The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models

Gilles Dutilh[1], Jeffrey Annis[2], Scott D. Brown[3], Peter Cassey[3], Nathan J. Evans[3], Raoul P. P. P. Grasman[4], Guy E. Hawkins[4], Andrew Heathcote[5], William R. Holmes[2], Angelos–Miltiadis Krypotos[6], Colin N. Kupitz[7], Fábio P. Leite[8], Veronika Lerche[9], Yi–Shin Lin[5], Gordon D. Logan[2], Thomas J. Palmeri[2], Jeffrey J. Starns[10], Jennifer S. Trueblood[2], Leendert van Maanen[4], Don van Ravenzwaaij[11], Joachim Vandekerckhove[7], Ingmar Visser[4], Andreas Voss[9], Corey N. White[12], Thomas V. Wiecki[13], Jörg Rieskamp[1], Chris Donkin[14]

[1], University of Basel, [2] Vanderbilt University, [3] University of Newcastle, [4] University of Amsterdam, [5] University of Tasmania, [6] Utrecht University, [7] University of California, Irvine, [8] Ohio State university, [9] University of Heidelberg, [10] University of Massachusetts Amherst, [11] University Groningen, [12] Syracuse University, [13] Brown University, [14] University of New South Wales

Abstract

Most data analyses rely on models. To complement statistical models, psychologists have developed cognitive models, which translate observed variables into psychologically interesting constructs. Response time models, in particular, assume that response time and accuracy are the observed expression of latent variables including 1) ease of processing, 2) response caution, 3) response bias, and 4) non–decision time. Inferences about these psychological factors, hinge upon the validity of the models' parameters. Here, we use a blinded, collaborative approach to assess the validity of such model-based inferences. Seventeen teams of researchers analyzed the same 14 data sets. In each of these two–condition data sets, we manipulated properties of participants' behavior in a two–alternative forced choice task. The contributing teams were blind to the manipulations, and had to infer what aspect of behavior was changed using their method of choice. The contributors chose to employ a variety of models, estimation methods, and inference procedures. Our results show that, although conclusions were similar across different methods, these "modeler's degrees of freedom" did affect their inferences. Interestingly, many of the simpler approaches yielded as robust and accurate inferences as the more complex methods. We recommend that, in general, cognitive models become a typical analysis tool for response time data. In particular, we argue that the simpler models and procedures are sufficient for standard experimental designs. We finish by outlining situations in which more complicated models and methods may be necessary, and discuss potential pitfalls when interpreting the output from response time models. *Manuscript submitted for publication on December 16, 2016.*

Introduction

In Experimental Psychology, we aim to draw psychologically interesting inferences from observed behavior on experimental tasks. Despite the wide variety of tasks to measure participants' performance in a range of cognitive domains, many assessments of performance are based on the speed and accuracy with which participants respond. It has long been recognized that the interpretation of data from such *response time tasks* is hampered by the ubiquitous speed–accuracy trade–off (Pew, 1969; Wickelgren, 1977): When people aim to respond faster, they are less accurate. Conversely, people can also slow down to increase their accuracy. To understand the implications of this trade–off, consider Mick J and Justin B, 73 and 22 years old respectively. Both perform a simple lexical decision task, where they press as quickly as possible one of two response buttons to indicate whether a string of letters represents a valid word. Now, averaged over many such lexical decisions, Justin turns out to be much quicker than Mick, but Mick has a slightly higher percentage of correct responses. What conclusions should we draw from these results? Is the younger person better at lexical decisions? Is the elderly person more conservative, or maybe just physically slower to press buttons?

To answer such questions, cognitive models have been developed that provide a better understanding of the behavior of participants in response time tasks. These cognitive models are now often used as a measurement tool, translating the speed and accuracy of responses into the latent psychological factors of interest, such as participants' *ability*, response *bias* and the *caution* with which they respond. In this article, we aim to study the validity of the inferences drawn from cognitive models of response time data. We do so by having 17 teams of response time experts analyze the same 14 real data sets, while being blinded to the manipulations, with the method of their choosing.

In the following, we first introduce the principle of cognitive modeling as an alternative to statistical modeling. We focus on evidence-accumulation models as the most popular class of cognitive models for response time data. We argue that the validity of inferences from cognitive models are threatened by a host of issues, including those that plague typical statistical analyses. We then present our collaborative project that we set up to test the validity of the inferences that are drawn using cognitive models for response time data.

Cognitive Models

The story of Mick J and Justin B is a simple example of the difficulty of a direct interpretation of the observed dependent variables as a measure of performance – being faster or more accurate at a task does not necessarily indicate superior performance. Arguably, ability affects observed data through an intricate series of psychological processes. Cognitive models were developed to provide an explicit account of such psychological processes. A cognitive model is a formalized theory that is intended to mimic the cognitive processes that give rise to the observed behavioral data. Such a formalization often describes a sequence of cognitive steps that are supposedly performed by a participant when performing a task. The precise formalization allows researchers to derive fine–grained predictions about the data that are observed when participants perform tasks that require the targeted cognitive process. Armed with a cognitive model, a researcher can reverse-engineer latent variables of interest, such as ability, from the observed behavioral data, e.g., the speed and accuracy of a response.

Perhaps the most-used formal cognitive model of human behavior is signal detection theory (Swets, 2014). Signal detection theory is a mathematical model for the decision over whether a

stimulus contains a signal or not. The model is popular because it allows the user to interpret observed responses - the probability of detecting a signal when present and when absent - in terms of psychologically-interesting latent variables, such as the ability to discriminate signal from noise, and the observer's bias when responding. Similarly, the cognitive models that are the focus of this study, evidence-accumulation models, are used to translate observed response times and choices into the psychologically interesting constructs *ease of information processing*, *response caution*, *response bias*, and the *time needed for non-decision processes*.

Cognitive models have two key benefits over statistical models that lack any psychological interpretation. First, mathematical cognitive models often capture key phenomena in behavioral data for which statistical models cannot account. For example, evidence-accumulation models offer a natural account of the relation between response speed and accuracy that is observed in response time experiments. Second, and more crucially for the current study, the parameters of cognitive models reflect the magnitude of assumed cognitive constructs, rather then content–free statistical properties. For example, for a researcher, it is much more interesting to learn that Mick J is twice as good at the task as Justin B, than it is to learn that Justin answers 5% more accurately than Mick, while being on average 150 milliseconds slower.

Cognitive Models of Response Time Data

In the last few decades, cognitive models have increasingly been used as measurement models for response time data. The most popular of these models is Ratcliff's (1978) diffusion model. Though originally proposed as an explanation for performance in memory tasks, the diffusion model is now commonly used to transform response time and accuracy data into latent constructs in a wide range of tasks and domains of study. For example, the diffusion model has been used to study implicit racial associations (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007), effects of aging on brightness discrimination (e.g., Ratcliff, Thapar, & McKoon, 2006), practice effects on lexical decisions (Dutilh, Krypotos, & Wagenmakers, 2011), the effect of attention deficit hyperactivity disorder on a conflict control task (Metin et al., 2013), and the effect of alcohol consumption on movement detection (van Ravenzwaaij, Dutilh, & Wagenmakers, 2012). Response time models have also been applied to inform the analysis of brain measures (e.g., Cavanagh et al., 2011; Forstmann et al., 2008; Mulder, Van Maanen, & Forstmann, 2014; Ratcliff, Philiastides, & Sajda, 2009).

The diffusion model is the prototypical example of an evidence-accumulation model. The model is illustrated in Figure 1. The Figure shows two hypothetical decisions between two options, responses A and B. The accumulation of evidence in favor of either response, as depicted by the solid grey lines, begins closer to the "A" than "B" boundary, indicating a slight bias to respond A. The dark grey line represents a trial on which evidence accumulates relatively quickly towards the "B" boundary, resulting in a fast B response. The light grey line, on the other hand, represents a slow A response. The time to respond is assumed to be the sum of this decision time, plus any time taken to encode the stimuli and execute the motor response (i.e., non-decision time).

The Four Key Components of Evidence-Accumulation Models. When used as measurement tools, evidence-accumulation models generally aim to measure four key components. These components are reflected in each of four parameters that are comparable across models.

Accumulation Rate. The average rate at which evidence accumulates towards the correct response boundary. The accumulation rate reflects the ratio of signal to noise provided by a stimulus, for a given observer. The accumulation rate parameter is interpreted as quantifying *the ease of responding*. Faster accumulation rates are associated with quicker and more accurate responding,
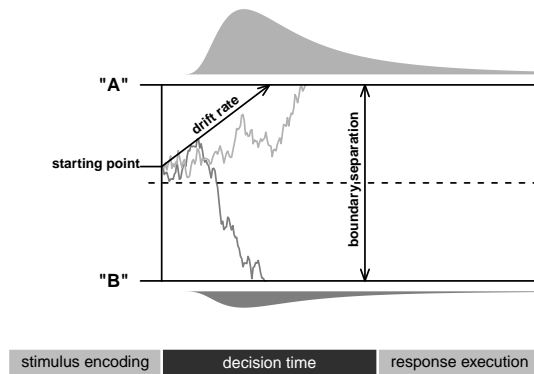
Figure 1. Graphical illustration of the diffusion model.

and slower accumulation rates with slower and less accurate responding. Relatively faster accumulation rates thus indicate participants who are performing better on a task, or stimuli that are easier to process.

Boundary Separation. The distance between the two response boundaries in Figure 1. Boundary separation defines the strength of evidence required for either of the response options to be initiated. The boundary separation parameter is interpreted as a measure of *response caution*. Small boundary separation is associated with quick and error–prone performance, while high boundary separation is associated with relatively slow, conservative performance.

Starting Point. The position on the evidence axis at which the information integration process begins. The starting point parameter quantifies the a priori *bias* towards either of the response options. A starting point that is equidistant from the two boundaries constitutes an unbiased decision process, such that only the sampled information from the stimulus determines the response. A starting point shifted in the direction of one response option (say, option "A" in Figure 1), makes that response occur more frequently and faster, on average. Thus, a shift in starting point signals eventual tendencies to press one button rather than the other, independent of the presented stimulus.

Non–decision time. The time added to the decision time (i.e., the result of the diffusion process in Figure 1) to yield the observed response time. The non-decision time parameter defines the location of the RT distributions and is assumed to reflect the time that is needed for *stimulus encoding and (motor) response execution*.

Interpretation of these four key components is the critical factor when using a response time model as a measurement tool. For example, when a researcher aims to study the difference in performance between 72 year old Mick J and 22 year old Justin B, she would fit a response time model to the response time and accuracy data of both participants. The researcher may find that the accumulation rate is slightly higher for Justin, whereas Mick has a larger boundary separation. This result would indicate that the younger participant was better at extracting lexical information from letter strings, but was also less cautious while making his decisions. Thus, instead of interpreting the raw RT and accuracy of both participants, the researcher draws inferences about the latent variables underlying performance, as reflected by the four key components of response time models.

Threats to the Validity of Cognitive Models

Researcher Degrees of Freedom. Cognitive models are also statistical models, and so presumably suffer from many of the same issues that burden any statistical analysis. For example, in recent years it has become widely acknowledged that most statistical analyses offer the researcher an excess of choices. These large and small choices lead the researcher into a "garden of forking paths" (Gelman & Loken, 2013), where, without a preregistered analysis plan, it is extremely difficult to prevent being biased by the outcomes of different choices. For example, Silberzahn et al. (2015) used a collaborative approach, similar to the approach we take, to demonstrate the degree to which researcher choices can influence the conclusions drawn from data. Their results, as well as many others, have shown repeatedly that such researcher degrees of freedom can threaten the validity of the conclusions drawn from standard statistical analyses.

In this light, it is important to note that the landscape of measurement models for response time data has expanded considerably over the last decade (e.g., Brown & Heathcote, 2005; Kvam, Pleskac, Yu, & Busemeyer, 2015; Usher & McClelland, 2001; Verdonck & Tuerlinckx, 2014). For example, thus far we have discussed but one variant of the diffusion model – the Simple Diffusion model. Another variant, the Full Diffusion model is more complex, assuming trial–to–trial variability in accumulation rate, starting point, and non-decision time (Ratcliff & Tuerlinckx, 2002). At the same time, exceedingly simple parameter estimation methods also exist for the diffusion model (EZ and EZ2 diffusion models: Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, Van der Maas, & Grasman, 2007).

One also need not use a diffusion model at all. A popular alternative model of response times is the Linear Ballistic Accumulator (LBA) model (Brown & Heathcote, 2008). The LBA model is similar to the diffusion model, in that it features the four key components in the diffusion model. However, the LBA differs from the diffusion model in two important ways. First, while the diffusion model assumes that evidence in favor of one response is counted as evidence against the alternative response, the LBA model assumes independent accumulation of evidence for each response. Second, the diffusion model assumes moment-to-moment variability in the accumulation process, while the LBA model makes the simplifying assumption that evidence accumulates ballistically, or without moment-to-moment noise.

Once a model is selected, researchers still face a number of choices regarding their analysis of response time data. For example, one must choose a method for estimating parameters from data. Though Frequentist approaches, such as chi-square (Ratcliff & Tuerlinckx, 2002), maximum-likelihood (Donkin, Brown, & Heathcote, 2011; Vandekerckhove & Tuerlinckx, 2007), or Kolmogorov-Smirnov (Voss & Voss, 2007) are and remain commonplace, Bayesian methods are becoming increasingly popular (Vandekerckhove, Tuerlinckx, & Lee, 2011; Wabersich & Vandekerckhove, 2014; Wiecki, Sofer, & Frank, 2013). Beyond estimation, there still also remains a choice over how statistical inference will be performed. Historically, null-hypothesis significance tests on estimated parameters were standard (Ratcliff, Thapar, & McKoon, 2004). However, based on growing skepticism over such techniques (e.g., Wagenmakers, 2007), model-selection based methods (Donkin, Brown, Heathcote, & Wagenmakers, 2011), and inference based on posterior distributions is now common (Cavanagh et al., 2011).

In summary, there are a very large number of modeler's degrees of freedom. Further, it is rare to see any strong motivation or justification for the particular choices made by different researchers. The extent to which the various factors influence the conclusions drawn from an

evidence-accumulation model analysis is under-explored (but see Donkin, Brown, & Heathcote, 2011; Lerche, Voss, & Nagler, 2016; van Ravenzwaaij & Oberauer, 2009).

Convergent and Divergent Validity.   We have argued that the key benefit of a cognitive model over a statistical model is the ability to make inferences about psychologically interesting latent variables. However, such an advantage relies on the parameters of the cognitive model being valid measures of their respective constructs. We must assume both convergent and discriminant validity, such that a manipulation of response caution, for example, should always affect boundary separation and boundary separation only. There exist already a number of studies which aimed to validate the inferences drawn from evidence accumulation models. We summarize their results and argue why the current large-scale validation study fills a timely place in the literature.

Voss, Rothermund, and Voss (2004) provided one of the earliest systematic empirical validation studies. In their study, participants judged the color of squares made up of green and orange pixels. Crucially, their study manipulated 4 factors across different blocks: the difficulty of the stimuli, participants' response caution, the probability of green vs orange stimuli, and the ease with which the motor response could be made. After fitting the Full diffusion model, Voss et al. (2004) observed evidence for the convergent validity of the interpretation of the model's parameters; for example, easier stimuli led to higher estimates of the accumulation rate parameter. Although the experimental manipulations influenced the intended parameters, the effects were not exclusive, speaking against the discriminant validity of the interpretation of diffusion model results. For example, increased caution was not only expressed in higher boundary separation, but also in larger non–decision time estimates.

Other empirical validation studies provide mixed support for convergent and discriminant validity of the Full Diffusion model. On the one hand, Ratcliff (2002) and Wagenmakers, Ratcliff, Gomez, and McKoon (2008) used brightness discrimination and lexical discrimination tasks, respectively, to demonstrate both forms of validity in accumulation rate, starting point, and boundary separation parameters. However, in a recent study, Arnold, Bröder, and Bayen (2015) found evidence for convergent, but not discriminant validity of the Full Diffusion model parameters in the domain of recognition memory. For example, they found that their manipulation of response caution had a large effect on the boundary separation parameter, but also on the accumulation rate and non-decision time parameters. It is important to note here that conclusions about divergent validity rely on the assumption of selective influence of the manipulations in a validation experiment: for example, a manipulation of response caution should affect response caution, and nothing else. In our results and discussion sections, we will cover this issue in more depth.

Though the majority of empirical validation studies have used only the Full Diffusion model (Ratcliff & Tuerlinckx, 2002), some work has also tried to validate alternative response time models. For example, Arnold et al. (2015) found that the EZ diffusion model, unlike the Full Diffusion model, showed both convergent and discriminant validity. Donkin, Brown, Heathcote, and Wagenmakers (2011) re–analyzed the lexical decision data set from Wagenmakers et al. (2008) using both the Full Diffusion and LBA models. With both models, they found the same results as the original authors. A similar re-analysis of the data set from Ratcliff et al. (2004) also indicated that the LBA and diffusion models yielded the same conclusion about the influence of aging on recognition memory tasks.

A Collaborative, Blinded Validation Study

Given the wide–spread application of evidence accumulation time models, the number of attempts to validate the inferences drawn from these models is small. Also, the results from existing validation studies are mixed. Further, we know of no attempt to investigate the extent and influence of researcher degrees of freedom in the application of cognitive modeling. We attempted to address these issues with a large-scale collaborative project.

In our study, we created 14 two–condition data sets and openly invited a range of modeling experts to analyze those data sets. 17 teams of experts committed to contribute. The contributors were asked to infer the difference between the two conditions in terms of four psychologically-interesting latent constructs – the ease of processing, caution, bias, and non-decision time. The 14 data sets, comprising human data, were created such that the two conditions could differ in terms of any combination of ease, caution, and bias. We did not manipulate the non–decision time because the component is not related to a clearly defined construct, but simply reflects the amount of time added to the decision process to yield a response time. Our collaborative approach offers two advantages over existing validation studies.

First, this approach allows us to assess the validity of a range of methods and models. In particular, since the expert collaborators chose their own methods to draw inferences, we have a sample of the validity of currently popular methods. This range of methods allows us to test the importance of the many choices that response time modelers face when analyzing response time data. This aspect of the project is critical, because most existing validation studies used just one method of analysis. For example, despite their comprehensive analysis, Voss et al. (2004) could draw conclusions about only the Full Diffusion model with Kolmogorov-Smirnov estimation performing null hypothesis tests on estimated parameters.

Second, the expert contributors to our study performed their analysis while "blinded". That is, the experts did not know the true manipulations that we made in the validation data sets. The use of this blinded analysis provides a strong advantage relative to existing validation studies. For example, in Donkin, Brown, Heathcote, and Wagenmakers (2011), the authors re-analyzed a data set that had already been analyzed with a Full Diffusion model. As such, Donkin et al. knew the conclusions drawn by the original authors, and so may have made choices that were biased to draw inferences consistent with previously published results. The model-based analysis of response time data is like any other data analysis, and there are many choices on offer to the researcher. Gelman and Loken (2013) argue that without pre-registration, such choices tend to be made so as to increase the likelihood of agreeable conclusions. Our blinded method guards against such potential biases. Indeed, a blinded analysis seems critical in an assessment of the impact of researcher degrees of freedom.

The Experiment

To create the 14 data sets that the teams of contributors analyzed, we collected data in a factorial design, in which we aimed to impose 1) two levels of difficulty (hard and easy trials) crossed with 2) two levels of response caution (speed and accuracy emphasis instructions), and 3) three levels of response bias (proportion of trials of each response type). Response caution and bias were manipulated across blocks of trials, whereas difficulty of the stimuli was manipulated within blocks. In this section, we describe the experiment, and show that the manipulations had their intended effects on behavior.

Method

Participants. Twenty psychology students (15 female, mean age 26.7, SD = 2.1) at the University of Basel participated in a single session for course credit. The entire session lasted slightly less than two hours. Participants were allowed to take a break between blocks (see below).

Materials. Participants performed a random dot motion task which was presented using the Psychopy package for Python (Peirce, 2007). In the random dot motion task, participants detect the direction (left or right) of the apparent motion constituted by a cloud of moving dots. Each stimulus consisted of 120 dots (each dot was 4 pixels wide), presented in a circular aperture (diameter = 400 pixels) on a 1680×1050, 60 Hz LCD screen. On the first frame of a trial, all dots were placed on random coordinates in the aperture. Then, for each subsequent frame, the dots were displaced according to the following rules. For the difficult stimuli, 10% of the dots moved 1 pixel every 6 frames in the target direction. The other 90% of dots were replaced randomly in the aperture. For the easy stimuli, 20% of the dots moved coherently in the target direction. Coherently moving dots traveled a maximum of 5 pixels, so that no single dot could be monitored to infer the correct direction of motion. Each stimulus had a maximum duration of three seconds. The interval between a response and the start of the next trial was a draw from a uniform distribution between 0.5 and 1 second.

Responses were registered as button presses on a computer mouse with high timing accuracy (1000 Hz). The left and right responses were given by pressing the right hand index and middle finger respectively. Participants were seated in front of a computer screen in a small room in the presence of an experimenter.

Design. The experiment used three manipulations.

Response caution (speed/accuracy instructions): In about half of the blocks, accuracy was emphasized (ac in Table 1). In these blocks, participants received a feedback message "error" on erroneous responses and no feedback on speed, except for "time is up" after three seconds. In the other half of the blocks, speed was emphasized (sp in Table 1). In these blocks, participants received feedback only on the speed of their response: Responses slower than 0.8 seconds resulted in a "too slow" message. No feedback on accuracy was given. All feedback messages lasted 1.5 seconds.

Bias: In about half of the blocks, left and right stimuli occurred equally often. In the rest of the blocks, stimuli in one direction occurred twice as often ($2/3$) as stimuli in the other direction ($1/3$). For one half of these *biased* blocks, the left stimulus occurred more often, for the other half, the right stimulus occurred more often.

Ease/difficulty: Each block consisted of 50% easy and 50% hard trials, randomly intermixed. Difficulty was manipulated within the blocks to rule out the possibility that participants adjust their response caution when difficulty changes.

Procedure. Trials were administered in blocks. The experiment started with 5 practice blocks of 80 trials each, totaling 400 trials. These practice blocks familiarized the participants both with the stimuli (presenting slightly easier stimuli in the first block) and the manipulations of response caution and bias. The main experiment consisted of 18 blocks consisting of 156 trials each (see Table 1).[1]

---

[1]The order of blocks was constant across participants. Although practice effects could potentially have influenced the results, various analyses can be found in the online supplementary materials (`https://osf.io/ktuy7/`) that convinced us that the extended training phase and the distribution of block types over the sessions guarded against such effects.

Before each block, instructions were shown concerning response caution and bias in the upcoming block. This screen informed the participant to either focus on accuracy or speed and the relative number of left and right stimuli in the upcoming block. For non–biased blocks, participants were informed that the relative frequency of left and right stimuli was 50–50; for biased blocks, participants were informed that one direction was twice as likely to occur as the other. It was stressed that this information was accurate and could thus be used to inform choices.

Table 1

Manipulations of response caution and response bias across blocks and descriptive statistics per block.

Block	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
speed/accuracy	sp	ac	sp	sp	ac	ac	sp	sp	ac	ac	sp	sp	ac	ac	sp	sp	ac	sp
bias	no	no	no	left	left	no	no	right	right	no	no	left	left	no	no	right	right	no
RT (ms)																		
.1 quantile	360	510	370	380	490	490	360	380	470	480	370	370	490	490	370	380	480	370
.5 quantile	490	670	490	500	640	640	470	490	610	610	480	480	630	630	470	480	610	480
.9 quantile	690	1040	660	670	970	990	640	660	920	900	660	630	960	980	630	650	910	650
accuracy	0.76	0.91	0.82	0.8	0.91	0.91	0.8	0.84	0.91	0.91	0.81	0.8	0.91	0.9	0.78	0.81	0.91	0.79

Note. Top section: the design of the experimental blocks in the experiment. Bottom section: Descriptive statistics for each experimental block. Note that behavior remains largely invariant over the course of the experiment. Response caution (speed: sp vs accuracy: ac) and response bias were manipulated across blocks. Difficulty was manipulated within blocks.

As can be seen in Table 1, from block 2 through block 17, each cell resulting from the factors *response caution* × *response bias* occurs 4 times over the experimental session. The first and 18th block contain two further blocks in which speed was emphasized and no response bias was implemented. In this cell of the factorial design, more trials were needed, as we explain below.

Behavioral Manipulation Checks

Naturally, the experimental manipulations were intended to have effects on participants' overt behavior. Indeed, the speed/accuracy emphasis instruction and difficulty manipulations had the expected effects on the participants' accuracy and response time. Looking first at proportion correct, we see that participants were more accurate in the accuracy-emphasis condition ($\hat{P}_C = 0.95$ correct, $SD = 0.04$) than the speed-emphasis condition ($\hat{P}_C = 0.85$, $SD = 0.11$) when the task was easy. The same pattern was observed when the task was harder, where participants were more accurate under accuracy emphasis ($\hat{P}_C = 0.86$, $SD = 0.07$) than speed emphasis ($\hat{P}_C = 0.75$, $SD = 0.10$). A 2 (Emphasis Instruction: Speed vs. Accuracy) x 2 (Difficulty: Easy vs. Hard) x 3 (Bias: Left, None, Right) Bayesian ANOVA on the arcsine-transformed proportion of correct responses confirmed these observations [2]. The data were best explained by a model with the two main effects of difficulty ($\omega^2 = 0.95$) and emphasis instruction ($\omega^2 = 0.63$), with a $BF = 2.11 \times 10^{24}$ compared to the null, intercept-only model. There was moderate evidence for the two main effects model over a model that also contained the interaction between difficulty and emphasis instruction ($BF = 4.6$). The two main effects model was also favored over a model that also included a main effect of the bias manipulation on overall accuracy ($BF = 7.45$).

The difficulty and emphasis instruction manipulations also influenced the speed of responding. When asked to respond relatively quickly, participants were only a little faster when the task

---

[2] All ANOVAs were performed using the BayesFactor package for R (Morey, Rouder, & Jamil, 2014). Effect sizes are ω^2 values, calculated based on output from the aov function in R.

was easy ($\hat{RT} = 0.49$s, $SD = 0.05$) compared to when the task was difficult ($\hat{RT} = 0.52$, $SD = 0.05$). However, when asked to be accurate, participants were much faster in the easy task ($\hat{RT} = 0.63$, $SD = 0.11$) than in the hard task ($\hat{RT} = 0.75$, $SD = 0.15$). The same 2 x 2 x 3 Bayesian ANOVA, but on the mean response times, confirmed the apparent interaction between difficulty and emphasis instruction. The model with both main effects, difficulty ($\omega^2 = 0.29$) and instruction ($\omega^2 = 0.70$), and their interaction ($\omega^2 = 0.75$), was the best model ($BF = 3.11 \times 10^{32}$ against the null model). There was strong evidence for the model with an interaction term over the model without an interaction between difficulty and emphasis instruction ($BF = 14.11$ in favor of the model with an interaction). Again, there was little evidence that the bias manipulation had any influence on the overall speed of responding. For example, adding the main effect of a bias manipulation to the best-fitting model reduces the evidence for the model ($BF = 14.36$ in favor of the model without a main effect of bias).

We use a signal detection analysis to determine whether the bias manipulation had its intended effect. We calculate the standard criterion value, C, using the equation $0.5(\Phi(H) + \Phi(FA))$, where H and FA are the proportion of 'left' responses to leftward and rightward moving dots, respectively, and Φ is the cumulative distribution function of a standard Normal distribution. When emphasizing speed, we see a clear effect of the bias manipulation (leftward bias: $\hat{C} = 0.14$; no bias: $\hat{C} = -0.025$; rightward bias: $\hat{C} = -0.26$). When participants were instructed to be accurate, however, there was no systematic influence of the bias manipulation (leftward bias: $\hat{C} = -0.10$); no bias: $\hat{C} = -0.13$; rightward bias: $\hat{C} = -.10$). A 2 x 2 x 3 Bayesian ANOVA on the criterion values indicates that a model with main effects of bias ($\omega^2 = 0.47$) and emphasis instruction ($\omega^2 = 0.20$) and their interaction ($\omega^2 = 0.75$) provides the best account of our data ($BF = 3.1 \times 10^8$ compared to the intercept-only model). The difficulty manipulation appeared to have no influence on the level of bias shown by participants ($BF = 4.84$ in favor of the best-fitting model, relative to a model that also included a main effect of difficulty).

These analyses show that all our manipulations worked as intended. This is important to know, because we now know there are effects on the overt behavior about which the response time models should in principle allow us to draw inferences. The only condition in which it is unclear whether there was a behavioral effect was the bias manipulation while accuracy was emphasized. We also note that these behavioral manipulation checks are flawed, since response time and accuracy are treated separately, and the analysis permits no inferences about latent factors such as caution and ability. Of course, such critiques are precisely the reason that response time models were proposed in the first place.

The Creation of the 14 Data Sets

We used the data from the factorial experiment described above to construct the 14 data sets that we asked the contributors to analyze. Each data set had two conditions (A and B), which differed on a different combination of effects, as shown in Table 2. The letters in the table indicate which of the two conditions has a higher value on the relevant component, "-" indicating no difference. Effects are defined as 1) ease: higher means easier, 2) caution: higher means more cautious, 3) bias to right: higher means stronger bias towards the right response option. These 14 data sets constitute the complete set of possible combinations of directions of effects.

Relabeling of Left and Right.   Note that we manipulated bias both favoring the right option (blocks 8, 9, 16, 17) and left option (blocks 4, 5, 12, 13). By doing so, we were able to create a balanced design for the participants: overall, the correct response was equally often left and right.

Table 2
Pseudo Experiments

exp	ease	caution	bias R	ndt	blocks cond. A	blocks cond. B
1	-	-	-	-	hard, speed, no bias	hard, speed, no bias
2	B	-	-	-	hard, speed, no bias	easy, speed, no bias
3	-	B	-	-	hard, speed, no bias	hard, accuracy, no bias
4	-	-	B	-	hard, speed, no bias	hard, speed, bias
5	B	B	-	-	hard, speed, no bias	easy, accuracy, no bias
6	B	-	B	-	hard, speed, no bias	easy, speed, bias
7	-	B	B	-	hard, speed, no bias	hard, accuracy, bias
8	A	B	-	-	easy, speed, no bias	hard, accuracy, no bias
9	A	-	B	-	easy, speed, no bias	hard, speed, bias
10	-	A	B	-	hard, accuracy, no bias	hard, speed, bias
11	A	B	B	-	easy, speed no bias	hard, accuracy, bias
12	B	A	B	-	hard, accuracy, no bias	easy, speed, bias
13	B	B	A	-	hard, speed, bias	easy, accuracy, no bias
14	B	B	B	-	hard, speed, no bias	easy, accuracy, bias

Each line shows for one data set which of the two conditions (A or B) was manipulated to have a higher value on each of the components: ease, caution, bias towards response B and non–decision time. "-" indicates no difference. Rightmost columns show from which conditions (see Table 1) the data in each of the two conditions originate.

Before constructing the data sets, however, we recoded the data such that the option favored by the bias manipulation was always labeled "right". For all blocks without a bias manipulation, we also flipped the labeling of both stimulus direction and response for the even half of the trials. We did so to assure that any natural bias of participants to press left or right was averaged out in the data. The analysts contributing to this study were not aware of any re-labeling. However, we did note in the description of the data set: "The coding of left and right stimuli and responses has in some cases been recoded to obfuscate the study's design. Practically, however, L and R [labels in the data set] should be treated simply as left and right."

Selection of Cells to Construct Pairs. For many of the combinations of effects shown in Table 2, more than one pair of cells from the factorial design could be chosen. For example, to create two sets that differ on only response caution, one could contrast condition A: *easy*, *accuracy stress*, *no bias* against condition B: *easy*, *speed stress*, *no bias*. One could, however, also contrast condition A: *difficult*, *accuracy stress*, *bias* against condition B: *difficult*, *speed stress*, *bias*. In all such situations where more than one pair could be constructed, we chose the pair with 1) no bias and 2) the highest expected number of errors. The latter condition was chosen to anticipate that many of the models applied to RT data rely on the presence of error responses.

The trials constituting the conditions in each data set were random draws from the relevant cells of the factorial design. The sampled trials could come from all blocks across the experiment, potentially raising issues with non-stationary data (e.g., due to practice or fatigue effects). Table 1 reports the speed of responding and overall accuracy for each block in the experiment. At the aggregate level, the speed and accuracy of participants' responses remains stable across the entire experiment. In an online supplementary material, we report a more comprehensively demonstration

that our decision to sample trials across blocks of a two-hour experiment did not compromise our ability to judge the accuracy of the inferences made by our modeling teams.

Note that Data Set 1 is a special case where there is no difference between the two conditions. To construct this data set, we needed two replications of the same cell in the design. Following the rules above, we chose the cell *difficult*, *speed stress*, *no bias* to constitute both conditions here. Because this cell should contain enough data to be split into two conditions, we needed to collect more data in this cell, hence the addition of block 1 and 18 to the testing session.[3]

Order of Data Sets, Conditions, and Participants.   To obfuscate the design of the study, we shuffled the order of the 14 data sets before we shared them with the modeling experts. Furthermore, the order of conditions A and B was flipped for half of the data sets[4] For each data set, the participant numbers were shuffled, and the modelers were made aware that participant numbers were not consistent across the 14 data sets.

Invitation of Contributors

A large number of response time data analysis experts were invited via email. The invitation letter was and is still available on the open science framework page of this project (`https://osf.io/9v5gr/`). The invitation letter contained a description of the goal of the project and a very limited description of the method. We carefully wrote this invitation such that the contributors could only know that 1) the data sets comprised real data, 2) the task was a random dot motion task, 3) each two–condition experiment should be treated as a separate data set, and 4) the order of trials was shuffled for all participants, so that sequential effects should be ignored.

The requirements to contribute to this study were as follows. First, contributors were asked to submit a table containing for each experiment the inferences about each of four components of response time performance: ease, response caution, bias, and non–decision time. For each decision, as for each experiment, we gave contributors the option to submit a degree of confidence in their inference[5]. Second, contributors were asked to submit a description of their methods. This description of method was asked to meet a reasonable level of reproducibility, describing at least: 1) outlier removal procedures, 2) the applied mathematical model (if one was used), 3) the method of estimation (if applicable), and 4) the rules applied to draw inferences. Finally, contributors were asked to submit a short summary of their method. Some of the descriptions provided by the experts required further clarification, and these discussions were carried out via email. All contributors in this project are naturally included as co–authors.

Contributed Methods of Analyses

Table 3 summarizes the different approaches contributed by each contributor or team of contributors. In what follows, we give a relatively coarse overview of the different approaches used. A full description of the methods used is given by the authors themselves in the supplementary materials that are available on the open science framework (`https://osf.io/ktuy7/`).

---

[3]The code creating the data sets for the modelers accidentally contained a line deleting trials quicker than 200 ms. This resulted in the deletion of 0.16% of the trials, and so unlikely has influence on our results.

[4]The analysts received the data sets in order: 4, 5, 12, 3, 13, 11, 10, 8, 6, 1, 7, 9, 2, 14.

[5]We chose not to analyze the confidence ratings given by contributors, owing to the heterogeneity in responses. However, the inferences including confidence ratings are available in the supplementary materials on the project's osf page (`https://osf.io/ktuy7/`).

Table 3

Methods Used by Contributors.

Contributors	Code	Model	Estimation	Inference
Grasman	GR	Simple Diffusion	EZ2	E (Quade test on Ind.)
Krypotos & Wiecki	KW	Simple Diffusion	HB	E (Population Post.)
van Ravenzwaaij	RA	Simple Diffusion	HB	E (Bayesian t–test on Pop. Post.)
Vandekerckhove & Kupitz	VK	Simple Diffusion	HB	M (Model indicator parameter)
White	WH	Simple Diffusion	χ^2	E (Bayesian t-test over Ind.)
Hawkins	HA	Full Diffusion[1]	HB	E (Population Post.)
Leite	LE	Full Diffusion	χ^2	H (Parameter Estimates)
Starns	ST	Full Diffusion[1]	χ^2	E (Bayesian t–test over Ind.)
Vandekerckhove	VA	Full Diffusion[1]	ML[2]	M (Wald Test)
Voss & Lerche	VL	Full Diffusion	KS	E (Frequentist t-test over Ind.)
Annis & Palmeri	AP	LBA	HB	M+E (wAIC + Population Post.)[3]
Cassey & Logan	CL	LBA	HB	E (Population Post.)[3]
Lin & Heathcote	LH	LBA[4]	ML	M+E (AIC/BIC + ANOVA)
Trueblood & Holmes	TH	LBA[5]	HB	E (Population Post.)
Visser	VI	LBA	ML	M (Stepwise regression)
Evans & Brown	EB	–	–	H (Summary Statistics)
van Maanen	MA	–	–	H (Summary Statitsics)

Note: HB = Hierarchical Bayes. χ^2 = Chi–Squared. ML = Maximum Likelihood. EZ2 = Method of Moments estimation, as implemented in EZ2. KS = Kolmogorov–Smirnov. E = Estimate–based. M = Model selection. H = Heuristic–based. Pop = Population. Post = Posterior. Ind = Individuals. [1] Variability parameters fixed across conditions. [2] Data treated as one participant. [3] Assumed just one manipulation per experiment, unless extremely strong evidence otherwise. [4] Both LBA and Full diffusion model were fit, but the best-fitting model was used, and this was always LBA. [5] Bias in accumulation rate parameters. Modelers AP, CL, ST, VI, and KW were allowed 2 extra weeks after the initial deadline to hand in their inferences.

Model Choice

Contributors were roughly equally divided among three options - the Full Diffusion model, in which all of the between-trial variability parameters are estimated; the Simple Diffusion model, containing no between–trial variability parameters, and the LBA model. However, note that even though different contributors may have used a particular class of model (e.g., the LBA), there were differences in the exact instantiation of the model. For example, some of the contributors using the

Full Diffusion model chose to fix the between-trial variability parameters across the two conditions of each experiment. Contributors using the LBA also differed in the way that they parameterized the model. Finally, note that two contributors (EB and MA) used no computational model, and are indicated in Table 3 by a "–". Both teams of contributors using heuristic methods based their inferences on summary statistics, but their choices were driven by known behaviors of evidence-accumulation models.

Estimation Type

Of those who fit computational models, about half of the contributors estimated the parameters of the model using Hierarchical Bayesian methods, wherein posterior distributions for population–level parameters were inferred. The other half of contributors fit their models separately to each of the 20 individual participants using a range of methods, including maximum likelihood, chi–squared, method of moments, and Kolmogorov–Smirnov estimation.

Inference

There was a wide range of methods used for drawing inferences about the theoretical differences across conditions. We encourage readers to examine each of the authors' own descriptions of their methods, but we attempt to summarize them as follows. There were roughly three approaches: estimate-based, model selection, and heuristic-based methods (or some combination thereof). In the "estimate-based" approach, the contributors used differences in the estimates of model parameters to draw inferences about differences across conditions. Most contributors had a unique way of doing their estimate-based inference, but the most popular approach was to subject the individual-participant parameter estimates to a hypothesis test (Frequentist or Bayesian). The groups who had used Hierarchical Bayes to estimate population-level parameter values generally used the posterior distribution of differences across conditions to draw their inferences, but differed on their exact method. Other contributors used model selection, such as a model-indicator parameter in a Hierarchical Bayesian model, Backwards stepwise regression, or a hybrid between estimate-based and model selection such as AIC, BIC, or wAIC combined with hypothesis tests. Finally, a number of groups used Heuristic-based approaches, based on either estimated parameters, or a range of summary statistics calculated from the raw data. For example, if the proportion of correct responses differed by more than 5% between two conditions, then they were assumed to differ in ease.

Results

The first result worthy of mention is that the 17 teams of contributors utilized 17 unique procedures for drawing their inferences. That is, despite all attempting to solve the same problem, no two groups chose identical approaches. This result highlights the garden of forking paths that cognitive modelers face.

The complete set of inferences made by each team of contributors is given in full in Table 5. The column labelled True shows, for each experiment, for each parameter, which, if any, manipulation was made. Under each team of contributors (columns), the colored letters indicate the effect inferred by each modeling group. Inferences that are green are in line with the "True" manipulation, blue inferences indicate misses, orange inferences reflect false alarms, and black inferences are cases in which an effect was inferred in the wrong direction.
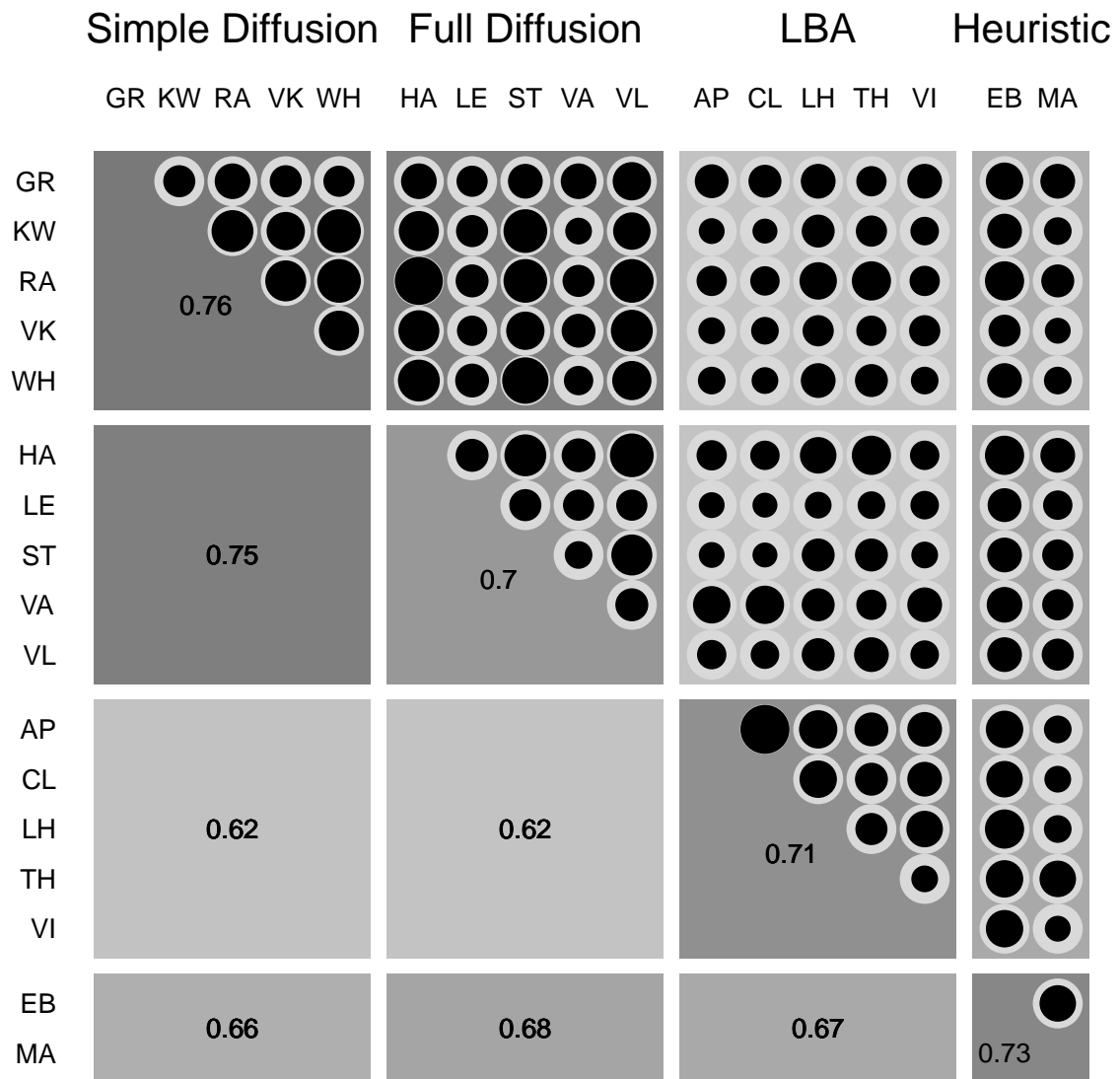
Figure 2. A visualization of the agreement between the different methods used. The radius of the black circles relative to the lighter colored background circles in the upper right of the matrix reflects the proportion of inferences shared between a pair of methods. The shade of the box underlying each set of points, and the numbers in the lower left of the matrix, depict the average of the proportion of shared inferences in each section. For example, the average proportion of shared inferences between all LBA and all simple diffusion models was 0.62.

Looking at Table 5, we see that many of the methods seem to make the same inferences, though there are also substantial differences. Figure 2 depicts the overall level of agreement between the inferences made by any pair of methods. The size of the circles in the upper diagonal of the matrix grow larger as the two methods yielded more similar inferences. For example, AP and CL differed only in terms of one inference (out of 56), and so the circle in that cell of the matrix is large.

The most notable pattern to come out of Figure 2 is that we see a lot of agreement within model classes. Diffusion models yielded similar inferences to one another, as did the LBA models. However, there is less agreement between the inferences coming from LBA and diffusion models. For example, the inferences from VL were similar to other diffusion model analyses, but were less consistent with the LBA or Heuristic methods. On a positive note, it is encouraging to see that although the models did sometimes reach different conclusions, there is some consensus in the inferences drawn. At worst, the overall agreement between model classes is 62%, indicating that even the methods that disagree most do complement each other reasonably well.

In what follows, we will unpack the reasons behind the patterns highlighted in Figure 2. We will focus on the performance of the different model classes, since we could identify clear differences between the patterns of inferences yielded by the different models. We did also look for systematic, aggregate-level influences of estimation or inference method, but could identify no strong, systematic effects. Any further attempt to look at select cases were compromised by the very many options available to researchers. Naturally, since the teams of contributors could select their own approach, we were unable to eliminate potential confounds between model choice, estimation, and inference method.

We begin with the analysis that was originally planned for this project, where we assume a selective influence of our experimental manipulations on the participants' behavior – for example, speed- and accuracy-emphasis instructions should influence only response caution. However, we also present the results of two alternative analyses that were based on alternative assumptions about what is the true effect of our manipulations. These alternative analyses were based on emails that we received from contributors before they submitted their inferences. As such, we note that these analyses are not entirely exploratory.

Planned Analysis of Validity

We start by considering the accuracy of the inferences under a selective influence assumption, in which difficulty affects ease, emphasis instructions affect caution, base-rates affect bias, and non-decision time was not manipulated in our experiments. The aggregate performance of all inferences is presented in Figure 3. The figure should be read as follows: Each row represents one of the 14 two-condition experiments. The four columns represent the four components of the response time task performance about which the collaborators drew their inferences. The grey letter to the left of each box shows the manipulated effect: an A indicates that condition A was manipulated to have a higher value than condition B on the component concerned; a B indicates that condition B had a higher value than condition A; a 0 indicates that both conditions had the same value. In other words, the grey letter indicates the "correct" inference. The size and location of the colored bars within the grey box indicate how many methods concluded in favor of inference A, 0, and B. The color of the bars indicate whether procedures concluded "correctly" (green), missed a manipulated effect (blue), detected an effect that was not manipulated (orange) or flipped the direction of a manipulated effect (black). Note that for each experiment, the largest bar thus indicates the inference made by the majority of approaches. This majority inference is not always the same as the "correct" inference.
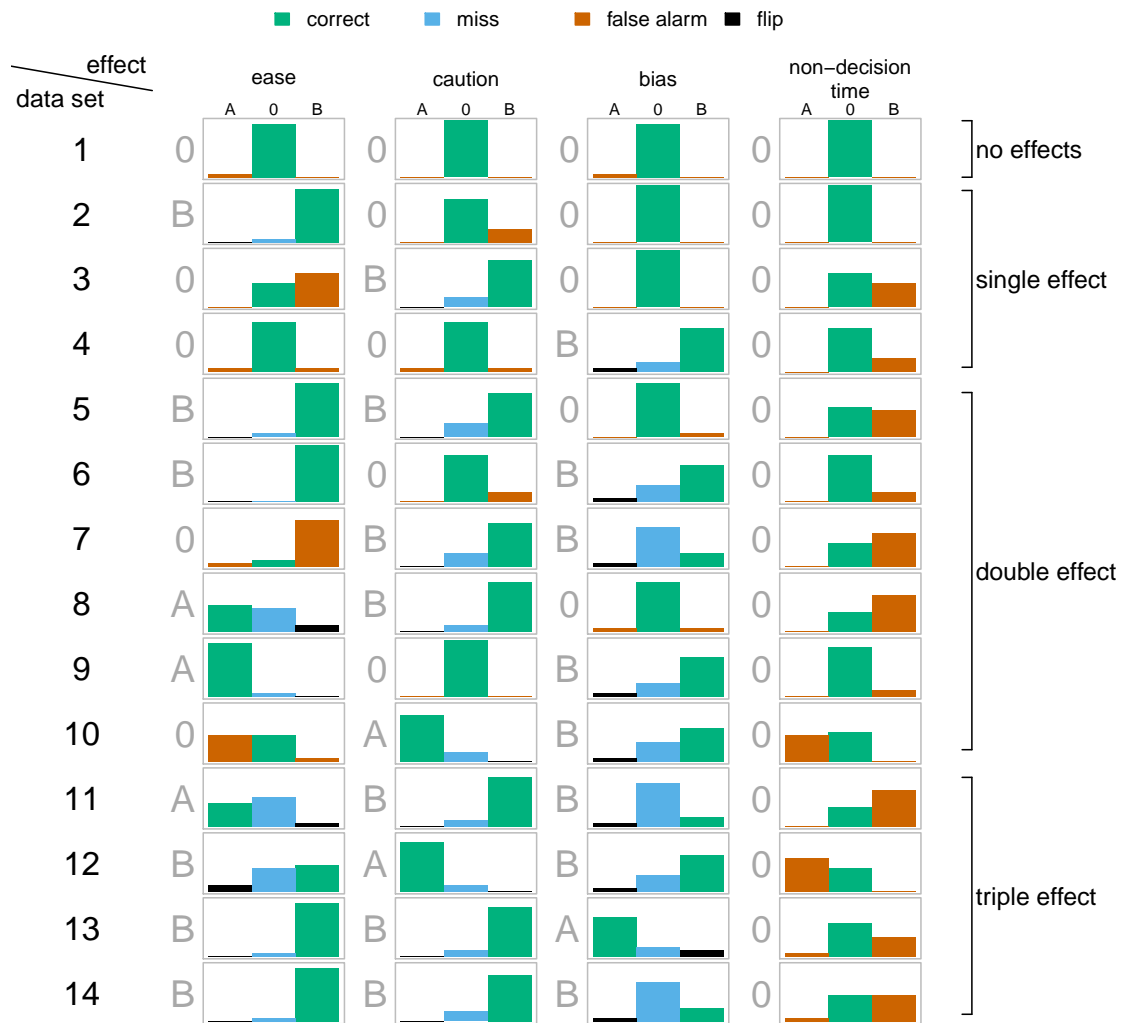
Figure 3. A summary of the inferences of all methods of analyses for all data sets. Grey letters in front of each box indicate the effect that was manipulated for the data set concerned. Bars indicate how many methods concluded for each of the options A, 0, and B. see text for details.

There is a lot of green in Figure 3, indicating that a large share of the inferences made by the contributors accurately reflect the manipulated effects. Importantly, in most cases there is a clear majority conclusion. Such agreement between methods is reassuring, given the issue of researcher degrees of freedom. That is, regardless of the choice of model, estimation method, or approach to inference, the conclusions drawn from the models tend to overlap substantially.

Figure 3 also reveals a fair number of incorrect inferences. Some of the erroneous inferences in Figure 3 are common to all methods. For example, the vast majority of methods inferred an effect of ease in Data Set 7, even though we did not make such a manipulation. In other cases, the incorrect inferences are not consensual. For example, in Data Set 7 the different contributors do not unanimously agree on whether there is a difference in non-decision time between the two conditions. The majority of what follows is an exploration of the systematic errors made by the different methods.

The overall accuracy of the inferences made by the teams of contributors are reported in Table 4. We now focus on the section labeled 'original', which reports the accuracy of the inferences when we assume selective influence. The method that performed best, according to the proportion of correct inferences, was submitted by GR (EZ2 estimation of simple diffusion model, 84% correct inferences). Out of 56 possible inferences, EZ2 yielded only 4 misses and 5 false alarms. Note that the overall accuracy of the EZ2 approach was more than two standard deviations better than the overall, average accuracy of 71%.

To get a sense of the accuracy for each model class, though potentially crude, we aggregate over all contributors using each model class. Averaging the accuracy of all contributors that used the full diffusion, we observe that 73% of their inferences were correct. The simple diffusion model performed similarly, with an average accuracy of 74%. The model-free approaches also performed relatively well (74% correct). Only the inferences based on LBA models were noticeably worse than average (66%), however, the LBA model did yield fewer false alarms (9.4%) than the simple and full diffusion model analyses (17.4% and 19.2%, respectively).[6]

The different models were systematic in the types of errors they produced. Consider first the incorrect inferences of the diffusion model. Whenever emphasis instructions were manipulated, the diffusion model incorrectly identified differences in non-decision time (cf. Voss et al., 2004 and Arnold et al., 2015). For illustration, consider Data Set 3 in Figure 3. Here, only caution was manipulated, and yet many methods inferred a manipulation of non-decision time. Though not clear from Figure 3, also looking at Table 5 indicates that all of these errors come about due to diffusion models. The same pattern is repeated in Data Sets 5, 7, 8, and 10-14: whenever caution was larger in one condition, the diffusion models tend to infer that non-decision time was also larger. Note that the EZ2 diffusion model was less likely to infer changes in non-decision time, and the lack of such errors is responsible for its superior performance.

The LBA model also made systematic errors in inference. The incorrect inferences made by LBA also relate to the manipulation of caution. Unlike the diffusion model, the LBA model often infers that manipulations of speed emphasis affect the ease of processing. For example, looking at Data Set 3 in Figure 3, we see although only caution was manipulated, many models inferred that ease was also manipulated. This issue is also present in Data Sets 7, 8, and 10-12, and in almost all

---

[6]Note that two of the contributors using the LBA, AP and CL, made the assumption that only one parameter would vary across the two conditions, necessarily lowering the false alarm rate. However, excluding their inferences does not change the overall conclusion that the LBA-based analyses yielded fewer false alarms. Indeed, throughout the paper, we looked at the impact of excluding AP and CL on the conclusions about model class and found it to have little impact.

cases, these incorrect inferences are produced by the LBA-based analyses.

The final source of systematic errors in inference are common to both LBA- and diffusion-based analyses, and relate to bias. We see in Data Sets 7, 11, and 14, that the bias manipulation is not detected by the majority of approaches. Unique to these experiments, both caution and bias were manipulated in the same direction. Returning to Table 2, we see that the data used to create these particular data sets involved a contrast between speed emphasis with no bias, and an accuracy emphasis condition with bias. Now, recall that our behavioral manipulation checks on hit and false alarm rates revealed no effect of the bias manipulation in the accuracy-emphasis condition. Therefore, the most likely reason that the models fail to detect a manipulation of bias in Data Sets 7, 11, and 14, is simply because the effect is not present in the data. It is worth noting explicitly that we expect no other issues with bias manipulations in any of the other data sets. In all other experiments in which bias was manipulated, our analysis of hit and false alarm rates suggested that there was a behavioral effect for the models to detect.

Post-hoc Analysis: Excluding non-decision time. Before we turn to the alternative methods for scoring the contributors' inferences, we first report an interesting post-hoc analysis of our results. If we simply exclude the inferences about non-decision time when assessing the accuracy of the different approaches, we see a dramatic increase in the performance of methods using the diffusion model. Looking at the final row of Table 4, we see that the full diffusion model used by ST draws correct inferences in 93% of cases. In fact, almost all diffusion model analyses are accurate ($> 80\%$) once non-decision inferences are ignored. The EZ2 diffusion model also fairs very well under this alternative scoring technique (86%). Of course, we chose to perform this analysis after having seen the data, and so this analysis is itself subject to issues with researcher degrees of freedom. Further, this analysis is only possible because we did not intend to manipulate non-decision time. As such, the results based on this post-hoc analysis should be taken with a grain of salt.

Alternative analysis I: Caution and Ease

A recent paper by Rae, Heathcote, Donkin, Averell, and Brown (2014) provided both empirical and model-based evidence that manipulations of speed- and accuracy-emphasis may influence both caution and the rate of evidence accumulation (see also Heitz & Schall, 2012; Starns, Ratcliff, & McKoon, 2012; Vandekerckhove, Tuerlinckx, & Lee, 2008).[7] Intuitively, when emphasizing accuracy, participants may also try harder to do the task, in addition to collecting more evidence before responding. As such, we now consider the accuracy of the inferences made by the teams of contributors when we assume that emphasis manipulations should affect both caution and ease of processing. For example, we manipulated only speed emphasis in Data Set 3, but under this alternative scoring scheme we now consider the correct inference to be that both ease and caution was larger in condition B. Note that in Data Sets 8, 11, and 12, the manipulations of caution and ease are in opposite directions, making it difficult to rescore those conditions, and so we exclude the analysis of these experiments in this section.

The second section of Table 4 reports the accuracy of the inferences under this alternative scoring method. First of all, note that the average performance of the methods increases from 75% under the original scoring, excluding Data Sets 8, 11, and 12, to 85% under these alternative scoring rules. Indeed, all models appear to do well under this scoring method (LBA: 76%, simple diffusion:

---

[7]The Rae et al. (2014) paper also found that emphasis instructions affected non-decision time, but since this was not the focus of the email that prompted this analysis, nor of the Rae et al. manuscript, we do not consider an analysis in which emphasis manipulation influences drift rate, boundary separation, and non-decision time.

78%, and full diffusion: 77%; model-free methods: 84%). Such an improvement in performance could be consensual evidence from the models to indicate that emphasis instructions do indeed influence both caution and ease.

The most accurate inferences under this alternative scoring scheme is the model-free method used by EB, whose inferences were accurate in 91% of cases. It is worth noting that their model-free method is based on years of experience with model-based methods. Interestingly, the EZ2 model remains among the better performers under this alternative scoring method, maintaining an accuracy of 80%. It should not be surprising that the LBA model performs better under this scoring scheme, since the major failing of the LBA model under the original scoring rule was that it would detect ease manipulations when caution was manipulated.[8]

Alternative analysis II: Caution and Non-decision time

Rinkenauer, Osman, Ulrich, Müller-Gethmann, and Mattes (2004) provide neuroimaging-based evidence that manipulations of speed- and accuracy-emphasis may affect both caution and non-decision time. Intuitively, when asked to be more accurate, participants may take some additional time when making their motor response, for example, checking that the button they intend to press is appropriate. We now consider the accuracy of the contributors' inferences based on the assumption that manipulations of speed- and accuracy-emphasis influence both caution and non-decision time. For example, in Data Set 3, it would be correct to identify that condition B showed both more caution and a slower non-decision time.

The overall accuracy of inferences in this alternative analysis is largely unchanged from the original analysis (72% and 71% for the alternative and original analyses, respectively). However, the average performance of all methods is misleading, since some methods perform much better under this alternative scoring, while other methods perform much worse. The simple diffusion model performs very well under this particular coding scheme, with an average performance of 83%. The full diffusion model also performs well, with an average accuracy of 78%. The model-free and LBA-based analyses, on the other hand, perform considerably worse under this alternative scoring (66% and 56%, respectively). It is worth noting that, unlike in the previous analyses, the EZ2 model performs relatively poorly under this alternative analysis (68% accuracy). Again, this pattern of results is not surprising, but simply reflects the advantage offered to models that detect non-decision time differences whenever caution is manipulated (i.e., the diffusion models, but not EZ2 or LBA models).

Discussion

Summary and Recommendations

In this project, we studied the validity of inferences that we draw from response time data when we apply various models. For this goal, seventeen teams of response time experts analyzed the data from 14 two-condition experiments. For each data set, the experts were asked to infer

---

[8]The two alternative scoring rules are based on the assumption that the models are telling us about regularities in the world – i.e., that emphasis instructions influence both ease and boundary separation. It is also possible that such inferences are due to correlations between model parameters. For example, the estimates of accumulation rate and threshold are correlated in an LBA model, while boundary separation and non-decision time are correlated in a diffusion model. It is impossible to disentangle such issues in our data, and so this alternative interpretation of our results is worth keeping in mind.

Table 4

Summary statistics for the quality of inferences drawn using each method. Statistics are shown for three different scoring keys (planned: assuming selective influence; alternative 1: assuming caution manipulations affected also ease; alternative 2: assuming caution manipulations affected also ndt) as well as for the planned key when ignoring non-decision time inferences. Methods are sorted by the applied RT model, from left to right: simple diffusion model, full diffusion, LBA, and model–free. See text for details.

key		Simple Diffusion					Full Diffusion					LBA					No Model	
		GR	KW	RA	VK	WH	HA	LE	ST	VA	VL	AP	CL	LH	TH	VI	EB	MA
planned	correct	0.84	0.73	0.73	0.66	0.75	0.73	0.71	0.75	0.77	0.71	0.66	0.64	0.70	0.62	0.68	0.77	0.70
	miss	0.07	0	0.07	0.12	0.04	0.09	0.05	0	0.20	0.09	0.29	0.30	0.12	0.23	0.07	0.12	0.20
	FA	0.09	0.25	0.20	0.21	0.21	0.18	0.20	0.25	0.04	0.20	0.05	0.05	0.12	0.14	0.11	0.11	0.09
alternative 1	correct	0.80	0.82	0.80	0.66	0.82	0.77	0.75	0.82	0.75	0.73	0.75	0.73	0.77	0.77	0.77	0.91	0.77
	miss	0.11	0	0.07	0.16	0.05	0.09	0.05	0	0.20	0.11	0.25	0.27	0.11	0.16	0.07	0.07	0.18
	FA	0.09	0.16	0.14	0.18	0.14	0.14	0.11	0.18	0.05	0.16	0	0	0.09	0.07	0.07	0.02	0.05
alternative 2	correct	0.68	0.86	0.89	0.82	0.91	0.89	0.70	0.91	0.64	0.84	0.50	0.48	0.64	0.68	0.52	0.68	0.64
	miss	0.23	0.02	0.07	0.12	0.04	0.09	0.12	0	0.34	0.11	0.45	0.46	0.23	0.29	0.23	0.25	0.30
	FA	0.09	0.11	0.04	0.05	0.05	0.02	0.11	0.09	0.02	0.05	0.05	0.05	0.07	0.04	0.11	0.07	0.04
no ndt	correct	0.86	0.83	0.86	0.76	0.88	0.86	0.74	0.93	0.71	0.86	0.55	0.52	0.67	0.64	0.57	0.74	0.71

which of four potential factors were manipulated between conditions. Of foremost importance is that inferences of this kind are not possible without a cognitive model. The contributors who used Heuristic methods also based their methods on years of experience with such models.

The first result worthy of mention is that no two teams spontaneously adopted the exact same approach to answering this question. Rather, we saw a variety of different models, estimation techniques, and inference methods across the different groups. However, despite the variety of methods, in general, we saw that the modeling teams reached a strong consensus over which manipulations we made, even when the inferences were not "correct".

Overall, regardless of the scoring method we used, inferences from the diffusion model tended to be accurate. Of the model-based analyses, the simple and full diffusion models have the highest accuracy across all four alternative scorings that we considered. Further, there appears to be considerable agreement between the inferences made by the different diffusion model analyses. Given the similarity between inferences drawn from the diffusion model, it appears that the conclusions drawn are robust to many of the choices available to researchers.

Many methods using the diffusion model detected effects on non–decision time where response caution was manipulated. This result is consistent with empirical evidence suggesting that manipulations to increase caution do also increase non–decision time (Rinkenauer et al., 2004), as well as previous validation studies (Voss et al., 2004). As such, we may want to apply a more careful interpretation of results such as those in Dutilh, Vandekerckhove, Tuerlinckx, and Wagenmakers (2009), in which both boundary separation and non-decision time were found to decrease with practice on a task. A more conservative interpretation of these results may be that only caution was changing with practice, and the effect manifested in both boundary separation and non-decision time parameters.

Both simple and full diffusion models tended to provide robust and valid inferences. Therefore, we do not find any evidence that the additional assumptions of the full diffusion model improve inferences about differences in the core latent variables of interest – ease, caution, bias, and non-decision time (see also van Ravenzwaaij, Donkin, & Vandekerckhove, Manuscript submitted

for publication). Similarly, we find that easy-to-implement estimation methods, such as the EZ2 method, tend to provide inferences that are as valid as the more complex estimation techniques. For those lacking the computational expertise to implement more complex approaches, the EZ2 method may be a suitable method for making inferences about processes giving rise to response time data (we hope, on the way to learning how to apply the models more generally).

We observed that, under assumptions of selective influence, the EZ2 method outperforms other methods that estimate the parameters of the simple diffusion model (cf. Arnold et al., 2015). This benefit likely comes about because the EZ2 method uses robust statistics, such as means and variances, to estimate parameters. Alternative methods estimate their parameters through more variable statistics, such as the minimum response time (or the 10th percentile of the response time distribution), and so are more likely to infer differences in parameters across conditions. The use of robust statistics might also partly underlie the relative success of the model–inspired, but heuristic approach followed by contributors EB, who based their inferences largely on median RT and accuracy.

The LBA model yielded accurate inferences under the first alternative scoring rule, where emphasis instructions were assumed to influence both caution and ease. Given that all models performed well under such a scoring scheme, it could be that this set of assumptions best reflects the true state of the world. That is, it may be that manipulating speed- and accuracy-emphasis really affects both caution and accumulation rate. This result is also in line with earlier findings, as described above (e.g., Arnold et al., 2015; Rae et al., 2014; Starns et al., 2012; Vandekerckhove et al., 2008), but raises the question of whether and how response caution can be manipulated selectively.

Systematic differences between inferences based on the various estimation or inference methods were difficult to extract. As such, no strong recommendations come out of our study. We recommend that choices over estimation and inference be based on their suitability for the research question at hand. For example, if one has data from multiple participants each contributing only a small number of trials, then hierarchical methods may be best (e.g., Krypotos, Beckers, Kindt, & Wagenmakers, 2015). Bayesian hypothesis testing allows for evidence in favor of a null hypothesis, which is relevant when no difference between conditions is of theoretical interest. Also note that researchers can use our Tables 5 and 4 to gauge the performance of any given method. For example, having decided to use a full diffusion model, we see that using chi-squared or Kolmogorov-Smirnov estimation, and Bayesian or Frequentist t-tests all led to accurate inferences.

Limitations of our study

It is important to make clear the boundary conditions on the generality of our conclusions. The most limiting factor of our study is that all of the experiments had just two conditions. We chose such a simple design because it is ideal for an initial large-scale validation study. However, the extent to which our results generalize to more complex designs, studies with different numbers of trials and participants, or different tasks, remains an exciting avenue for future research. It remains to be seen whether the simple diffusion model can maintain its relatively good performance for more complex, factorial designs. We note that our entire factorial data set is available from the open science foundation archive, and so provides an open data set for researchers interested in pursuing such issues.

We expect that more complex models, estimation techniques, and inference methods would do better under more complex experimental designs. For example, in a factorial experiment in which a single parameter can be mapped naturally onto changes in one factor, the parameters that explain

differences in the other factor should be more readily identifiable. This notion has become folklore among response time modelers (e.g., Donkin, Brown, & Heathcote, 2011, 2009). Therefore, in the same way that a signal detection analysis is rarely the ultimate analysis, the simple diffusion model may not always be the best available method. That said, a simple diffusion model analysis is vastly superior to any model-free method should one wish to make claims about latent processes.

Some of the contributors indicated that their choice of analysis was influenced by the special setting of this collaborative project. First, they had a rather short time window of 3 months to perform their analyses. Some contributors indicated that with more time, they would maybe have invested time reconsidering various model choices. Second, the comparative nature of the project may have encouraged the use of relatively novel approaches from some contributors. It is possible that under more typical settings, the range of estimation and inference methods may be less variable. The researcher degrees of freedom on display in our project may be more variable than what would be typically encountered in the literature.

For reasons outlined earlier, we chose to keep the modeling teams blind to the experimental manipulations we made. Such blinding has desirable properties for validity studies, but creates a rather difficult and maybe artificial inferential process for modelers. Indeed, because the analysts were completely blind to the manipulations, their analyses were entirely exploratory. In "real applications", where researchers have well–informed prior expectations to guide their inference, we expect inferences to be more accurate (though also likely more biased). For example, in non-blinded applications we would not often see inferences in which the direction of an effect were "flipped". Potentially, prior information about experimental manipulations can correct for some of the issues we discussed earlier. For example, in a Bayesian analysis, such information may help inferences by setting informative priors to convey sensible expectations about non-decision time. Alternatively, non-decision time parameters may be fixed across conditions in which they are not expected to change.

Application Beyond Response Time Data Inference

Beyond this article's direct importance for response time data analysis, this project illustrates the strength of blinded analyses in a many–lab collaboration. The blinding allows us to draw much *stronger* conclusions than would have been possible if the same researchers who designed the experiments had also analyzed the data. The collaborative nature allows us to draw much *broader* conclusions than would have been possible if only one lab had designed and applied a method of analysis. In particular for studying research questions where some answers are more desirable than others, this approach is invaluable. Even in the case of validating response time models, some results are more desirable than others, and so the blinded analysis presumably protects against bias. The inspiring study by Silberzahn et al. (2015) applied "statistical crowdsourcing" to address the question whether soccer players with a dark skin tone receive more red cards than players with a light skin tone. In such controversial cases, a blinded analysis would be superior to the original unblinded approach. For this reason, we hope that future validation studies will adopt this approach and improve the reliability and generalizability of their results.

Conclusions

This study offers a better understanding of the validity of inferences that we draw when using cognitive models to analyze response time data. Indeed, the knowledge gained from this project

will still grow when analysts apply new or improved methods to the openly available data sets and contribute the results to the open science framework project page[9]. The current results justify the increasing popularity of response time modeling techniques to a large extent. Future efforts should lie an even more solid basis for the advancement of response time modeling techniques throughout psychological science.

References

Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter–estimation methods. *Psychological research*, *79*, 882–898.

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117–128.

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

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience*, *14*, 1462–1467.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*, 140–151.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus Linear Ballistic Accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin and Review*, *18*.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*, 1129–1135.

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task–related vs. stimulus–specific practice: A diffusion model account. *Experimental Psychology*, *58*, 434–442. doi: 10.1027/1618-3169/a000111

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036.

Forstmann, B. U., Dutilh, G., Brown, S. D., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*, 17538–17542.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p–hacking" and the research hypothesis was posited ahead of time.* (Department of Statistics, Columbia University in the City of New York. Retrieved from http://www.stat.columbia.edu/ gelman/ research/unpublished/p_hacking.pdf)

Grasman, R. P., Wagenmakers, E.-J., & van der Maas, H. L. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, *53*, 55–68.

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*, 616–628.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion–model analysis. *Journal of Personality and Social Psychology*, *93*, 353–368.

Krypotos, A.-M., Beckers, T., Kindt, M., & Wagenmakers, E.-J. (2015). A bayesian hierarchical diffusion model decomposition of performance in approach–avoidance tasks. *Cognition and Emotion*, *29*, 1424–1444.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences*, *112*, 10645–10650.

---

[9] `https://osf.io/jy7pi/`. An example of such contribution is already submitted post hoc by contributor LE, who corrected a typo, thereby improving his inferences considerably.

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.

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology*, *27*, 193-200.

Morey, R. D., Rouder, J. N., & Jamil, T. (2014). Bayesfactor: Computation of Bayes factors for common designs. *R package version 0.9*, *7*. Retrieved from `http://CRAN.R-project.org/package=BayesFactor`

Mulder, M., Van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences–a model-based review. *Neuroscience*, *277*, 872–884.

Peirce, J. W. (2007). PsychoPy–psychophysics software in Python. *Journal of neuroscience methods*, *162*, 8–13.

Pew, R. W. (1969). The speed–accuracy operating characteristic. *Acta Psychologica*, *30*, 16–26.

Rae, B., Heathcote, A., Donkin, C., Averell, 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*, 1226–1243.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 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*, 278–291.

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial–to–trial variability of the eeg. *Proceedings of the National Academy of Sciences*, *106*, 6539–6544.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*, 408–424.

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

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

Rinkenauer, G., Osman, A., Ulrich, R., Müller-Gethmann, H., & Mattes, S. (2004). On the locus of speed–accuracy trade-off in reaction time: inferences from the lateralized readiness potential. *Journal of Experimental Psychology: General*, *133*, 261–282.

Silberzahn, R., Uhlmann, E., Martin, D., Anselmi, P., Aust, F., Awtrey, E., . . . others (2015). *Many analysts, one dataset: Making transparent how variations in analytical choices affect results.*

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal–variance and dual–process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*, 1–34.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Pshychological Review*, *108*, 550–592.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*, 1017–1025.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011–1026.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. (2008). A Bayesian approach to diffusion process models of decision–making. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1429–1434).

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two–choice response times. *Psychological methods*, *16*, 44–62.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (Manuscript submitted for publication). The EZ-diffusion model provides a powerful test of simple empirical effects.

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.

Verdonck, S., & Tuerlinckx, F. (2014). The ising decision maker: A binary stochastic network for choice response time. *Psychological review*, *121*, 422–462.

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

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

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior research methods*, *46*, 15–28.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*, 779–804.

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.

Wagenmakers, E.-J., Van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ–diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift–diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, doi:10.3389/fninf.2013.00014.

Table 5

Performance of the different methods. Column "True" shows for each experiment, for each component the manipulation in the data set. Letters indicate inferences made by the analysts; green indicate correct, blue misses, orange false alarms, black cases where there direction of the effect was flipped. Methods are sorted by the applied RT model, from left to right: simple diffusion model, full diffusion, LBA, and model–free.

Data Set	Effect	True	GR	KW	RA	VK	WH	HA	LE	ST	VA	VL	AP	CL	LH	TH	VI	EB	MA
			Simple Diffusion					Full Diffusion					LBA					No Model	
1	ease	0	0	0	0	0	0	0	A	0	0	0	0	0	0	0	0	0	0
	caution	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	bias	0	0	0	0	0	0	0	0	0	A	0	0	0	0	0	0	0	0
	ndt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	ease	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0	B
	caution	0	0	0	0	B	0	0	0	0	0	0	0	0	B	0	B	B	0
	bias	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ndt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	ease	0	0	B	B	0	B	0	0	B	0	0	B	B	B	B	B	B	0
	caution	B	B	B	B	B	B	B	B	B	B	B	0	0	B	0	B	B	B
	bias	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ndt	0	0	0	B	B	B	B	0	B	0	B	0	0	0	B	0	0	0
4	ease	0	0	0	0	0	0	0	B	0	0	0	0	0	0	0	A	0	0
	caution	0	A	B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	bias	B	B	B	B	0	B	B	B	B	0	B	B	0	B	B	A	B	B
	ndt	0	B	0	0	0	0	0	0	B	0	B	0	0	0	0	0	0	B
5	ease	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0
	caution	B	B	B	B	B	0	B	B	B	B	B	0	0	0	B	B	B	B
	bias	0	0	0	0	0	0	0	B	0	0	0	0	0	0	0	0	0	0
	ndt	0	0	B	B	B	B	B	0	B	0	B	0	0	B	0	0	0	0
6	ease	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
	caution	0	0	B	0	B	0	0	0	0	0	0	0	0	0	0	B	0	0
	bias	B	B	B	B	0	B	B	B	B	0	0	0	0	B	B	A	B	B
	ndt	0	B	0	0	0	0	0	0	B	0	0	0	0	0	0	0	0	B
7	ease	0	B	B	B	B	B	B	A	B	0	B	B	B	B	B	B	B	0
	caution	B	B	B	B	B	B	B	B	B	B	B	0	0	0	0	B	B	B
	bias	B	0	B	0	B	B	0	A	B	0	0	0	0	0	0	0	0	0
	ndt	0	0	B	B	B	B	B	0	B	B	B	0	0	B	B	0	0	0
8	ease	A	A	A	A	A	A	0	A	A	0	A	0	0	B	0	B	0	0
	caution	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0	B	B	0
	bias	0	0	B	0	0	0	0	A	0	0	0	0	0	0	0	0	0	0
	ndt	0	0	B	B	B	B	B	B	B	0	B	0	0	0	B	0	B	B
9	ease	A	A	A	A	A	A	A	A	A	A	A	A	A	A	0	A	A	A
	caution	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	bias	B	B	B	B	0	0	B	0	B	0	B	B	B	B	B	A	B	B
	ndt	0	B	0	0	0	0	0	0	0	0	B	0	0	0	0	0	0	0
10	ease	0	0	A	0	0	A	0	B	A	0	0	A	A	A	0	A	A	0
	caution	A	A	A	A	A	A	A	A	A	A	A	0	0	A	0	A	A	A
	bias	B	0	B	B	0	B	B	B	B	0	B	0	0	A	B	0	B	B
	ndt	0	0	A	A	A	A	A	0	A	0	A	0	0	0	A	0	0	0
11	ease	A	A	A	0	A	A	0	A	A	0	A	0	0	0	0	B	0	0
	caution	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0	B	B	0
	bias	B	0	B	0	0	B	0	0	B	0	0	0	0	0	0	0	0	A
	ndt	0	0	B	B	B	B	B	B	B	0	B	0	0	0	B	0	B	B
12	ease	B	B	B	B	B	B	B	0	B	0	B	0	0	A	0	A	0	0
	caution	A	A	A	A	A	A	A	A	A	A	A	A	A	A	0	A	A	0
	bias	B	B	B	B	0	B	B	B	B	0	0	0	0	B	B	A	B	B
	ndt	0	0	A	A	A	A	A	A	A	0	A	0	0	0	A	0	0	A
13	ease	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0
	caution	B	B	B	B	B	B	B	B	B	B	B	0	0	B	B	B	B	B
	bias	A	A	A	A	A	A	A	B	A	0	A	0	0	A	A	B	A	A
	ndt	0	0	B	B	B	B	B	A	B	0	0	0	0	0	0	0	0	0
14	ease	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	0
	caution	B	B	B	B	B	B	B	B	B	B	B	0	0	0	B	B	B	B
	bias	B	0	A	0	0	B	0	B	B	B	0	0	0	0	0	0	0	0
	ndt	0	0	B	B	B	B	B	A	B	0	B	0	0	B	0	0	0	0