doi.org/10.3758/s13423-020-01742-7

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101.

Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2014). *Psychological Review*, *121*(4), 679–688.

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. https://doi.org/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. https://doi.org/10.1016/j.visres.2004.01.002

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *B64*(4), 583–639.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.

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

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.

Tran, M.-N., Scharth, M., Gunawan, D., Kohn, R., Brown, S. D., & Hawkins, G. E. (2020). Robustly estimating the marginal likelihood for cognitive models via importance sampling. *Behavior Research Methods*. https://doi.org/https://doi.org/10.3758/s13428-020-01348-w

Tran, N. H., van Maanen, L., Heathcote, A., & Matzke, D. (2021). Systematic Parameter Reviews in Cognitive Modeling: Towards a Robust and Cumulative Characterization of Psychological Processes in the Diffusion Decision Model. *Frontiers in psychology*, *11*, 267–14.

Tuerlinckx, F., & Boeck, P. D. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650.

Turner, B. M., Maanen, L. v., & Forstmann, B. U. (2015). Informing Cognitive Abstractions Through Neuroimaging: The Neural Drift Diffusion Model. *Psychological Review*, *122*(2), 312–336. https://doi.org/10.1037/a0038894

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. https://doi.org/10.1177/0145721709355835.Continuous

Turner, B. M., Wang, T., & Merkle, E. C. (2017). Factor analysis linking functions for simultaneously modeling neural and behavioral data. *NeuroImage*, *153*, 28–48. https://doi.org/10.1016/j.neuroimage.2017.03.044

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356.

van Vugt, M., Simen, P., Nystrom, L. E., Holmes, P., & Cohen, J. D. (2012). EEG oscillations reveal neural correlates of evidence accumulation. *Frontiers in Neuroscience*, 6:106. https://doi.org/10.3389/fnins.2012.00106

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208–256.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/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. https://doi.org/10.1016/j.jmp.2014.06.004

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. https://doi.org/10.1016/j.jmp.2009.09.004

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 1–20.

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., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, *142*(2), 536–559.

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*(14), 470.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: an R Package Providing Distribution Functions for the Wiener Diffusion Model. *The R Journal*, *6*, 49–56.

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671. https://doi.org/10.1080/09541440802205067

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. https://doi.org/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. https://doi.org/10.1016/j.jml.2007.04.006

Weigard, A., Huang-Pollock, C., Brown, S., & Heathcote, A. (2018). Testing formal predictions of neuroscientific theories of ADHD with a cognitive model-based approach. *Journal of Abnormal Child Psychology*, *127*(5), 529–539. https://doi.org/10.1037/abn0000357

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. https://doi.org/10.1080/02699931.2013.858028

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00014

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for snp association studies. *The Annals of Applied Statistics*, *4*, 1342–1364.

Wrinch, D., & Jefferys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*(249), 369–390.

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. https://doi.org/10.1037/xlm0000064