Doctoral thesis submitted to
the Faculty of Behavioural and Cultural Studies
Heidelberg University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Dr. phil.)
in Psychology

Title of the publication-based thesis

Parameter Estimation in Diffusion Modeling: Guidelines on Requisite Trial Numbers and Estimation Procedures

presented by
Veronika Lerche

year of submission
2016

Dean: Prof. Dr. Birgit Spinath
Advisor: Prof. Dr. Andreas Voß

Table of Contents

Acknowledgments

First, I would like to thank Prof. Dr. Andreas Voß for the great supervision during my dissertation project. I am very grateful for your fantastic support, Andreas, during all phases of the dissertation. I learned a lot from you in these past years, about diffusion modeling as well as about research designs, programming, and data evaluation. Thank you also for giving me the possibility of pursuing other research projects and teaching activities at the same time. By the way, I have to correct you on one thing you said during my job interview: Diffusion models are not such an "exotic" topic anymore! ☺

I am greatly indebted to Dr. Jan Rummel who immediately agreed to review this thesis. Thank you also a lot for introducing the excellent "Studienportal" right at the moment when I planned to start my first data collection. It was and still is extremely useful!

Many thanks further go to Dr. Andreas Neubauer and Hayley Jach for your excellent proof reading of this thesis. Besides, I want to thank our research assistants Anne Bülow, Sarah Hladik, Julia Karl, Stephanie Schuchhardt, and Jasmin Stein for your great support during data collection.

Furthermore, I am very grateful to my great colleagues and friends Alica Bucher, Dr. Tanja Kollei, Ulf Mertens, Markus Nagler and Dr. Andreas Neubauer. Eating wraps, constructing paper bananas, playing ping-pong and sophisticated versions of Cyberball, and discussing about the silliest future research projects has been the perfect distraction from my otherwise very serious activities as a diffusion modeler. I also want to thank my lovely "Fremdessen"-appointments Shahrzad Izadpanah, Dr. Cornelia Schwarze, Dr. Rosalux Falquez, Ivan Marevic and Lena Steindorf for the great moments at the various Heidelberg locations.

Inoltre vorrei ringraziare Francesco Mazzella per i bei momenti vissuti insieme durante questi anni del mio dottorato di ricerca. Sono particolarmente grata per la tua pazienza e comprensione per i miei orari di lavoro che—non raramente—si estendevano anche nel fine settimana. Finally, I am deeply grateful to my parents who were of invaluable support (certainly not only) during these last years.

List of Scientific Publications of the Publication-Based Dissertation

Manuscript 1

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

Manuscript 2

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*(336). doi: 10.3389/fpsyg.2015.00336

Manuscript 3

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. doi: 10.3758/s13428-016-0740-2

Manuscript 4

Lerche, V., & Voss, A. (2016). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24. doi: 10.1007/s00426-016-0770-5

Manuscript 5

Lerche, V., & Voss, A. (2016). Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious. *Frontiers in Psychology*, *7*(1324). doi: 10.3389/fpsyg.2016.01324

1 Introduction

Within the last 15 years, across diverse psychological subfields, a formal mathematical model of response times and accuracy called the *diffusion model* has become increasingly popular (Voss, Nagler, & Lerche, 2013). The diffusion model (sometimes also termed *Ratcliff diffusion model* or *drift diffusion model*) was originally mainly used to investigate the basic cognitive processes underlying memory and simple perceptual decision-making (e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999). While there remains substantial implementation of the diffusion model in these areas (e.g., Ratcliff & McKoon, 2015; Ratcliff, Thompson, & McKoon, 2015; Starns, Ratcliff, & White, 2012), also researchers of other fields, such as clinical psychology, neuropsychology, or social psychology, are beginning to adopt the model to address their specific research questions (e.g., Aschenbrenner, Balota, Gordon, Ratcliff, & Morris, 2016; Germar, Schlemmer, Krug, Voss, & Mojzisch, 2014; Weigard & Huang-Pollock, 2014). There are two main reasons for this development. Firstly, more researchers are coming to appreciate the diffusion model’s facility to disentangle the cognitive processes involved in binary decision tasks, thereby allowing these researchers to suggest and investigate candidate cognitive mechanisms to explain the data that they observe. Secondly, in the last decade, several software solutions were developed (e.g., Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2007) that made the implementation of the diffusion model simpler and more streamlined. In the past, the diffusion model has generally only been able to be used by researchers with ample experience in mathematical modeling and with rich programming skills. However, with the development of these new programs, it became simpler for less technically-minded researchers to reap the benefits of the diffusion model’s differential equations.

The increase in use of the diffusion model is a positive development as, by means of the application of the model to diverse fields, more profound knowledge about the aspects underlying decision processes can be gained. However, this development also goes along with a certain risk: Researchers with restricted knowledge about the diffusion model might feel tempted to apply the model to their data without being sufficiently informed of the reliability of different approaches. Unfortunately, a lot of knowledge about procedures of diffusion modeling that experts have gained throughout the years has not been written down (or at least not in an easily comprehensible way), and is thus not accessible to newcomers. In addition, even experts differ largely in their strategies of modeling response time (RT) data as the variety of approaches taken by different research teams in a recent Open Science project demonstrates (Dutilh et al., 2016; see Chapter 8.2 for a description of this project). As the varieties of modeling approaches used increases alongside the number of published studies of the diffusion model, it is imperative that best-practice recommendations are developed for the future implementation of the diffusion model in experimental paradigms. In this thesis, I aim to provide a first set of guidelines towards this purpose. These guidelines are intended both for researchers who are new to diffusion modeling, as well as for experienced researchers who are interested in ascertaining whether their approach truly leads to reliable parameter estimates.

One core element that I tackle in my thesis concerns the minimum acceptable number of trials required for an experiment. Earlier studies were often based on a small number of participants and a large number of trials. For example, a study by Ratcliff (2002) was conducted with only three participants, who each worked on more than 10,000 trials (see also, for example, Ratcliff, 1981; Ratcliff & Rouder, 1998; Ratcliff et al., 1999). In contrast, now it is becoming more common to employ many fewer trials and many more participants (e.g., 100 trials and 120 participants in a study by Metin et al., 2013). Indeed, for many research questions, trial numbers of several hundreds or even thousands are not feasible. For example, clinical populations may not have the capacity to participate in a response time task for an hour or more. In addition, cognitive processes might change with the time spent on the task (e.g., Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009). For example, attention might decrease, and people might become more distracted or more practiced. Moreover, in certain cases, it is difficult to generate a sufficient number of stimuli, since the material is specific and limited. In sum, it might not always be possible or beneficial to use very high trial numbers. In this thesis, I systematically tackle the question of whether such high trial numbers are truly necessary for reliable parameter estimation.

As I previously mentioned, there are a variety of modeling methods, and there are already a few studies that have compared the performance of some of these different methods (e.g., Ratcliff & Childers, 2015; van Ravenzwaaij & Oberauer, 2009). However, these studies bear one major limitation: They compared methods that are implemented in different programs. Accordingly, in these studies, the results cannot be clearly attributed to the method applied, but might also be a result of program specifications. To avoid this potential confound, we employ a single program—*fast-dm* (Voss & Voss, 2007, 2008)—to analyze all different methods of implementing the diffusion model. In order to analyze a variety of different approaches, we added two further methods to this program: specifically, *fast-dm-30* (Voss, Voss, & Lerche, 2015) now includes the Kolmogorov-Smirnov approach (KS;

implemented also in the former versions of fast-dm), as well as a chi-square (CS) and a maximum likelihood based approach (ML). CS is probably the method that has been used most frequently in the diffusion model literature (e.g., Ratcliff & McKoon, 2008; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; White, Ratcliff, Vasey, & McKoon, 2009). KS has also been applied in a number of studies (e.g., Aschenbrenner et al., 2016; Bowen, Spaniol, Patel, & Voss, 2016; Horn, Bayen, & Smith, 2011). ML, on the other hand, has been used only very rarely so far (for an exception, see Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007). In my thesis, I compare the estimation performance of these three different optimization criteria.

Diffusion modelers do not only use different optimization criteria, but also models of different complexity. Whereas the so-called *basic diffusion model* consists of only four parameters, the *full diffusion model* further includes intertrial variabilities of three of the four main diffusion model parameters. In most applications, the full diffusion model is used (e.g., Allen, Lien, Ruthruff, & Voss, 2014; Dutilh et al., 2009; Ratcliff, Thapar, & McKoon, 2004; Ratcliff & Van Dongen, 2009). However, two of the intertrial variabilities are estimated very inaccurately as the results from simulation studies demonstrate (e.g., van Ravenzwaaij & Oberauer, 2009; Vandekerckhove & Tuerlinckx, 2007). This finding suggests that it might be better to fix these parameters to improve the estimation of the psychologically more interesting parameters. Accordingly, in some more recent studies, these intertrial variabilities have been fixed at zero (e.g., Germar, Albrecht, Voss, & Mojzisch, 2016; Hartanto & Yang, 2016; Schubert, Frischkorn, Hagemann, & Voss, 2016). The necessity of the estimation of the intertrial variabilities has not yet been examined systematically. In my thesis, I therefore compared the estimation performance of models of different complexity (e.g., with fixations of intertrial variabilities) assuming that the data were generated based on a full diffusion model. That is, I analyzed whether *false fixations* (false because the true model included substantial intertrial variabilities) can lead to equally or even more reliable estimates of the four main diffusion model parameters.

Thus, the focus of this thesis is on (a) the deduction of requisite trial numbers, and (b) a comparison of the performance of different estimation procedures (most importantly, different optimization criteria and models of different complexity) depending on the trial number. The estimation performance can be evaluated by means of different criteria, such as deviations or correlations between true (i.e., data generating) and re-estimated parameters in simulation studies. In addition to the more common simulation studies, I also present an approach of using test-retest data for the evaluation of estimation accuracy. The aim of these analyses is to develop a set of guidelines on how to estimate diffusion model parameters reliably.

The thesis is structured in the following way: First, a short introduction to the diffusion model is given which is complemented by Manuscript 1—an introductory article for newcomers to diffusion modeling. In the subsequent chapter, the extension of the program, *fast-dm-30*, is presented (Manuscript 2). After that, I will give an overview of criteria that can be used to evaluate the performance of parameter estimation procedures. This is followed by a presentation of the main results from simulation studies (Manuscript 3) and test-retest studies (Manuscript 4). A reanalysis of data from Manuscripts 3 and 4 using models of different complexity is presented in the subsequent chapter (Manuscript 5). More detailed information regarding the topics of these chapters can be found in the five manuscripts that are attached to the thesis. Finally, in the discussion, apart from summing up the main guidelines from this thesis, I will present ideas for future research projects aimed at extending the guidelines, including also some first new findings.

2 Introduction to Diffusion Modeling (Manuscript 1)[1]

The diffusion model is a model of two-choice decision-making. It is part of the class of sequential sampling models (see Ratcliff & Smith, 2004; Ratcliff, Smith, Brown, & McKoon, 2016, for a comparison of different sequential sampling models). In this thesis, I will refer to the diffusion model that has been proposed by Roger Ratcliff for memory retrieval (Ratcliff, 1978) and that has been greatly influenced by earlier work by Laming (1968) and Link and Heath (1975). In the almost 40 years since the influential article by Roger Ratcliff, the literature on diffusion modeling has grown rapidly, especially in the last 15 years, as the citation rates of this article in the PsycINFO database demonstrate (see Figure 1, Voss et al., 2013). Currently, the citations have reached a total number of 1,060[2]. Manuscript 1 is an introductory paper addressed primarily to newcomers to diffusion model analyses. In the remainder of this chapter, a short introduction to diffusion modeling is given.

The diffusion model is applicable to tasks that are comprised of two response alternatives, which is a very common type of task in psychology. Typical binary tasks that have been analyzed with the diffusion model are recognition memory tasks (e.g., Ratcliff, 1978; Ratcliff, Thapar, & McKoon, 2011; Spaniol, Madden, & Voss, 2006), color discrimination tasks (e.g., Germar et al., 2016; Voss, Rothermund, & Voss, 2004), brightness discrimination tasks (e.g., Ratcliff, 2002; Ratcliff, Hasegawa, et al., 2011; Ratcliff, Thapar, & McKoon, 2003), numerosity discrimination tasks (e.g., Ratcliff, 2014; Ratcliff, Love, Thompson, & Opfer, 2012; Ratcliff & Van Dongen, 2009), motion discrimination tasks (e.g., Herz, Zavala, Bogacz, & Brown, 2016; Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; Ratcliff & McKoon, 2008), and lexical decision tasks (e.g., Dutilh et al., 2009; Ratcliff, Thapar, Gomez, & McKoon, 2004; Rummel, Kuhlmann, & Touron, 2013).

The diffusion model is based on the assumption that information is accumulated continuously until one of two thresholds is reached. This accumulation process is also referred to as *decisional process* and is illustrated in Figure 1. In this example, one threshold is associated with the response "word" and the other threshold with the response "non-word" (i.e., the response options of a lexical decision task). In addition to the decisional process, the

[1] Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

[2] This search was conducted on July 27, 2016. Even if certainly not all studies applied the diffusion model, probably most of them did.

diffusion model includes time that the participant requires, for instance, to encode information and execute motor responses (e.g., a key press). One major advantage of the diffusion model is the fact that it does not only use the accuracy rate or mean RT of correct responses, but the distributions of both correct and error responses. This high information quantity allows for disentangling parameters that map different cognitive processes. For example, one problem that researchers often encounter in RT paradigms is speed-accuracy trade-offs. One group of participants (e.g., young adults) might respond faster than the other group (e.g., older adults), but might at the same time yield lower accuracy rates. Such findings are difficult to interpret based on only the behavioral data (i.e., mean RT or accuracy rate). For example, is the young adult group superior at the task, as they respond faster? Or alternatively, is the older adult group superior, as their accuracy rate is higher? The diffusion model provides both a parameter that maps speed-accuracy settings and a parameter that maps cognitive speed. Thus, the application of the diffusion model supplies a more process-pure measure of speed of information processing, not confounded by speed-accuracy settings. Interestingly, in several studies, it has been shown that older adults do not differ from young adults in their cognitive speed, but—amongst others—in their speed-accuracy settings featuring a more conservative criterion (e.g., Ratcliff, Thapar, Gomez, et al., 2004; Ratcliff, Thapar, & McKoon, 2004; but see Thapar, Ratcliff, & McKoon, 2003).

The *basic diffusion model* comprises four main parameters: drift rate (v), threshold separation (a), starting point (z_r) and nondecision time (t_0). The drift rate (v) is a measure of the speed and direction of information accumulation. Easier tasks correspond to higher (absolute) drift rates than more difficult tasks as information is accumulated faster for these tasks (e.g., Voss et al., 2004). Similarly, more intelligent people have been shown to manifest higher drift rates than their less intelligent counterparts (e.g., Ratcliff, Thapar, & McKoon, 2010). People can differ also in their decision criterion. If an individual wants to be certain about a decision, he or she will accumulate more information before taking a decision—resulting in slower but more accurate responses—and will consequently have a higher threshold separation (a). Parameter a is therefore a measure of speed-accuracy trade-off.
If an individual has a decisional bias for one of the two options (e.g., because one category appears more often; see for example Arnold, Bröder, & Bayen, 2015), they will be fast and correct in trials in which this option is the correct response. On the other hand, they will be slower in responding correctly and more prone to errors when the correct response is the non-favored response option. Such a data pattern is reflected by the starting point (z, or the relative starting point $z_r = z/a$). If, for example, people are biased in favor of the response “non-word”,

their starting point is shifted toward the non-word threshold. Finally, the time needed for extra-decisional processes—like encoding of information and execution of the motoric response—is termed *nondecision time* (t_0).

The four main diffusion model parameters are generally considered to be valid measures of psychological processes. Their validity is supported both by experimental validation studies that used several different binary tasks (e.g., Arnold et al., 2015; Gomez, Ratcliff, & Childers, 2015; Voss et al., 2004; Wagenmakers et al., 2008; see also chapter 8.4 for an example of an experimental validation design) and by correlational studies that found relationships between, for instance, drift rate and intelligence (e.g., Ratcliff et al., 2010; Schulz-Zhecheva, Voelkle, Beauducel, Biscaldi, & Klein, 2016).

In addition to the four main diffusion model parameters, the diffusion model includes one *intra*trial variability parameter and three *inter*trial variability parameters. The *intra*trial variability parameter (also termed *diffusion constant*, or *scaling parameter*) is set to a constant value (typically, 0.1 or 1) in the common applications of the model. It is a measure of the random noise of the Wiener process. Due to this noise, the decision process does not end at the same threshold after the same interval of time in each trial (even if the same stimulus was shown again). The three *inter*trial variabilities, namely the variability of starting point, drift rate (Ratcliff & Rouder, 1998) and nondecision time (Ratcliff & Tuerlinckx, 2002), reflect the assumption that processes may vary from trial to trial due to variability in the attention of the participant or heterogeneity of the stimulus material. The diffusion model, including all parameters described above, can be termed the *full diffusion model*[3].

By now, several software solutions for parameter estimation are available that make the model more accessible also to researchers with restricted programming experience. Examples of such programs are *EZ* (Wagenmakers, van der Maas, & Grasman, 2007), *DMAT* (Vandekerckhove & Tuerlinckx, 2007, 2008), *HDDM* (Wiecki, Sofer, & Frank, 2013), and *fast-dm* (Voss & Voss, 2007, 2008). The command-line program fast-dm has been recently extended; this extension is described in more detail in the subsequent chapter.

[3] Note that more recently a further parameter has been introduced that maps a response-execution bias (Voss, Voss, & Klauer, 2010).

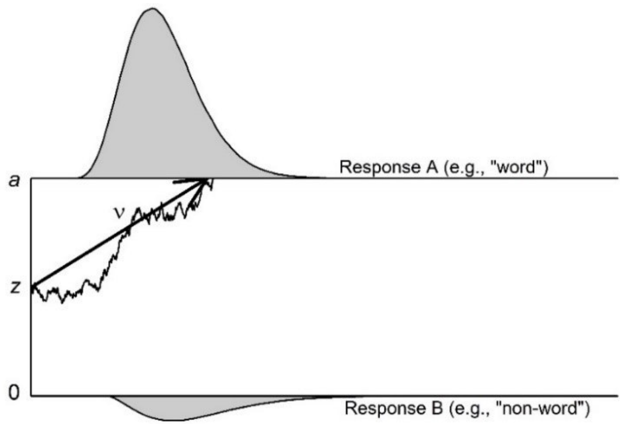


Figure 1. Illustration of a single trial of a decision process in a lexical decision task. The starting point *z* is here centrally positioned on the threshold distance *a*, indicating no bias for either of the two options (here, “word” or “non-word”). Information is accumulated with drift rate *v*. The decision process ends as soon as one of the two thresholds (in this case, the upper threshold) is hit. The nondecision time and the intertrial variabilities are not depicted in this illustration.

3 Extension of a Program for Parameter Estimation: *fast-dm-30* (Manuscript 2)[4]

Fast-dm (Voss & Voss, 2007, 2008) is a command-line program (implemented in C) that estimates diffusion model parameters based on the optimization criterion KS. Manuscript 2 introduces *fast-dm-30*, which is an extended version of fast-dm. The manuscript includes both information on the novel aspects of the program and a user's manual. Fast-dm-30 contains two major new features: First, it is now possible to estimate one further parameter, *d*, which is a measure of a response-execution bias (Voss et al., 2010). The second new feature is in the focus of this thesis. In the latest version of fast-dm, the user has the possibility to choose between one of three optimization criteria: KS (already implemented in the former versions of fast-dm), ML, and CS. The user can specify the optimization criterion he or she wishes to apply in an external control file, which also contains other model settings (e.g., fixation of parameters to constant values). Apart from giving the user the possibility to choose between the three criteria, fast-dm-30 also allows a better comparison of the performance of these criteria, because they can now be contrasted without confounding factors (i.e., differences in individual program specifications). In this chapter, the optimization criteria KS, ML and CS will be shortly described.

All of the three optimization criteria KS, ML, and CS have previously been used for diffusion modeling. Nevertheless, there are clear differences in the frequencies of their usage: CS is the by far most commonly applied method (e.g., Ratcliff & McKoon, 2008; Ratcliff & Van Dongen, 2009; Wagenmakers et al., 2008; White et al., 2009). In the CS approach, the RT distributions of correct and error responses are each divided into a number of bins. A very common procedure is the use of six bins with the outer two bins each comprising 10 % of the data and the inner bins 20 % each (Ratcliff & Tuerlinckx, 2002). Over all bins, a chi-square value is computed, thereby comparing the expected number of trials of each bin to the empirically observed number. Parameters are adjusted in order to minimize the chi-square value. Notably, due to the binning, the ratio scale of the RTs is converted into a nominal scale. Thus, CS does not use the full information that has been collected.

A criterion that—since the availability of *fast-dm*—has also been increasingly used, is KS (e.g., Aschenbrenner et al., 2016; Bowen et al., 2016; Boywitt & Rummel, 2012; Germar et al., 2014; Horn et al., 2011). This criterion is based on the maximum absolute vertical

[4] 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*(336). doi: 10.3389/fpsyg.2015.00336

distance between the expected and the empirical cumulative distribution functions (CDFs). Parameters are adjusted in order to minimize this maximum distance. In contrast to CS, the level of measurement of the RTs is not reduced. However, KS is based on only one single value, namely the maximum distance between the two curves.

Finally, the ML criterion, utilizes the information provided on each trial. The density values predicted for each empirical RT are logarithmized and summed up. The parameters are adjusted in order to maximize the sum of logarithmized densities. As each single RT affects the ML criterion, the degree of information utilization is the highest of the three optimization methods. However, this is also accompanied by the possibility of a higher influence of contaminants. By *contaminants*, we mean responses that result from sources other than a diffusion process (e.g., guessing; Ratcliff & Tuerlinckx, 2002). Especially fast outliers have a major influence on parameter estimates, because the nondecision time estimate must be at least as small as the smallest RT observed; otherwise, the estimated density of the smallest RT observed would be zero. KS, on the other hand, is less sensitive to contaminants as single trials should have limited influence on the maximum distance between the CDFs.

In Manuscripts 3-5, fast-dm-30 was employed in order to compare the performance of the three optimization criteria. In the presence of uncontaminated data (i.e., data with all trials resulting from a diffusion process), we expected ML to score best in parameter estimation, followed by KS and CS. However, in the presence of contaminated data (i.e., data with not all trials emanating from a diffusion process), ML would theoretically be substantially negatively impacted and should, therefore, provide less accurate parameter estimates than KS or CS. Before turning to a presentation of the main findings from Manuscripts 3-5, in the subsequent section, different evaluation criteria of estimation performance will be outlined. These criteria were used in the manuscripts for a comparison of estimation reliability of the different methods.

4 Criteria for Evaluation of Estimation Performance

Different approaches can be used to assess the performance of estimation procedures. These approaches are not specific to diffusion modeling, but are also applicable in other forms of mathematical modeling. One very common strategy is the usage of simulation studies (e.g., Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009). As in simulated data sets the true (i.e., data generating) parameters are known, these can be compared to the re-estimated parameters, allowing an assessment of the accuracy of parameter recovery. For the comparison of true with re-estimated parameters, different statistics can be appropriate, depending on the aim of the researcher. First, *correlations between the true and the re-estimated parameter values* can be computed. Evidently, higher correlation coefficients indicate better parameter estimation. A correlational criterion can be useful if the aim of the researcher is the uncovering of relationships between diffusion model parameters and external criteria. For example, a researcher might be interested in the relationship between drift rate and intelligence. Arguably, however, this criterion is only valuable if there is substantial variance in the generating parameter values. Moreover, this criterion has a major disadvantage: It might conceal *estimation biases,* which are deviations between the true and re-estimated values. Imagine, for example, two individuals, Chris and Steve. Chris has a drift rate of $\nu = 1.5$, and Steve has a higher drift rate of $\nu = 3$. If Chris was estimated to have a drift rate of 2, and Steve a drift rate of 3.5, this bias—an overestimation of the true drift rate by 0.5—would not be detected by means of a correlational criterion. However, it would be detected if deviations between true and re-estimated values were computed.

One might argue that an over- or underestimation does not constitute a severe problem, assuming that the deviation is the same for all participants. But this is very unlikely as, for example, the number of errors influences the accuracy of parameter estimation with more errors resulting in more reliable parameter estimates (e.g., Voss et al., 2004; White, Ratcliff, Vasey, & McKoon, 2010). Most problematic is a scenario in which there is a substantial negative relationship between the true parameter and its estimation bias. Using the above example, imagine that for Chris a drift rate of 2 was estimated, whereas the drift rate estimate for Steve was 2.5. Although Chris and Steve differ considerably in their true drift values ($diff_{true} = 1.5$), this difference is not so apparent if the estimated parameters are compared ($diff_{estimated} = 0.5$). Whereas negative relationships between the true values of a parameter and the estimation bias lead to an underestimation of the true effect, positive relationships lead to an overestimation. For researchers striving for significant results, such a positive relationship might be seen as advantageous. However, if, for example, one aims at assessing the value of an intervention program, it is certainly important to get correct estimates of effect size. Thus, both positive and negative relationships between the true parameters and their respective biases can pose severe problems.

Another criterion for evaluating the estimation performance is the *power of detecting differences in parameters between groups or conditions*. For example, Wiecki et al. (2013) analyzed the power to detect differences for a within-subjects design and van Ravenzwaaij, Donkin, and Vandekerckhove (2016) for a between-subjects design. Here, inaccuracies in parameter estimation can be compensated by an increase in the number of participants. If, however, the aim of the diffusion model analysis is not the detection of significant results, but the diagnostic assessment of one single individual, any inaccuracy in parameter estimation is problematic. As I mentioned earlier, in previous studies, relationships were found between drift rate and intelligence (e.g., Ratcliff et al., 2010; Schulz-Zhecheva et al., 2016). Accordingly, the drift rate could—in the long term—be a candidate for intelligence assessment. In fact, a diffusion model account might bare some advantages over traditional intelligence assessment, such as a restricted influence of training and a more restrained use of resources (e.g., less effort on the part of the investigator is required, since participants can work on the task independently, and the data collection can also be conducted in groups).

Thus, for the diffusion model to be applied as diagnostic tool, it is important that parameters be estimated accurately. If, as a measure of estimation precision, averaged biases across participants were computed, inaccuracies in parameter estimation might remain concealed because positive and negative biases can cancel each other out. Imagine, for example, that for half of the participants a parameter was overestimated and for the other half it was underestimated by similar amounts, a mean bias close to zero would result. Thus, an alternative way to analyze estimation precision is the use of either *absolute or squared deviations between true and re-estimated parameter values*.

Note that the diffusion model parameters have quite different scales and ranges. Therefore, the comparability of biases or absolute/squared deviations between parameters is restrained. For example, an absolute deviation of 0.1 would be small for drift rate estimates (given that the diffusion coefficient is set to 1), but would be large for nondecision time. To address this problem, from our simulation studies in Manuscript 3 we deducted *parameter accuracies*. This means that for each parameter, we estimated the smallest possible deviation that can be reached given optimal conditions (i.e., a high number of 5,000 trials, no contaminants, ML estimation; for more details, please see Manuscript 3). Next, we standardized each parameter's squared deviation, dividing it by its parameter accuracy.

Whereas simulation studies are the common method of assessing the estimation performance in diffusion modeling, to my knowledge, no one has thus far considered the use of test-retest studies. Generally, few diffusion model studies analyze data from test-retest designs (for two exceptions, see Schubert et al., 2016; Yap, Balota, Sibley, & Ratcliff, 2012) and in none of these were different estimation procedures compared. In test-retest designs, the retest correlation coefficients (i.e., the correlations between the estimated parameters of Session 1 and Session 2), can be used as criteria of estimation performance. Certainly, the size of these correlations depends on the stability of the respective parameter. However, for the assessment of estimation performance, the absolute values of the correlations are of less interest than the differences in coefficients between methods (e.g., the size of the retest coefficient based on ML estimates compared to the size of the retest coefficient based on CS estimates). An advantage of empirical test-retest studies is that they have a high ecological validity. In contrast, simulation studies, for data generation require assumptions about, amongst others, the parameter ranges and type and amount of contamination. These assumptions might not always be realistic.

The evaluation criteria presented above were used in the Manuscripts 3-5. In particular, Manuscript 3 is based on simulation studies and analyzes the data in terms of the correlational criterion, the bias criterion, the power of detection of within-subject differences and the estimation precision criterion. In Manuscript 4, test-retest studies were conducted and test-retest correlation coefficients served as measure of estimation performance. Finally, in Manuscript 5, data from Manuscript 3 and Manuscript 4 were reanalyzed to examine the influence of model complexity.

5 Evaluation of Estimation Performance: Simulation Studies (Manuscript 3)[5]

In the past, it was common to conduct experiments on the basis of very few participants that performed a task for several sessions, comprising in total several hundreds or even thousands of trials (e.g., Ratcliff, 1981, 2002; Ratcliff & Rouder, 1998; Ratcliff, Thapar, Gomez, et al., 2004; Ratcliff et al., 1999). In recent years, however, often the diffusion model is applied to data sets with substantially fewer trials (e.g., Klauer et al., 2007; Metin et al., 2013; Moustafa et al., 2015; Mueller & Kuchinke, 2016). Notably, there has been relatively little research on the influence of trial numbers on parameter estimation in diffusion modeling. The studies that analyzed different trial numbers indicate that—as expected—higher trial numbers go along with better parameter estimation (e.g., Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007; Wiecki et al., 2013). However, so far, no one has given indications on the minimum number of trials that can reasonably be used for diffusion modeling, or whether, at some point, the costs of a further increase in trial numbers might outweigh its benefits. In Manuscript 3, we tackled these issues using fast-dm-30. In contrast to other studies that have compared different programs (Ratcliff & Childers, 2015; van Ravenzwaaij & Oberauer, 2009), this is—to our knowledge—the first comparison of KS, CS and ML within the same program. In the remainder of this chapter, the method and results of Manuscript 3 will be summarized.

For our simulations, we first generated parameter sets, using a uniform distribution for each parameter, assuming no intercorrelations between parameters and relying on parameter ranges typically observed in empirical studies. Next, on the basis of these parameter sets, data sets were generated using *construct-samples*, which is integrated in the fast-dm software. The data sets were composed of either 24, 48, 100, 200, 500, 1,000 or 5,000 trials. We also compared models of different complexity. One approach for the analysis of the influence of model complexity was the use of a one-drift and a two-drift design. A one-drift design can be applied when the two stimulus types (e.g., “orange” and “blue” in a color discrimination task) have identical (or, very similar) drift rates (in our simulation study, we assumed identical drift rates). In such cases, it is a common approach to collapse across the two stimulus types, so that one threshold is associated with correct responses and the other threshold with error

[5] 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. doi: 10.3758/s13428-016-0740-2

responses. The one-drift model can also be used if separate models are estimated for the two stimulus types. The two-drift design is based on two stimuli with different drift rates. That is, the drift rate for the stimulus at one threshold (e.g., “new” in a recognition memory paradigm) is higher (in absolute value) than the drift rate for the stimulus at the other threshold (e.g., “old” in the recognition paradigm). The influence of model complexity was also tackled by means of models with different parameter settings. In particular, in the full diffusion model (in this text, also termed the seven-parameter model[6]), for the generation of the data sets all parameters, including the three intertrial variability parameters, could vary. In the slightly more restricted six-parameter model, the starting point was fixed at the center between the two thresholds (i.e., $z_r = .5$). In the four-parameter model, additionally, the intertrial variabilities of drift rate, starting point and nondecision time were fixed at zero. Finally, the three-parameter model differed from the four-parameter model in its fixed starting point (like in the six-parameter model, $z_r = .5$). In all models, the parameters that were fixed for the generation of data sets were also fixed for re-estimation (e.g., in the six-parameter model the starting point was fixed both for generation of data and for re-estimation of the parameters).

In empirical data, there will always be a certain amount of contaminant trials. Accordingly, another independent variable was the type of contaminants. We used one condition without contaminants, one with 4 % of fast contaminants and one with 4 % of slow contaminants. The fast contaminants were responses with random accuracy that were positioned partly outside and partly overlapping with the leading edge of the RT distribution (simulating guesses). The slow contaminants were responses situated between 1.5 and 3 interquartile ranges above the third quartile (simulating responses that are slow due to temporary distraction from the task).

Parameters were re-estimated using the three optimization criteria implemented in fast-dm: KS, ML, and CS. In addition, we used a Bayesian method (the nonhierarchical approach of the software HDDM, see Wiecki et al., 2013) and, for the data of the three-parameter model, EZ (Wagenmakers et al., 2007). EZ was applied solely to this restricted model because it is only able to estimate three parameters: threshold-separation, drift rate, and nondecision time. In contrast to fast-dm, EZ does not use an optimization procedure, but computes parameters by means of closed-form equations. Like CS and KS, both EZ (e.g., Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007; van Ravenzwaaij, Dutilh, &

[6] Note that, in theory, the diffusion model comprises eight parameters. Parameter *d* that measures a bias in the response execution (Voss et al., 2010) was fixed at zero in all models.

Wagenmakers, 2012; van Vugt & Jha, 2011) and HDDM (e.g., Dunovan, Tremel, & Wheeler, 2014; Herz et al., 2016; Jahfari, Ridderinkhof, & Scholte, 2013) have already been employed in several diffusion model studies.

Parameter estimation performance was assessed in terms of correlations between true and re-estimated parameters, parameter biases and, for the two-drift design, the power to detect a drift rate difference between the two conditions. Our most important criterion was the estimation precision (the squared, standardized deviations between true and re-estimated parameter values; see Chapter 4 and the section on evaluation criteria in Manuscript 3, for a more detailed description of the criteria). This criterion was used for the deduction of requisite trial numbers. More specifically, we defined (arbitrary) critical values that must be reached for low or high precision. The *requisite trial numbers* (that is, the trial numbers at which these critical values were reached) were computed for all conditions. Notably, our critical values seem to have been rather strict, as the high correlations between true and re-estimated parameters at the requisite trial numbers demonstrated.

Most importantly, our findings lend support to the use of more limited trial numbers. Especially for the more restricted models (three- and four-parameter models), reliable parameter estimates can often be achieved with even fewer than 100 trials. More complex models (six- and seven-parameter models), on the other hand, require more trials. However, in many cases reliable estimates necessitate still fewer than 300 trials[7]. The higher trial numbers of the more complex models seem to be attributable mostly to two parameters that are recovered very poorly: the intertrial variabilities of drift rate and starting point. This issue was investigated further in Manuscript 5.

In the condition with no contaminants, ML and HDDM showed the best estimation precision and CS performed worst. However, ML and HDDM suffered the most from fast contaminants. Here, often our criteria for the required trial numbers could not be reached (not even with 5,000 trials). Slow contaminants, on the other hand, did not cause severe problems. KS and EZ were quite robust to the presence of fast contaminants (however, EZ was affected by the presence of slow contaminants). While in the past mostly CS was used for parameter estimation, our results suggest that this is not the best strategy, as CS usually requires the highest trial numbers, and also clearly suffered from the addition of contaminants.

[7] For detailed information on the requisite trial numbers, see tables 4-6 in Manuscript 3.

Importantly, the comparison of the estimation methods showed very similar results for all different evaluation criteria (correlational criterion, biases, power of difference detection and estimation precision). Here, I will shortly describe two additional findings that concern the bias measure. The most notable result was a consistent negative correlation between starting point and its bias (for all methods and trial numbers), indicating that it is difficult to discover significant differences in starting points between conditions as the true effect may often be underestimated. Another finding was that the number of trials had an influence not only on estimation precision, but also on biases. Often, biases decreased with an increase in the trial numbers. This contradicts the hypothesis stated by van Ravenzwaaij and Oberauer (2009) that trial numbers do not influence biases. The pattern that we found might not yet have been observed in all earlier studies, because in these studies mostly high trial numbers were analyzed and our results indicate that biases get stable from around 500 trials onward.

Whereas Manuscript 3 was exclusively based on simulated data sets, in Manuscript 4 we examined empirical data from test-retest studies. In addition, the manuscript includes a simulation study in which data were generated relying on the parameter ranges observed in the experiments.

6 Evaluation of Estimation Performance: Test-Retest Studies (Manuscript 4)[8]

Manuscript 4 extends the findings from Manuscript 3 by using empirical data (accompanied by a simulation study). For empirical data, one single point in time cannot provide reliable information on the accuracy of parameter estimation as, unlike with simulated data sets, the true, underlying parameter values are not known [9]. However, repeated measures designs, allow a computation of test-retest correlation coefficients. The size of these coefficients depends on the stability of the measure and on error variances. Procedures with small estimation errors lead to higher test-retest correlation coefficients and give a better estimate of the trait characteristics. In Manuscript 4, we conducted two test-retest studies based on large samples (Study 1: $N = 105$ and Study 2: $N = 128$), both with a test-retest interval of one week. In Study 1, the participants worked on a lexical decision task and a recognition memory task. In Study 2, the participants had to perform a lexical decision task based on associative priming.

For parameter estimation, the three optimization criteria implemented in fast-dm-30 were used, in addition to the EZ method (Wagenmakers et al., 2007). For parameter estimation with fast-dm-30, we employed a five-parameter model in which the intertrial variabilities of drift rate and starting point were fixed at zero. We decided to use this approach because the intertrial variabilities of these two parameters were estimated very poorly in the studies of Manuscript 3 (see also Chapter 7 for a further discussion). The parameter estimation was based on different trial numbers (the first 32, 48, 100, 200, and 400 trials). In addition to the analysis of empirical data (Studies 1 and 2 of the manuscript), in Study 3 we simulated data sets based on the parameter means, standard deviations and intercorrelations from Study 1, allowing us to test whether a different simulation strategy (using a multivariate

[8] Lerche, V., & Voss, A. (2016). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24. doi: 10.1007/s00426-016-0770-5

[9] If data from only one single point in time are collected, parameter estimation performance can only be assessed via the analysis of model fit between the empirical and estimated distributions. For example, the *p*-value of the KS or CS statistic gives an indication of model fit. Note, however, that this *p*-value depends on the number of conditions and trials (see also Voss et al., 2013, p. 398, for a more detailed discussion). Furthermore, due to model mimicry (specifically, when different parameter combinations produce similar predicted RT distributions), a satisfactory fit value may result even if the parameters of the true (data generating) model were not correctly recovered (see Wagenmakers, Ratcliff, Gomez, & Iverson, 2004, for a procedure of quantifying model mimicry).

normal distribution and parameter values based on empirical data of specific experiments) results in similar findings as the simulation studies reported in Manuscript 3. In addition, we obtained estimates of maximal values of retest coefficients, because we simulated data sets for both “sessions” based on the same parameter sets (i.e., without state influences).

The data yielded by the empirical studies showed a consistent pattern regarding the performance of the different parameters. In all three paradigms, both drift rate and threshold separation had acceptable test-retest correlations (*r*s > .70), whereas nondecision time and starting point featured lower coefficients.[10] We also found that ML and EZ consistently outperformed the other methods in all of the three tasks, and akin to Manuscript 3, the simulation studies revealed that using more than 400 trials did not notably improve the reliability coefficients. Finally, the comparison of reliability coefficients from the simulated data (i.e., the maximal possible retest coefficients when there are no state influences) with the empirical data allowed an approximate estimation of state and trait proportions of parameters. These analyses indicated that drift rate is the parameter with the highest trait proportion. This is also corroborated by current analyses from a test-retest study with a longer interval of 8 months (Schubert et al., 2016).

While the focus of Manuscripts 3 and 4 is on the comparison of different optimization criteria and trial numbers, Manuscript 5 addresses more explicitly the issue of model complexity. Manuscript 3 has shown that the intertrial variabilities of starting point and drift rate cannot be recovered well. This suggests that the basic diffusion model parameters might be recovered better if these parameters are not estimated, but fixed to constant values. This comparison of models of different complexity is the topic of Manuscript 5.

[10] The worse performance of starting point and nondecision time might be attributable to a lack of interindividual variance or a lack of temporal stability in the specific paradigms used. In fact, in the priming paradigm, the nondecision time showed a better reliability than in the other paradigms. It is plausible that the handling of the prime influenced the encoding of the target stimulus. To an open-framed question, participants responded having used different strategies to deal with the prime (e.g., counting the primes, “ignoring” them, or explicitly paying attention to them) that might have been used consistently at both sessions.

7 Evaluation of the Influence of Model Complexity (Manuscript 5)[11]

The four parameters of the basic diffusion model (i.e., drift rate, threshold separation, nondecision time, and starting point) have proven to be valid measures of psychological processes (e.g., Voss et al., 2004) and one or more of these parameters have been in the focus of probably all diffusion model studies. The intertrial variabilities, on the other hand, are—from a psychological point of view—typically of less interest. These parameters are mostly included in order to improve model fit and to account for specific patterns of speed of correct and error trials. More specifically, faster errors than correct responses are mapped by the intertrial variability of starting point, while slower errors than correct responses are mapped by the intertrial variability of drift rate (Ratcliff & Rouder, 1998). Finally, the intertrial variability of nondecision time produces a more gradual rise of the leading edge of the RT distribution (Ratcliff & Tuerlinckx, 2002).

To date, in most diffusion model studies, the full diffusion model (i.e., including all three intertrial variabilities) is estimated (see Germar et al., 2016; Hartanto & Yang, 2016; Schubert et al., 2016, for some exceptions). However, is this procedure actually justified? As the results from Manuscript 3 revealed, the intertrial variabilities of starting point and drift rate cannot be recovered well, not even for an optimal estimation condition (5,000 trials, no contaminants, ML estimation). In this condition, correlations of estimates with true values were still smaller than .50. The intertrial variability of nondecision time, on the other hand, was estimated much better ($r = .97$) and so were the main diffusion model parameters (with all $rs \geq .99$). The poor recovery of the intertrial variabilities of drift rate and starting point is also in line with findings from previous simulation studies (van Ravenzwaaij & Oberauer, 2009; Vandekerckhove & Tuerlinckx, 2007).

The studies from Manuscript 3 further indicated that parameter estimation was better for less complex models: Three- and four-parameter models that do without the intertrial variabilities required smaller trial numbers than more complex models with intertrial variabilities. Note, however, that in Manuscript 3 we made assumptions that will not always be met. For example, in the case of the four-parameter model, we assumed that there is no variability across trials in drift rate, starting point, and nondecision time. In contrast, in Manuscript 5, we explicitly assumed the presence of substantial intertrial variabilities and

[11] Lerche, V., & Voss, A. (2016). Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious. *Frontiers in Psychology*, *7*(1324). doi: 10.3389/fpsyg.2016.01324

examined the impact of false fixations. By false fixations, we mean the fixation of parameters to values that are different from the data generating values. To examine the influence of false fixations, in Manuscript 5, we reanalyzed data sets from the Manuscripts 3 and 4.

In Manuscript 3, the data were always generated and estimated in the same manner (e.g., generated and estimated with a seven-parameter model or generated and estimated with a four-parameter model). In Manuscript 5 (Study 1), on the other hand, we reanalyzed the data from Manuscript 3 that were generated from the seven-parameter model, using models of different complexity. In particular, we compared the full diffusion model estimation to five-, four- and three-parameter models. In the five-parameter model, the intertrial variabilities of starting point and drift rate were fixed at zero; and in the four-parameter model, the intertrial variability of the nondecision time was also fixed at zero. Finally, in the three-parameter model we further fixed the starting point at the center between the two thresholds.

For the data from Manuscript 4 (see Study 2 in Manuscript 5), we applied the same procedure. Again, we estimated parameters using three-, four-, five-, and seven-parameter models. In contrast to the simulation studies of Manuscript 3, here we did not know for sure whether there were substantial intertrial variabilities in the data (even though the estimation with the seven-parameter model indicated that there might have been substantial intertrial variabilities). However, since we analyzed paradigms that have often been used in diffusion model studies (lexical decision task and recognition memory task), the ecological validity of these analyses is high. In addition, we also reanalyzed the data from the accompanying simulation study (Study 3 in Manuscript 4). Importantly, these data sets had been generated based on the means, standard deviations and intercorrelations of the parameter estimates from the seven-parameter model estimation. Thus, the simulated data sets featured substantial intertrial variabilities.

The most important finding from our reanalyses is that in most cases, a less complex model outperformed the full diffusion model. Interestingly, sometimes it took 5,000 trials for the seven-parameter model to outperform the other models (and even at this very high trial number, it did not always do so). For ML and CS, the five-parameter model won most often and for KS the four-parameter model showed the best performance. Note that the results by van Ravenzwaaij et al. (2016) go in a similar direction: They found EZ to be superior to a full diffusion model estimation even if the data were generated on the basis of the full diffusion model. Consistent with this, especially in Manuscript 4 of this thesis, EZ also provided very good parameter estimates. Thus, not only false fixations of the intertrial variabilities can improve parameter estimation. Sometimes, also a further fixation of the starting point might be advised. Note that in our test-retest studies on lexical decision and recognition memory we did not expect large deviations from a centered starting point. If a more biased decision process is to be expected, the fixation of the starting point to the center between the thresholds might essentially distort parameter estimates.

Finally, I want to point out one further interesting finding of Manuscript 5, in regards to the condition with fast contaminants. While in Manuscript 3, ML was highly sensitive to fast contaminants in all models, this was not the case for the five-parameter model applied in Manuscript 5. The inclusion of the intertrial variability of the nondecision time seemed to have absorbed the negative influences of these contaminants. This is plausible as the nondecision time leads to a more gradual rise of the leading edge of the estimated RT distribution and can thereby “capture” fast contaminants. If the intertrial variability of nondecision time was not included, the nondecision time estimated by ML would have to be adjusted to the smallest RT observed (see Chapter 3 for a more detailed explanation). Note that the five-parameter model showed a good performance for the condition with fast contaminants, whereas the seven-parameter model failed, even if it also includes the intertrial variability of nondecision time. It seems that here, the negative influence of the poorly estimated variabilities of starting point and drift was too large. To sum up this point, whereas in the past, it was recommended against using ML for parameter estimation due to its sensitivity to fast contaminants (e.g., Ratcliff & Tuerlinckx, 2002), our recent findings suggest that in combination with a five-parameter model, ML can supply reliable parameter estimates.

8 Discussion

In this section, we will first outline some important guidelines for diffusion modeling that are based on the main results emanating from the manuscripts of this thesis. Second, I will present a recent validation project by Dutilh et al. (2016), accompanied by reanalyses of the data from their project. Third, I will discuss limitations of this thesis and sketch ideas for future studies. One of these ideas will be presented in more detail in the last chapter of this discussion. Specifically, I will outline an idea for an extension of the diffusion model to slower RT tasks. This extension is possible because—as the studies of this thesis demonstrate—small- to medium-sized trial numbers can be sufficient for diffusion modeling.

8.1 Guidelines for Diffusion Modeling

In the last 15 years, the diffusion model (Ratcliff, 1978) has risen steadily in popularity (Voss et al., 2013). By analyzing whole distributions of RT data (not only accuracy rates and mean RTs), the model can disentangle different processes involved in binary decision tasks. There are various methods for the estimation of diffusion model parameters. We extended fast-dm (Voss & Voss, 2007, 2008)—based on the KS optimization criterion—to include two further criteria, ML and CS. Accordingly, fast-dm-30 (Voss et al., 2015) gives the user the choice between these three criteria. The implementation in the same program also makes it possible to compare the estimation performance of the three criteria without confounding factors (i.e., program specifics). Using both simulation studies and empirical test-retest studies, we analyzed the performance of the three optimization criteria and of models of different complexity, using different trial numbers. Incorporating the knowledge gained from these studies, we developed a set of guidelines that can serve both as an assessment of the reliability of published findings, and as an aide when planning future diffusion model studies. In addition to the three fast-dm methods, in several analyses, we included two further estimation methods: the Bayesian approach HDDM (Wiecki et al., 2013) and the method EZ (Wagenmakers et al., 2007).

First, we could clearly see that as the number of trials increases, the parameter estimation becomes more precise (Lerche & Voss, 2016b; Lerche, Voss, & Nagler, 2016). This has also been demonstrated by other studies (e.g., Ratcliff & Childers, 2015; Wiecki et al., 2013). More importantly, however, in our studies we observed substantial differences between optimization criteria regarding the relationship of trial number and precision of parameter estimation. Namely, ML and KS had a steeper course, resulting in reliable parameter estimation even for small- to medium-sized trial numbers, and reached an asymptote earlier than CS. This is because CS reduces the information supplied, as the RTs are grouped into a number of bins. Thus, more trials are required to attain the amount of information that other methods reach already for smaller trial numbers. Accordingly, I recommend against using CS for small- to medium-sized trial numbers.

As I mentioned in the introduction of this thesis, in the past mostly CS and KS have been used for diffusion modeling. Importantly, CS has almost always been applied in studies based on data sets with very high trial numbers. For these trial numbers, all three approaches yield satisfying results and thus, findings from CS-based studies are expected to be reliable. However, in recent years, often substantially lower trial numbers have been used in diffusion model analyses. For example, Moustafa et al. (2015) conducted a study based on only 160 trials and still applied the CS criterion. In this case, ML or KS would probably have been a better choice, as these methods yield reliable results also for lower trial numbers (in some cases even for fewer than 100 trials; for more detailed information on requisite trial numbers, see tables 4-6 from manuscript 3).

The prevalence of CS in the diffusion model literature has likely been fostered by Ratcliff and Tuerlinckx (2002) who, on the basis of several simulation studies, came to the conclusion that CS is the method of choice. More specifically, they argued for the use of CS with a correction for contaminant trials and the inclusion of the intertrial variability of nondecision time (in addition to the other two intertrial variabilities). Note, however, that this CS approach only performed well in the presence of contaminants if 1,000 trials per condition (and, thus a total of 4,000 trials) were used. Critically, the authors did not test the performance of ML for this condition, arguing that it would have taken too long to conduct the parameter estimation. Thus, a direct comparison of ML and CS for this condition is not available in their article. Importantly, for 250 trials per condition (i.e., a total of 1,000 trials), the performance of CS was “very poor” (p. 467). Despite this incomplete comparison of ML and CS, the authors concluded that CS is the method of choice and many articles have later referred to this article (e.g., Moustafa et al., 2015; Ratcliff, 2008; Ratcliff & McKoon, 2008; Wagenmakers et al., 2008) [12].

Whereas Ratcliff and Tuerlinckx cautioned against using ML, in the studies of the present thesis, ML (and HDDM which is also based on ML estimation) not only performed very well for uncontaminated data, but also for data with slow contaminants (Lerche et al., 2016). Only in the presence of fast contaminants were there problems for the ML approach. However, importantly, this problem could be counteracted if a model with one freely-varying intertrial variability—nondecision time variability—was included. This is because this parameter helps to capture fast contaminants. The other two intertrial variabilities, on the other hand, generally cannot be estimated well, and often even deteriorate the estimation of the main diffusion model parameters (Lerche & Voss, 2016a). Therefore, it will mostly be better to fix them at constant values.

To sum up, while I advise against using CS for parameter estimation (especially, for smaller trial numbers), both KS and ML can supply very good results. Interestingly, often more restricted models with false fixations (in particular, of s_{zr} and s_v) can produce more reliable results. This finding is also in line with the good performance of EZ, a three-parameter model with threshold separation, drift rate and nondecision time (Wagenmakers et al., 2007), in the test-retest studies (Lerche & Voss, 2016b) and in power studies (van Ravenzwaaij et al., 2016).

Especially in earlier diffusion model studies, often extremely high trial numbers were used (e.g., Ratcliff, 1981, 2002; Ratcliff & Rouder, 1998; Ratcliff, Thapar, Gomez, et al., 2004; Ratcliff et al., 1999). For example, in the study by Ratcliff (2002) participants completed more than 10,000 trials. First, such high trial numbers are expensive, often requiring several sessions of data collection, which may limit the number of participants able to complete the study (e.g., only three in the study by Ratcliff, 2002). Obviously, in such studies, interindividual differences cannot be examined. Second, such high trial numbers might actually be detrimental to the validity of study results, as individuals may become more bored or distracted, resulting in higher percentages of contaminants and, even more critically, a change of processes (e.g., learning effects). Third, it is not possible to use high trial numbers

[12] Note that in the references to the article by Ratcliff and Tuerlinckx (2002) important details are often omitted. For example, it is only stated that CS "provided the best balance between accurate recovery of parameter values […] and robustness to contaminant RTs" (Ratcliff & McKoon, 2008, p. 885). However, in the presence of contaminants, CS in fact only performed well when the additional parameters (intertrial variability of nondecision time and contaminant correction parameter) were included and very high trial numbers were used.

for all paradigms, because the stimulus material is restricted. Finally, clinical populations might not be capable of undergoing long and frequent sessions. In sum, there are several reasons why the use of very high trial numbers might not be possible or suitable. Importantly, the results from our studies (Lerche & Voss, 2016b; Lerche et al., 2016) reveal that using trial numbers above 500 is mostly not necessary, as parameter estimation improves only marginally (or indeed, it might even deteriorate if more contaminants occur).

To sum up, the three top guidelines of this thesis are the following:

(1) CS should not be used for small- to medium-sized trial numbers.

(2) It is advisable to fix the intertrial variabilities of starting point and drift rate (e.g., at zero) to obtain more reliable estimates of the four main diffusion model parameters (in particular, for smaller trial numbers). If such a "five-parameter model" is used (i.e., a model based on the four main diffusion model parameters plus the intertrial variability of nondecision time), different from previous indications (Ratcliff & Tuerlinckx, 2002), ML can also be applied to data with fast contaminants.

(3) Increasing trial numbers to more than 500 is of limited advantage for the precision of parameter estimation.

8.2 Collaborative Project on Validity of Results from RT Analyses

In 2015, Gilles Dutilh and Christopher Donkin started an interesting validation project (Dutilh et al., 2016). They contacted various researchers from the field of RT modeling and invited them to analyze data from "pseudo experiments", with the task of detecting eventual differences in parameters between two conditions (left vs. right moving dots in a dot motion task). Seventeen teams (each including one or two researchers) participated in the data analyses. Among the 14 pseudo experiments, there was one in which the authors did not use any manipulation. In three experiments, one single parameter was manipulated. For example, in Condition A, participants had a more difficult task than in Condition B, which would theoretically lead to a difference in drift rates between the two conditions. Furthermore, there were six experiments with manipulations of two parameters at the same time (e.g., drift rate and threshold separation). Finally, in four experiments, the three parameters of drift rate, threshold separation, and starting point were all manipulated simultaneously. The authors did not manipulate nondecision time and did not inform the teams about any details of their manipulations. In particular, they did not give any information on which parameter was manipulated in which experiment.

The project gives, in general, quite an optimistic view on analyses based on RT data. Specifically, for two of the four main diffusion model parameters, the results obtained by the different groups of researchers were very accurate, with 71 % and 86 % of correct classifications for drift rate and threshold separation, respectively. However, the percentages were lower for starting point (68 %) and nondecision time (62 %).

Interestingly, the different research teams employed diverse analysis approaches, demonstrating that there is not one single way of analyzing RT data. For example, some groups used the diffusion model while some the linear ballistic accumulator model (LBA; Brown & Heathcote, 2008); and there were more sophisticated approaches based on hierarchical Bayesian analyses, as well as model-free approaches relying on summary statistics. Interestingly, only half of the diffusion modelers employed the full diffusion model. Notably, the team that showed the highest percentage of correct classifications based their inferences on EZ2 (Grasman, Wagenmakers, & van der Maas, 2009), which is an extension of EZ that additionally permits the estimation of the starting point. Generally, teams that used the full diffusion model did not reach better results than teams that based their analyses on simplified versions of the model. Also of interest is that the hierarchical Bayesian analyses did not outperform the non-hierarchical analyses (in fact, they even performed slightly worse). Furthermore, there was no clear “winner” regarding the comparison of LBA and diffusion model—if at all, the LBA model seemed to have more serious difficulties.

One problem with this project is that the differences found cannot be clearly attributed, because the approaches taken by the different teams vary in several aspects at the same time. For example, different estimation methods were used (e.g., KS, CS, and ML), hierarchical and non-hierarchical analyses were conducted, and the statistical inferences were based on different strategies. Therefore, I reanalyzed the experimental data, varying only the optimization criterion used for parameter estimation (KS, ML, or CS) and the model complexity (four-, five-, or seven-parameter models). I estimated parameters separately for each participant and for the two conditions. For the drift rate, I computed v_{total} as a measure of the overall speed of information accumulation ($v_{\text{total}} = v_{\text{upper threshold}} - v_{\text{lower threshold}}$). The inferences were based on a *t* test (two-sided, with an alpha level of .05).

The results are presented in Figure 2. Two main findings emerged: (1) The detection of parameter differences in threshold separation, drift rate, and starting point was very effective, and (2) in line with the findings from Manuscript 5, the seven-parameter model was not superior to the less complex models. The four- and five-parameter models often outperformed the seven-parameter model. In fact, using ML or KS, the seven-parameter model did not show better performance than the less complex models for any of the parameters.

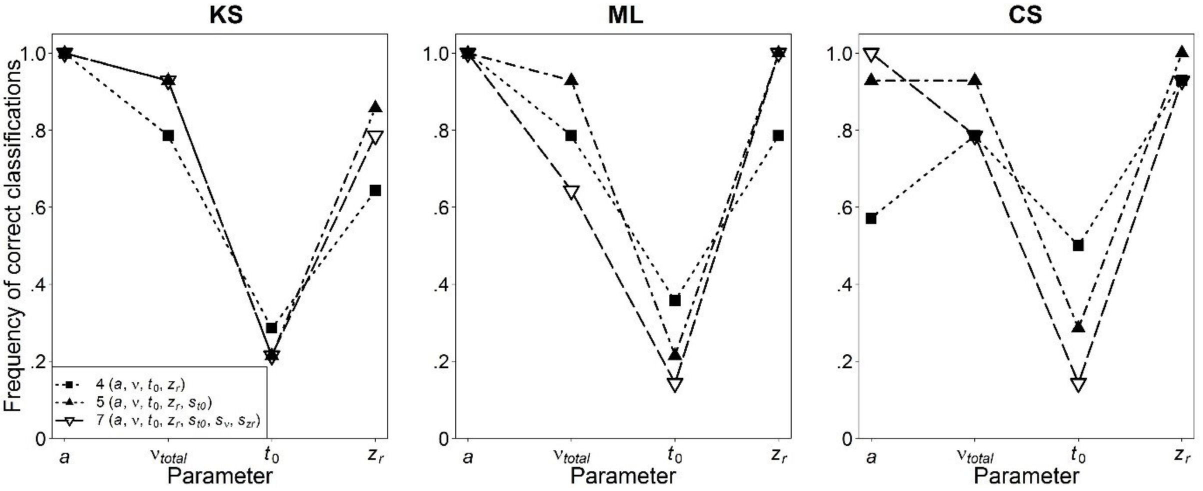


Figure 2. Frequencies of correct classifications depending on the optimization criterion, parameter and model complexity.

Striking are the low percentages of correct classifications for the nondecision time. Remember that Dutilh and Donkin did not explicitly manipulate the nondecision time in their experiments (i.e., “no effect” is the correct response for all experiments). However, it is plausible that they implicitly manipulated this parameter. In their manipulation of the threshold separation, they employed speed-accuracy instructions. In the accuracy blocks, participants were informed about erroneous responses, whereas in the speed blocks they got the feedback “too slow” if their response took longer than 0.8 s. It is possible that in the speed blocks, participants did not only speed up their decision process (which should result in a lower threshold separation), but also their motor response (and possibly the encoding of information, as participants may have been responding before completely encoding all information). In the accuracy blocks, on the other hand, they might have taken more time to execute the actual key press (and encoding). A more detailed analysis of the results reported by Dutilh et al. (2016) revealed the following pattern: If there was a manipulation of speed-accuracy settings, on average 62.7 % of the teams found an effect on nondecision time (in the direction of the speed-accuracy manipulation). In contrast, only 12.9 % of the teams detected an effect on nondecision time if the speed-accuracy settings had *not* been manipulated. These findings are in line with results from previous experimental studies (Arnold et al., 2015; Rinkenauer, Osman, Ulrich, Müller-Gethmann, & Mattes, 2004; Voss et al., 2004; but see Ratcliff, 2006). In these studies, speed-accuracy instructions also influenced both threshold separation and nondecision time.

The fact that several research teams found unexpected differences in nondecision time could be due to parameter estimation problems of the methods employed, or it could be attributable to a lack of discriminant validity of the experimental manipulation. To further explore this issue, I conducted a simulation study with two conditions, each with a different threshold separation (with $d_z = 0.35$, using a similar simulation strategy as in the two-drift model of Manuscript 3). One thousand parameter sets with 400 trials each (200 trials per condition, like in the study by Dutilh et al., 2016) were generated and the parameters were re-estimated. More specifically, parameters were estimated separately for the two conditions using the ML criterion and a five-parameter model. If I found no effect on nondecision time, this would support the idea that the lack of discriminant validity observed in the experimental validation studies was merely an effect of the type of manipulation. If, on the other hand, there was (also) an effect on nondecision time, this would suggest that there might (also) be a trade-off in the estimation of these two parameters.

Turning to the results, I found an effect size of $d_z = 0.26$ for threshold separation and an effect size of $d_z = 0.12$ for nondecision time (the effect sizes of the other parameters were at maximum 0.07). Thus, even if there was no "true effect" in nondecision time, a small difference still appeared. This supports the view that the unexpected effects in nondecision time observed in several studies (e.g., Arnold et al., 2015; Voss et al., 2004) may have been at least partly a problem of the estimation procedure, and not only of the experimental manipulation. The positive message here is that the effect on threshold separation was still larger than the effect on nondecision time.

If future experimental manipulation studies were accompanied by simulation studies based on characteristics of the empirical studies (e.g., in terms of trial numbers and parameter ranges), one could better disentangle manipulation problems from estimation problems. To estimate the influence of these problems, one could, for example, compare ratios of effect sizes between simulated and empirical data (such as the effect size of threshold separation compared to the sum of absolute effect sizes of threshold separation and nondecision time). If, for instance, this ratio was essentially smaller than 1 for both types of data, and at the same time clearly larger for the simulated data than for the empirical data, this would indicate that there is both a lack of discriminant validity of the manipulation and a trade-off issue in parameter estimation. In addition, one could simulate data with an effect only in nondecision time (and none in threshold separation) to find out more about trade-offs between these two parameters.

The approach presented in the preceding section could also be used for the analysis of other unexpected findings. For example, it has been found that if speed is highly emphasized, participants will feature lower drift rates compared to an accuracy condition (e.g., Rae, Heathcote, Donkin, Averell, & Brown, 2014; Starns, Ratcliff, & McKoon, 2012). Accordingly, one could analyze whether the influence of speed-accuracy instructions on the drift rate is a problem of the manipulation or the parameter estimation. In the above simulation study, I used a threshold separation of 1.65 (speed-condition) vs. 2 (accuracy-condition) and a standard deviation of 1, representing a small to medium sized effect size, and found no relevant drift rate difference between the two conditions ($d_z = 0.02$). However, the pattern might be different if the threshold separation in the speed condition was even smaller. A systematic analysis of different mean threshold separations and effect sizes could be a topic for future studies. In addition, it would be interesting to analyze the influence of the trial number on possible parameter trade-offs. Generally, trade-offs will likely be smaller for higher trial numbers. However, as far as I know, the number of trials required to clearly separate different effects has not yet been systematically examined.

8.3 Limitations and Ideas for Future Research

The greatest limitation of the studies of this thesis may be the restriction to relatively simple designs. To test the generalizability of the results and to extend the set of guidelines, analyses of additional experimental designs are necessary. To give an example, in the two-drift model of Manuscript 3, we generated a difference in drift rates between two conditions. Similarly, one could generate differences in other parameters. In Chapter 8.2, I demonstrated this approach for the threshold separation parameter. Furthermore, whereas in Manuscript 3, in parameter re-estimation, we only let drift rate vary between conditions, one could also let more parameters vary, or estimate parameters separately for the different conditions. In addition, for data generation, one might vary more than one parameter between conditions simultaneously and/or have more than two conditions. Indeed, in probably most diffusion model studies, the authors analyze which out of two or even more diffusion model parameters is influenced by the experimental manipulation. Therefore, it is important to know how many trials are required for the disentangling of influences on different parameters.

So far, we only analyzed three types of contamination (no contaminants, 4 % of fast or slow contaminants). However, it is plausible that also a combination of fast and slow contaminants can occur, and that the percentages will sometimes be higher than 4 %. This will depend, for example, on the type of task with more exhausting, long-lasting tasks featuring higher percentages of contaminants. It is an open question whether KS would continue to be so robust to contamination in this condition. Whereas Ratcliff and Tuerlinckx (2002) assumed that data with more than 5 % of contaminants are “unusable for model fitting” (p. 462), this still needs to be investigated systematically.

One further limitation of the studies of this thesis is that, until now, we only analyzed non-hierarchical models. It would be interesting, for the future, to also compare the performance of non-hierarchical analyses to hierarchical (Bayesian) analyses. Even though the collaborative project by Dutilh et al. (2016) provides some first insight that more complicated analyses may not be necessary, this still needs to be analyzed in a more systematic way because as I already noted in the previous chapter, the comparability of the approaches taken by the different research teams was limited.

One notable result from the project by Dutilh et al. (2016) was the strong performance of the very simple method EZ2 (Grasman et al., 2009), which performed best of all approaches. It would be interesting to further examine the performance of EZ2 that up to date has been rarely applied in diffusion model studies (for some exceptions, see Lee & Chabris, 2013; Schmittmann, van der Maas, & Raijmakers, 2012; Whitson et al., 2014). EZ, in Manuscript 3, was found to be robust to the presence of fast contaminants, but was clearly affected by slow contaminants (likely because slow contaminants have a greater effect on the mean RT, which is used for parameter computation). In my eyes, it would be interesting to apply EZ2 to the data from the studies of this thesis and to new studies with more complex designs to test its performance under different conditions of model complexity and contamination.

Both EZ and EZ2 do not allow an estimation of intertrial variabilities. Thus, these variabilities are implicitly fixed at zero. In contrast, in fast-dm, the intertrial variabilities can be fixed at user-specified values. In Manuscript 5, we explicitly fixed the intertrial variabilities at zero even though data had been generated on the basis of a full diffusion model. This often resulted in better estimates of the basic diffusion model parameters than the full diffusion model estimation. Note that it is possible that the fixation to specific values other than zero might result in even better estimates. For example, the intertrial variabilities could be fixed at values that have been observed in other diffusion model studies. In line with this possibility, I reanalyzed the data sets from the no contamination condition of the one-drift model of Manuscript 3, fixing intertrial variabilities of starting point, drift rate and nondecision time at 0.25, 0.5 and 0.1, respectively. These values are each half of the maximum values used for the generation of the data sets. Whereas with fixation at zero, models with restrictions (three-, four- or five-parameter models) performed best in 61 % of all conditions, for the new fixation strategy the percentage increased to 76 %. Thus, as expected, the performance of the more restricted models improved with the different fixation strategy. Certainly, the choice of fixation values was, in this case, informed by knowledge about the true parameter ranges, and thus the increase in best-fitting models to 76 % presents an upper limit. For empirical data with unknown true values, one has to make guesses that can be more or less appropriate. The guesses could be based on previous studies with the same or a similar paradigm and, ideally, very high trials numbers (so that the estimates of the intertrial variabilities are reasonably reliable).

In summary, there are plenty of possible extensions of the approaches taken in this thesis, and our analyses can be seen as first important step in the developing of more general guidelines for diffusion modeling. Generally, I am in favor of the combination of empirical studies with simulation studies. First, simulation studies can be useful to assess fit values of empirical studies (see Manuscript 1). For example, the *p*-value of the KS statistic depends on the number of trials, the number of conditions, and the parameter ranges. Thus, the general use of .05 as exclusion criterion is questionable. Rather, for each empirical study, the specific fit criterion should be deduced based on the distribution of fit values of a simulation study (with the simulation study based on the characteristics of the empirical study, e.g., using the same trial number, number of conditions, and parameter ranges). In addition, the combination of empirical and simulation approaches can help to disentangle estimation problems from other influences. In Manuscript 4 of this thesis, data were simulated based on the empirical test-retest data observed. This allowed a disentangling of estimation problems from state influences of parameters, thereby allowing us to see that drift rate is a particularly stable parameter. A further field of application has been outlined in more detail in Chapter 8.2. Here, the combination of empirical studies and simulation studies was suggested to disentangle influences of the experimental manipulation from estimation trade-offs between parameters (e.g., the influence of speed-accuracy instructions on nondecision time).

Finally, I would like to add a more general, reflective point. First, I think that the availability of different software solutions for diffusion modeling and the newly implemented choice of different criteria within fast-dm-30 is a great improvement. In the past, researchers had to write their own code for modeling RT data with a diffusion model, which is no longer necessary now, thereby expanding the usership and the fields of application of the model. However, the amount of choices that users are now faced with also introduces complications. Carrying out additional analyses with, for example, ML instead of KS, requires the user of fast-dm-30 to modify solely two letters in the control file. In addition, the computation time required by fast-dm-30 is very low for any of the three optimization criteria (even ML estimation usually takes much less than 1 hour per data set). Therefore, it might become somehow tempting to researchers to try out all three different optimization criteria (and, possibly different complex models) and report the “best” results.

Ideally, effects are very stable and different methods come to the same conclusions as was the case, for example, for threshold separation and drift rate in most experiments of the project by Dutilh et al. (2016). If effects are large (like in their project), they will probably be correctly detected in most cases. However, if effects are small- to medium-sized, trade-offs between parameters can play a critical role, and such effects should thus be interpreted cautiously. In a future project, empirical data of diffusion model studies already published could be reanalyzed using different optimization criteria and models of different complexity to test the stability in face of different estimation procedures.

Certainly, the abundance of choices that researchers are faced with as a result of user-friendly software solutions is not an issue specific to diffusion modeling, but of more advanced mathematical models or statistical analyses in general. One important approach of counteracting the “fishing” for the best results is preregistration (e.g., Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Additionally, in my eyes, it is very important that software solutions be accompanied by well-founded guidelines on how to use the software. As the variety of approaches of the project by Dutilh et al. (2016) demonstrates, even experts do not agree on which method to use. The more that interest in diffusion modeling grows, the more important it becomes to further investigate the reliability and validity of diffusion model parameters. If guidelines on the use of the diffusion model (and also of other RT models) exist and, importantly, are known by a wide community, both researchers and reviewers can make better informed decisions.

8.4 Extension of the Diffusion Model to “Slow” RT Tasks

Ratcliff has repeatedly stressed that the diffusion model should only be applied to the analysis of very fast RT tasks that require on average at maximum 1.5 s per trial (e.g., Ratcliff & Frank, 2012; Ratcliff & McKoon, 2008; Ratcliff, Thapar, Gomez, et al., 2004). This “1.5 s rule” clearly restricts the field of applications and the external validity of the diffusion model as in many tasks, decisions will take more time than 1.5 s (e.g., Dummel, Rummel, & Voss, 2016; Johnson, 1996; Kahane et al., 2012). Besides, the value of 1.5 is arbitrary and cannot, to my knowledge, be justified by any empirical finding.

If large numbers of trials (e.g., several thousands) were required to reliably fit the data of a participant, the diffusion model would be restricted to very fast response time tasks. For paradigms with longer RTs, the experiment would take too long, potentially resulting in changes in processes (e.g., owing to fatigue). However, as the results of this thesis demonstrate, small to moderate trial numbers can already supply satisfactory parameter estimates. This renders the application of the model to slower RT tasks possible from a practical point of view. However, what about the theoretical reasoning behind the 1.5 s rule?

On principle, the diffusion model formula could be used for any period (whether for milliseconds or for years). However, it is questionable whether the theoretical assumptions underlying the diffusion model are still met if periods are longer. Two critical assumptions are that (a) there is a single stage of processing, and (b) that parameters are constant over time.

For example, imagine a very basic task that has been frequently used for diffusion model analyses: a lexical decision task. In this task, participants have to judge whether the presented letter string is a word or not. Here, the assumption of a single-stage process (in comparison to a multiple-stage process) and of constant parameter values is very plausible. In a slightly different task, the letter strings presented might be adjectives each describing an individual that has to be judged according to their likeability. This decision will probably take longer and the single-stage assumption might be violated in some trials. For instance, an adjective like “pious” might be judged differently depending on the context. The participant might think of a good friend who is very pious and likeable in their eyes, and might want to decide in favor of “positive”. Thus, he or she might have almost already reached a threshold. Then, however, he or she might think of another context, such as religious fanatics, and might, in the end, press the “negative” key. Other examples of ambiguous words might be “sensitive” or “talkative”. Thus, in this simple example, there might already be more than one processing step.

Imagine further a task that requires the judgment of likeability of a person that is described by not one single adjective, but by a number of different adjectives. There will be terms described that are part of the everyday language of the participant, and can thus be assessed easily; and others that have a lower familiarity. There will be adjectives that in the eyes of the participant are clearly positive or negative, and others that are more ambiguous. In other words, the adjectives will have a different informative quality with adjectives with a higher quality going along with faster information accumulation (i.e., higher drift rates) than adjectives that are less familiar or more ambiguous. In this task, the assumption of one single constant drift rate is not very likely. In such tasks with a number of simultaneously presented stimuli, it is also very plausible that participants do not encode all stimuli before starting the decision process. One might illustrate such tasks by a chain of diffusion models (e.g., one for each adjective) that can have, for example, different drift rates and nondecision components. To sum up, as these examples illustrate, in more realistic and complex tasks, the assumptions of single-stage processing and constancy of parameters could be regularly violated.

In Manuscript 5, we demonstrated that a simplified modeling of the underlying processes (here, in terms of the false fixation of intertrial variabilities) can result in good or even better estimates of the main diffusion model parameters. Similarly, one might wonder whether a violation of the assumptions of single-stage processing and parameter constancy poses a severe problem for model fitting. In fact, the diffusion model has already been applied to tasks that took longer than 1.5 seconds and model fit remained satisfactory (e.g., Aschenbrenner et al., 2016). In addition, in a simulation study by Ratcliff (2002), data were generated so that drift rates increased with the time, thus violating the assumption of parameter constancy. Interestingly, the diffusion model with one constant drift still displayed a good performance. Thus, there is already preliminary support that the violation of assumptions does not necessarily have negative effects on model fitting.

The analysis of model fit of empirical data or simulation studies that explicitly make “false” assumptions are ways of analyzing conditions of the applicability of the diffusion model. We recently took a further approach and conducted an experimental validation study. Whereas there are several experimental validation studies based on fast RT tasks (e.g., Arnold et al., 2015; Voss et al., 2004), as far as I know, none has been conducted with slower RT tasks. In our study, participants had to work on a figural task. Eight rectangles were presented in each trial, with half of them surrounded by a blue and red border, respectively. The participants had to mentally sum up the sizes of the rectangles separately for each color and decide which size was larger. The task had a mean trial duration of 7.43 s (in the baseline condition), and can thus clearly be considered a “slow” RT task. In this task, very different strategies can be taken, as also the responses to an open-framed question demonstrated. For example, one common strategy was the search for pairs of similar-sized rectangles. Another was the mental collapsing of same-colored rectangles. In this task, it is also very likely that participants changed strategy during a trial, or that they used one strategy but restarted the decision process to check the accuracy of their decision (e.g., building different pairs). Thus, it is likely that the task violates basic assumptions of the diffusion model.

We experimentally manipulated the four main diffusion model parameters, using a similar approach as Voss et al. (2004). Threshold separation was addressed by means of speed-accuracy instructions. Drift rate was supposed to be affected by differences in the difficulty of the task. For the manipulation of nondecision time, participants had to press a key not only once, but three times in a row. Finally, the starting point was manipulated by the use of an asymmetric payoff matrix. Both convergent and discriminant validity of drift rate, threshold separation and nondecision time were comparable to those in experimental validation studies with fast RTs. The manipulation of the starting point had an effect on drift rate (more specifically, the drift criterion) instead of starting point. This was probably mainly due to the fact that the manipulation only had an influence on the percentages of key presses (with the favored key pressed more often), but not on RT. Of interest are also the results of a study by Leite and Ratcliff (2011). They compared the effects of the manipulation of stimulus frequency and a payoff matrix (for a fast RT task). The study revealed that a manipulation of stimulus frequency affected the starting point, whereas an asymmetric payoff matrix affected the drift rate. This is in line with our finding (but in contrast with the study by Voss et al., 2004). Thus, our finding for the starting point manipulation does not seem to challenge our hypothesis that the diffusion model is also applicable to slower RT tasks. For a subsequent study, we could use a manipulation of stimulus frequency, which might then be more likely to influence the starting point instead of the drift criterion.

To sum up, this thesis has rebutted one “myth” regarding diffusion modeling, namely, the assumption that very high trial numbers are necessarily required for diffusion modeling. Thereby, it has opened the field for other types of diffusion model tasks and for the challenge of another possible “myth”: The assumption that the diffusion model can only be applied to very fast response time tasks. The first results from our experimental validation study based on a figural task are promising and suggest that the model might also be applicable to tasks that require more than 1.5 s, even if assumptions might be violated. Nevertheless, it is still necessary to investigate whether this also generalizes to different types of tasks (e.g., numerical tasks or judgement tasks like the one presented as introductory example) and to even longer trial durations (e.g., of up to 30 seconds per trial).

9 Conclusions

To date, substantiated knowledge about requisites of diffusion modeling is rather sparse. Indeed, not even experts agree on which method to use for parameter estimation. In this thesis, I compared different estimation procedures and trial numbers in order to deduce a set of guidelines for newcomers as well as for more experienced diffusion modelers. The guidelines are intended both for assessing the reliability of previous findings and for planning new experiments and analyses.

Most importantly, our studies based on fast-dm-30, the newest version of fast-dm, could show that reliable parameter estimates can often be attained on the basis of small- to medium-sized trial numbers. In fact, it is even likely that with very high trial numbers—of, for example, more than 10,000 per participant, as in the study by Ratcliff (2002)—the motivation of the participants decreases substantially, resulting in higher percentages of contaminants. Additionally, parameters might change over time. Whereas only three participants took part in the study by Ratcliff (2002), nowadays, researchers are often interested in interindividual differences in diffusion model parameters, and thus apply research designs with more participants, but significantly fewer trial numbers. Additionally, in the future, the diffusion model might also be applied to slower RT tasks (i.e., with RTs > 1.5 s) that—owing to their longer trial durations—will often be limited in trial numbers. Importantly, as the studies of this thesis demonstrate, for such small- to medium-sized trial numbers, CS—which used to be the standard procedure for diffusion modeling—is not recommended. Alternatively, ML and KS can supply reliable estimates even for lower trial numbers.

In this thesis, I also argue for the use of less complex models, in particular of models with fixations of intertrial variabilities (specifically, of the intertrial variabilities of starting point and drift rate). In making this guideline, I do not question the reasons for which these parameters have been introduced (e.g., for modeling differences in speed between correct and erroneous responses). However, in particular for lower trial numbers, the full diffusion model that includes all three intertrial variabilities cannot be estimated reliably. The “false” fixation of the intertrial variabilities of starting point and drift rate can improve the estimation of the main diffusion model parameters (threshold separation, drift rate, starting point, and nondecision time) that are in the center of interest of probably all diffusion model studies.

To conclude, even if the guidelines offered by this thesis are still limited in generalizability (e.g., based on rather simple designs), they do provide a first step and will, hopefully, be extended within the course of the next years.

References

Allen, P. A., Lien, M.-C., Ruthruff, E., & Voss, A. (2014). Multitasking and aging: Do older adults benefit from performing a highly practiced task? *Experimental Aging Research, 40*(3), 280-307. doi: 10.1007/s00426-014-0608-y

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*(5), 882-898. doi: 10.1007/s00426-014-0608-y

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology, 30*(2), 225-238. doi: 10.1037/neu0000222

Bowen, H. J., Spaniol, J., Patel, R., & Voss, A. (2016). A Diffusion Model Analysis of Decision Biases Affecting Delayed Recognition of Emotional Stimuli. *PLoS ONE, 11*(1), 1-20. doi: 10.1371/journal.pone.0146769

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition, 40*(1), 70-82. doi: 10.3758/s13421-011-0128-6

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

Dummel, S., Rummel, J., & Voss, A. (2016). Additional information is not ignored: New evidence for information integration and inhibition in take-the-best decisions. *Acta Psychologica, 163*, 167-184. doi: 10.1016/j.actpsy.2015.12.001

Dunovan, K. E., Tremel, J. J., & Wheeler, M. E. (2014). Prior probability and feature predictability interactively bias perceptual decisions. *Neuropsychologia, 61*, 210-221. doi: 10.1016/j.neuropsychologia.2014.06.024

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., . . . Donkin, C. (2016). A collaborative project on the validity of response time data inference. *Manuscript in preparation.*

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review, 16*(6), 1026-1036. doi: 10.3758/16.6.1026

Germar, M., Albrecht, T., Voss, A., & Mojzisch, A. (2016). Social Conformity is due to Biased Stimulus Processing: Electrophysiological and Diffusion Analyses. *Social Cognitive and Affective Neuroscience*. doi: 10.1093/scan/nsw050

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin, 40*(2), 217-231. doi: 10.1177/0146167213508985

Gomez, P., Ratcliff, R., & Childers, R. (2015). Pointing, looking at, and pressing keys: A diffusion model account of response modality. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1515-1523. doi: 10.1037/a0039653

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53*(2), 55-68. doi: 10.1016/j.jmp.2009.01.006

Hartanto, A., & Yang, H. (2016). Disparate bilingual experiences modulate task-switching advantages: A diffusion-model analysis of the effects of interactional context on switch costs. *Cognition, 150*, 10-19. doi: 10.1016/j.cognition.2016.01.016

Herz, Damian M., Zavala, Baltazar A., Bogacz, R., & Brown, P. (2016). Neural Correlates of Decision Thresholds in the Human Subthalamic Nucleus. *Current Biology, 26*(7), 916-920. doi: 10.1016/j.cub.2016.01.051

Horn, S. S., Bayen, U. J., & Smith, R. E. (2011). What can the diffusion model tell us about prospective memory? *Canadian Journal of Experimental Psychology, 65*(1), 69-75. doi: 10.1037/a0022808

Jahfari, S., Ridderinkhof, K. R., & Scholte, H. S. (2013). Spatial Frequency Information Modulates Response Inhibition and Decision-Making Processes. *PLoS ONE, 8*(10), e76467. doi: 10.1371/journal.pone.0076467

Johnson, A. T. (1996). Comprehension of Metaphors and Similes: A Reaction Time Study. *Metaphor and Symbolic Activity, 11*(2), 145-159.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience, 7*(4), 393-402. doi: 10.1093/scan/nsr005

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*(3), 353-368. doi: 10.1037/0022-3514.93.3.353

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

Lee, J. J., & Chabris, C. F. (2013). General cognitive ability and the psychological refractory period: Individual differences in the mind's bottleneck. *Psychological Science, 24*(7), 1226-1233. doi: 10.1177/0956797612471540

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*(7), 651-687.

Lerche, V., & Voss, A. (2016a). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology, 7*(1324). doi: 10.3389/fpsyg.2016.01324

Lerche, V., & Voss, A. (2016b). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24. doi: 10.1007/s00426-016-0770-5

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. doi: 10.3758/s13428-016-0740-2

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *psychometrika, 40*(1), 77-105. doi: 10.1007/BF02291481

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27*(2), 193-200. doi: 10.1037/a0031533

Moustafa, A. A., Kéri, S., Somlai, Z., Balsdon, T., Frydecka, D., Misiak, B., & White, C. (2015). Drift diffusion model of reward and punishment learning in schizophrenia: Modeling and experimental data. *Behavioural Brain Research, 291*, 147-154. doi: 10.1016/j.bbr.2015.05.024

Mueller, C. J., & Kuchinke, L. (2016). Individual differences in emotion word processing: A diffusion model analysis. *Cognitive, Affective, & Behavioral Neuroscience, 16*(3), 489-501. doi: 10.3758/s13415-016-0408-5

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience, 32*(7), 2335-2343. doi: 10.1523/JNEUROSCI.4156-11.2012

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*(5), 1226-1243. doi: 10.1037/a0036801

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88*(6), 552-572. doi: 10.1037/0033-295X.88.6.552

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

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53*(3), 195-237. doi: 10.1016/j.cogpsych.2005.10.002

Ratcliff, R. (2008). Modeling aging effects on two-choice tasks: Response signal and response time data. *Psychology and Aging, 23*(4), 900-916. doi: 10.1037/a0013930

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 870-888. doi: 10.1037/a0034954

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*(4), 237-279. doi: 10.1037/dec0000030

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-Based Decision Making in Corticostriatal Circuits: Mutual Constraints by Neurocomputational and Diffusion Models. *Neural Computation, 24*(5), 1186-1229. doi: 10.1162/NECO_a_00270

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Childers, R., Smith, P. L., & Segraves, M. A. (2011). Inhibition in Superior Colliculus Neurons in a Brightness Discrimination Task? *Neural Computation, 23*(7), 1790-1820. doi: 10.1162/NECO_a_00135

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child Development, 83*(1), 367-381. doi: 10.1111/j.1467-8624.2011.01683.x

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

Ratcliff, R., & McKoon, G. (2015). Aging Effects in Item and Associative Recognition Memory for Pictures and Words. *Psychology and Aging*. doi: 10.1037/pag0000030

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

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review, 111*(2), 333-367. doi: 10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences, 20*(4), 260-281.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278. doi: 10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65*(4), 523-535. doi: 10.3758/BF03194580

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*(4), 408-424. doi: 10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*(3), 464-487. doi: 10.1037/a0023810

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115-136. doi: 10.1016/j.cognition.2014.12.004

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481. doi: 10.3758/bf03196302

Ratcliff, R., & Van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review, 16*(4), 742-751. doi: 10.3758/PBR.16.4.742

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261-300. doi: 10.1037/0033-295X.106.2.261

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*(2), 261-282. doi: 10.1037/0096-3445.133.2.261

Rummel, J., Kuhlmann, B. G., & Touron, D. R. (2013). Performance predictions affect attentional processes of event-based prospective memory. *Consciousness and Cognition: An International Journal, 22*(3), 729-741. doi: 10.1016/j.concog.2013.04.012

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology - General, 136*(3), 414-429. doi: 10.1037/0096-3445.136.3.414

Schmittmann, V. D., van der Maas, H. L., & Raijmakers, M. E. (2012). Distinct discrimination learning strategies and their relation with spatial memory and attentional control in 4-to 14-year-olds. *Journal of experimental child psychology, 111*(4), 644-662. doi: 10.1016/j.jecp.2011.10.010

Schubert, A.-L., Frischkorn, G., Hagemann, D., & Voss, A. (2016). Trait Characteristics of Diffusion Model Parameters. *Journal of Intelligence, 4*(3), 7. doi: 10.3390/jintelligence4030007

Schulz-Zhecheva, Y., Voelkle, M., Beauducel, A., Biscaldi, M., & Klein, C. (2016). Predicting Fluid Intelligence by Components of Reaction Time Distributions from Simple Choice Reaction Time Tasks. *Journal of Intelligence, 4*(3), 8. doi: 10.3390/jintelligence4030008

Spaniol, J., Madden, D. J., & Voss, A. (2006). A Diffusion Model Analysis of Adult Age Differences in Episodic and Semantic Long-Term Memory Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 101-117. doi: 10.1037/0278-7393.32.1.101

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*(1–2), 1-34. doi: 10.1016/j.cogpsych.2011.10.002

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion Model Drift Rates Can Be Influenced by Decision Processes: An Analysis of the Strength-Based Mirror Effect. *Journal of Experimental Psychology. Learning, Memory & Cognition, 38*(5), 1137-1151. doi: 10.1037/a0028151

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging, 18*(3), 415-429. doi: 10.1037/0882-7974.18.3.415

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2016). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, 1-10. doi: 10.3758/s13423-016-1081-y

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology, 219*(4), 1017-1025. doi: 10.1007/s00213-011-2435-9

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*(6), 463-473. doi: 10.1016/j.jmp.2009.09.004

van Vugt, M. K., & Jha, A. P. (2011). Investigating the impact of mindfulness meditation training on working memory: A mathematical modeling approach. *Cognitive, Affective, & Behavioral Neuroscience, 11*(3), 344-353. doi: 10.3758/s13415-011-0048-8

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

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

Journal of Mathematical and Statistical Psychology, 63(3), 539-555. doi: 10.1348/000711009x477581

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: a tutorial based on fast-dm-30. *Frontiers in Psychology, 6*(336). doi: 10.3389/fpsyg.2015.00336

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*(1), 28-50. doi: 10.1016/j.jmp.2003.11.004

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*(1), 140-159. doi: 10.1016/j.jml.2007.04.006

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*(1), 3-22. doi: 10.3758/bf03194023

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. doi: 10.1177/1745691612463078

Weigard, A., & Huang-Pollock, C. (2014). A diffusion modeling approach to understanding contextual cueing effects in children with ADHD. *Journal of Child Psychology & Psychiatry, 55*(12), 1336-1344. doi: 10.1111/jcpp.12250

White, C. N., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition and Emotion, 23*(1), 181-205. doi: 10.1080/02699930801976770

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*(1), 39-52. doi: 10.1016/j.jmp.2010.01.004

Whitson, L. R., Karayanidis, F., Fulham, R., Provost, A., Michie, P. T., Heathcote, A., & Hsieh, S. (2014). Reactive control processes contributing to residual switch cost and mixing cost across the adult lifespan. *Frontiers in Psychology, 5*, 383. doi: 10.3389/fpsyg.2014.00383

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics, 7*, 14. doi: 10.3389/fninf.2013.00014

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

List of Figures

Appendix A 1

Manuscript 1: Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402.

Reprinted by permission from *Experimental Psychology*

Diffusion models in experimental psychology:
A practical introduction

Andreas Voss, Markus Nagler, and Veronika Lerche
Ruprecht-Karls-Universität Heidelberg

Author Note

Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Markus Nagler, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany.

This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant Nr. VO1288/2-1).

Correspondence concerning this article should be addressed to Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstrasse 47-51, D-69117 Heidelberg, Germany (email: andreas.voss@psychologie.uni-heidelberg.de).

ABSTRACT

Stochastic diffusion models (Ratcliff, 1978) can be used to analyze response time data from binary decision tasks. They provide detailed information about cognitive processes underlying the performance in such tasks. Most importantly, different parameters are estimated from the response time distributions of correct responses and errors that map (1) the speed of information uptake, (2) the amount of information used to make a decision, (3) possible decision biases, and (4) the duration of non-decisional processes. Although this kind of model can be applied to many experimental paradigms and provides much more insight than the analysis of mean response times can, it is still rarely used in cognitive psychology. In the present paper, we provide comprehensive information on the theory of the diffusion model, as well as on practical issues that have to be considered for implementing the model.

Keywords: Diffusion Model, Mathematical Model, Response Times, fast-dm, EZ, DMAT

Diffusion models in experimental psychology: A practical introduction

Experimental research in cognitive psychology is often based on speeded response time tasks: In most paradigms, participants have to classify stimuli according to category membership, like valence (positive vs. negative), lexical status (word vs. non-word), or familiarity ("old" vs. "new" words in a memory experiment), according to superficial stimulus properties (e.g., color or location), or according to stimulus identity (e.g., in the Eriksen flanker task, Eriksen & Eriksen, 1974). Performance in such tasks is then compared between conditions (e.g., primed vs. non-primed words), stimulus types (e.g., high frequency vs. low frequency words), or between groups of participants (e.g., younger vs. older adults).

Depending on research tradition either mean response time (RT) or accuracy of responses is used as measure of performance. This traditional approach to data analysis has two major drawbacks: Firstly, there is the problem of a missing common metric for performance (Spaniol, Madden, & Voss, 2006; Wagenmakers, 2009) and, secondly, the degree of information usage is poor. We will discuss both problems below before introducing the diffusion model approach (Ratcliff, 1978) and its special advantages.

No Common Metric for Performance

The problem of a (missing) common metric refers to the fact that the performance in response time tasks can be measured in terms of response times or in terms of accuracy (or using both measures). As mentioned above, research traditions differ whether mean latencies or accuracy is considered the most important dependent variable. For example, sequential priming effects are more often analyzed in terms of response times, whereas in memory research typically the percentage of correct responses is used. The availability of two measures of performance poses two problems: On the one hand, there is the risk of the accumulation of Type I error probability: It might be tempting to interpret and report a significant effect on one metric (e.g., mean latencies), and ignore non-significant results on the alternative metric (e.g., accuracy), without making *a priori* predictions (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). On the other hand, statistical power might also be reduced, whenever differences in performance spread over the two metrics, possibly resulting in non-significant effects for both mean RTs and accuracy. Researchers tried to respond to this latter problem by introducing response window techniques (e.g., Greenwald, Draine, & Abrams, 1996) that force the complete effect of an experimental manipulation on the accuracy dimension. However, the extreme time pressure introduced by a response window might change the ongoing cognitive strategies and processes, thus impairing comparability of tasks with and without response windows and endangering external validity of experiments. Finally, the most important problem of the two metrics is based on the possibility of speed-accuracy trade-offs. Whenever directions of effects on RTs and accuracy diverge (i.e., responses in one condition are faster but less accurate or vice versa), it is no longer possible to interpret results in terms of overall performance. In this case, the manipulation (or group membership, etc.) influences decisional style rather than performance (see Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010, for a neural explanation of the speed-accuracy trade-off).

Poor Degree of Utilization of Information

The second weakness of using only mean response time or only accuracy as dependent measure for performance is the relatively poor degree of utilization of the available information. The performance of a participant working through several hundred trials of a response time task is described poorly by one single number (i.e., the mean RT). The available data comprise two RT *distributions* for the two alternative responses that are characterized by their position (e.g., mean), by their specific forms (e.g., standard deviations, skewness) and by their relative sizes representing the percentage of each response. If we use all this information we might get a better understanding of what is going on while the participant performs the experimental task. That is, the information from RT distributions can help to disentangle not only whether performance differs between conditions, but also in what ways it differs, and how this difference in performance can be explained in cognitive terms.

The Diffusion Model as a Theory for Binary Choice Decisions

To utilize the full information provided by position, shapes, and sizes of empirical response time distributions of a speeded response time task, it is essential to draw upon a theory explaining the composition of response time distributions from such tasks. Such a theory is provided by the diffusion model (e.g., Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999). The basic assumptions of the diffusion model approach are that during a binary decision information accumulates continuously and that this accumulation of information can be described by a Wiener diffusion process. This Wiener diffusion process is characterized by a constant systematic component, the so-called *drift*, and by normally distributed random noise. The drift rate determines the average slope of the information accumulation process, that is, the speed and direction of information accumulation. The assumption of random noise explains that the processing of the same stimulus—or the same type of stimulus—results in different response times, and sometimes in different (i.e., erroneous) responses. Most important, the diffusion model can explain the skew that is typically found in empirical response time distributions (Ratcliff, 2002).

Advantages of the diffusion model approach

The full diffusion model is characterized by several parameters that are discussed below. In a diffusion model analysis, values for these parameters are estimated from empirical response time distributions. Although there are different ready-to-use software solutions for diffusion model analyses (Vandekerckhove & Tuerlinckx, 2007b; Voss & Voss, 2007), analyses are still more complex compared to simply entering mean response times into an ANOVA analysis. Therefore, the question arises which benefits come along with the costs of doing this kind of analysis.

The major advantage of the diffusion model approach is that different cognitive processes are mapped on different psychological meaningful parameters. Therefore, the diffusion model provides a solution to the above mentioned problem of the missing common metric of traditional analyses of response time tasks: Effects do not longer spread over different measures. For example, the influence of differences in performance is disentangled from the influence of decisional styles, because these processes are mapped on separate parameters.

The estimation of different process-pure measures for different cognitive processes makes it possible to test specific theories. We get information not only *whether* participants are slower (or less accurate) in an experimental condition, but also *why* this is so. Imagine—for example—that there is a significant difference in mean RTs between two experimental conditions. This deceleration of responses can be explained (1) by slowdown of information uptake or processing, (2) by a more conservative response criterion, or (3) by a delayed (motoric) response execution. With the diffusion model, it is possible to distinguish empirically between these alternative explanations. In a recent study from our own lab (Voss, Rothermund, Gast, & Wentura, 2012) we found, for example, priming effects for associative and affective priming tasks that were nearly identical in terms of response times (about 10 ms). However, diffusion model analyses revealed that these effects were based on completely different mechanisms: While semantically associated primes caused a faster identification of the target word, affectively matching primes increased the speed of response execution.

Besides the fact that the diffusion model provides more specific measures, it will in many cases also provide more valid measures for specific research questions. It can also be argued that parameter estimates are less noisy measures compared to response time means, which could also improve reliability. However, this latter question needs to be addressed empirically.

The Prevalence of Diffusion Model Analyses in Psychological Research

The diffusion model approach was introduced as a tool for analyzing data from speeded response time tasks three and a half decades ago by Roger Ratcliff (1978). Nonetheless, the usage of this kind of modeling was restricted for quite a long time to a small number of researchers who invested a lot of effort in programming their own software solution. Recently, however, different tools were published simultaneously that allow applying the diffusion model without extensive programming skills. These tools comprise the *EZ*-diffusion model (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008; Wagenmakers, van der Maas, & Grasman, 2007), *DMAT* (Vandekerckhove & Tuerlinckx, 2007a, 2008), and *fast-dm* (Voss & Voss, 2007, 2008). We will discuss these programs below in the section on parameter estimation procedures. The availability of programs for diffusion model analyses answers to a strong increase in interest for diffusion model analyses in different fields of psychology. This increase of interest is reflected by an exponential increase in the number of citations of the original publication introducing the diffusion model to psychology (Figure 1). We hasten to add that obviously not all articles citing Ratcliff (1978) are concerned with the diffusion model; however, clearly the vast majority of them will be.

Although the interest in diffusion modeling has grown considerably, this method is far from being a standard method in cognitive psychology. The aim of the present article is to introduce the possibilities and limitations of this form of analysis to a broader audience of researchers that are not yet experts in this field.

The Rationale of the Diffusion Model

In this section, we start with a description of a simplified four-parameter model before introducing several model extensions, followed by a short discussion on how to model performance in different experimental conditions simultaneously.

The Simple Diffusion Model

The diffusion model approach assumes that while performing a binary choice task information accumulates continuously. The accumulated information is represented by an internal counter which is driven in opposite directions by information supporting the different decisional outcomes. For example, in an evaluation task the counter might be increased by positive information and decreased by negative information. The change of the counter over time is modeled as a diffusion process that runs in a corridor between two thresholds. As soon as the upper or lower threshold is hit the decision is reached and response A or B, respectively, is initiated.

Originally, the diffusion model was introduced as a four parameter model (Ratcliff, 1978). In this model, performance is described by the average slope of the diffusion process (drift rate: *v*), threshold separation (*a*), starting-point (*z*), and duration of non-decisional processes (t_0). The basic model is depicted in Figure 2. The Figure shows three sample paths for the diffusion process. The course of these paths varies from trial to trial—even if identical information is available—because of random noise[1]. This variability of process paths leads to different process durations and different process outcomes. Thus, it is possible to predict decision time distributions for both possible responses from the model parameters (Figure 2).

The most important question in a diffusion model analysis is what psychological processes are mapped by the parameters. There are straightforward interpretations for all parameters of the diffusion model: The drift rate (*v*) maps the speed of information uptake and thus provides a measure of performance. In the comparison of conditions the drift reflects task difficulty (with more difficult tasks represented by smaller drift rates). In the comparison of participants drift is a measure for individual cognitive or perceptual speed of information processing (Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007).

Threshold separation (*a*) is defined by the amount of information that is considered for a decision. A conservative decisional style that is characterized by slow but accurate responding leads to large estimates for *a*, while liberal responding implies small threshold separations. Different studies have shown that the parameter *a* is sensitive to speed vs. accuracy instructions (e.g., Voss, Rothermund, & Voss, 2004). Additionally, there is a large body of research showing that age related slowing in response time tasks can be partially explained by more conservative styles of responding (e.g., Ratcliff, Spieler, & McKoon, 2000; Ratcliff, Thapar, & McKoon, 2006, 2010, 2011).

The third parameter of the simple diffusion model, the starting point (*z*), can map *a priori* biases in the decision thresholds. Since *z* can only be interpreted in its relation to *a*, we prefer reporting the relative starting point $z_r=z/a$. If *z* differs from *a/2* (i.e., $z_r \neq 0.5$), different amounts of information are required before deciding on option A or B. Such differences might reflect situations with asymmetric pay-off matrices: For example, Voss et al. (2004) showed that the starting point is moved towards a response threshold when the corresponding

[1] The amount of noise is determined by the so-called diffusion constant (*s*) which is a scaling parameter that cannot be estimated but has to be chosen. Fast-dm (Voss & Voss, 2007) uses a diffusion constant of *s*=1 while Roger Ratcliff usually uses a diffusion constant of *s*=0.1. Estimates for *a, z*, and *v* depend on the chosen diffusion constant. These parameters can be transformed to the case of *s*=1 by dividing the estimated values by the diffusion constant used for the estimation procedure.

response leads to greater rewards. Similarly, in the domain of motivated perception, it has been found that the starting point is closer to the "positive" threshold than to the "negative threshold" in an evaluation task, even when expectancy values for both responses were symmetric (Voss, Rothermund, & Brandtstädter, 2008).

Finally, diffusion model analyses take into account the duration of non-decisional processes (t_0 *or* t_{err}). Such processes may comprise basic encoding processes, the configuration of working memory for a task, and processes of response execution (i.e., motor activity). The estimated duration of these processes is added to the decision times predicted by the diffusion process, resulting in a shift of the predicted RT distributions. A common finding is that extra-decisional processes are slowed in elder participants (e.g., Ratcliff et al., 2000). Recently, Schmitz and Voss (2012) showed that task switching costs are partially mapped onto t_0, at least when task-switches cannot be predicted. In this case, obviously, the working memory has to be configured for the actual task, before the decision process can start.

Inter-Trial Variability

To accurately accommodate different shapes of RT distributions of correct responses and errors, Ratcliff suggested extending the model to allow for so-called inter-trial variability in performance (e.g., Ratcliff & Rouder, 1998). This extension permits variability of the parameters of the simple model across trials of an experiment. For example, the drift might not be exactly the same for each trial of one condition of an experiment, either because of fluctuations of the participant's attention, or because of differences in stimuli.

Specifically, it has been proposed to model inter-trial variability of drift, starting point, and of the non-decisional parameter. The drift is assumed to be normally distributed with mean v and standard deviation s_v (or η). For the sake of simplicity, for starting point and non-decisional component, uniform distributions around z and t_0 with the width s_z and s_{t0} are adopted. Recently Ratcliff (2012) showed that these distributional assumptions still lead to valid results if the true distributions differ.

For most applications of the diffusion model, inter-trial variability will be of minor interest. However, sometimes adopting these parameters increases model fit notably. For example, Ratcliff and Rouder (1998) showed that large variability of drift can explain slow errors and large variability of starting point can explain fast errors. Nonetheless, the influence of s_v and s_z on predicted response time distributions is rather limited and thus can only be estimated with any reliability from huge data sets. This is different for s_{t0}, which shows a greater effect on the shape of response time distributions (i.e., reducing the skewness).

Differences in Speed of Response-Execution

Another suggestion to extend the diffusion model relates to the non-decisional component (Voss, Voss, & Klauer, 2010). Typically, it is assumed that t_0 is equal for both responses. However, this assumption might be wrong whenever motor-response programs differ in level of pre-activation. For example, when one response occurs more frequently, is more urgent, or is more likely in a given situation, it is highly plausible that this response will be executed with more vigor, resulting in a faster motor response. Voss et al. (2012) showed that categorical priming (e.g., affective priming) can be explained by a faster execution of the primed response. According to this argument, the prime stimulus (pre-) activates a specific response program, which leads to a faster execution of this response, when it is finally triggered by the target stimulus.[2]

Complex Models: Mapping Different Conditions

Experimental RT-paradigms typically comprise different stimulus types or experimental conditions. When modeling such data with the diffusion model, the researcher has to decide whether completely independent models should be estimated for each condition, or whether certain parameters are restricted to be equal across conditions. Especially when data sets are small to medium size (i.e., trial numbers below 200) models will be more stable when all data is fitted simultaneously.

Imagine, as an example, the case of a lexical decision task. In this case, you have—minimally—two types of stimuli (i.e., words and non-words) that require opposite responses. For this task the upper and lower thresholds of the model represent the responses “word” and “non-word”, respectively. Obviously two different drift rates are necessary, because for word stimuli the diffusion process will mostly rise to the upper threshold (positive drift), while non-words will have a negative drift. Also, inter-trial variability of drift may vary when stimuli from one class are more similar than those from the other class. However, it is unlikely that the remaining parameters of the model differ between stimulus types, because participants have no information on the next stimulus before it is presented, and consequently cannot adopt starting point or threshold separation to the stimulus of the next trial.

If, however, a task is considered in which participants do have information about the upcoming trial, it is possible that other parameters but the drift also have to be estimated for each condition separately: For example, in a task switching paradigm, participants may be aware that switching trials are more difficult to process. Therefore, the threshold separation might be increased, if task switches can be predicted (Schmitz & Voss, 2012).

[2] The possibility to map differences in t_0 between responses will be included in the forthcoming version of *fast-dm* (Voss & Voss, 2007), which will be published soon.

Finally, there are cases where—even in the presence of different stimulus types—simple models can be adopted that do not at all differ between different stimuli. To make this possible, data has to be recoded in terms of accuracy: Then, the upper threshold reflects correct responses, and the lower threshold corresponds to error responses. Because there cannot be an *a priori* bias for or against the correct response, the starting point should be fixed to *a*/2 in this case. Recoding your data in terms of accuracy allows for a more parsimonious model (with only six parameters) at the price of the implicit assumptions that (1) drift rates for different stimulus types are identical in absolute magnitude and (2) that there is no bias in starting point. This assumption can be assumed to be met when both stimulus types lead to the same performance, that is, accuracy, as well as position and shape of RT distributions are similar.

Theoretical Assumptions and Prerequisites of Diffusion Model Analyses

To decide whether a task is suited for diffusion model analyses, it is important to explicitly review the theoretical assumptions and task prerequisites that have to be met. We will address all important prerequisites in the following section.

Binary Decisions

Firstly, the applicability of the diffusion model is limited by the fact that it is a model of *binary* decision making. Therefore, the diffusion model as presented here cannot account for performance of multiple responses (for a similar multiple response approach see Brown & Heathcote, 2008; Donkin, Brown, Heathcote, & Wagenmakers, 2011). However, even in the case of a multiple choice task, the diffusion model might be applied: When it is reasonable to assume that the same processes underlie the different responses, it is appropriate to recode responses as correct (upper threshold) vs. error (lower threshold). Consider, for example, the Stroop task (Stroop, 1935). Although there are multiple responses (one for each color), one might try to model accuracy data with the diffusion model, allowing for different drift rates for congruent and incongruent trials. This approach averages performance over different color responses. This procedure is obviously only valid if performance (i.e., response times and accuracy) is similar across trials demanding different responses for both congruent and incongruent trials (see above).

Continuous Sampling

A second basic assumption is that decisions are based on a continuous sampling process. This assumption is obviously plausible for ambivalent visual stimuli that contain information supporting both possible responses, like fields of pixels with different colors that have been used in brightness or color discrimination tasks (Ratcliff, 2002; Ratcliff, Thapar, & McKoon, 2003; Spaniol, Voss, Bowen, & Grady, 2011; Voss et al., 2008; Voss et al., 2004). However, the successful fitting of data from lexical decision tasks (Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, Gomez, & McKoon, 2004) demonstrates that even the identification of words is no sudden insight but is based on a continuous—albeit rapid—increase in familiarity. In this case the information entering the decision process comes from long-term memory rather than from an extern visual stimulus. The same argument applies to diffusion model accounts on recognition memory (Ratcliff, 1978; Ratcliff, Thapar, & McKoon, 2004; Spaniol et al., 2006; Spaniol, Voss, & Grady, 2008): Here, after encoding the stimuli, memory is the source of information driving the decision process. To sum up, we are confident that in many simple decision tasks information sampling can be conceived as a continuous process.

Single Stage Decisions

More problematic might be the implicit assumption that decisions are based on a single-stage processing. Whenever participants adopt more complex strategies (e.g., double-check their solutions with an alternative strategy after an initial threshold is reached), the decision process might be divided in multiple steps with cognitive processes varying between steps. Such multiple-stage decisions cannot be mapped with a simple diffusion model as presented in this paper, although more complex models including diffusion processes for separate stages of information processing have been developed, for example, for the visual search paradigm (Guided Search 4.0, Wolfe, 2007) or for the flanker task (Hübner, Steinhauser, & Lehle, 2010; White, Ratcliff, & Starns, 2011).

Constancy of Parameter Values over Time

Another crucial prerequisite that is related to the single-stage assumption is the assumption of constancy of parameter values across time. Imagine, for example, a very difficult task where a stimulus contains little or no useful information. In this case participants might be tempted to reduce threshold separation when the information accumulation did not reach the *a priori* set thresholds after a couple of seconds. Also, in some tasks drift might not be constant over time, for example, when the decision process starts prior to fully encoding a stimulus. In this case, the process might start with a small drift rate that increases in later stages of information processing. Therefore, paradigms with stimuli that are easy to encode are optimal for a diffusion model analysis.

Decision Times

From the assumptions described above it is often derived that diffusion model analyses apply primarily for tasks with latencies below one second (e.g., Ratcliff & Rouder, 1998). If decisions take notably longer, information processing might comprise qualitatively different stages or parameter values might differ over time. However, this arbitrary limit artificially restricts the area of application for diffusion models. There is no empirical evidence stating that diffusion models cannot be successfully applied for longer decisions (e.g., RT ≈ 10s). For such applications obviously test for model fit and parameter validity are of crucial importance.

Numbers of Trials and Percentage of Errors

Next to the theoretic assumptions summed above that might restrict applicability there is one practical limitation of the diffusion model analysis: To get reliable estimates for seven—or more, in case of multiple conditions—diffusion model parameters a high number of decisions per participant is essential. For example, in a recent experiment by Leite and Ratcliff (2011, Exp. 2), participants had to complete 5 sessions of 64 blocks with 36 trials each (i.e., 11,520 trials per participant). However, diffusion models have been applied successfully to experiments with notably smaller number of trials: Klauer, Voss, Schmitz, and Teige-Mocigemba (2007, Exp. 2) report diffusion model results that are based on 72 trials per participant.

As will be discussed below, the required trial number depends strongly on the estimation procedure (i.e., the adopted optimization criterion), the percentage of errors, and the complexity of the model. Rough estimates of the minimum number of trials that we consider essential for a sound analysis are presented in Table 1.

Parameter Estimation

In a diffusion model analysis, typically data from each participant are modeled separately, resulting in separate sets of estimates for all parameters, which subsequently can be entered in inferential statistical analyses. It is also possible to collapse data from all participants or from groups of participants with similar performance (so-called “super subjects“; e.g., Ratcliff, Perea, Colangelo, & Buchanan, 2004) for the analyses to increase the data base for parameter estimation.

The estimation procedure is based on a multidimensional search for the parameter estimates that lead to an optimal fit of predicted and empirical response time distributions. This search can be computationally costly because of the high number of parameters and because the calculation of predicted distributions takes some time, even for modern high-speed computers.

Figure 3 shows the predicted response time distributions from 8 different parameter sets. Model A (with a=1, z_r=0.5, v=2, t_0=0.5, s_z=0, s_v=0, s_{t0}=0) serves as a comparison standard, and the following 7 panels show how the distributions change when the value of a single parameter is modified (Panel B: increased threshold separation; C: increased starting point; D: increased drift; E: increased non-decisional parameter; F: increased variability of starting point; G: increased variability of drift; H: increased variability of the non-decisional parameter). To facilitate comparison, predicted distributions from the comparison model (A) are presented in each panel as hatched areas. A closer look on Figure 3 might help to understand the problems of parameter estimation. Firstly, there is the problem of model mimicry: If you compare, for example, the models with increased starting point (panel C) and increased drift rate (panel D), it is evident that predictions are fairly similar. The main difference lies in the prediction of somewhat faster error responses in the high-drift model. Therefore, you need a high number of error responses in the empirical data to be able to reliably differentiate between these two models.

A second problem is that some parameters have only minor influence on the predictions. This is especially true for inter-trial variabilities of starting point and drift rate: Although quite extreme values were chosen for the figure (the starting point follows a uniform distribution from 0.2 to 0.8 in panel F and drift follows a normal distribution with mean 2 and standard deviation 2 in panel G), the influence on the RT distributions is limited. Obviously, one needs large empirical distributions (large trial numbers) to estimate these parameters with any accuracy. In the case of small to medium trial numbers one might decide to fix these parameters to 0 to make the model more parsimonious and to enhance stability of the estimation procedure.

Comparison of Optimization Criteria

For the multidimensional search for the optimal vector of parameter values, an optimization criterion has to be defined that quantifies the fit between predicted and empirical RT distributions. Because the choice of these criteria has influence on the speed, precision and robustness of the estimation procedure (Ratcliff & Tuerlinckx, 2002), we will briefly discuss three different approaches in the following sections.

Maximum Likelihood. Mathematically most efficient are *maximum likelihood* (ML) approaches (cf. Klauer et al., 2007, for an example adopting this method). For this approach, the logarithmized density of predicted RT distributions is summed over all responses, and this sum is maximized in the search. The drawback of this method is that results can be strongly biased by single outliers: For example, when using a ML approach, a single fast guessing response (that does not result from a diffusion process) might force t_0 to be very small, because otherwise this fast response would lead to a density of zero, rendering the total likelihood to be zero as well (log-likelihood is no longer defined in this case). Another disadvantage is that calculation can be very slow in case of large trial numbers. Consequently, we recommend using a ML based search only if data sets are so small that other optimization criteria fail and if a careful outlier analysis was conducted.

Chi Square. Most frequently used are searching algorithms based on the *χ^2 statistic* (e.g., Ratcliff & McKoon, 2008; Ratcliff & Tuerlinckx, 2002). This procedure uses quantiles from the empirical RT distributions to define bins on the RT axis. Ratcliff suggests using 6 bins that are defined by the .1, .3, .5, .7, and .9 quantiles of the empirical RT distributions. The outer (open) bins contain 10 percent of data each, while all inner bins comprise 20% of trials. From the predicted cumulative distribution function it is calculated how many trials are predicted for each bin (by multiplying the portion of the predicted distribution for each bin by the total number of trials). Then, a χ^2 value is calculated from the numbers of observed and predicted responses from the 12 bins (6 for the upper and 6 for the lower threshold):

$$\chi^2 = \sum \frac{\left(n_{observed} - n_{predicted}\right)^2}{n_{predicted}}$$

Advantages of the χ^2 approach are the fast calculation (independent of trial numbers), and the robustness against outliers. Since the first and last bin are open bins, only the numbers of responses within these bins are important, not the actual latencies of each response. Therefore, even an outlier of 0 ms would not distort results dramatically. However, these advantages come at a cost as well: Due to the binning, information is lost, and in case of small trial numbers, the identification of empirical quantiles might be inaccurate. This is especially problematic in experiments with high accuracy and hence few error responses. Therefore, we recommend using a χ^2 approach only for studies with large trial numbers (i.e., at least 500 trials), and enough error trials (the smaller distribution should have at least 20 responses for each participant, so that the first and last bin comprises at least 2 responses). If there are fewer errors, results of the estimation procedure tend to depend strongly on the handling of these (e.g., ignoring error information, collapsing all errors in one bin, or nonetheless using six error bins).

Kolmogorov-Smirnov. A third possibility for the optimization criterion is based on *the Kolmogorov-Smirnov (KS) statistic* (Voss et al., 2004; Voss & Voss, 2007). This statistic is the maximum vertical distance between predicted and empiric cumulative response time distributions. In the case of the parameter search for the diffusion model there are always two empirical and two theoretic distributions (as there are two thresholds) that have to be compared simultaneously. This problem has been solved by Voss et al. (2004) by multiplying all RTs from the lower threshold by -1. Thus, both distributions can be combined on a single dimension without overlapping each other (Figure 4). The KS approach can be considered as a compromise between the highly efficient ML method and the more robust χ^2 method. The KS approach—like the χ^2 method—provides robust estimates in the presence of outliers and simultaneously—like the ML method—considers the exact shape of the response time distribution without categorizing responses.

Table 1 sums the strengths and weaknesses of the three optimization criteria. Efficiency reflects the ability to accurately recover true parameter values from small data sets, robustness reflects the stability of estimates in the presence of outliers, and calculation speed points to the duration of the complete parameter estimation procedure.

It is very difficult to provide a recommendation for a minimum number of trials that is required for robust parameter estimations. Generally, estimations will be more precise for data sets comprising a high percentage of error responses (i.e., when there are distributions of reasonable size at both thresholds). Secondly, the estimation procedure tends to be more robust when the number of free parameters is reduced. Especially fixing the starting point to $z=a/2$ notably increases the stability of results. Most important, when there are participants with virtually no error responses, fixing the starting point is indispensable, because then the distance from starting point to the lower threshold is no longer defined. Finally, the necessary trial number depends on the number of experimental conditions, that is, models might be more stable when different conditions are modeled simultaneously, while some parameters are fixed across conditions.

Although the exact dependency of the required number of trials on these factors is still unclear, we decided to give rough recommendations for the minimum trial numbers we consider to be necessary for an acceptable diffusion model analysis, because this is one of the most frequent questions posed by cognitive researchers who think about applying the diffusion model to their data. Note however, that larger trial numbers are strongly recommended.

Existing Software Solutions

In order to facilitate the application of diffusion model analyses different software solutions have been developed allowing parameter estimation also to researchers with limited programming experience. Important differences between these programs regard the degree of information used for parameter estimation and the number of model parameters that can be estimated.

For the *EZ-diffusion model* (Grasman et al., 2009; Wagenmakers et al., 2008; Wagenmakers et al., 2007) a JavaScript, an Excel sheet, R code, and a MATLAB implementation are available (see http://www.ejwagenmakers.com/papers.html for links to the respective implementations). The *EZ-diffusion model* makes use of a limited degree of information of the observed RT distributions. Only the mean and variance of the correct responses and the accuracy rate are used for parameter estimation. The estimation procedure is—as the name of the program implies— *easy* as parameter estimates can be immediately obtained by entering the three calculated values (mean, variance, and accuracy) into three equations. Thereby parameter estimates for the simple diffusion model can be obtained, that is, for drift rate, threshold separation and duration of non-decisional processes. The parameter calculation via these closed-form equations is very fast as no time-consuming iterative optimization process has to be applied. However, the restrained use of information (especially about the error trials—only the percentage of observed errors is considered) do not allow the estimation of inter-trial variabilities. These are implicitly assumed to be equal to zero. In the standard *EZ* model the starting point cannot be calculated either. It is fixed to the mid-point of the threshold separation ($z = a/2$) while *EZ2* (Grasman et al., 2009)—a more recent, extended version of *EZ*—allows the calculation of an estimate for the starting point. Furthermore *EZ2*, in contrast to *EZ*, allows the estimation of different values for a parameter depending on diverse conditions (e.g., one drift rate for words, another for non-words) while at the same time the other parameters are held constant over conditions. Another extension of *EZ*, so called *robust-EZ* (Wagenmakers et al., 2008) deals with contaminant data.

In contrast to *EZ*, the *Diffusion Model Analysis Toolbox* (DMAT: Vandekerckhove & Tuerlinckx, 2007a, 2008) and *fast-dm* (Voss & Voss, 2007, 2008) utilize more information from the RT distributions to draw conclusions about the decision processes. *DMAT* is a MATLAB toolbox which is available from the website http://ppw.kuleuven.be/okp/software/dmat/. *Fast-dm* is a command-line program implemented in C that can be downloaded from the website http://www.psychologie.uni-heidelberg.de/ae/meth/fast-dm/. While *EZ* only requires the mean and variance of correct responses and the accuracy rate, for the use of *fast-dm* and *DMAT* all correct and error response times have to be supplied to the program. A file with at least two columns is needed: one coding the accuracy of the response (error vs. correct response), the other containing the response times. Considering accuracy rate and distribution of correct as well as error responses *DMAT* and *fast-dm* allow the estimation of all of the diffusion model parameters, i.e. inclusive of starting point and inter-trial variabilities. Like *EZ2* both *DMAT* and *fast-dm* comprise the option of restricting parameters across conditions while letting other parameters vary between these conditions.

Fast-dm and *DMAT* use the parameters *v*, *a* and t_0 as estimated by *EZ* as starting points for an iterative optimization routine. The principal difference between *DMAT* and *fast-dm* lies in the optimization criterion used for parameter estimation. While *DMAT* is based on the χ^2 statistic *fast-dm* uses the Kolmogorov-Smirnov (KS) approach. Therefore, *fast-dm* is characterized by the usage of the complete distributional information, while DMAT draws upon the number of RTs in different "bins" on the RT axis. As outlined above (see Comparison of Optimization Criteria) the higher degree of information usage of the KS method implies longer calculation times but on the same time might lead to higher efficiency of parameter estimation.

To compare performance and results of different diffusion model programs it has to be considered that *fast-dm* uses a diffusion constant of s=1 while *DMAT* and *EZ* fix *s* to 0.1. Therefore *DMAT* and *EZ* estimates for all parameters except for t_0 *and* st_0 have to be divided by 0.1 to establish comparability with *fast-dm* results (see Footnote 1). First simulation studies in order to compare the parameter recovery of *fast-dm*, *DMAT* and *EZ* have been conducted by van Ravenzwaaij and Oberauer (2009). Regarding the correlation between the true parameter values (on which the simulated data were based) and the estimated parameter values *EZ* and *fast-dm* emerged as superior to *DMAT*. *EZ* and *DMAT* performed better than *fast-dm* in terms of the recovery of the mean true values. However, more studies systematically varying trial numbers, parameter ranges, and contamination by outliers are necessary to determine which algorithms are superior for which data and which research questions.

Typical Experimental Paradigms for Diffusion Model Analyses

In the following section, we will present short overviews of three experimental paradigms that have been frequently and successfully used for diffusion model analyses. Specifically, these typical diffusion model paradigms comprise brightness- or color-discrimination tasks, recognition memory tasks, and the lexical decision task. For studies employing the diffusion model approach to analyze general principals of information processing (e.g., age related differences in information processing: Ratcliff et al., 2000; Ratcliff, Thapar, Gomez, et al., 2004; Ratcliff, Thapar, & McKoon, 2001; Ratcliff et al., 2003; Ratcliff, Thapar, & McKoon, 2004; Spaniol et al., 2006; Thapar, Ratcliff, & McKoon, 2003) these well-tested paradigms are recommended, because a good validity of parameters can be assumed.

Brightness, Color, or Numerosity Discrimination

Diffusion model analyses have been applied in numerous studies requiring the classification of ambiguous stimuli with respect to brightness (Ratcliff, 2002; Ratcliff & Smith, 2010; Ratcliff et al., 2003, 2006) or color (e.g., Spaniol et al., 2011; Voss et al., 2008; Voss et al., 2004). In these experiments, participants see squares that are composed of a random pixel pattern of two different brightness or hue values (e.g., black vs. white or orange vs. blue). The task is to classify stimuli according to the dominating color, that is, to judge which kind of pixels appears in greater number. Across trials the frequency of occurrence of colors is varied (e.g., 44%, 48%, 52%, or 56% of pixels are white). Structurally very similar are numerosity discrimination tasks that require participants to judge whether a high or low number of stimuli (e.g., asterisks) is presented (e.g., Leite, 2012; Ratcliff, 2008).

Such color, brightness, or numerosity discrimination tasks are very well suited for diffusion model analyses because they meet the theoretical assumptions of the model particularly well: It is highly plausible that performance in these tasks is based on a one-stage continuous information accumulation process, where drift is determined, for instance, by the ratio of presented colors. Another advantage is that stimuli are artificial and initially meaningless, thus making *a priory* biases unlikely. Therefore, stimuli can be assigned with values or meanings experimentally to study biases in decision making (Voss et al., 2008). Finally, such discrimination tasks are very easy to learn, making it a flexible tool for many research questions and even rendering possible the application in animal research (Ratcliff, Hasegawa, et al., 2011; Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007).

Recognition Memory

Recognition memory is the paradigm for which the diffusion model was originally conceived (Ratcliff, 1978) and where it is still frequently applied (e.g., McKoon & Ratcliff, 2012; Ratcliff, Thapar, & McKoon, 2004; Spaniol et al., 2006; Spaniol et al., 2008; White, Ratcliff, Vasey, & McKoon, 2010). The task comprises a study-test paradigm, with stimuli—usually words—being presented once or more in an acquisition phase. In a later recognition phase participants have to judge whether presented stimuli are “old” or “new”. Other memory tasks can be used as well: For example, participants can be asked to decide whether a pair of stimuli has been shown together before in the same way ("intact pair") or whether it has been rearranged. For recognition memory, like for the brightness discrimination paradigm, the theoretical assumptions of the model can be considered to be well met, in that there is a single fast binary decision, and presumably an isolated, recently formed trace in memory which supplies the evidence that is accumulated.

In recent applications, Starns, Ratcliff, and White (2012) and Starns, White, and Ratcliff (in press) use the diffusion model to decide between accounts of the strength-based mirror effect, that is, better recognition performance for stronger memories (e.g., from repeatedly presented stimuli) compared to weaker memories (e.g., stimuli presented only once). Starns and colleagues show that empirical findings are more consistent with a change in the drift *criterion*, than with a differentiation account according to which stronger memories produce at the same time weaker familiarity of “lure” items, which would predict changes in drift rates for both targets and lures (Criss, 2009, 2010). This is particularly relevant for future applications of the diffusion model as it demonstrates that the drift rate can be affected by decision processes.

For the effect of age on recognition memory Ratcliff, Thapar, et al. (2011) and McKoon and Ratcliff (2012) find that in item recognition *non-decision time* and *boundary separation* increase with age, whereas *drift* remains fairly constant (but see Spaniol et al., 2006). In contrast, for associative recognition—i.e., the ability to tell whether a particular pair of word stimuli was presented jointly previously—*drift rate* decreased with age. Some studies also investigate the influence of intelligence on diffusion model results from recognition memory: Drift rate is generally positively associated with intelligence, but in the associative memory task this was much less the case for elder compared to younger participants (McKoon & Ratcliff, 2012).

Lexical Decision

Several studies used the diffusion model to analyze performance in lexical decision tasks (e.g., Ratcliff, Gomez, et al., 2004; Ratcliff, Thapar, Gomez, et al., 2004). This task requires participants to decide quickly whether presented letter strings are valid words. The diffusion model account of the lexical decision task is silent on the details of lexical access and offers a—poorly defined—concept of “wordiness” in its stead (Norris, 2009; Ratcliff, Thapar, Gomez, et al., 2004). Following this concept, words (and non-words) differ in their degree of wordiness, that is, in their typicality for the category “word”. One main finding of the diffusion model approach to the lexical decision task is that many aspects of stimuli reliably map onto *drift rate*. For example, high frequency words show larger drift rates compared to low frequency words, and random letter strings have stronger (negative) drift rates compared to word-like non-words (Ratcliff, Gomez, et al., 2004).

Recently, the lexical decision task has also been applied in diffusion model studies analyzing different forms of sequential priming (Voss et al., 2012; Yap, Balota, & Tan, 2012). Voss et al. (2012) show that semantic (associative) priming increases *drift rate* (i.e., facilitates lexical access), while categorical priming reduces the *non-decisional component* (i.e., speeds response execution). Yap, Balota, and Tan (2012) obtain a more complicated pattern of results for priming: priming mapped onto *non-decisional time* when stimuli were presented clearly visible and influenced *drift* and *non-decisional time* when stimuli were degraded (cf. also Gomez, Perea, & Ratcliff, in press).

Yap, Balota, Sibley, and Ratcliff (2012) found that participants' ability is reflected in *drift* and that those participants who do well in the lexical decision task show less pronounced effects of lexical variables such as length/structure, neighborhood and frequency/semantics. For reading impaired children diffusion model analyses revealed a difference both in *drift* (lower rates for the impaired participants) and a smaller effect on *non-decision time,* in that the reading impaired use a larger *boundary separation* (Zeguers et al., 2011).

A Practical Guide to the Application of Diffusion Models

In the following section we give some important practical advice on how to conduct a state-of-the-art diffusion model study. Specifically, we will address issues of experimental design, data pre-treatment, model specifications, and tests of model-fit and parameter validity.

Experimental Design: Which task should be implemented?

In a first step, an adequate experimental response time paradigm must be chosen. It is always preferable to draw upon tasks that have already been tested for diffusion model analyses before. If a new task has to be used, the theoretical assumptions of the model have to be considered carefully, and the validity of the model should be analyzed empirically in elaborate pre-studies that should be conducted independently of an application question; we will address the issue of empirical parameter validation below (see Voss et al., 2004, for an example of a parameter validation study).

Once an apt task has been identified, the researcher has to decide how many conditions or stimulus types should be used in the experiment. In our experience, the inclusion of different stimulus types that differ in task-difficulty (e.g., four types: easy and difficult stimuli requiring response A or B) often increase robustness of the estimation procedure.

Finally, the number of trials has to be chosen. Generally large trial numbers (e.g., *N*>200) are preferable. However, such massive testing is not only costly in terms of time and money, but extensive practice might also change the underlying cognitive processes in the later blocks or sessions of an experiment (Dutilh, Krypotos, & Wagenmakers, 2011). Therefore a compromise between highly accurate parameter estimation and the practical possibilities has to be adopted. Generally, trial numbers need to be high when the model test is of key importance (i.e., when new paradigms are tested) or when parameters need to be estimated with high precision (e.g., for correlative research). Lower trial numbers might be sufficient when parameters are tested for group differences and when simplified versions of the model are applied.

Analyzing your data

In this paper, we cannot give a comprehensive tutorial on the different computer programs for diffusion model analysis (see section Existing Software Solutions). However, we will present an overview of the typical procedural steps and associated decisions that have to be made.

Data Pre-Treatment. Although the diffusion model is designed to predict the complete response time distribution, the removal of outliers from the individual response-time distribution is highly recommended. For the analyses, fast outliers (e.g., fast guesses) are generally more problematic than slow outliers. Since χ^2 and KS based estimation procedures are relatively robust, liberal criteria for lower and upper outliers (e.g., *fast outliers*: RT<200 ms; *slow outliers:* RT>5000 ms) will often be sufficient. For the ML method, stricter criteria that are derived from individual response time distributions are preferable. For example, all responses 1.5 interquartile distances below the first quartile or above the third quartile of the individual RT distributions might be excluded (outliers sensu Tukey, 1977). Finally, Ratcliff and Tuerlinckx (2002) suggest a highly sophisticated procedure to remove fast outliers: They suggest starting with a fixed upper limit for fast outliers (e.g., 300 ms) and increasing this limit continuously until performance rises above change, that is, until more correct than erroneous responses are made. All trials with RTs below the so found limit are discarded. However, this procedure is only feasible for easy tasks with generally low numbers of errors.

Grouping of Data. Typically, data is modeled for each participant separately. This will require saving data in separate date files. If the individual data sets are too small for a sound parameter estimation, one might consider collapsing data across the complete sample or across so-called super-subjects (Ratcliff, Perea, et al., 2004), that is, across participants with similar response time distributions and error percentages. Grouping has always the disadvantage that it is unclear whether the estimated parameters are valid measures, because—possibly—cognitive processes differ between participants, even if overall performance is similar. Another problem is that subsequent statistical comparisons are impossible or lack power.

Mapping Actual Responses vs. Accuracy. For the diffusion model analysis, responses have to be linked to thresholds. In a binary choice task, this can be done by maintaining the actual responses (e.g., upper threshold = “word”, lower threshold = “non-word” in a lexical decision task), or by linking thresholds to accuracy (i.e., upper threshold = correct response, lower threshold = error). In the first case, one has to use separate drift rates for alternate stimulus types, and drift for the stimuli requiring the response linked to the lower threshold will be negative. To compare speed of information uptake between stimulus types, *absolute values* of the drift have to be checked against each other. In the second case—that is when accuracy data is modeled—one drift rate is sufficient, and implicitly the assumption is made that performance is equal across stimulus types.[3] When thresholds are linked to accuracy the starting point should always be fixed to $z=a/2$, because—logically—there cannot be an a priori bias towards (or against) the correct response.

Varying Parameters between Stimuli or Conditions. Often, diffusion model analyses are employed to test empirically which parameters account for the effect of an experimental manipulation. To answer this question, *different* parameters must be allowed to vary between conditions or stimulus types. If only one parameter is allowed to vary, trivially any present effect will map on this parameter. One possibility is to split the data and estimate completely independent models for each condition.[4] Thus, all parameters could possibly account for an effect of the manipulation. Often more parsimonious models can and should be chosen. For example, it might make sense to assume that threshold separation and starting point are constant across trials. Also, for the sake of simplicity, it might often be helpful to fix variability parameters across conditions and stimulus types.

Selection of an Optimization Criterion. As discussed in the section Comparison of Optimization Criteria, the criteria have different advantages (see Table 1 for a comparison): We recommend using the ML approach for small, the KS approach for medium, and the χ^2 approach for large trial numbers. Until now, these criteria are implemented in different

[3] Obviously, different drift rates could be used in the model on accuracy data as well. However, then recoding will not make the model more parsimonious and parameter estimation will be less stable because the distribution at the lower (error-) threshold will be rather small.

[4] If all parameters are allowed to vary between stimulus types it is highly recommended to use separate data files and thus separate runs of the estimation program. Theoretically *fast-dm* or *DMAT* could estimate all parameters for all conditions in one search. However, the multidimensional search procedure (the SIMPLEX algorithm, Nelder & Mead, 1965) has problems finding the optimal solution when too many parameters are optimized simultaneously.

software solutions. In the forthcoming version of *fast-dm* it will be possible to choose between the three optimization criteria discussed above (ML, KS, or χ^2).

Interpretation of Parameter Values. In the last step, the estimated parameter values are entered as dependent variables into statistical analyses. Thus, it is possible to check which parameters account for differences in performance between groups (e.g., younger vs. elder participants, Ratcliff et al., 2000), stimulus types (e.g., high frequency vs. low frequency words in lexical decisions, Ratcliff, Gomez, et al., 2004), experimental manipulations (e.g., speed vs. accuracy instructions, Voss et al., 2004), and so on. Another strategy is to use parameter estimates for correlational analyses (e.g., predicting intelligence scores, Schmiedek et al., 2007). However, there are two caveats that should be considered prior to the interpretation of diffusion model results. Results can only be considered valid if, firstly, the chosen model fits the data well, and, secondly, if one can be sure about the psychological meaning of the parameters. Both issues will be discussed in the following sections on Model Fit and on Empirical Validation Studies, respectively.

Model Fit

A crucial precondition for the interpretation of diffusion model results is an acceptable model fit. Only if the model can recover response time distributions and accuracy rates adequately, results might reflect the ongoing cognitive processes. Different strategies have been developed to assess model fit.

Statistical Tests of Model Fit. Firstly, it is possible to assess model fit via statistical tests: The χ^2 statistic as well as the KS statistic can be directly translated into *p*-values from the corresponding statistical test for the comparison of predicted vs. empirical response time distributions. Small values of *p* (e.g., p<.05) indicate that the diffusion model cannot account for the data. However, the interpretation of these *p*-values has several problems: (1) Firstly, because the shapes of the predicted RT distributions are fitted flexibly to the empirical distributions the tests will tend to be too conservative, that is, models will be rejected too seldom (D'Agostino, 1986). (2) Secondly, if several conditions are fitted simultaneously (i.e., if at least one parameter is free to vary between conditions), *fast-dm* reports the product of all *p*-values from the different conditions. Therefore, the displayed *p*-values might be very small in case of multiple conditions. (3) Thirdly, results from statistical tests depend strongly on the number of trials: In case of small or medium trial numbers, on the one hand, the power of both χ^2 test and KS tests are small and— consequently—misfits will often not be detected. One the other hand, an *exact* model fit cannot be expected, because—albeit being quite sophisticated—diffusion models as any theory propose a simplified model of reality.

Therefore, statistically significant misfits are to be expected in case of large trial numbers and might be considered rather unproblematic.

Graphical Displays of Model Fit. Because of the problems related to the statistical tests of model fit, graphical displays of concordance of predictions with data have been proposed. A good way to do this is to present a display of fit for each person and each condition separately by plotting the predicted and empirical cumulative distribution functions (CDF) in the same graph.[5] Another possibility is provided by so-called quantile probability plots that display quantiles of the RT distributions as a function of response probabilities for different conditions or stimulus types (e.g., Ratcliff & Smith, 2010). However, in many studies there are too many participants to present separate figures for each model. In this case it might be helpful to average CDFs across participants (Schmitz & Voss, 2012, Appendix A). Another possibility to present model fit for many participants simultaneously is to display scatter plots plotting empirical values (x-axis) against predicted values (y-axis) for accuracy, and quartiles of the RT distributions (Voss et al., 2012, Appendix B). The main problem of all types of graphical display of model fit is the ambiguity of interpretation: There is no clear criterion on how much deviance of data from predictions is acceptable.

Monte-Carlo-Simulations. Monte Carlo simulations provide a highly sophisticated possibility of overcoming the discussed biases of *p*-values from statistical model tests (Clauset, Shalizi, & Newman, 2009). For this purpose, many (e.g., 1,000) data sets have to be simulated from the diffusion model, matching the characteristics of the empirical data. That is, parameter values for the simulation should be based on the estimated parameter values, and the numbers of trials, conditions, etc. should be equivalent to those used in the experiment. To accomplish this, parameter values for the simulation study might be generated following the multivariate normal distribution defined by the mean values and the variance-covariance matrix from the estimated parameters (e.g., using the *mvrnorm* routine from the MASS R-package). For the simulation, the construct-samples routine from *fast-dm* (Voss & Voss, 2007) can be used.[6] These simulated data sets are then re-analyzed with the diffusion model. From the results a distribution of fit-values (e.g., *p*-values provided by *fast-dm*) can be obtained, and the 5% quantile of this distribution can be taken as a critical value to assess model fit of the empirical models.

[5] Predicted CDFs can be computed, for example, by the plotCDF routine from the *fast-dm* software (Voss & Voss, 2007).

[6] If there are multiple conditions, data sets have to be simulated separately for each condition, and afterwards combined into one file.

Interpretation of Model Fit. If a low percentage of models (e.g., less than 5 percent) shows suspicious fit-indices, it can be assumed that the diffusion-model generally describes data well. Obviously, a good model fit does not prove that the diffusion model is the "correct" model. Especially for small samples, non-significant results have to be interpreted with great caution. In any case, it is recommended to discard data from participants with bad model fit before running further analyses.

If data from a substantially larger portion of participants cannot be fitted, the diffusion model in the applied form has to be discarded. Possibly, a stricter exclusion of outliers or a more complex model with fewer restrictions can fit the data in this case. Often, however, bad model fit indicates that the theoretical assumptions of the model are not met.

Empirical Validations

Since it is difficult to assess the adequacy of the diffusion model with an analysis of model fit alone, empirical validations of the parameters are often a useful addition. For such validation studies, face-valid experimental manipulations have to be adopted for each parameter of the diffusion model. Then, independent models are estimated for each experimental condition, where manipulations theoretically could map on all parameters. If each manipulation successfully and exclusively influences the expected parameter(s), the validation can be considered successful. An example for an elaborated validation study for the color-discrimination task is provided by Voss et al. (2004): In a series of experimental manipulations with high face-validity the authors showed that it is possible to manipulate single parameters in the proposed way: For example, task-difficulty exclusively mapped on the drift parameter while speed-accuracy instructions did influence threshold separation.

We recommend employing such validation studies whenever the diffusion model is applied to a new task. A successful validation might even be considered to be more important than the demonstration of an excellent model fit, because model fit will often be good in case of small trial numbers.

Conclusions and Perspectives

The diffusion model is not a new approach in psychology. However, in the first two decades after its initial proposal (Ratcliff, 1978), it was used rather rarely. This slow increase in popularity has several reasons: Firstly, the implementation of this method was difficult, before software solutions like *fast-dm* or *DMAT* were published. Secondly, access to computational power, which is necessary for this kind of analyses, was limited. And finally, there is still a general lack of knowledge about the possibilities and problems of diffusion model analyses.

With the present paper we hope to address this last hurdle, and help to clear the way to diffusion model analyses for a broad community of cognitive researchers.

In recent years, applications of the diffusion model have greatly increased both in number as well as in broadness of addressed research areas. Such new fields of application comprise, for example, research on intelligence (Ratcliff et al., 2010; Schmiedek et al., 2007), or clinical psychology (White, Ratcliff, Vasey, & McKoon, 2009; White et al., 2010). We are confident that future diffusion model analyses will provide interesting insights in many other fields of psychology as well.

References

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences, 33*(1), 10-16. doi: 10.1016/j.tins.2009.09.002

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

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev., 51*(4), 661-703. doi: 10.1137/070710111

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59*(4), 297-319. doi: 10.1016/j.cogpsych.2009.07.003

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 484-499. doi: 10.1037/a0018435

D'Agostino, R. B. (1986). Tests for the normal distribution. In R. B. D’Agostino & D. A. Stephens (Eds.), *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review, 18*(1), 61-69. doi: 10.3758/s13423-010-0022-4

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice: A diffusion model account. *Experimental Psychology, 58*(6), 434-442. doi: 10.1027/1618-3169/a000111

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143-149.

Gomez, P., Perea, M., & Ratcliff, R. (in press). A diffusion model account of masked vs. unmasked priming: Are they qualitatively diferent? *Journal of Experimental Psychology: Human Performance and Perception*.

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*(2), 55-68.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science, 273*(5282), 1699-1702. doi: 10.1126/science.273.5282.1699

Hübner, R., Steinhauser, M., & Lehle, C. (2010). A dual-stage two-phase model of selective attention. *Psychological Review, 117*(3), 759-784. doi: 10.1037/a0019471

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*(3), 353-368. doi: 10.1037/0022-3514.93.3.353

Leite, F. P. (2012). A comparison of two diffusion process models in accounting for payoff and stimulus frequency manipulations. *Attention, Perception, & Psychophysics, 74*(6), 1366-1382. doi: 10.3758/s13414-012-0321-0

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*(7), 651-687.

McKoon, G., & Ratcliff, R. (2012). Aging and IQ effects on associative recognition and priming in item recognition. *Journal of Memory and Language, 66*(3), 416-437. doi: 10.1016/j.jml.2011.12.001

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308-313.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review, 116*(1), 207-219. doi: 10.1037/a0014259

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

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

Ratcliff, R. (2008). Modeling aging effects on two-choice tasks: Response signal and response time data. *Psychology and Aging, 23*(4), 900-916. doi: 10.1037/a0013930

Ratcliff, R. (2012). Parameter Variability and Distributional Assumptions in the Diffusion Model. *Psychological Review*. doi: 10.1037/a0030775

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review, 111*(1), 159-182. doi: 10.1037/0033-295x.111.1.159

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Childers, R., Smith, P. L., & Segraves, M. A. (2011). Inhibition in Superior Colliculus Neurons in a Brightness Discrimination Task? *Neural Computation, 23*(7), 1790-1820. doi: 10.1162/NECO_a_00135

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97*(2), 1756-1774. doi: 10.1152/jn.00393.2006

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

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition, 55*(2), 374-382. doi: 10.1016/j.bandc.2004.02.051

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

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*(1), 70-94. doi: 10.1037/a0018128

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin and Review, 7*(1), 1-25.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278. doi: 10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16*(2), 323.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65*(4), 523-535. doi: 10.3758/BF03194580

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*(4), 408-424. doi: 10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review, 13*(4), 626-635. doi: 10.3758/BF03193973

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*(3), 464-487. doi: 10.1037/a0023810

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaching to dealing with contaminant reaction and parameter variability. *Psychonomic Bulletin and Review, 9*(3), 438-481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261-300. doi: 10.1037/0033-295X.106.2.261

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414-429.

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. doi: 10.1037/a0026003

Spaniol, J., Madden, D. J., & Voss, A. (2006). A Diffusion Model Analysis of Adult Age Differences in Episodic and Semantic Long-Term Memory Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 101-117. doi: 10.1037/0278-7393.32.1.101

Spaniol, J., Voss, A., Bowen, H. J., & Grady, C. L. (2011). Motivational incentives modulate age differences in visual perception. *Psychology and Aging, 26*(4), 932-939. doi: 10.1037/a0023297

Spaniol, J., Voss, A., & Grady, C. L. (2008). Aging and emotional memory: Cognitive mechanisms underlying the positivity effect. *Psychology and Aging, 23*(4), 859-872. doi: 10.1037/a0014218

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1137-1151. doi: 10.1037/a0028151

Starns, J. J., White, C. N., & Ratcliff, R. (in press). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition*.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662.

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging, 18*(3), 415-429. doi: 10.1037/0882-7974.18.3.415

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Weasley.

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*(6), 463-473. doi: 10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007a). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2007b). Fitting the Rateliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44*(4), 1048-1056. doi: 10.1016/j.jesp.2007.10.009

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2012). Cognitive Processes in Associative and Categorical Priming: A Diffusion Model Analysis. *Journal of Experimental Psychology: General*. doi: 10.1037/a0029459

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response tendency and decision biases: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*, 539-555.

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. doi: 10.1080/09541440802205067

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15*(6), 1229-1235. doi: 10.3758/pbr.15.6.1229

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*(1), 3-22. doi: 10.3758/bf03194023

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. doi: 10.1177/1745691612463078

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

White, C. N., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition and Emotion, 23*(1), 181-205. doi: 10.1080/02699930801976770

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*(1), 39-52. doi: 10.1016/j.jmp.2010.01.004

Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated models of cognitive systems.* (pp. 99-119). New York, NY US: Oxford University Press.

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

Yap, M. J., Balota, D. A., & Tan, S. E. (2012). Additive and Interactive Effects in Semantic Priming: Isolating Lexical and Decision Processes in the Lexical Decision Task.

Journal of Experimental Psychology: Learning, Memory, and Cognition. doi: 10.1037/a0028520

Zeguers, M. H. T., Snellings, P., Tijms, J., Weeda, W. D., Tamboer, P., Bexkens, A., & Huizenga, H. M. (2011). Specifying theories of developmental dyslexia: A diffusion model analysis of word recognition. *Developmental Science, 14*(6), 1340-1354. doi: 10.1111/j.1467-7687.2011.01091.x

Table 1
Comparison of Optimization Criteria

	Optimization Criterion		
	Maximum Likelihood	Chi-Square	Kolmogorov-Smirnov
Efficiency	High	Low	High
Robustness	Low	High	High
Computational Speed	Low	High	Low
Required Number of Trials	Small (*N*>40)	Large (*N*>500)	Medium (*N*>100)

Note. The values for the required numbers of trials are just rough estimates for the lower bound of acceptable trial numbers. See text for further explanations.

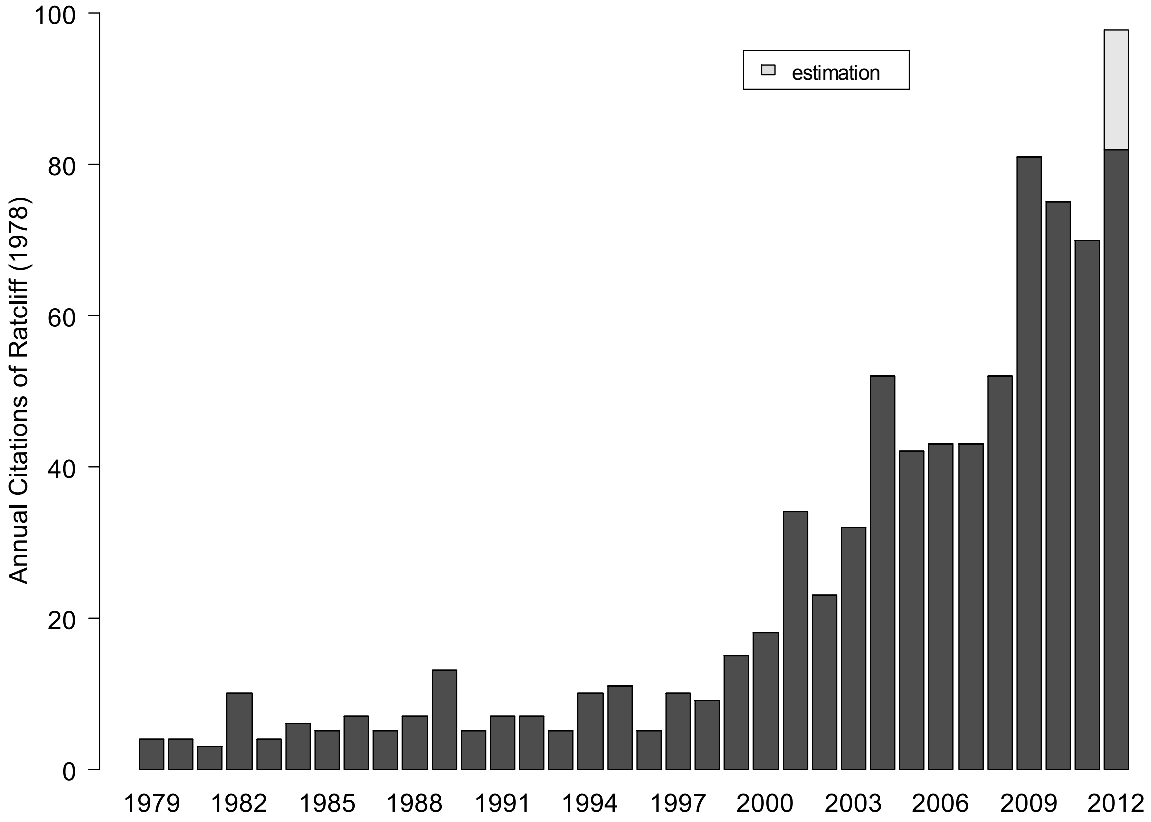


Figure 1. Number of annual citations of the original publication introducing the diffusion model account in psychology (Ratcliff, 1978). Data includes all papers listed in PsycINFO Database until October 2012 (Light grey bar: Estimation for Nov. and Dec. 2012).

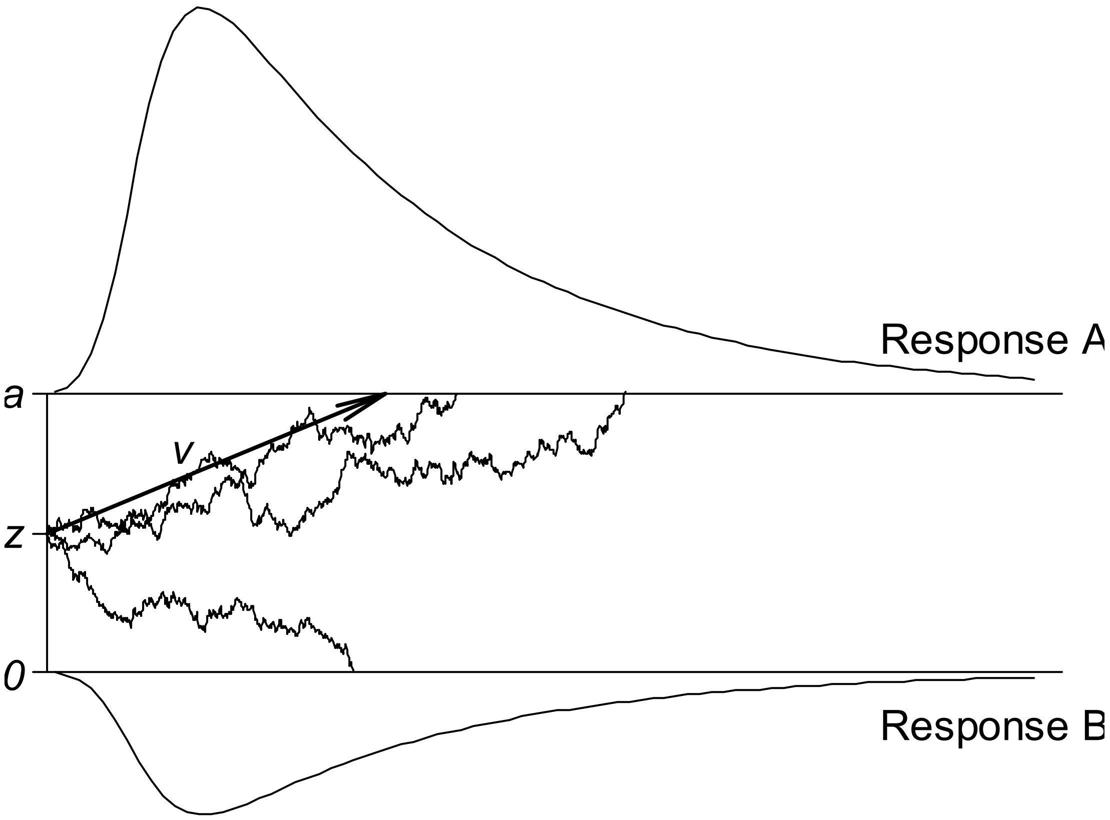


Figure 2. Simplified version of the diffusion model: An information accumulation process starts at starting point *z* and runs over time with the mean slope *v* until it hits an upper (*a*) or lower (0) threshold. Because of random noise, the process durations and outcomes vary from trial to trial. Outside the thresholds decision-time distributions are shown.

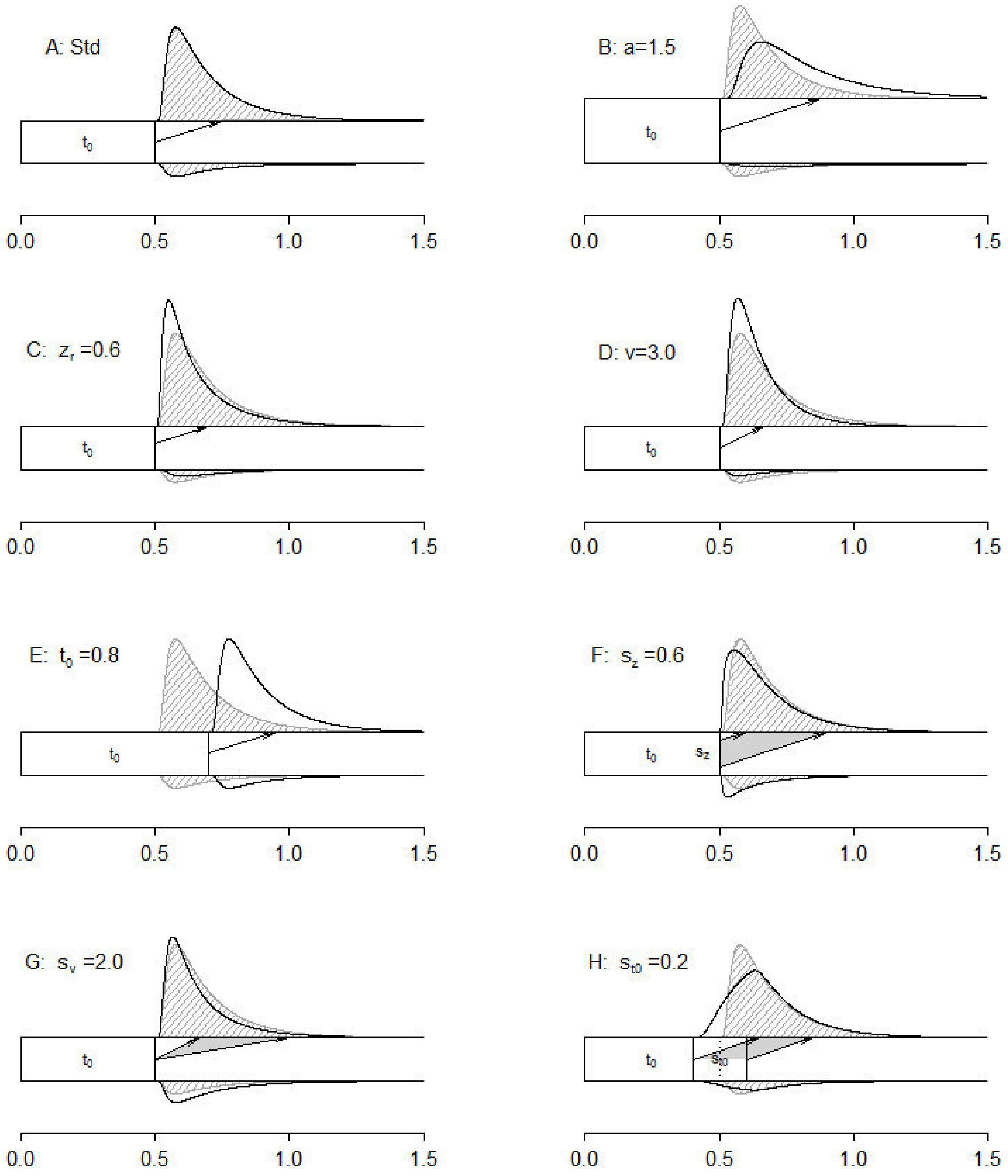


Figure 3. Predicted RT distributions from different parameter sets. Panel A shows predictions from a comparison model with a=1, z_r=0.5, v=2, t_0=0.5, s_z=0, s_v=0, s_{t0}=0. In the following panels, always one parameter is increased. To facilitate comparison, the distributions from Panel A are presented as hatched shapes in each display.

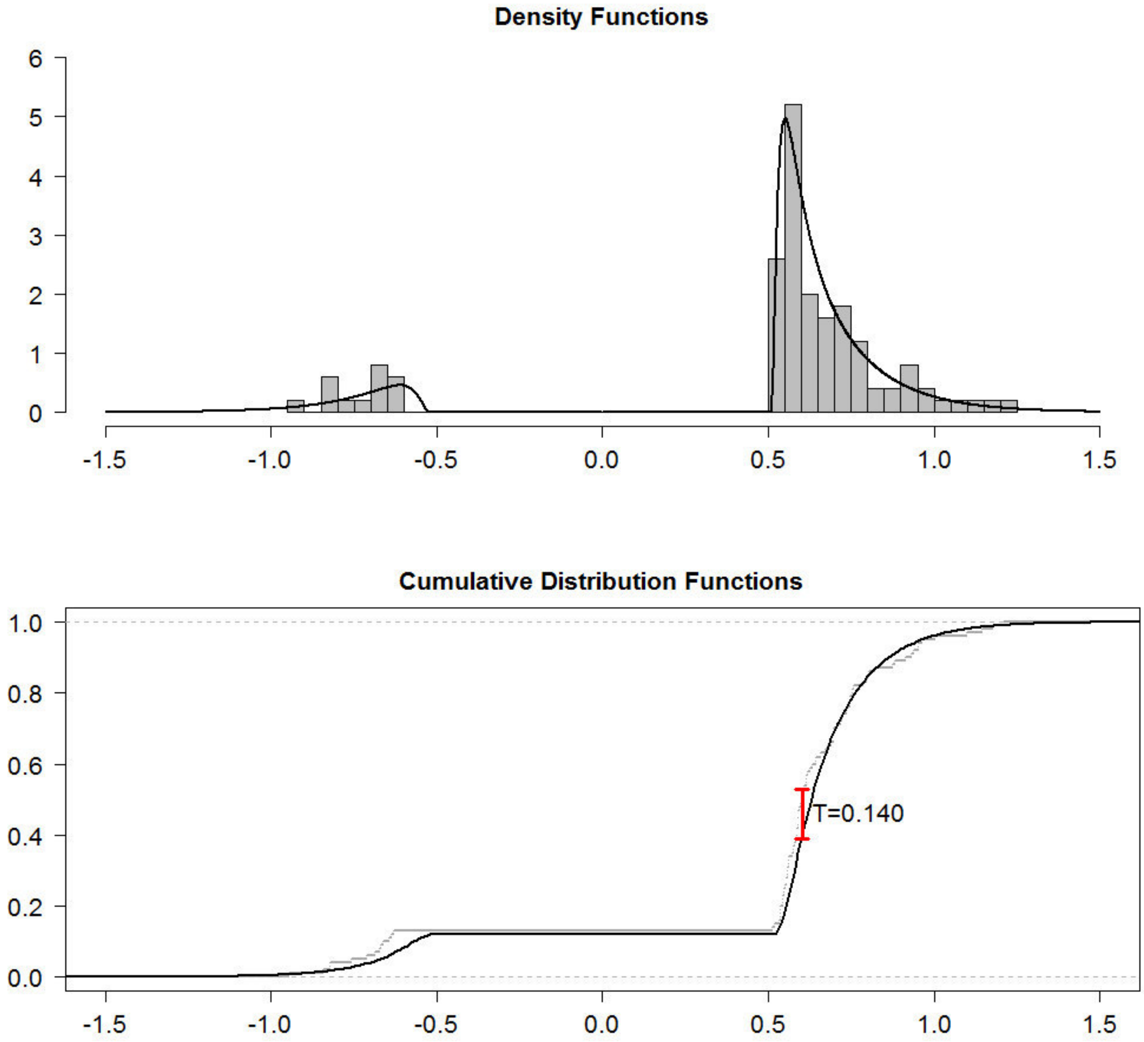


Figure 4. Illustration of the Kolmogorov-Smirnov approach. Distributions from both thresholds are combined in one distribution function by multiplying all times from the lower threshold by -1. The upper panel shows the comparison of empirical (histogram) and predicted (lines) density functions. The lower panel shows the cumulative distribution functions (CDF; data: grey line; predictions: black line). In the parameter search, the maximum vertical distance (*T*) between both CDFs is minimized.

Appendix A 2

Manuscript 2: 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*(336).

doi: 10.3389/fpsyg.2015.00336

Assessing cognitive processes with diffusion model analyses: a tutorial based on fast-dm-30

Andreas Voss[1], Jochen Voss[2], & Veronika Lerche[1]

[1]Ruprecht-Karls-Universität Heidelberg
[2]University of Leeds

Author Note

Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Jochen Voss, University of Leeds, Leeds, UK; Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany.

This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant No. VO1288/2-1).

Correspondence concerning this article should be addressed to Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstr. 47-51, D-69117 Heidelberg, Germany (email: andreas.voss@psychologie.uni-heidelberg.de).

ABSTRACT

Diffusion models can be used to infer cognitive processes involved in fast binary decision tasks. The model assumes that information is accumulated continuously until one of two thresholds is hit. In the analysis, response time distributions from numerous trials of the decision task are used to estimate a set of parameters mapping distinct cognitive processes. In recent years, diffusion model analyses have become more and more popular in different fields of psychology. This increased popularity is based on the recent development of several software solutions for the parameter estimation. Although these programs make the application of the model relatively easy, there is a shortage of knowledge about different steps of a state-of-the-art diffusion model study. In this paper, we give a concise tutorial on diffusion modelling, and we present *fast-dm-30*, a thoroughly revised and extended version of the *fast-dm* software (Voss & Voss, 2007) for diffusion model data analysis. The most important improvement of the *fast-dm* version is the possibility to choose between different optimization criteria (i.e., Maximum Likelihood, Chi-Square, and Kolmogorov-Smirnov), which differ in applicability for different data sets.

Keywords: fast-dm, diffusion model, parameter estimation, response time distribution

Assessing cognitive processes with diffusion model analyses: a tutorial based on fast-dm-30

Six years ago, we published *fast-dm-26* (Voss & Voss, 2007). Since then, applications of diffusion models have thrived in different domains of psychology (Voss, Nagler, & Lerche, 2013): Although diffusion models are still far from being a standard method in the cognitive sciences, they are now successfully applied by many different researchers addressing a wide variety of research questions. Different aims of the application of diffusion models can roughly be grouped into three groups.

A first type of diffusion model studies is interested in the development of cognitive models, and—specifically—in demonstrating that the diffusion model adequately describes the ongoing cognitive processes (e.g., Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2004). For such studies, the key objective is demonstration of a good model fit, because a satisfactory model fit supports the assumption that actual cognitive processes are similar to the processes presumed by the model.

Secondly, diffusion models can be used to test predictions from psychological theories (e.g., Voss, Rothermund, Gast, & Wentura, 2013). For such studies the validity of the diffusion model for the applied task should be undisputed. The application of the diffusion model aims at getting valid measures for specific cognitive processes, which then are entered into further statistical analyses as dependent variables. With this technique it becomes possible to explain *why* response latencies are shorter in one condition compared to another condition. As detailed below, in the diffusion model framework, faster responses can be based on (1) fast information processing, (2) low response thresholds, or (3) fast response execution.

Recently, a third—related—type of question has been addressed by diffusion model accounts: These are studies that use diffusion models as a diagnostic tool (e.g., Schmiedek, Oberauer, Wilhelm, Suess, & Wittmann, 2007; White, Ratcliff, Vasey, & McKoon, 2010). Diffusion models provide valid criteria for cognitive processes which, subsequently, can be related to other measures. For example, speed of information processing might be a proxy for intelligence (Ratcliff, Schmiedek, & McKoon, 2008; Ratcliff, Thapar, & McKoon, 2010; Schmiedek et al., 2007) and a low response threshold might predict impulsive behaviour.

In parallel to these applications of the model, many important theoretical and methodological advances helped to promote diffusion modelling in psychology. Most important, the development of user-friendly software solutions cleared the path for this kind of analyses. Available solutions comprise the *EZ*-method (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008; Wagenmakers, van der Maas,

& Grasman, 2007), *DMAT* for Matlab (Vandekerckhove & Tuerlinckx, 2007, 2008), *fast-dm* (version 29: Voss & Voss, 2007, 2008), and, most recently, two Bayesian implementations for (hierarchical) diffusion models (Vandekerckhove, Tuerlinckx, & Lee, 2011; Wiecki, Sofer, & Frank, 2013). All these programs have special advantages, because they differ in (a) the mathematical methods used for parameter estimation (e.g., optimization criteria), (b) their flexibility to adapt to different complex data (e.g., experiments with multiple conditions), and (c) the usability and handling of the programs.

With this paper we want to introduce a new and extended version of *fast-dm* (*fast-dm-30*). The new developments regard the following points:

(1) The new version allows the user to choose between different optimization criteria (Kolmogorov-Smirnov, Chi-Square, and Maximum Likelihood). This allows optimizing parameter estimation for different data, because optimization criteria differ in robustness and efficiency depending on characteristics of data.

(2) A new parameter measuring so-called response-execution biases has been implemented (Voss, Voss, & Klauer, 2010). This parameter allows for the non-decisional component to differ between the two possible responses.

(3) The code was optimized and minor bugs have been removed, including problems of using command-line options on Windows systems.

(4) The tools to simulate data (*construct-samples*), and to calculate predicted CDFs and density functions (*plot-cdf* and *plot-density*) have been improved and are now better documented.

In the following we will provide a short introduction to diffusion model analysis followed by a discussion of advantages and disadvantages of different optimization criteria. Then, we give a step-by-step tutorial how to run a diffusion model project. The paper concludes with a description of the handling of *fast-dm-30* and its accompanying tools.

The Basics: A Short Introduction to Diffusion Modelling

Diffusion models are a formal model of decision making, that is, they provide a mathematical framework to understand decisional processes. They belong to the continuous sampling models (Ratcliff & Smith, 2004): These models assume that information is continuously sampled during a decision phase until evidence is sufficiently clear. As soon as one of two thresholds is reached, a response is initiated. The information sampling is described by a Wiener Diffusion Process which is characterized by a constant systematic drift (v) and Gaussian noise. The drift determines the average slope of the diffusion process and can be interpreted as the speed of information uptake. The standard deviation of the random noise (diffusion constant) is a scaling parameter in diffusion model analyses: It has to be fixed to a specific value that defines the scale for all other diffusion model parameters[1]. *Fast-dm* uses a diffusion constant of s=1, while other researchers prefer to use s=0.1. To make solutions comparable, it is essential to transform estimates for drift (v) , threshold separation (a), starting point (z), and the so-called intertrial variability of drift and starting point (s_v and s_z) by the following equation:

$$p_{new} = \frac{s_{new}}{s_{old}} p_{old}, \tag{1}$$

where p_{new} and p_{old} are the transformed and the original estimates, and s_{new} and s_{old} are the diffusion constants.

A second characteristic of standard diffusion models is the assumption that the diffusion process runs in a corridor between two thresholds, and it is terminated when one of them is hit. These thresholds represent two alternative outcomes of the decision process; depending on which threshold is hit, different responses are executed. By convention the lower threshold is positioned at 0 on the decision dimension and the upper threshold at a. Thus, a gives the amount of information that separates both possible decisional outcomes. Larger threshold separations lead—on average—to longer durations of the decision process. At the same time, an increasing distance between thresholds renders it more unlikely that random influences drive the process to the threshold opposite of the drift; that is, decision errors become rarer.

Sometimes one decisional outcome might be preferred over the other. To reach the preferred decision less information might be needed than for the non-preferred decision. Such a bias is often denoted in psychology as *response bias* (e.g., in Signal Detection Theory, Green & Swets, 1966) to emphasize that this kind of bias is independent of the quality of information processing (or sensitivity). However, we prefer here the term *decisional bias* because this bias is also unrelated to processes of response *execution*. In diffusion modelling, such a decisional bias is mapped on the starting point (z), which is positioned between 0 and a on the decision dimension. The closer the starting point is positioned to one threshold, the less information is needed to decide for the associated option. The new version of *fast-dm* uses the relative starting point (z_r) for input and output. The relative starting point is defined as z_r=z/a (range: 0 to 1; z_r=0.5 indicates unbiased decisions).

Obviously, the diffusion process as described so far cannot account for the total chain of information processing. Depending on the task, there will be additional processes of preparing for a task and encoding of stimuli that take place before a decision phase starts. After

[1] Strictly speaking, it is also possible to use a different parameter (a or v) to define the scale; in this case, intra-trial variability of the drift (s) can be estimated as a free parameter (Donkin, Brown, & Healthcote, 2009).

the decision is reached, motor processes have to be executed. The diffusion model sums the duration of all extra-decisional processes into one additional parameter, denoted as non-decisional component t_0 (or sometimes T_{er}, for time of encoding and response) measuring the total duration of those processes. Total response time is assumed to be the sum of the duration of the decisional processes (mapped the diffusion process) and the non-decisional processes (t_0).

The new version of *fast-dm* allows for different durations of motor processes for both outcomes (Voss et al., 2010). This might be relevant if one response is pre-activated (e.g., by response priming, Voss, Rothermund, et al., 2013), or if it is executed more (or less) frequently (e.g., in rare target search). In the implementation of the two execution times in *fast-dm*, a common t_0 parameter is used, giving the average duration of non-decisional processes, and a difference parameter *d*, giving the difference of duration of non-decisional processes for the responses connected to the lower vs. upper threshold. These parameters can be re-transformed into separate t_0 parameters with

$$t_0(upper\ threshold) = t_0 - 0.5 \cdot d \tag{2a}$$

$$t_0(lower\ threshold) = t_0 + 0.5 \cdot d. \tag{2b}$$

Most diffusion model analyses also take into account trial-to-trial fluctuations in cognitive components. For example, it is implausible to assume that participants' attention is equal throughout an experiment of several hundreds of trials; thus speed of information uptake (i.e., the drift) might differ slightly from trial to trial. Fluctuations in drift may also arise from different stimuli that are employed in different trials of an experiment. Similar points can be made for the inter-trial variability of starting point and of duration of non-decisional processes. For these reasons, most applications of the diffusion model allow for inter-trial variability of the drift (v), starting point (z), and non-decision constant (t_0). Specifically, the actual drift is assumed to follow a normal distribution with mean v and standard deviation s_v. Starting point and non-decisional constant follow uniform distributions with mean z and width s_z, and mean t_0 and width s_{t0}, respectively. As for the starting point, *fast-dm-30* uses a relative measure for inter-trial-variability of starting points, with s_{zr}=s_z/a.

The complete diffusion model as described above decomposes the decision process into 8 parameters (Table 1). Of course, models need not to include all of these parameters. Sometimes it might be better to make models more parsimonious by fixing parameters to given values. This regards specifically the starting point that can be fixed to $z_r = 0.5$ when no decision bias is expected (especially, when responses coded as false vs. correct), the response-time difference d that should be fixed to $d = 0$ when there is no reason to expect differences in speed of response execution, and the inter-trial variability parameters, that can be fixed to $s_v = s_{zr} = s_{t0} = 0$ when trial numbers are too small to allow for a robust estimation of these parameters.

On the other hand, diffusion models often comprise more than the 8 parameters described above: Typically, different values for one parameter are estimated for different types of stimuli or different experimental conditions.

How to Estimate Parameters: A Comparison of Different Optimization Criteria

A diffusion model analysis is based on the multi-dimensional search for an optimal set of estimates for all free parameters, so that there is a close fit between predicted and observed response time distributions. Since the RT distribution is split into two parts—for responses connected to the upper and lower threshold—the probability of responses (e.g., the error rate) is implicitly contained by the RT distributions. For the parameter search an optimization criterion has to be defined that quantifies the match between predicted and observed distributions. The most important improvement of *fast-dm-30* is that the user can now choose between three different optimization criteria: In addition to the Kolmogorov-Smirnov (KS) criterion that was used exclusively in *fast-dm*-29, we now implemented the commonly used Chi-Square (CS) approach and a Maximum Likelihood (ML) based algorithm. Because all algorithms have specific advantages, we will consider each of them below. Further information on the technical implementation of the algorithms is given in the section on technical details.

Maximum Likelihood (ML)

ML algorithms are highly efficient and are broadly applied to optimization problems for different models. In the case of diffusion models, the natural logarithms of density values (*g*)—calculated from predicted RT-distributions—are summed over all trials *i* (with response time RT_i and response k_i):

$$LL = \sum ln\big(g(RT_i, k_i)\big) \tag{3}$$

To make the algorithm more robust, a minimum value for density of g=10^{-6} is used in *fast-dm*, that is, g is set to 10^{-6}, when the predicted density is smaller than this value. The parameter search procedure then maximizes the resulting log-likelihood value. Because the ML procedure is highly efficient, it is especially useful in the case of small trial numbers. With the ML method parameters of parsimonious models may be estimated accurately from only 50 trials or less (Lerche, Voss, & Nagler, 2014). However, ML methods are especially sensitive to (fast) outliers. Even if only one (or very few) responses are added at the lower edge of the RT distribution, the accuracy of results will be derogated dramatically.

An additional advantage of the ML approach is that it allows the calculation of information criteria to compare different models. For example, the Bayesian Information Criteria (BIC) could be used here (Fific, Little, & Nosofsky, 2010):

$$BIC \ = -2 \ ln(L) + \ P \ \cdot \ ln(M), \tag{4}$$

where P is the number of free parameters and M is the number of observations (i.e., trials).

Chi-Square (CS)

Th CS criterion has been frequently used in diffusion model approaches (Ratcliff & Tuerlinckx, 2002). The main advantages are the very fast calculation and its robustness against outliers. The computed CS value is based on the comparison of the number of observed and predicted responses in so-called bins of the RT-distributions. The borders of these bins are defined by convention by the .1, .3, .5, .7, and .9 quantiles of the empirical response time distributions, separately for the upper and lower threshold[2]. Thus, the optimization criterion is calculated across the 2 x 6 bins as

$$CS = \sum \frac{(o_i - p_i)^2}{p_i}, \tag{5}$$

with o_i and p_i being the observed and predicted, respectively, number of responses in bin i. The parameter search minimizes the CS value. If more experimental conditions are fitted simultaneously, CS values are added over conditions as well. Next to the advantages of fast calculation and its robustness, the CS approach comes with the benefit that the CS value can be taken as a test statistic for model fit. The degrees of freedom are then given by

$$df = K(N - 1) - P, \tag{6}$$

with K conditions of an experiment, N bins per condition (N=2·6=12), and P free diffusion model parameters (White et al., 2010). A significant CS value indicates substantial misfit of the diffusion model. However, with large trial numbers, significant deviations are to be expected and other strategies of model tests might be preferable (Voss, Nagler, & Lerche, 2013).

Generally, CS based parameter estimations are only feasible for medium to large trial numbers (minimum 200 trials). It is especially problematic if empirical response distributions are small at one of the thresholds (e.g., less than 12 trials). In this case, the borders of bins are defined very unreliably. Unfortunately, this is often the case in diffusion model applications, where typically very easy tasks are used (e.g., lexical decision) and few errors occur. If in one

[2] Strictly speaking, this is not an exact implementation of a chi-square criterion because bins are defined by the data (and not by predicted distributions). However, the resulting values approximate nonetheless a chi-square distribution, and parameter estimates do not differ substantially (Ratcliff & Childers, 2014), while computation is much faster.

experimental condition one response is given in less than 12 trials, *fast-dm-30* ignores these responses for the calculation of the CS value.

Kolmogorov-Smirnov

Previous versions of *fast-dm* only implemented the KS criterion (Voss & Voss, 2007). We originally opted for this approach because its characteristics can be seen as a compromise between ML and CS based methods: On the one hand, the KS method is efficient, because it is not based on binning responses but utilizes the complete distribution; on the other hand, the KS criterion is not as sensitive to outliers as is the ML criterion (Lerche et al., 2014).

The KS criterion is defined as the maximum absolute vertical distance between the empirical and the predicted cumulative density functions (CDF) of the response time distributions. Over *n* responses of an experiment, it can be computed as

$$KS = \max_{i=1\ldots n} |\mathrm{eCDF}(\mathrm{RT}_i) - \mathrm{pCDF}(\mathrm{RT}_i)|, \tag{7}$$

where *RT$_i$* is the response latency in trial i, and *eCDF* and *pCDF* are the empirical and predicted CDFs, respectively. In diffusion modelling there are always two empirical distributions to be compared with their predicted counterparts (i.e., the distributions linked to the two responses). In *fast-dm* this problem is solved by combining both distributions into one. This is achieved by multiplying all RTs from responses linked to the lower threshold with -1 (Voss, Rothermund, & Voss, 2004; Voss & Voss, 2007). *Fast-dm* transforms KS-values in associated *p* values (with *df=number of responses*), which are then maximized (Voss & Voss, 2007). In case of multiple experimental conditions, the product of all *p* values from the different conditions is maximized.

Simulations from our lab (Lerche et al., 2014) show that—for uncontaminated data—the KS method tends to be slightly less efficient compared to the ML method but reveals notably more accurate results compared to the CS approach. For contaminated data, KS performs best in most cases.

Some Technical Details

The Calculation of Cumulative Density Functions (CDF)

The optimization routines based on the KS or CS statistics require the calculation of predicted CDFs. For the basic diffusion model (without inter-trial variabilities) the CDF for decision time *t* for responses at the upper threshold can be calculated as the solution of the following partial differential equation (PDE; see Voss & Voss, 2008):

$$\frac{\partial}{\partial t}F_+(t,z) = \frac{1}{2}\frac{\partial^2}{\partial z^2}F_+(t,z) + v\frac{\partial}{\partial z}F_+(t,z) \tag{8a}$$

with boundary conditions

$$F_{+}(t,0) = 0, F_{+}(t,a) = 1, \text{for all } t > 0 \tag{8b}$$

and initial condition

$$F_{+}(0,z) = \begin{cases} 0 & if\ 0 \leq z < a \\ 1 & if\ z = a \end{cases} \tag{8c}$$

It is possible to derive an explicit solution to this PDE that allows the direct calculation of the CDF (Blurton, Kesselmeier, & Gondan, 2012; Ratcliff, 1978). However, a numerical solution of the PDE introduced by Voss and Voss (2008) proved to be much faster while yielding the same accuracy, especially if inter-trial variability of starting point and non-decisional component are included in calculations. The PDE is solved numerically using a finite difference scheme, by discretizing the ranges of the starting point *z* and decision time *t* (see Press, Teukolsky, & Vetterling, 1992, Ch. 19, for an introduction to numerical solutions of PDEs). The accuracy of the solution depends on discretization step sizes for *z* and *t*. In *fast-dm* a "precision" parameter allows to control step sizes used in the calculation of CDFs (see below).

The Calculation of Density Functions

For the ML approach, density functions have to be calculated. For the basic diffusion model (without inter-trial variabilities) there are two different representations of the density g_{+} for the first-passage time *t* of a diffusion process with starting point *z* and threshold separation *a* (Navarro & Fuss, 2009; Van Zandt, Colonius, & Proctor, 2000; Voss et al., 2004):

$$g_{+}(t,z,a,v) = \frac{exp[(a-z)v - 0.5v^2 t]}{\sqrt{2\pi t^3}} \sum_{n=-\infty}^{\infty} exp\left(-\frac{[(1+2n)a-z]^2}{2t}\right) \cdot [(1+2n)a - z] \tag{9a}$$

and

$$g_{+}(t,z,a,v) = \frac{\pi}{a^2} \exp[(a-z)v] \sum_{n=1}^{\infty} n \cdot \sin\left(\frac{\pi(a-z)n}{a}\right) exp\left[-0.5\left(v^2 + \frac{\pi^2 n^2}{a^2} t\right)\right]. \tag{9b}$$

Navarro and Fuss (2009) show that Equation (9a) converges quickly for small *t* and Equation (9b) converges quickly for large *t*. In *fast-dm-30*, we implemented this finding and calculate densities always with the equation that converges faster. The numbers of terms used to approximate the infinite series are chosen to keep a maximum error bound of 1e-6 (Navarro & Fuss, 2009).

The value *t* in Equations (9a) and (9b) is the decision time. The non-decision parameter t_0 has to be subtracted from all empirical response times, before the densities are computed (t = RT-t_0). The density of the distribution at the lower threshold (g_{-}) can be easily obtained by replacing *v* with -*v* and *z* with *a-z*, respectively. To include inter-trial variabilities, g_{+} has to be integrated over *v*, *z*, and t_0.

$$g'(t,z,a,v,s_v) = \int_{-\infty}^{\infty} g_{+}(t,z,a,v') \cdot \frac{1}{\sqrt{2\pi s_v^2}} e^{-\frac{(v'-v)^2}{2s_v^2}} dv' \tag{10}$$

$$g''(t, z, a, v, s_v, s_z) = \int_{z-0.5s_z}^{z+0.5s_z} \frac{g'(t, z', a, v, s_v)}{s_z} dz' \quad (11)$$

$$g'''(t, z, a, v, s_v, s_z, s_{t0}) = \int_{t-0.5s_{t0}}^{t+0.5s_{t0}} \frac{g''(t', z, a, v, s_v, s_z)}{s_{t0}} dt' \quad (12)$$

The integral of Equation (10) can be solved analytically. Equations (11) and (12) are computed numerically in *fast-dm*; the discretization step size is controlled again by the precision settings (minimum number of steps is 4). Thus, the precision settings take influence on the results (and calculation time) for the ML method only if inter-trial variability of z and/or t_0 is greater than 0.

Optimization Routine

The optimization procedure is based on a multidimensional search for the optimal set of parameters that maximizes *p*(KS) or minimizes CS or -LL. For this procedure, we use an implementation of the SIMPLEX downhill algorithm (Nelder & Mead, 1965). This method is based on a simplex that comprises of n+1 vectors of parameter values when n parameters are optimized. For the starting simplex, we use results from the *EZ*-method (Wagenmakers et al., 2007) for the first vector (with $z_r = 0.5$, and $s_v = s_z = s_{t0} = 0$), and variations where values for one parameter are increased by a small amount for the remaining vectors.

In our implementation of the simplex, we use two criteria simultaneously. Firstly, we penalize theoretically impossible parameter constellations (e.g., z_r<0, z_r>1, a<0, etc.). For these cases, the optimization criteria cannot be calculated; solutions with penalty are always assumed to fit worse than any solution without penalty. The second criterion is the optimization criterion (*p*(KS), CS, or -LL). This second criterion is only used when no penalty is assigned to a solution. In case of KS, the corresponding *p*-value is minimized to allow the optimization of multiple experimental conditions.

Because the simplex algorithm is known to be unreliable in case of multidimensional search, we repeat the simplex search three times with different starting points and consecutively stricter stopping criteria.

Planning, Running, and Interpreting Diffusion Model-Analyses: A Step-by-Step Guide

The following sections describe some important steps in a typical diffusion model analysis and provide some help on crucial choices that have to be made. An excellent general introduction in cognitive modelling is provided by Heathcote, Brown, and Wagenmakers (in press). Specific advices on fitting parameters to the related ballistic-accumulator model can be found in other tutorials (Donkin, Averell, Brown, & Heathcote, 2009; Donkin, Brown, & Heathcote, 2011).

Step 1: Choosing an Experimental Paradigm

If a study aims at a general investigation of cognitive processes (e.g., cognitive aging or practice effects), it is often to choose between different paradigms (i.e., experimental tasks) for a study. If this is the case, paradigms should be selected that have already been validated for diffusion model analyses. Such well-tested paradigms comprise—for example—recognition memory tasks (e.g., Ratcliff, 1978; Spaniol, Madden, & Voss, 2006), numerosity or colour-judgment tasks (e.g., Ratcliff, 2002; Voss, Rothermund, & Brandtstädter, 2008; Voss et al., 2004), and lexical decision tasks (e.g., Ratcliff et al., 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

Sometimes, however, it may be the aim of a project to investigate whether a specific (new) paradigm is apt for diffusion modelling. If this paradigm has not yet been validated for a diffusion model analysis before, it should be verified first that all theoretical prerequisites and assumptions of the model are met. We will explicate these assumptions below. Secondly, it needs to be shown empirically that model fit is satisfactorily (see Step 6), and finally, an empirical validation of model parameters is essential (Voss, Nagler, & Lerche, 2013). For example, in such a validation study it can be tested whether face-valid manipulations map on single parameters as expected (see Voss et al., 2004, for an example of an empirical validation).

Theoretical prerequisites of diffusion models are often neglected or addressed only implicitly. In the following, we give a short overview of basic assumptions (Voss, Nagler, & Lerche, 2013): Firstly, diffusion models assume a *continuous sampling of information*. This makes the model more suitable for tasks using stimuli containing conflicting information. A prototypical example is a field of pixels with two different colours. Here, it can be argued that colour information is continuously sampled. In recognition tasks, not the stimuli itself are ambiguous; rather the familiarity (or the absence of familiarity) can be assumed to cumulate until a response is made.

Secondly, diffusion models require typically *binary decision* tasks. Optimally, diffusion model tasks should comprise two response keys that are linked in the analyses to the upper vs. lower threshold. It is also possible to recode responses as correct (upper threshold) vs. incorrect (lower threshold). However, this mapping requires some attention: (a) One needs to be sure that drift rates do not differ between stimulus types; (b) there should be no decision bias, and the relative starting point has to be fixed to 0.5; and (c) it has to be considered that results might be less robust in case of low error numbers (because then the distribution of responses at the lower threshold is absent or small). If these requirements are not met, the linear ballistic accumulator model should be preferred because it allows mapping data with multiple responses (Donkin, Averell, et al., 2009).

A third prerequisite refers to the assumption of *constancy of parameter values over time*. The Wiener diffusion model as described in this paper assumes that drift and threshold separation are constant over the time of a decision (and independent of the accumulated amount of evidence). The assumption of constant threshold separation might be violated when sparse information is present or decision times are long. In this case, shifts in criterion are highly plausible. However, the direction of such shifts remains rather unclear: It could be argued that participants will set more liberal criteria when they notice that they do not reach the conservative criteria after several seconds. On the other hand, it is possible that threshold separation is increased to avoid errors when the decision is really difficult. The assumption of constant drift could be violated when a stimulus changes over time (e.g., a hidden stimulus is continuously unmasked), or when it is removed from screen before a decision is reached (the drift might be stronger while the stimulus is present and weaker when it is only remembered).

Changes of drift rate over time might also occur in interference tasks like the stroop task or the flanker task, when distracting information has to be inhibited. The inhibition of irrelevant information might take some time, which results in an increase of drift rate during the decision phase.

A fourth assumption regards the required components of a task. The diffusion model is apt only for relatively simple *single-stage decisions*. More complex tasks that are composed of different steps (or insights) might again challenge the assumptions of continuous information sampling and constant drift.

Step 2: How many Trials should be used?

The number of trials of an experiment determines the accuracy of parameter estimation: The more data are entered into an analysis the more accurate all parameters can be estimated. In a recent set of simulation studies, Lerche et al. (2014) found that for parsimonious models with few parameters reasonably accurate estimations were possible with only 48 trials or sometimes even less. In most situations, a good accuracy is reached with 200 trials.

The recommended trial number depends on several aspects of the present data (Lerche et al., 2014): Firstly, if data is contaminated by trials in which participants do not use a continuous information sampling (but, e.g., a guess), more data are required. This is even the case, when these contaminants are no outliers in a statistical sense. Imagine, for example, a participant that uses a diffusion-like information sampling strategy in 95% of all trials, but bases his responses on guesses in the remaining 5%. Because guesses involves other (and probably faster) cognitive processes, the RT distribution from the guess trials will differ from the RT distribution of the judgement trials. If, however, both distributions overlap it will not be possible so identify the guessing trials on basis of RTs.

A second determinant of the required trial number lies in the scientific question that is addressed: If estimates need to have a high reliability (e.g., because inter-individual differences are in the focus of a study) larger trial numbers might be necessary. Thirdly, if data are mapped as correct vs. incorrect (see above) the absence of error responses will make a precise estimation of parameters difficult. Therefore, enough trials should be used so that each participant makes several errors. Finally, one has to consider that some parameters are more difficult to estimate than others: While, for example, the duration of non-decision-times can be estimated with high accuracy from medium trial numbers (n≈100), very large trial numbers (n>1000) are often required to estimate the inter-trial-variability parameters of drift and starting point with satisfactory accuracy.

Step 3: Data Pre-Treatment

Results from diffusion model analyses can be biased strongly when data is contaminated (Lerche et al., 2014; Ratcliff & Tuerlinckx, 2002). Especially fast outliers have a strong impact and should be removed. Because of the positive skew of RT distributions fast outliers might be missed with typical procedures (e.g., inspecting box plots or *z*-scores). Therefore, RTs should be log transformed before an outlier analysis (for the diffusion model analyses, of course, the untransformed data has to be used). Another possibility is to find a point at the lower edge of the RT distributions where performance rises above chance level (Ratcliff & Tuerlinckx, 2002).

A careful outlier analysis is of special importance when the parameter estimation is based on a maximum-likelihood procedure; on the contrary, the KS method proved to be very robust (Lerche et al., 2014).

Step 4: Defining your Model: Choosing free Parameters

The degree of complexity of a model depends on several factors. On the one hand, a model should not oversimplify reality: When important parameters are neglected (i.e., fixed to a specific but wrong value), effects will be forced on other parameters and thus results become invalid. Imagine, for example, a situation where there is a decision bias but the relative starting point is fixed to z_r=0.5 (indicating an absence of a decision bias). The decision bias would make responses at the preferred threshold faster; to account for this, the drift for “preferred” ("unwanted") stimuli would be overestimated (underestimated). Thus, results from the restricted model would erroneously indicate a bias in terms of information processing.

On the other hand, models should be defined as parsimonious as possible, because many free parameters might lead to overfitting and make results unstable, especially if not enough trials are used. For example, model fit might be excellent no matter if you allow for a decision bias (i.e., asymmetric starting points) or for a perceptual bias (i.e., different drift for different stimulus types). In our experience, for small and medium trials numbers (<500) setting inter-trial-variability of drift (s_v) and starting point (s_{zr}) to zero makes the estimation of the remaining parameters more robust, even if there is an inter-trial-variability in data. Note that this is not the case for inter-trial-variability of non-decision time (s_{t0}). Because s_{t0} has a great impact on the shape of the RT-distribution it is often harmful to neglect this parameter. Additionally, the difference in non-decision time for upper and lower threshold (*d*) can usually not be estimated simultaneously with starting point (Voss et al., 2010); therefore you should set either d=0 or z_r=0.5, whatever seems theoretically more plausible (large trial numbers might allow to estimate both parameters simultaneously).

Decisions of model complexity get more complicated when different types of stimuli or different experimental manipulations are compared. In this case, the researcher has to decide which parameters are allowed to vary between conditions. If, for example, an experiment comprises "easy" and "difficult" trials, it is plausible that this affects the drift, and different drift parameters should be estimated for different trial types. However, decisions on this matter need careful consideration, because false fixations will again lead to invalid results. Whenever it is the aim of a study to check on which parameters a manipulation maps, we recommend to model data from the different conditions completely independently (allowing for all parameters to vary between conditions). A disadvantage of estimating completely independent models for all conditions is than not all available information is used, and power to find relevant differences might be reduced. A discussion of this problem is given by Donkin and colleagues (Donkin et al., 2011; Donkin, Tran, & Nosofsky, 2014).

Step 5: Choosing an Optimization Criterion

On this step, the researcher needs to decide which software or algorithm to use for the parameter estimation. This decision may depend on the number of trials and the quality of data (Lerche et al., 2014). For large data sets (>500), always robust procedures (like KS or CS) are recommended. For small data sets (<100) chi-square based approaches will not work properly, and maximum likelihood procedures may be a good option if one is confident that data are not contaminated, and the Kolmogorov-Smirnov approach should be used, when a more robust procedure is required.

Step 6: Assessing Model Fit

In diffusion model application, the assessment of model fit should be a mandatory step. It is problematic to use the standard statistical tests associated with the chi square or Kolmogorov Smirnov criteria here, because results strongly depend on numbers of trials: For small data sets the power is too small to reliably detect misfit, and for large data sets deviations will nearly always be significant. Therefore, either graphical inspection or Monte Carlo simulations provide better alternatives.

Graphical inspection can be done for each individual by so-called quantile-probability plots (e.g., Ratcliff & Smith, 2010). These graphs show different quantiles of the empirical and predicted RT distributions as a function of the probability of correct (or erroneous) responses (for different stimulus types). If an experiment comprises data from many participants, we recommend using scatter plots that plot predicted values against empirical values for the 25, 50 and 75 quantiles of the RT distributions and for accuracy of responses (e.g.,Voss, Rothermund, et al., 2013, Appendix B). When all data points are positioned near the main diagonal, a good fit can be assumed.

The assessment of model fit with Monto Carlo simulations has the advantage that it leads to a clear criterion for which participants there is a satisfactory model fit. To this end, a critical value for an acceptable fit has to be determined. This critical value will depend on the number of trials, conditions, and parameters, on the estimation procedure, and possibly as well on the observed range of parameter values. Therefore, data sets have to be simulated that match the empirical data sets as closely as possible. It is recommended to draw at least 1,000 parameter sets from a multidimensional normal distribution defined by the covariance matrix of the estimated parameter values. This can be accomplished, for example, by the *mvtnorm* library from the *R* environment. Then, for each of the 1,000 parameter-sets one data set is simulated. The *construct-sample* tool of *fast-dm* can be used for this purpose (see below; note that each condition must be simulated separately and combined later into one file). In the next step, simulated parameter sets are entered into a diffusion model analysis with the same settings as used for the analysis of empirical data. From the results, only the fit indices are of importance: The 5% quantile of the distribution of fit indices is then used as critical value to assess fit of empirical results: All data-sets performing worse than this 5% criterion should be regarded as bad fitting. If notably more than 5% of data sets show bad fit, it should be questioned critically whether the diffusion model is suitable for the task.

Step 7: Interpretation of Results

The last step of the diffusion model analysis is the interpretation of results. Typically, parameters are estimated for each individual; in this case estimates can be entered as dependent measures into statistical analyses (e.g., ANOVA) to check for differences between conditions. Alternatively, it is possible to compare model fit (e.g., BIC) between models with different restrictions to see which restrictions lead to a notable decrease of model fit.

Using fast-dm-30: A User's Manual

Overview

When *fast-dm* is started, it reads commands from an external *control file* (named by default *experiment.ctl*). Commands in the *control file* control program settings, specify parameters that are estimated or fixed to given values, and set file names for input and output. *Fast-dm* can be started by double clicking on the program icon; in this case, the control file *experiment.ctl* will be read from the directory in which *fast-dm* is started. If no such file exists, *fast-dm* terminates immediately. Generally, we recommend starting *fast-dm* from a command console.[3] Otherwise, error or warning messages can be lost because these are presented only on the screen in a window that closes as soon as *fast-dm* terminates. From a command window, the program is started by typing "fast-dm" (within the correct directory). You can add the file name of a control file as command line option: For example, "fast-dm exp1.ctl" will start *fast-dm* with the control file *exp1.ctl*.

Generally, the following steps are necessary to use *fast-dm*.

(1) Create a directory for your analyses.
(2) Save all data files and a copy of *fast-dm* in this directory
(3) Create a control file with a text editor (see below)
(4) Start *fast-dm* (optimally from a command window)
(5) Read results into your favourite statistics software for further analysis

License, Source Code, and Compiled Binaries

Fast-dm is free software; you can use, redistribute it and or modify it under the terms of the GNU General Public License. Details are given in the file *COPYING* that is included in the download archives. In the Downloads section of the *fast-dm* homepage (http://www.psychologie.uni-heidelberg.de/ae/meth/fast-dm/index.html) we provide three different zip-files. The first, labelled as "Windows Binaries", contains the precompiled exe-

[3] Windows users can open a command window by typing "cmd" in the start menu.

cutable files for Microsoft Windows systems. Specifically, we provide the programs *fast-dm.exe* (for parameter estimation), *construct-samples.exe* (for the simulation of data samples), *plot-cdf.exe* (for generating a cumulative distribution function from a set of parameter values), and *plot-density.exe* (for the generation of the density function from a set of parameter values). You may need to install the Microsoft Visual C++ Redistributable package for Visual Studio 2012 (http://www.microsoft.com/en-us/download/details.aspx?id=30679) to get these programs running.

Secondly, we provide the complete C source code of *fast-dm* in the "source" archive. Together with the source code files, this archive contains short instructions (file INSTALL) on compiling *fast-dm* on Unix-like systems (e.g., Linux and MacOS), and a short manual (file MANUAL).

Finally, we provide a Visual Studio 2012 Project (including source code files and reasonable project setting) for Windows users who want to modify the software. To make use of this, Microsoft Visual Studio 2012 needs to be installed, which is freely available in the Express edition (http://www.microsoft.com/en-us/download/details.aspx?id=34673).

Data Files

Data is read from plain text files. Each line of a *data file* contains information from one trial, and data columns have to be separated by blanks or tabs (see Figure 1 for an example of a *data file*). Lines starting with a hash mark (#) are considered as comments and are ignored. Each data file needs to comprise at least two columns: One column—referred to as "RESPONSE" column in the *control file*—contains information about responses coded as 0 and 1 for the lower and upper threshold, respectively. The second required column—labelled as "TIME" column in the control file—gives response times in seconds. Optionally, further columns can be added containing information about stimulus types (e.g., "word" vs. "non-word") and/or the experimental conditions (e.g., "speed instruction" vs. "accuracy instruction"). In these additional columns either words or numbers can be used for coding different conditions.

Fast-dm estimates parameters independently for separate *data files*. Usually, each *data file* will contain data from one participant. However, sometimes it may be a good idea to split data from one participant into separate files, so that independent models are estimated for different conditions.

Control Files

To run *fast-dm,* a *control file* is required containing commands that specify settings for the parameter estimation process. This *control file* is a plain text file that can be constructed with any text editor (see Figure 2 for an example of a *control file*). Each line of a control file contains a *fast-dm* command and additional values specifying the chosen settings (separated by blanks). As in *data files*, lines starting with a hash mark (#) are ignored. Table 2 gives an overview of all commands with explanations and examples. Some commands are required (*format*, *load*, and *save* or *log*), while others are optional. In the command file, the definition of the model (*depends* and *set* commands) have to precede the *format* command, and *load* and *save/log* commands must come after. All other commands can be placed anywhere in the *control file*.

The *method* command specifies the optimization criterion. Possible values are “ml” for Maximum Likelihood, “ks” for Kolmogorov-Smirnov, and “cs” for Chi-Square. Depending on the chosen method, the appropriate criterion is given in the output. If no method is specified, KS is chosen by default.

The *precision* command controls the accuracy of calculation of predicted CDFs (for the KS and CS method) or DFs (for the ML method). Any positive real numbers can be used as arguments, with higher precision values leading to a higher accuracy and longer duration of calculation. Reasonable values range from about 2.0 to 5.0. We tuned the calculation routines to achieve an error in calculated values that is approximately $\varepsilon=10^{-\text{precision}}$ (however, we cannot guarantee that this bound is always strictly observed). The command is optional; if no precision is specified, a default value of *precision = 3* is used.

With the *set* command, parameters are fixed to given values. The command requires 2 arguments (separated by blanks): First, the name of the parameter is given (see Table 1 for the *fast-dm* notation for all parameters), followed by the desired value. For example, “*set zr 0.5*” fixes the relative starting point to 0.5, that is, the process starts at $0.5 \cdot a$ and is thus assumed to be unbiased. Parameters that are fixed to a value are not estimated by *fast-dm*. Generally, we recommend fixing either d to 0 or z_r to 0.5 because it is difficult to estimate both parameters simultaneously (Voss et al., 2010). In case of small trial numbers, it often makes sense to make a model as parsimonious as possible. For this purpose it might help to additionally fix s_z and s_v to 0 because these parameters have only minor impact on the predicted distributions and can only be reliably estimated from huge data sets (Voss, Nagler, & Lerche, 2013). The *set* command is optional; by default all parameters are estimated. The *set* command can be used repeatedly to fix different parameters.

With the *depend*s command parameters can be specified that are estimated separately for different types of stimuli or different experimental conditions. The *depends* command must be followed by a parameter name and by user-chosen labels for the conditions. Parameters can depend on different factors (e.g., type of stimulus and block of the experiment); in this case, labels for each factor are specified one after another (separated by blanks). For each parameter that can vary between conditions, a separate *depends* command must be specified. All condition labels that are used in any *depends* command must be specified as a column in the data file(s) with the *format* command (see below). The *depends* command is optional. By default, all parameters are assumed to be equal across all experimental conditions.

The *format* command defines the columns of the data file(s). The labels RESPONSE and TIME are mandatory (capital letters are required for these). Additionally, all factor labels used in *depends* commands have to be named here as well (capitalization must be identical in the *format* command and the *depends* commands). Columns that shall be ignored by *fast-dm* can be assigned with any new name or with an asterisk (*). The *format* command is required and needs to be placed after all *set* and *depends* commands but before *load*, *save*, and *log*.

The *load* command specifies the file name(s) of data files. *Fast-dm* tries to load data from the directory in which it is started, unless a path is given. File names may contain asterisks (e.g., “participant_*.dat”); in this case, the asterisk is a wildcard character that can be replaced by any number of characters. Any matching files within the chosen directory will be loaded. The *load* command is required.

To save results, the *save* or the *log* command (or both) have to be used. With the *save* command, separate output files are generated for each data file. When the data file name as specified in the *load* command contains an asterisk, an asterisk is also required in file name defined in the *save* command, so that multiple file names for output can be generated. With the *log* command, one common output file is generated that contains estimated parameter values as a table that can be read from any statistical software for further analyses of results.

Output

The output of the estimation procedure is shown directly in the console (Figure 3). First, the name of the *control file* and central characteristics of the estimation procedure are presented (precision, method of estimation, format of data files, estimated and fixed parameters). Then, parameters that are estimated within each condition of an experiment are listed (numbers represent fixed parameters). For parameters that depend on conditions the labels of conditions as found in the appropriate columns of the data files are attached to the parameter identifier. At the end of these lines, the number of observed responses at lower and upper threshold (coded with 0 or 1 in the data file, respectively) within each condition are presented.

Following the model specifications, fit values resulting from each of the three consecutive runs of the parameter search are displayed. If the KS criterion has been selected, the (combined) *p*-values of the KS distances will be presented. We warn not to take these *p*-

values as a direct indicator of significant model misfit (Voss, Nagler, & Lerche, 2013): Firstly, if multiple conditions are used the presented p-value is the product of p-values from all conditions, which may lead to very small combined values, even if the single KS statistics from all conditions are not significant (e.g., $p = 0.10 \cdot 0.35 \cdot 0.12 \cdot 0.60 = 0.0025$). On the other hand, p-values would be too liberal, if—as is done here—the forms of predicted functions are fit to the empirical functions before the KS statistic is determined, which may possibly prevent statistical significance. If the ML method is chosen the presented fit index is –LL. Small values indicate a good fit. Finally, when selecting CS as optimization criterion the chi-square values will be displayed; here smaller values again indicate better fitting. If no valid model is found (e.g., if the likelihood for at least one RT is zero), a penalty value is presented instead of the fitting index.

After the third run of the parameter search is finished the resulting estimates for all parameters are shown. If multiple data sets are processed, the estimates will be presented one after the other. Finally, the total computation time is presented.

If the user wrongly defines a command (e.g., a condition is named in the *depends* command which has not been assigned to a data column in the *format* command) an error message appears and the program is aborted. Furthermore, a warning message will be presented and the estimation process stopped if the number of trials is not sufficient for parameter estimation. For the ML and KS methods, for each experimental condition at least 10 trials are required (no matter whether responses vary between trials or not). For the estimation with CS as optimization criterion at least 12 trials sharing the same response are required (i.e., 12 trials with all 12 responses at the upper threshold would be ok, while 20 trials with 10 responses at each threshold cannot be analysed using the CS method).

Besides the output on the screen the results are also saved in files, either separately for each data file (using the *save* command in the control file; see Figure 4) and/or in one summary file including the estimates of all data files (using the *log* command; see Figure 5).

Additional Tools

Construct-samples, plot-cdf, and plot-density are command-line tools which can be downloaded from the "*fast-dm* Downloads" section (archive "Windows binaries"; source code is also available in the "source" archive). The programs need to be started from the command console, and all settings are entered directly as command line arguments.

Making Simulations with *construct-samples.* *Construct-samples* allows simulating data sets for a given parameter set. This is useful (1) to evaluate the quality of parameter recovery of *fast-dm* and (2) to get a distribution of fit-values that allows assessing the fit of models estimated from empirical data. For these purposes data sets have to be simulated from known parameter values. Then, *fast-dm* is applied to the simulated data sets and the estimated parameter values are compared to the true values from the simulation.

If this tool is started by just typing *construct-samples* into the command line, all default settings are used (see Table 3). Typically, however multiple command line options will be entered at starting *construct-samples*. Options start with a minus sign, followed by a letter and in most cases by an additional argument, typically a number (exceptions: -r has no additional argument and -o needs a string determining the file name).

Command-line options are used to set parameter values for the simulation. Please note that notation differs here slightly from the usual *fast-dm* labels. This is because only one-letter commands can be used here. Therefore, "-z" is used for z_r, "-t" for t_0, and capital letters "-Z", "-V", and "-T", for the intertrial variabilities s_{zr}, s_v, and s_{t0}, respectively. The "-r" argument ensures that random samples are generated. This is what normally is needed for simulations. If "-r" is not present, a deterministic data set is calculated, where response times reflect directly the quantiles of the predicted distributions. With "-p" the precision of calculation can be adapted as in *fast-dm*. The number of trials within each simulated data set is set by "-n", and the number of data sets is defined by "-N". The file name(s) for output are determined with the "-o" command. If multiple data sets are generated, it is necessary to include "%d" in the name, which is then replaced by a different number for each data set (from 0 to *N*-1). If "-o" is not used, results are presented in the console only. Results always comprise two columns: The first is coding simulated responses (0 vs. 1) and the second gives the response times in seconds. Finally, a short help page can be opened by typing "construct-samples –h".

For example, *construct-samples* could be started by typing the following command:

```
construct-samples -a 2 -z 0.5 -v 3 -t 0.5 -r -n 250 -N 1000 -o %d.sim
```

With this command, 1,000 data sets named *0.sim* to *999.sim* are generated containing random samples of 250 trials simulated from parameter values a=2, z_r=0.5, v=3, and t_0=0.5 (for d and intertrial variabilities the default values of 0 are assumed).

Often, you will need to simulate data sets for more complex situations. Imagine, for example, that multiple conditions with different parameter values should be simulated. To do so, you need to simulate data separately for each condition and then combine data sets into common files. This can be done automatically—for example—using *R*. The application of *construct-samples* (and *fast-dm*) from the *R* environment is illustrated in the examples that can be downloaded from the *fast-dm* website.

Plotting (combined) CDFs with *plot-cdf and plotting DFs with plot-density*

Plot-cdf can be used to calculate values of predicted CDFs of a certain parameter set. This can be useful to demonstrate model fit graphically: If predicted and empirical CDFs are plotted in the same diagram, it is possible to assess whether both curves match sufficiently well, and—if not—where the main differences are (see Voss et al., 2008, for an example of this strategy). Note that *plot-cdf* generates so-called combined CDFs, where distributions from lower and upper threshold are merged by multiplying all RTs from the lower threshold by -1 (Voss et al., 2004; Voss & Voss, 2007).

Command line options are very similar to those of *construct-samples* (see Table 3). The only differences are that -r, -n, and -N cannot be used with *plot-cdf*. For example,

```
plot-cdf -a 2 -z 0.5 -v 3 -t 0.5 -o cdf.dat
```

generates values for a predicted CDF with a=2, z_r=0.5, v=3, and t_0=0.5 and saves these values into a file named “cdf.dat”. Output consists of two columns: The first contains the reaction times (with negative values indicating responses at the lower threshold). The second column displays the cumulative probability values. For graphic diagrams output from *plot-cdf* has to be entered in other programs like *R* or *Excel*.

The *plot-density* tool can be used to get values for the density functions at upper and lower threshold. Command line options are identical to those in *plot-cdf*. Therefore,

```
plot-density -a 2 -z 0.5 -v 3 -t 0.5 -o density.dat
```

will save the density functions for the same settings as used in the CDF example. Here, the output comprises three columns that contain values for predicted response times and density functions at upper and lower threshold (densities at the lower threshold get a negative sign here).

Concluding Remarks

After six years of using *fast-dm*, several optimizations have been made improving the performance and functionality of the program. The most important extension is the inclusion of different optimization criteria (Maximum Likelihood, Kolmogorov-Smirnov, and Chi-Square). This can potentially improve results from diffusion model analyses greatly, because all criteria have different advantages and shortcomings, and now the criterion that is best for a given data set can be chosen. Obviously, the number of trials is an important factor for this choice. Often, ML will outperform the other methods at small data sets. Secondly, purity of data will influence quality of results as well: When RTs are contaminated, ML can be strongly biased (Ratcliff & Tuerlinckx, 2002), while both other methods will probably be more robust.

Further factors, like the number of estimated parameters, the number of experimental conditions, the task difficulty (i.e., percentage of errors) will also influence the accuracy of parameter recovery. However, it is less clear how these factors influence performance of the different criteria. Future simulation studies are essential to allow an informed choice of the best criterion.

In the development of *fast-dm* we did not (yet) program a graphical user interface. We are aware that this might be seen by some as a barrier to the application of the program. The main reason for us to develop *fast-dm* without graphical user interface was to ensure that the program can be compiled within any operating system. We hope that many users of *fast-dm* find it usable and helpful and that *fast-dm* thus helps to promote diffusion model analyses as a powerful method to infer cognitive processes.

References

Blurton, S. P., Kesselmeier, M., & Gondan, M. (2012). Fast and accurate calculations for cumulative first-passage time distributions in Wiener diffusion models. *Journal of Mathematical Psychology, 56*(6), 470-475. doi: 10.1016/j.jmp.2012.09.002

Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods, Instruments & Computers, 41*(4), 1095-1110.

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*(2), 140-151.

Donkin, C., Brown, S. D., & Healthcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16*(6), 1129-1135.

Donkin, C., Tran, S. C., & Nosofsky, R. (2014). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception, & Psychophysics, 76*(7), 2103-2116.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117*(2), 309-348. doi: 10.1037/a0018526

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*(2), 55-68.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford England: Wiley.

Heathcote, A., Brown, S. D., & Wagenmakers, E. J. (in press). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer.

Lerche, V., Voss, A., & Nagler, M. (2014). How Many Trials are Required for Robust Parameter Estimation in Diffusion Modeling? Comparison of Different Estimation Algorithms. *Manuscript submitted for publication.*

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology, 53*(4), 222-230. doi: 10.1016/j.jmp.2009.02.003

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308-313.

Press, W. H., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical Recipes in C (2nd Ed.).* Cambridge Cambridge University Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

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

Ratcliff, R., & Childers, R. (2014). Individual Differences and Fitting Methods for the Two-Choice Diffusion Model. *Manuscript in preparation*.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review, 111*(1), 159-182. doi: 10.1037/0033-295x.111.1.159

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence, 36*(1), 10-17. doi: 10.1016/j.intell.2006.12.002

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review, 111*(2), 333-367. doi: 10.1037/0033-295X.111.2.333

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*(1), 70-94. doi: 10.1037/a0018128

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481. doi: 10.3758/bf03196302

Schmiedek, F., Oberauer, K., Wilhelm, O., Suess, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414-429.

Spaniol, J., Madden, D. J., & Voss, A. (2006). A Diffusion Model Analysis of Adult Age Differences in Episodic and Semantic Long-Term Memory Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 101-117. doi: 10.1037/0278-7393.32.1.101

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7*(2), 208-256. doi: 10.3758/BF03212980

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44*(4), 1048-1056. doi: 10.1016/j.jesp.2007.10.009

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, 142*(2), 536-559. doi: 10.1037/a0029459

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response tendency and decision biases: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*, 539-555.

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*(1), 140-159. doi: 10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15*(6), 1229-1235. doi: 10.3758/PBR.15.6.1229

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*(1), 3-22. doi: 10.3758/bf03194023

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*(1), 39-52. doi: 10.1016/j.jmp.2010.01.004

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 1

Parameters of the Diffusion Model, typical ranges of values, and cognitive interpretation

Parameter	Fast-dm	Typical Range	Interpretation
Drift	v	-4 to +4	average speed of information uptake
Threshold Separation	a	0.6 to 2	response caution
Starting Point	z_r	0.4 to 0.6	decision bias
Non-Decisional Constant	t_0	0.2 to 1.0	duration of non-decisional processes
Difference in Non-Decisional Constant	d	-0.1 to +0.1	response preparation / response inhibition
Intertrial Variability of Drift	s_v	0 to 1	differences in stimulus properties or fluctuations in attention
Intertrial Variability of Starting Point	s_{zr}	0.0 to 0.5	differences in expectations
Intertrial Variability of Non-Decisional Constant	s_{t0}	0 to 1	differences in speed of response execution

Note. Ranges for parameters refer to a diffusion constant of s=1.

Table 2

Commands of the control file

Command	Description	Examples
method *CRITERION*	determines the optimization criterion (ml=Maximum Likelihood; ks=Kolmogorov-Smirnov; cs=Chi-Square); default setting: method=ks	• method ml • method cs
precision *VALUE*	defines the precision of the calculation; default setting: precision=3	• precision 2.5 • precision 5
set *PARAMETER VALUE*	fixes a parameter to a specific value; default setting: no fixations	• set d 0 • set zr 0.5 • set szr 0
depends *PARAMETER CONDITION*	denotes that a parameter may vary between different conditions; default setting: parameters do not depend on conditions	• depends t0 block • depends v stimulus difficulty
format *CONDITION* ...	defines columns of the data file(s). The command requires the variables RESPONSE and TIME	• format RESPONSE TIME • format RESPONSE TIME stimulus difficulty • format * RESPONSE TIME
load *FILE_NAME*	declares the names of the input files	• load participant_1.dat • load participant_*.dat
save *FILE_NAME*	defines the names of separate output files (one output file for each data set)	• save parameters_participant_1.dat • save parameters_participant_*.dat
log *FILE_NAME*	defines the name of a common output file (one output file for all data sets)	• log all_participants.dat

Note. Commands in **bold font** are required while the other commands are optional.

Table 3

Command-Line Options for construct-samples, plot-cdf and plot-density

Option	Description	Default
-a *VALUE*	*VALUE* is assigned to parameter a	1
-z *VALUE*	*VALUE* is assigned to parameter z_r	0.5
-v *VALUE*	*VALUE* is assigned to parameter v	0
-t *VALUE*	*VALUE* is assigned to parameter t_0	0.3
-d *VALUE*	*VALUE* is assigned to parameter d	0
-Z *VALUE*	*VALUE* is assigned to parameter s_{zr}	0
-V *VALUE*	*VALUE* is assigned to parameter s_v	0
-T *VALUE*	*VALUE* is assigned to parameter s_{t0}	0
-p *VALUE*	the computational precision is set to *VALUE*	4
-n *VALUE* [a]	*VALUE* defines the trial number per data set	100
-r [a]	a random data set is generated	a deterministic data set is generated
-N *VALUE* [a]	*VALUE* defines the number of random data sets	1
-o *FILE_NAME*	the generated data is not presented in the console but saved to *FILE_NAME*	the generated data is presented in the console window but not saved

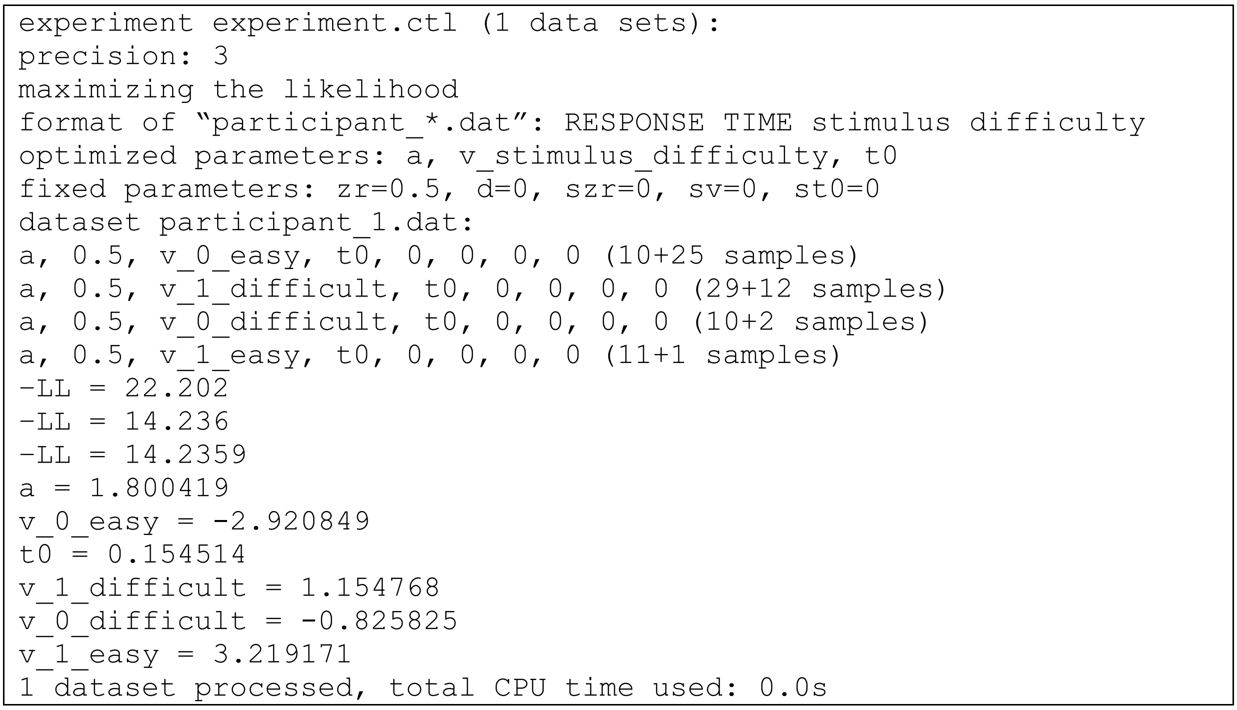
Note. [a] expression cannot be applied for *plot-cdf* and *plot-density.*

```
# RESPONSE TIME stimulus difficulty
 0    0.424  0   easy
 1    0.667  1   difficult
 0    0.598  0   easy
 1    0.713  1   difficult
 1    0.701  1   difficult
 1    0.655  0   easy
 1    0.452  1   difficult
 0    0.577  0   easy
```

Figure 1. Example of the first lines of a data file. Lines starting with “#” are ignored. The RESPONSE (0=”lower threshold”, 1=”upper threshold”) and TIME column (response time in seconds) are mandatory. Further columns can be added to give information about the stimulus (e.g., 0 = “word” vs. 1 = “non-word”) or experimental condition (e.g., “easy” vs. “difficult”).

```
method ml
precision 2.5
set d 0
set zr 0.5
set szr 0
set sv 0
depends v stimulus difficulty
format RESPONSE TIME stimulus difficulty
load participant_*.dat
save parameters_participant_*.dat
log D:/fast-dm/all_participants.dat
```

Figure 2. Example of a control file. The maximum likelihood criterion is used with (reduced) precision 2.5. Four parameters (d, z_r, s_{zr}, s_v) are fixed to given values. Drift is free to differ depending on *stimulus* and *difficulty*. If both conditions have 2 values (see Figure 1), 2 x 2 = 4 different values for the drift will be estimated, whereas for the remaining parameters (a, t_0, s_{t0}) one value is estimated for all conditions (resulting in 7 free parameters). The remaining commands specify the format of data files, and file names for data, save, and log files.



```
experiment experiment.ctl (1 data sets):
precision: 3
maximizing the likelihood
format of “participant_*.dat”: RESPONSE TIME stimulus difficulty
optimized parameters: a, v_stimulus_difficulty, t0
fixed parameters: zr=0.5, d=0, szr=0, sv=0, st0=0
dataset participant_1.dat:
a, 0.5, v_0_easy, t0, 0, 0, 0, 0 (10+25 samples)
a, 0.5, v_1_difficult, t0, 0, 0, 0, 0 (29+12 samples)
a, 0.5, v_0_difficult, t0, 0, 0, 0, 0 (10+2 samples)
a, 0.5, v_1_easy, t0, 0, 0, 0, 0 (11+1 samples)
-LL = 22.202
-LL = 14.236
-LL = 14.2359
a = 1.800419
v_0_easy = -2.920849
t0 = 0.154514
v_1_difficult = 1.154768
v_0_difficult = -0.825825
v_1_easy = 3.219171
1 dataset processed, total CPU time used: 0.0s
```

Figure 3. Example of the console output. First, information on the selected control file, the precision and method of estimation, the format of the data files and the estimated and fixed parameters are given. Any parameter depending on a condition is indexed with the name of the condition variable(s). Estimated parameters and numbers of responses at lower and upper threshold are presented separately for each condition. The three “-LL” values result from the three consecutive runs of the simplex algorithm. In the following lines the estimated values for all parameters are displayed. Finally, the number of processed data sets and the required computational time is given.

```
a = 1.800419
v_0_easy = -2.920849
t0 = 0.154514
v_1_difficult = 1.154768
v_0_difficult = -0.825825
v_1_easy = 3.219171
precision = 3.000000
method = ML
penalty = 0.000000
fit index = -14.235880
time = 0.047000
```

Figure 4. Example of a save file. When the save command is used for each data file a separate output file is generated containing a short version of the screen output (see Figure 3).

```
dataset a v_0_easy t0 v_1_difficult v_0_difficult v_1_easy  penalty fit time method
1 1.8004 -2.9208 0.1545 1.1548 -0.8258 3.2192   0.0000 -14.2359 0.0470 ML
2 2.0001 -2.5112 0.1953 1.3414 -0.7276 2.850    0.0000 -17.332  0.0320 ML
```

Figure 5. Example of the beginning of a log file. When the log command is used, one common file containing the estimates from all data files is generated. This is especially convenient for further statistical analyses.

Appendix A 3

Manuscript 3: 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.

doi: 10.3758/s13428-016-0740-2

How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria

Veronika Lerche, Andreas Voss, and Markus Nagler
Ruprecht-Karls-Universität Heidelberg

Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Markus Nagler, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany.

This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant Nr. VO1288/2-1).

Correspondence concerning this article should be addressed to Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstrasse 47-51, D-69117 Heidelberg, Germany, email: veronika.lerche@psychologie.uni-heidelberg.de, telephone: +49-6221-54-7322.

ABSTRACT

Diffusion models (Ratcliff, 1978) make it possible to identify and separate different cognitive processes underlying responses in binary decision tasks (e.g., the speed of information accumulation vs. the degree of response conservatism). This becomes possible because of the high degree of information utilization involved. Not only mean response times or error rates are used for parameter estimation but also the response time distributions of both correct and error responses. In a series of simulation studies, the efficiency and robustness of parameter recovery were compared for models differing in complexity (i.e., in numbers of free parameters) and trial numbers (ranging from 24 to 5,000) using three different optimization criteria (maximum likelihood, Kolmogorov-Smirnov, and chi-square) that are all implemented in the latest version of *fast-dm* (Voss, Voss, & Lerche, 2015). The results revealed that maximum likelihood is superior for uncontaminated data, but in the presence of fast contaminants, Kolmogorov-Smirnov outperforms the two other methods. For most conditions, chi-square-based parameter estimations lead to less precise results than the other optimization criteria. The performance of the fast-dm methods was compared to the *EZ* approach (Wagenmakers, van der Maas, & Grasman, 2007) and to a Bayesian implementation (Wiecki, Sofer, & Frank, 2013). Recommendations for trial numbers are derived from the results for models of different complexities. Interestingly, under certain conditions even small numbers of trials ($N < 100$) are sufficient for robust parameter estimation.

Keywords: diffusion model, fast-dm, mathematical models, reaction time methods

How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria

The diffusion model was introduced almost four decades ago by Roger Ratcliff (1978) as a model for cognitive processes in memory retrieval. Since then, it has been shown that the model can map cognitive processes from a multitude of different cognitive tasks that require fast binary decisions, including—for example—color or numerosity classifications, or lexical decision tasks (see Voss, Nagler, & Lerche, 2013, for a recent review). Thus, the diffusion model can be seen as a generic model for binary decisions. Why have accumulator models like the diffusion model become so popular in recent years? The advantage over traditional analyses of response time (RT) means (or error rates) is that different aspects of cognitive processing can be measured separately. Imagine a study on cognitive aging that analyzes the stability (or decline) of cognitive performance in a specific task at high age. If the mean RT is used as the dependent measure, you cannot be sure whether the longer RTs are really based on slower information processing, because older adults may be more cautious—that is, they may respond only if they are really sure about the correct response (e.g., Forstmann et al., 2011; Ratcliff, Thapar, Gomez, & McKoon, 2004). Additionally, older adults may be slower in motor response execution (e.g., Ratcliff, Thapar, & McKoon, 2004). Thus, it is important to get a valid measure for speed of information processing that is not confounded by speed-accuracy settings or the speed of motor response. In the diffusion model framework, the drift parameter provides such a measure of cognitive speed (see Voss, Rothermund, & Voss, 2004, for an experimental validation study).

In the first three decades following its introduction in psychological research, Ratcliff's diffusion model (1978) was used primarily by researchers with a profound interest in, and knowledge of, mathematical psychology. In recent years, however, the diffusion model has increasingly attracted the attention of researchers from various other fields of psychology. Examples indicating the wide range of applications for the diffusion model include analyses of cognitive processes in such typical experimental paradigms as the lexical decision task (e.g., Yap, Balota, & Tan, 2013), sequential priming paradigms (e.g., Voss, Rothermund, Gast, & Wentura, 2013), task switching (Schmitz & Voss, 2012, 2014), or prospective memory paradigms (e.g., Boywitt & Rummel, 2012). Other applications encompass social cognitive research (e.g., Germar, Schlemmer, Krug, Voss, & Mojzisch, 2014; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Voss, Rothermund, & Brandtstädter, 2008), cognitive aging (e.g., McKoon & Ratcliff, 2013; Spaniol, Madden, & Voss, 2006), cognitive processes related to psychological disorders (e.g., Metin et al., 2013;

Pe, Vandekerckhove, & Kuppens, 2013; White, Ratcliff, Vasey, & McKoon, 2010b), and other fields of psychology.

So far, the diffusion model has often been applied for the detection of differences between groups or conditions (e.g., Boywitt & Rummel, 2012). More recently, the correlations between diffusion model parameters and external criteria have also constituted a research field (e.g., between drift rate and general intelligence; see Ratcliff, Thapar, & McKoon, 2010). On the basis of such observations, lately, the idea has been expressed that the diffusion model might also be used as diagnostic tool (e.g., Aschenbrenner, Balota, Gordon, Ratcliff, & Morris, 2016; Ratcliff & Childers, 2015).

These different types of applications of the diffusion model go along with different requirements regarding parameter estimation accuracy. For example, for the detection of differences between conditions, biases in parameter estimation are not necessarily a problem. Imagine an estimation procedure that results in a systematic overestimation of the drift rates in both of two conditions. If the estimation bias is similar over conditions, it will not affect the power of difference detection. If, however, the estimation bias depends on the experimental condition (e.g., via the number of error responses), the power to detect differences between the conditions might be affected. In another scenario, there might be no systematic estimation bias, but imprecise measurement could lead to large average deviations between the true and reestimated parameter values. The increased error variance would directly diminish the power of difference detection. In this case, an increase in the number of participants can reestablish the power to detect any effects on parameters. Finally, if the diffusion model is applied for the diagnosis of interindividual differences in cognitive functioning, it is important that the relevant parameter be estimated very accurately (i.e., reliably) for each single individual. Thus, depending on the aim of the researcher, more or less strict criteria would have to be applied.

One important methodological factor that directly influences the precision of results is the number of trials. There has been a huge variation in the numbers of trials used for previous diffusion model experiments, ranging from less than 100 to several thousands of trials per participant. The choice of trial numbers typically seems to be rather arbitrary. It is remarkable that the required trial numbers have rarely been analyzed systematically so far (see Lerche & Voss, 2016b; Ratcliff & Childers, 2015; Wiecki et al., 2013, for some exceptions).

The main aim of the present article is to provide well-founded recommendations regarding the requisite trial numbers for robust diffusion modeling. As we discussed above, the question of requisite trial numbers is closely related to the precision that is necessary for a specific research question. In a series of simulation studies, we tested the precision of parameter estimation procedures for very small to very high trials numbers. This allowed us to derive conclusions for minimally required trial numbers (i.e., a number below which diffusion modeling becomes virtually meaningless), as well as “maximum” trial numbers (above which increases in precision become negligible).

A factor influencing the required number of trials is the efficiency of the applied estimation procedure. Accordingly, a second objective of this article is the comparison of the efficiency of different optimization criteria for the parameter search procedure for diffusion modeling. The simulations in this article were carried out using *fast-dm-30* (Voss et al., 2015), which is the newest version of *fast-dm* (Voss & Voss, 2007, 2008). Besides the Kolmogorov-Smirnov criterion that was implemented in former versions, *fast-dm-30* now includes implementations of the chi-square and maximum likelihood criteria. The implementation of these within the same program facilitates comparisons of the criteria’s performance. Thus, in contrast to studies that have compared different programs (e.g., van Ravenzwaaij & Oberauer, 2009), we can exclude the possibility that any differences between optimization criteria are due to program specifics.

Finally, a third focus is the influence of model complexity on the required numbers of trials. Typically, diffusion model analyses allow for intertrial variations of diffusion model parameters (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). Although estimates of intertrial variability seldom allow meaningful psychological interpretations, they often do improve model fit. However, it remains unclear how this increase in model complexity would influence the precision of estimates for the more meaningful diffusion model parameters. To investigate the influence of model complexity, we analyzed four differently complex models. Note that in the present article only models for simple experimental designs are considered. If data from more complex designs with different conditions were mapped, models would probably be more stable (and hence, the requisite trial numbers lower) if it were known on which parameters the manipulation would map; if not, the increasing number of model parameters might make the model even more unstable.

In the following sections, we first give a short introduction to diffusion modeling. This is followed by the presentation of the main properties of the different optimization criteria (i.e., chi-square, maximum likelihood, and Kolmogorov-Smirnov). In the subsequent section, the available computer programs for diffusion model analyses are briefly presented. After this, we go into the main research issues, giving an overview of initial simulation studies comparing different optimization criteria and different trial numbers. Finally, we outline and discuss the methods and results of our simulation studies.

The rationale of the diffusion model

Researchers dealing with data from binary decision tasks often use either the percentage of correct responses or the mean RTs as dependent measures. However, some research questions cannot be properly addressed on the sole basis of (one of) these measures. For instance, different speed-accuracy settings can make it difficult to interpret an observed difference in mean RTs between two groups or conditions. Are the longer RTs in one of these conditions due to slower information uptake, or rather the result of a conservative response style? The diffusion model helps solve this problem, because it maps speed-accuracy settings and the speed of information processing on independent parameters. This decomposition becomes possible by taking into account the complete distributions of both correct and error responses (and thus, implicitly, also the error rate). Thereby, several cognitive components are identified that have clear psychological interpretations (Voss et al., 2008; Voss et al., 2004). This makes it possible to answer not only the question of *whether or not* people (or tasks) differ in their performance in a cognitive task, but also to determine *in what way* they differ (e.g., *why* one person is faster than another). Note that several mathematical models allow such a separation of the different components involved in decision tasks. One prominent example is the linear ballistic accumulator model (Brown & Heathcote, 2008). In this article, we focus on the diffusion model (Ratcliff, 1978).

The basic assumption of the diffusion model is that decisions are based on a continuous information-sampling process that is described by a Wiener diffusion process (i.e., a diffusion process with constant drift) running in a corridor between two thresholds (see Figure 1). The current information drives the decision process toward the upper or the lower threshold, representing two possible decisional outcomes. As soon as the upper or lower threshold is hit, the decision is reached, and a corresponding motor program is initiated. Because the diffusion process is a stochastic (i.e., noisy) process, durations and outcomes may vary from trial to trial, even if identical stimulus information is presented.

In the following paragraphs, we shortly present the parameters of the diffusion model. The *drift rate* (parameter *v*) indicates the average speed (and direction) of information uptake. High (absolute) drift rates lead to fast responses and few errors, whereas a drift around zero indicates chance performance with long RTs. Thus, high drift rates indicate higher cognitive speed or easy tasks.

A second model parameter is the *distance between the two thresholds* (parameter a). This parameter defines how much information is considered before a decision is made. A large threshold separation means that a lot of information needs to be sampled before the decision is made, which will result in large RTs with a low error rate. Thus, conservative decision-makers will have large threshold separations, and liberal decision-makers small ones.

The *starting point* (parameter z) defines the position at which information accumulation begins. If z is not centered between the thresholds, there is a decision bias in favor of the threshold that is closer to the starting point. To reach this “preferred” threshold, the process needs less information, and the corresponding responses will therefore be more frequent and faster. Instead of the absolute value z, often the relative starting point $z_r = z/a$ is reported (e.g., Voss et al., 2015), with $z_r = .5$ reflecting an unbiased decision process. Note that decision bias mapped by the starting point is conceptually similar to response bias in the signal detection framework. We prefer the term *decision* bias because it influences the decision process, and not merely response execution.

In addition to the decision times, in the analysis of RT data the duration of *nondecisional processes* (parameter t_0 or T_{er}, not shown in Figure 1) also needs to be considered. These nondecisional processes can temporally precede (e.g., encoding of information) or follow (motoric execution of the response) the accumulation process.

Furthermore, the diffusion model can also explain trial-to-trial fluctuations in performance that arise—for example—from variability in the stimulus information or in the attention of the participant. For this purpose, intertrial variability parameters have to be included (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002; see also Laming, 1968). Specifically, it is assumed that the drift across trials follows a normal distribution with mean v and standard deviation s_v. The starting point and nondecision time are assumed to be normally distributed, with means z_r and t_0 and widths s_{zr} and s_{t0}, respectively. More recently, the diffusion model has been expanded to include a response bias parameter (parameter d) that maps differences in the duration of nondecisional processes between the two responses (Voss, Voss, & Klauer, 2010; Voss et al., 2015).

Finally, the diffusion model includes the diffusion coefficient—that is, the amount of noise in the diffusion process (sometimes called the *intratrial variability of drift*). The diffusion coefficient is typically not estimated but instead used as a scaling parameter (theoretically, either v or a could be used as the scaling parameter, and then the diffusion coefficient could be estimated). We set the diffusion coefficient to $s = 1$ (and thus held it constant across conditions; see Donkin, Brown, & Heathcote, 2009, for a different suggestion). If another value is used, all diffusion parameters (except t_0 and s_{t0}) are rescaled by the factor s (e.g., Ratcliff usually sets s to .1 in his applications).

Optimization Criteria

One common aim in diffusion model analysis is to find a set of parameters that optimally describes the empirical data. To achieve this, deviations between the observed data and the data predicted from a certain set of parameters are minimized by adjusting the parameter values. For this purpose, different optimization criteria quantifying the goodness of fit between the observed and expected data have been used in the diffusion model literature. In the following discussion, we present three criteria that have frequently been applied in the context of diffusion modeling: chi-square (CS), maximum likelihood (ML), and Kolmogorov-Smirnov (KS) (see also Table 1).

Chi-Square

The CS criterion has often been used for parameter estimation (Ratcliff & McKoon, 2008; Ratcliff & Tuerlinckx, 2002; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). To calculate CS, responses are grouped into bins according to latency. This is done separately for the responses at the upper and lower thresholds. The borders of the bins are based on quantiles of the RTs observed. Ratcliff and Tuerlinckx (2002) proposed the use of six bins, with the two outer bins each comprising 10 % of the observed RTs, and the other four bins 20 % each. Accordingly, the borders of the RT bins are defined by the .1, .3, .5, .7, and .9 percentiles of the empirical RT distributions. These bins are then applied to the predicted distributions. From the deviations between the numbers of predicted and observed responses for each bin, a CS value is computed.[1] In an iterative parameter search process, this CS sum is minimized. Because the predicted (cumulative) distributions only need to be evaluated at the borders of the bins, computation is fast and independent of the number of trials.

Maximum likelihood

The ML criterion is used in various mathematical modeling approaches. In contrast to the CS approach (e.g., Ratcliff & Tuerlinckx, 2002), the ML approach uses every RT, and no binning is necessary. A set of parameters is sought to maximize the likelihood of the empirical data. For technical reasons (see Ratcliff & Tuerlinckx, 2002), typically the sum of logarithmized density values is maximized rather than the product of densities. Unlike CS, the

[1] Strictly speaking, the resulting value is not exactly a chi-square value, because the borders of the bins are determined from the empirical distributions and not from the predicted distributions (Speckman & Rouder, 2004). However, the values approximate a chi-square distribution (Fific, Little, & Nosofsky, 2010). In diffusion model analyses, CS is usually calculated in this way, because it is computationally much easier and faster.

computation time required by ML depends strongly on the number of trials per data set, because the predicted density has to be computed for each trial.

Kolmogorov-Smirnov

The KS criterion was introduced as an optimization criterion for diffusion modeling by Voss et al. (2004) and has been applied in numerous studies (e.g., Horn, Bayen, & Smith, 2011; Metin et al., 2013; Voss, Rothermund, et al., 2013). The criterion is based on the cumulative distribution functions (CDFs) of RTs. To calculate the KS criterion, the distributions at the upper and lower thresholds are combined by multiplying all RTs from the lower threshold by -1 (a procedure first proposed by Voss et al., 2004; see also Voss & Voss, 2007). This creates a cumulative density function for the whole data set. The KS criterion is the maximum absolute vertical distance between the observed and predicted CDFs. Accordingly, for each observed RT the distance between the two CDFs needs to be computed to identify the maximum. The iterative search for parameter estimates then aims at minimizing this maximum distance.

Estimation Programs

For many years researchers had to develop parameter search implementations of their own for the diffusion model analyses. In recent years, several programs were published for this purpose. Among them is the *EZ-diffusion model* (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008; Wagenmakers et al., 2007), which is available as JavaScript, R code, a MATLAB implementation, and an Excel spreadsheet. In comparison with search procedures based on the three optimization criteria presented in the last section, EZ uses a more limited amount of information. In the original version of EZ (Wagenmakers et al., 2007), parameters were estimated from error rates and the mean and variance of the correct responses. Closed-form equations are utilized for the parameter calculation. In this way, estimates for three parameters can be obtained (a, v, t_0). In the extended versions of EZ (Grasman et al., 2009; Wagenmakers, van der Maas, et al., 2008) further parameter options are available, such as estimation of parameter z and the consideration of contaminant data.

The *Diffusion Model Analysis Toolbox* (DMAT; Vandekerckhove & Tuerlinckx, 2007, 2008) is a MATLAB toolbox. In DMAT, the CS method is implemented. Furthermore, the toolbox offers the possibility of using quantile maximum probability estimation (see also Heathcote & Brown, 2004; Speckman & Rouder, 2004).

A third program, *fast-dm* (Voss & Voss, 2007, 2008), is a command-line program. In fast-dm-29 and all earlier versions, parameter search was generally based on KS as the optimization criterion. The newest version, *fast-dm-30* (Voss et al., 2015), offers a choice between a KS, ML, and CS approaches.

The last few years have seen the advent of software solutions for hierarchical diffusion model analyses. Vandekerckhove, Tuerlinckx, and Lee (2011) proposed a plug-in to the WinBUGS software. A platform-independent solution HDDM (for *hierarchical drift diffusion model*) has been presented by Wiecki et al. (2013). HDDM is a toolbox based on Python and uses a Bayesian method for parameter estimation. It can be used either for fitting a hierarchical model or for fitting parameters for each individual subject. Recently, another platform-independent software option was introduced by Wabersich and Vandekerckhove (2014).

Literature on comparison of optimization criteria

There is a lack of systematic research on the performance of different optimization criteria for diffusion modeling. One exception is the study by van Ravenzwaaij and Oberauer (2009), who compare the performance of EZ, fast-dm*,* and DMAT (using the multinomial log-likelihood function, MLF). They find KS to be superior to MLF in terms of the correlations between the true and recovered parameter values. However, MLF performed better than KS in recovering the mean true values. Since the comparison of the optimization criteria KS and MLF was based on different software solutions, however, program details may have been the factor behind the resulting differences, which were not necessarily based on the different optimization criteria. Interestingly, EZ performed very well in this study. Especially in the event of a reduction in the number of trials (80 instead of 800), EZ outperformed fast-dm in the correlations (DMAT could not even be applied in this condition, since it needs a minimum of 11 errors in each RT quantile). However, EZ (even in the more recent versions) does not allow for the estimation of intertrial variabilities, so full comparability with KS and MLF cannot be established. Note that the estimation of intertrial variabilities (especially s_z and s_v) posed serious problems for fast-dm and DMAT. This may have had a negative influence on the recovery of the other parameters. EZ, on the other hand, circumvents the estimation difficulties associated with intertrial variabilities by providing estimates for only three parameters (a, v, t_0). Besides, no contaminated trials were included in the simulation studies. *Contaminants* are responses resulting from sources other than a diffusion process. In several simulation studies, Ratcliff (2008) demonstrated the sensitivity of EZ to the presence of contaminants (but see Wagenmakers, van der Maas, et al., 2008).

Ratcliff and Tuerlinckx (2002) compared ML, CS, and a weighted least squares (WLS) fitting method, both with and without the inclusion of contaminants. They showed that for a model consisting of eight parameters (a, t_0, four drift rates, s_z, and s_v; z was assumed to be centered between the thresholds) and data without contaminants, ML outperformed CS and WLS. When contaminants were added, ML's performance deteriorated dramatically. CS was also impaired by the presence of contaminant trials, whereas the performance of WLS deteriorated only slightly. Ratcliff and Tuerlinckx (2002) counteracted the deterioration of ML and CS by explicitly modeling the contaminants with a uniform distribution. Consequently, parameter recovery improved. Furthermore, they included the intertrial variability of t_0 into the model (s_{t0}), and with both this additional parameter and the modeling of contaminants, CS resulted in precise and unbiased estimation when 1,000 trials per condition were used. For 250 trials per condition the performance was significantly worse. The authors recommended using CS with the correction for contaminant trials and s_{t0} included in the model. Note, however, that for 250 trials per condition, the performance of CS in this model was "very poor" (p. 467). Subsequently, many researchers have used CS, referring to the studies by Ratcliff and Tuerlinckx (2002) stating that CS "provides the best balance between robustness and the ability to recover parameter values" (Wagenmakers, Ratcliff, et al., 2008, p. 146).

Recently, the performance of newly developed hierarchical diffusion models has been tested. Wiecki et al. (2013) compared CS- and ML-based algorithms to HDDM, their software solution for a hierarchical Bayesian estimation of parameters. Their work revealed the superiority of HDDM, especially for small trial numbers. Besides, ML often outperformed CS.

Ratcliff and Childers (2015) ran a series of simulation studies in which they compared eight different estimation methods and programs. DMAT cut a poor figure, and EZ did not perform very well in the presence of contaminants. However, CS (based on either ten or six bins), ML, and KS generally recovered the parameters quite well. Some of the findings for HDDM were inconsistent. For example, in Simulation Study 1, in one design (with four drift rates) HDDM featured high correlations between the true and re-estimated parameter values even for small trial numbers, outperforming the other methods. In another design (with two drift rates) for smaller numbers of trials, it performed worse than most of the other methods. Besides, in another simulation study (Simulation Study 2), unexpectedly, high biases were found for a large trial number, whereas the biases for a small trial number were smaller than those in the other methods.

With the availability of fast-dm-30 (Voss et al., 2015), the three criteria CS (based on six bins), ML, and KS can be compared to each other independently of confounding factors (i.e., program specifics). As we outlined in the section on optimization criteria, CS, ML, and KS differ in the amounts of information used for the fitting process. Whereas CS reduces the available information by dividing the distributions into bins, ML and KS consider the exact value of each RT observed. This is why we expected our simulation studies to reveal that the number of trials required for efficient parameter estimation was higher for the CS criterion than for ML and KS. KS requires calculation of the vertical distance for each RT observed; the criterion itself, however, is based on only one of these distances (the maximum distance). Accordingly, ML may be a more efficient estimator than KS.

However, we also expected the optimization criteria to differ in terms of robustness in the presence of "contaminants".[2] Because the (log-)likelihood can be strongly influenced by single RTs, we assumed that results from the ML method would be most strongly biased when RT distributions were contaminated, whereas the CS and KS criteria were expected to be more robust. Therefore we expected ML to require the lowest number of trials, followed by KS and CS, with uncontaminated data. In the presence of contaminants, however, ML should perform worse than KS.

Number of trials required

Is diffusion modeling restricted to experimental designs with more than 1,000 trials per participant? After receiving regular inquiries from researchers greatly interested in diffusion modeling but uncertain about the number of trials required for robust analysis, we decided to address this issue systematically. Conventionally, high numbers of trials are used for diffusion modeling. For instance, Ratcliff, Thapar, Gomez, et al. (2004) used 2,100 trials in their experimental session (see also Leite & Ratcliff, 2011; Ratcliff & Rouder, 1998; Ratcliff & Van Dongen, 2009; Wagenmakers, Ratcliff, et al., 2008; but see Klauer et al., 2007). Although generally a large data base makes the fitting of mathematical models more stable, obviously using extraordinarily large trial numbers can cause problems of its own. First, the experimental sessions require more time and effort. More importantly, psychological effects may change over time due to practice effects, and after several hundred trials, some effects of interest may be diminished or even disappear completely. Additionally, it may often be difficult to find sufficient stimuli, if they are not supposed to be repeated.

One interesting approach to addressing the issue of trial numbers by way of experimental design has been proposed by White, Ratcliff, Vasey, and McKoon (2009), who used filler trials (see also White et al., 2010b) to achieve higher accuracy in parameter

[2] We consider an estimation procedure to be "robust" when its results are not biased by contaminants.

estimation. Some parameters (response criteria and non-decisional processes) were estimated on the basis of both target and filler trials. In this way, the authors could use several hundred trials for the parameter estimation, resulting in more stable estimates for the drift rates, which were the actual focus of their studies. Although this approach addresses the problem of sparse stimulus material, the authors were still using several hundred trials, and the question remains unanswered whether these high trial numbers are actually necessary.

There is a general consensus that higher trial numbers lead to higher accuracy in parameter estimation. This has been confirmed by several simulation studies (e.g., Ratcliff & Tuerlinckx, 2002; Vandekerckhove & Tuerlinckx, 2007). In these studies, however, trial numbers were manipulated only in a limited range. A more systematic comparison of different trial numbers was done by Wiecki et al. (2013; see also Ratcliff & Childers, 2015). They varied the number of trials from 20 to 150 per condition (in a design with two drift rates and s_t and s_z fixed at zero) and analyzed the mean absolute errors of the single parameters and the probability of detecting a significant difference between the two drift rates. Their results revealed an improvement in parameter estimation when the number of trials was increased.

Although these studies clearly demonstrated that parameter estimation improves with the number of trials, they did not focus on the inference of guidelines for the trial numbers required.

Method

To compare the performance of different parameter estimation methods (CS, KS, ML) and programs (HDDM and EZ) and to infer guidelines for the numbers of trials necessary for efficient and robust parameter estimation, a set of simulation studies was carried out. In the following sections, we first describe the design of these studies, proceeding from there to present our criteria for evaluating the performance of the optimization criteria.

Design

In our studies, we tackled two different designs, in which one drift rate or two drift rates were estimated. Diffusion models with one drift rate are mostly used to analyze data that has been coded as correct (e.g., upper threshold) vs. error (e.g., lower threshold). This kind of analysis allows collapsing data across different stimulus types. Alternatively, one-drift models might be applied for subsets of data based on the same stimulus types. The one-drift design was used in a first series of simulations. Both for simulation and parameter reestimation, we used models that differed in the number of free parameters. The seven-parameter model was composed of all seven parameters typically used in diffusion model analyses (a, v, t_0, z_r, s_v, s_{t0}, and s_{zr}); in the six-, four-, and three-parameter models, certain parameters were fixed at constant values. In the six-parameter model, the relative starting point z_r was fixed to .5 (the process starts centered between the two thresholds); in the four-parameter model the three intertrial variabilities (s_v, s_{t0}, and s_{zr}) were fixed at zero; and in the three-parameter model both the intertrial variabilities and the starting point were fixed. For each model (i.e., three-, four-, six-, and seven-parameter models) 1,000 random parameter sets were generated with typical parameter values observed in previous applications. The parameter values were drawn from uniform distributions with the minimum and maximum values shown in Table 2. Subsequently, for each parameter set, seven random data sets with different numbers of trials (24, 48, 100, 200, 500, 1,000, and 5,000) were simulated with *construct-samples*[3], resulting in a total number of 4 (models) × 1,000 (parameter sets) × 7 (trial numbers) = 28,000 simulated data sets.

Whenever performance differs between the stimulus types, or when there is an a priori bias in favor of one of the responses, a more complex model using two drift rates is needed. Typically, thresholds are associated with responses for the two stimulus types, and drift rates are estimated separately for each stimulus type in one model (e.g., White, Ratcliff, Vasey, & McKoon, 2010a; Yap, Balota, Sibley, & Ratcliff, 2012). This results in a drift with positive sign for the stimulus at the upper threshold and a drift with negative sign for the stimulus at the lower threshold. In our simulations, this procedure was mapped by a “two-drift design”. In particular, we simulated data sets with two stimulus types, using one positive and one negative drift that were allowed to vary in absolute values (i.e., difficulty). The drift values were drawn from a multivariate normal distribution. They were generated to represent a difference of d_z = 0.35 (Cohen, 1988).[4] All other parameters were equivalent to those in the one-drift design. We also used the same numbers of trials as in the one-drift design, with, for example, 24 trials composed of 12 trials of one stimulus and 12 trials of the other.[5] A total of 1,000 data sets were constructed for each of the four models (with different numbers of parameters) and for each trial number, resulting in another 28,000 data sets (4 models × 7 trial numbers × 1,000 parameter sets). In the remainder of this study, we will refer to the models as the *three-*, *four-*, *six-* and *seven-parameter models*, so as to use the same terms as in

[3] *construct-samples* is part of the fast-dm software. It simulates response time data by applying a random-walk with very small time steps.

[4] Effect size formula: $d_z = \frac{M_1 - M_2}{SD_{diff}}$, with $d_z = 0.35$, $M_1 = 2.00$, $M_2 = 2.35$, $SD_1 = SD_2 = 1$, and $r = .50$.

[5] CS, as implemented in fast-dm, only allows for parameter estimation if in each condition at least 12 trials are observed for one of the two responses. Accordingly, in the condition with 24 trials, the comparability of the CS method to the other estimation methods is limited, because not all data sets met this precondition.

the one-drift design, even if each model actually contains one further parameter (i.e., the second drift rate).

We also performed robustness tests for both the one-drift and two-drift designs. In particular, 4 % of the trials of each data set were randomly chosen and substituted for by either fast or slow contaminants. Fast contaminants were used to simulate fast guesses, that is, trials in which participants respond quickly without processing the target stimulus. Since fast-guess response is situated on the level of chance (Swensson, 1972), the response of each selected trial was randomly set to either 0 (*lower threshold*) or 1 (*upper threshold*). In terms of RTs, we used latencies at the left-hand edge of the original distribution for fast contaminants, thus ensuring that these values cannot be easily identified as outliers; real "statistical" outliers farther from the original distribution would bias the result more severely but would at the same time be easier to detect prior to analysis. More specifically, latencies for fast contaminants were drawn randomly from a uniform distribution with a range of 200 ms centered around the fastest theoretically possible time for each parameter set (i.e., using an interval from t_{min} - 100 ms to t_{min} + 100 ms, with $t_{min} = t_0 - s_{t0}/2$). Secondly, we were interested in the influence of slow contaminants resulting from temporary distraction of participants from the task in hand. In this condition, only the RTs of the selected trials were changed, but not the respective types of response. Latencies for these types of contaminants were randomly chosen from a uniform distribution ranging from 1.5 to 5 interquartile ranges above the third quartile of the original data.

Parameter estimation

Parameter values were recovered using fast-dm-30 (Voss et al., 2015) from uncontaminated and contaminated data sets.[6] This was done using each of the three optimization criteria (CS, KS, ML). Furthermore, we analyzed all data sets with HDDM (Wiecki et al., 2013) and the data sets of the three-parameter model additionally with the EZ method (Wagenmakers et al., 2007).[7] As with our settings for HDDM (version: 0.5.3), we used 2,000 samples, a 20-sample burn-in and the proportion of outliers was fixed at zero.[8]

[6] Fast-dm was executed with the precision parameter set to 3. Setting the precision to 4 significantly slows down the estimation process without having any relevant positive impact on the parameter recovery achieved.

[7] EZ cannot be applied to data sets with an accuracy rate of 0 %, 50 % or 100 %. For data sets with an accuracy of 100 %, we applied an edge correction method that has also been used by Wagenmakers et al. (2007): accuracy = $1 - \frac{1}{2 \times n}$, with n being the number of trials. We used similar approaches for 0 % (accuracy = $\frac{1}{2 \times n}$) and 50 % (accuracy = $0.5 + \frac{1}{2 \times n}$) accuracy rates. In the two-drift design, EZ was applied separately to the trials of each response type. We then computed the means over the two threshold separations and the two nondecision components.

[8] The prior distributions in HDDM are a Gamma distribution (threshold separation, nondecision time), a normal distribution (drift rate, starting point), a half normal distribution (intertrial variability of drift rate and nondecision time) and a Beta distribution (intertrial variability of starting point; see also Wiecki et al., 2013).

In total, for both the one-drift and the two-drift designs we used 28,000 data sets (4 models × 1,000 parameter sets × 7 trial numbers) with three types of contamination (none, fast, slow) analyzed with four methods (CS, KS, ML, and HDDM) requiring 672,000 runs (+ 21,000 runs of EZ) of the parameter estimation procedure.[9]

As we mentioned before, the number of parameters reestimated was equivalent to those of the parameter model on which the simulation was based. For instance, in the case of the three-parameter model in the one-drift design only the three parameters *a*, *v*, and t_0 were estimated, whereas the remaining four parameters were each fixed at the correct constant value ($s_{t0} = s_{zr} = s_v = 0$, $z_r = .5$). The four- and seven-parameter models includes the estimation of starting point z_r. This is only possible if there are two distributions of responses (at the upper and lower thresholds). If no data are available for one of the two thresholds, the distance from the starting point to the "empty" threshold is not defined. Accordingly, for the models that estimated a starting point, we excluded all data sets in which the smaller distribution (response 0 or response 1) comprised fewer than 4 % of all trials (i.e., at least one trial at each threshold in the smallest data sets with 24 trials). The number of remaining data sets ranged from 689 to 801 out of 1,000 for the different conditions in the one-drift design. In the two-drift design, only in one condition did one data set have to be excluded due to the "4 % criterion". Because in the three- and six-parameter models the starting point is fixed (and thus the distance from a threshold without trials to the starting point is also defined), the estimation was carried out for all data sets.

Evaluation Criteria

The evaluation of parameter estimation performance was based on four main criteria: (1) correlations between the true and reestimated parameter values, (2) parameter estimation biases (i.e., deviations of the reestimated from true parameter values), (3) the numbers of participants required for detection of a drift rate difference in the two-drift design, and (4) the estimation precision, assessed as squared deviations of the reestimated from the true parameter values. In the following discussion, we present the rationale for the choice of these criteria (see also Table 3 for a summary) and give details on the computation. Additionally, (5) we evaluated the computation time required for parameter estimation.

Note that these distributions differ from the one (uniform distribution for all parameters) that we used to create the parameter values, which could deteriorate the performance of HDDM.

[9] The estimations with fast-dm and HDDM were carried out using the computational resource bwUniCluster, funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

First, parameter recovery performance was assessed by the *correlations of each parameter's true values with the reestimated values*. This criterion is of relevance if the focus of the researcher lies in the detection of relationships between diffusion model parameters and external criteria (e.g., the relationship between the drift rate and general intelligence; e.g., Ratcliff et al., 2010). One weakness of correlation coefficients is that they fail to reveal systematic biases in parameter recovery. Often, such a systematic bias might be unproblematic, because it does not invalidate the interpretation of the results. However, there might be cases in which an estimation bias is related to the true parameter values (e.g., an estimation bias might be stronger when fewer error data are present—i.e., when drift is strong). In such a case, biased parameter estimation might challenge the internal validity of results.

Thus, our second criterion was a *measure of parameter biases*. For each parameter, we computed the differences between the estimated and the true parameter values. Accordingly, a positive value indicated that the parameter was overestimated, whereas a negative value showed a parameter underestimation. Besides, we computed the mean bias for each parameter quartile (i.e., the mean of all data sets lying in the first, second, third and fourth quartiles of the true parameter values), to graphically depict possible dependencies between the parameter values and biases. We also computed Pearson correlation coefficients between the true parameter values and the respective biases.

Note that a parameter might be estimated without bias, but still with low precision. For some participants the parameter might be overestimated, and for some underestimated, with no clear pattern. This can be a problem for difference detection, due to higher variability of the values within groups/conditions. Using a higher number of trials is one way to enhance the power of a statistical test, as parameters are estimated with less of a noise variance. Another way is to enhance the number of participants. Our third criterion was the number of participants required for the detection of a drift rate difference between two conditions. Specifically, for the two-drift model we calculated the effect sizes resulting from the recovered drift rates. Using *pwr.t.test* from the pwr R package (Champely, 2012; R Development Core Team, 2014) for the observed effect sizes between the two drift estimates, we obtained the numbers of participants required for a power of 80 % (in a two-sided paired *t* test with a significance level of 5 %). If the drift parameters were estimated perfectly (i.e., with no deviations of the estimated from the true parameter values), 66 participants would be required to detect this difference with a power of 80 % (two-sided testing).

Although an increase in the sample increases the power to detect differences between conditions, no such compensation of low precision is possible, when the aim of the researcher lies in a diagnostic application of the diffusion model (Aschenbrenner et al., 2016; Ratcliff & Childers, 2015). For this purpose, it is of great importance that parameters be estimated precisely for all individuals, thus minimizing deviations between the true and estimated values. Accordingly, our fourth evaluation criterion was the precision of parameter estimates, calculated as the *squared deviations of the reestimated from the true parameter values*. In contrast to the bias measure, here we do not differentiate between over- and underestimation of a parameter (using squared values, any deviation would contribute equally). Note that the diffusion model parameters have quite different scales, and the accuracy of recovery varies appreciably between parameters. Whereas, for example, t_0 can be estimated very precisely (e.g., to the third decimal place), the deviation of true and recovered values is often much greater for the drift. Accordingly, to enhance the comparability of parameters, we standardized each parameter's bias by its respective "possible accuracy". These "possible accuracies" were deducted from an optimal parameter recovery condition—that is, from the parameter reestimations using the ML approach for data sets in the one-drift design with 5,000 trials, a minimum of 4 % of trials at each threshold, and no contaminants. From the results of these analyses, the 95 % quantiles of the absolute differences between the true and estimated parameter values were used as the "possible accuracies" for all parameters (see Table 2 for each parameter's "possible accuracy").

Finally, our last criterion—of minor importance relative to the four evaluation criteria previously presented—was *computation time*, which was the time required for the estimation process. An efficient optimization criterion should not only recover the true parameter values with high efficacy, but also require only a short time for the estimation process.

Results

In the following sections, we report our results structured by our five evaluation criteria: (1) correlations between the true and reestimated parameter values; (2) parameter estimation biases; (3) the number of participants required for detection of drift rate differences; (4) estimation precision—that is, squared deviations of the reestimated from the true parameter values; and (5) computation time.

Evaluation Criterion 1: Correlations between true and reestimated parameter values

Figure 2 shows the results obtained for our first evaluation criterion—that is, the correlations between the true and reestimated parameter values. The dependent variable in the figure is the mean correlation averaged across all parameters of the respective model using

Fisher's Z transformation. The figure shows that—as expected—with higher trial numbers, higher correlation coefficients were reached. Two main aspects emanate from these correlational analyses: (1) The CS estimation criterion mostly showed lower correlation coefficients than the other estimation methods, and (2) the six- and seven-parameter models performed worse than the more restrained three- and four-parameter models. Responsible for the latter finding is the poor parameter recovery of the intertrial variability parameters s_{zr} and s_v, which generally cannot be recovered well. Even under the "optimal" condition (no contamination, 5,000 trials, and ML as the optimization criterion), for s_{zr} and s_v only moderate correlations of .31 and .47, respectively, are found. The performance of s_{t0} (.97) in this optimal condition is much better and, most importantly, the correlation coefficients of parameters a (1.00), t_0 (.99), v (1.00), and z_r (.99)—which are usually of greater interest than the intertrial variabilities because of their high psychological validity (Voss et al., 2004)—are excellent.

Evaluation Criterion 2: Parameter estimation biases

Second, we analyzed parameter estimation biases. Figures 3, 4, 5, and 6 present results of the one-drift design for the four psychologically most interesting diffusion model parameters a, v, t_0 and z_r, respectively[10]. We will sum up the main findings from the figures, always starting with the mean bias of each parameter (indicated by the large symbols connected by lines) passing on to an examination of the relationship between the true parameter values and the biases.

As can been seen in Figure 3, CS clearly overestimated parameter a. This overestimation decreased with the number of trials and, in the condition with no contaminants, becomes negligible at about 200 trials in the three- and four-parameter models, and at approximately 500 trials in the six- and seven-parameter models. The biases of the other methods were smaller and—akin to CS— became stable from around 200 to 500 trials on. In the case of slow or fast contaminants, often a notable bias in threshold separation remained even at large trial numbers. An interesting finding is observed for ML and HDDM for the condition with fast contaminants in the three- and four-parameter models. Whereas the biases of the other methods decreased with the number of trials, their biases increased (again, getting stable from around 200 to 500 trials on). This reveals that the absolute number of fast contaminants (the relative frequency was stable, with 4 % for all trial numbers) has an influence on the recovery of parameter a. We want to anticipate that a similar pattern emerged for parameter t_0, which was systematically underestimated by ML and HDDM, with this bias

[10] We also analyzed biases for the two-drift design. The findings were very similar to those from the one-drift design.

increasing with the number of trials. This makes sense, because these methods try to account for all RTs and adapt t_0 to the smallest observed time. With the inclusion of s_{t0}—as in the six- and seven-parameter models—the biases were much smaller, because s_{t0} helps to explain very fast RTs.

Next, we analyzed whether and how the bias depends on the true value of the parameter. For the condition with no contaminants, there were at maximum small relationships with no clear pattern ($|r| < .30$). For the condition with slow contaminants, however, the relationship of true parameter value *a* to the bias increased with the number of trials. For example, for the three-parameter model estimated by ML the correlation rose from $r = -.07$ to $r = .89$ for $n = 24$ and $n = 5{,}000$, respectively. This increase with the number of trials was less pronounced for the more complex models (e.g., for the seven-parameter model and ML: $r = .15$ at $n = 24$ and $r = .46$ at $n = 5{,}000$). In the condition with fast contaminants, the pattern was less clear-cut. For KS and CS, there were mostly no relationships or very small relationships. For ML and HDDM, on the other hand, especially in the three- and four-parameter models, a (negative) relation of the true value and the bias increased with the trial number (e.g., for the four-parameter model and HDDM: $r = -.08$ at $n = 24$ and $r = -.64$ at $n = 5{,}000$).

Akin to parameter *a,* for the drift rate, biases (mostly overestimation, especially in the six- and seven-parameter models) got stable at approximately 200-500 trials. Relationships between the true drift value (with negative true values were transformed into positive values[11]) and the respective bias were very small for data with no contaminants in basically all conditions. For data with slow contaminants, the (negative) relationship increased with the number of trials, especially in the three- and four-parameter models (e.g., for the four-parameter model and ML: $r = -.30$ for $n = 24$ and $r = -.95$ for $n = 5{,}000$). A similar increase was observed for the condition with fast contaminants for the three-parameter model, and for ML and HDDM in the six-parameter models. In the four- and seven-parameter models, the relationships were mostly positive, with a smaller influence of the number of trials.

The nondecision time was estimated quite precisely in the conditions with no or slow contaminants. Again, stability of the biases was reached at 200-500 trials. In the condition with fast contaminants, we observed a systematic underestimation, which is plausible given the added fast outliers. As we mentioned before, for ML and HDDM in the three- and four-parameter models, this bias increases essentially with the number of trials. Importantly, there

[11] This transformation was used so that the four quartiles would span a range from a very slow to a very high speed of information accumulation. Note that the results were very similar if positive and negative true drift rates were analyzed separately.

was no relevant relationship between the true value of t_0 and the sign and size of the bias (with the exception of HDDM showing negative correlations of at maximum $r = -.28$; these relationships decreased with the number of trials).

Finally, the pattern for z_r revealed that this parameter was more often under- than overestimated. More importantly, we found a negative relationship between the size of z_r and the bias present for almost all conditions. There was also no clear improvement with the number of trials. Sometimes the relationship decreased in absolute value (e.g., for no contaminants in the seven-parameter model and HDDM: $r = -.70$ for $n = 24$ and $r = -.10$ for $n = 5{,}000$); often, however, it did not change or even increased (e.g., for no contaminants in the seven-parameter model and KS: $r = -.29$ for $n = 24$ and $r = -.38$ for $n = 5{,}000$). The method with the smallest absolute correlation was ML. However, in the condition with fast contaminants, all methods featured essential negative relationships (e.g., for the seven-parameter model and ML: $r = -.45$ for $n = 5{,}000$).

Evaluation Criterion 3: Number of participants required for detection of a drift rate difference

Figure 7 shows the numbers of participants required for detecting a difference in drift rates ($d_z = 0.35$) in a two-sided paired *t* test with a power of .80 conditional on the number of trials. If parameters were recovered perfectly (i.e., the estimated drift rates were equivalent to the true drift rates), 66 participants were needed for the detection of this difference (represented by the horizontal line in the figure). Obviously, parameters are estimated less precisely from small than from higher trial numbers. Thus, more participants are required in order to compensate for the inflated error variance. Figure 7 shows that an increase of trial numbers above 200-500 did not further reduce the required sample size. In most conditions, ML outperformed the other methods. Interestingly, even for data sets with fast contaminants, ML showed a good performance. Furthermore, HDDM failed to outperform the non-Bayesian ML approach in either condition. A further finding is that the performance of EZ was generally very good.

Evaluation Criterion 4: Estimation precision—squared deviations of reestimated from true parameter values

Akin to the mean correlation coefficient over all parameters, we also computed an average measure for the squared deviations. Figure 8 shows the 95 % quantiles of these mean squared deviations for each condition, depending on the number of trials in the one-drift design. The use of the 95 % quantiles makes it possible to compare the worst cases for each condition, since deviations are smaller for most data sets. If the parameters are to be used for diagnostic purposes, it is important that the parameters be estimated accurately for all individuals.

One central aim of this article is to provide guidelines on the numbers of trials required for diffusion model analyses. Because the squared deviation criterion is the strictest criterion, we used this criterion for the definition of required trial numbers.

In the subsequent section, we first specify our procedure for identifying the trial numbers required. Then we report the results for data sets without contaminants, followed by the results observed in the conditions with contaminants. Whereas Figure 8 shows the "mean deviations" (averaged over all parameters of the respective model), in the following sections we present results separately for the four main diffusion model parameters (i.e., *a*, *v*, t_0, and z_r).

Criteria for trial numbers required

As can be seen from Figure 8, for the one-drift design, the higher the number of trials, the better the estimation usually is[12]. For uncontaminated data, the relation of deviations to trial numbers is mostly described well by power functions[13]. To find the requisite trial numbers, the fitted power functions were used whenever they fitted well (i.e., when the adjusted R^2 was at minimum .80); otherwise, linear interpolation was used.

We defined a squared deviation of 15 as a criterion for the minimal number of trials required*,* indicating that the 95 % quantiles of the deviations should be no more than 15 times as large as in the "optimal" condition. This value is obviously quite high (allowing for large deviations) and at least in part arbitrary. For interpretation, one has to bear in mind that 95 % of data sets would fit better (i.e., have a squared deviation below 15). We further determined the number of trials at which the stricter criterion of deviations of 5 was reached for 95 % of the data sets, thereby deriving guidelines on the trial numbers needed for low (15) and high (5) precision.

As the asymptotic courses of the fitted functions describing the relation of trial numbers to mean deviations in Figure 8 illustrate, adding further trials is very helpful when the number of trials is small, but for higher trial numbers further increases bring only marginal gains in accuracy. Accordingly, we also defined the number of trials above which a

[12] Some exceptions have been found. We observed that the performance of KS deteriorated from the condition with 1,000 to that with 5,000 trials in the four- and seven-parameter models. So did the performance of HDDM in the six- and seven-parameter models. We had similar findings using the two-drift design.

[13] $D = b_0 \cdot n^{b_1}$, where *D* is the 95 % quantile of squared deviations and *n* is the number of trials.

further increase had only a minimal impact on the quality of parameter recovery. As a criterion, we used the point at which the functions describing the relation of deviations to trial numbers had a slope of -0.01. The trial numbers required for low and for high precision and the limit at which a further increase made little sense are presented in Table 4 (one-drift model; at least 4 % of trials at each threshold), Table 5 (one-drift model; less than 4 % of trials at one threshold) and Table 6 (two-drift model; at least 4 % of trials at each threshold). Trial numbers are given separately for the four main diffusion model parameters (*a, v*, t_0, and z_r), depending on the complexity of the parameter model (three-/four-/six-/seven-parameter models), the type of contamination (none/fast/slow) and the estimation method (KS/ML/CS/HDDM/EZ).

Trial numbers required for uncontaminated data

In the three- and four-parameter models of the one-drift design, using ML or HDDM, low precision could be reached with fewer than 60 trials, and high precision with fewer than 160 trials. KS also performed well, with fewer than 200 trials for low precision. EZ applied to the three-parameter model was competitive with ML, HDDM and KS in terms of drift rate estimation, with approximately 70 trials for low and 200 trials for high precision. Parameters *a* and t_0, on the other hand, were estimated worse. CS showed the poorest performance requiring still fewer than 290 trials for low precision. The comparison of the different parameters reveals that with the exception of EZ, the nondecision time required the least number of trials, followed by the drift rate and the threshold separation. Parameter z_r was estimated very well by ML and HDDM (requiring fewer than 40 trials for low precision), whereas KS and CS required more trials (< 170).

In the six- and seven-parameter models, more trials were required than in the three- and four-parameter models. Again, the lowest trial numbers were always needed for the nondecision time, and the highest numbers were usually required for the threshold separation. The drift rate was estimated best by KS with fewer than 200 trials for low precision in both models. CS, on the other hand, required more than 700 trials in the six-, and more than 400 trials in the seven-parameter model. In fact, CS is usually applied with such trial numbers (or even higher trial ones), and should thus give reliable results. However, our results also show that other methods can supply satisfying reliability already with smaller trial numbers.

For data sets with fewer than 4 % of trials at one of the two thresholds, the estimation of parameters *a* and *v* requires higher trial numbers (see Table 5). Only t_0 was estimated with a performance similar to that for the other data sets.

The numbers of trials required by the two-drift design are depicted in Table 6, analogous to Table 4 for the one-drift design.[14] The parameter that suffered most from the more complex design was the threshold separation. The drift rate was also estimated worse, whereas there was no deterioration (and sometimes even an improvement) for the nondecision time. Besides, the starting point was estimated better in the two-drift design. As for the comparison of the different optimization criteria, the pattern was similar to the one observed for the one-drift design. Most importantly, HDDM in sum showed the best performance, followed by ML, KS, and CS. EZ also performed very well, even beating ML for the estimation of the drift rate.

For a better understanding of the precision of the results for trial numbers derived from the criteria for low and high precision, we also calculated the *correlations of the true and recovered parameters* at these points. Toward this aim, power functions[15] or linear interpolation (if the adjusted R^2 was beneath .80) were used. Importantly, the correlations were generally very high (most of them above .90), and therefore do not imply that even stricter criteria should be applied for the trial numbers required. The correlation coefficients were lowest for parameter t_0 (ranging from .79 to .97) and z_r (.76 - .96). Note that the trial numbers required for t_0 were often very low (even $n < 24$). Because usually many more trials will be used, even higher correlation coefficients may be reached.

Besides the requisite trial numbers, Tables 4, 5, and 6 also show the maximum trial numbers based on the slope criterion. For instance, in the three- and four-parameter models with uncontaminated data and at least 4 % of trials at each threshold, HDDM and ML reached this criterion for all parameters after fewer than 300 trials in the one-drift design, and after fewer than 600 trials in the two-drift design. In the six- and seven-parameter models, the criterion was reached after fewer than either 700 trials (one-drift design) or 1,000 trials (two-drift design). Generally, the criterion was reached earlier for nondecision time, drift rate, and starting point than for threshold separation.

Trial numbers required for contaminated data

Up to this point, we have only presented results for the condition without contaminant trials. The middle and right columns of Figure 8 show the mean deviations of the recovered parameters from contaminated data. As can be seen in Tables 4 (one-drift design) and 6 (two-drift design), in the condition with *slow contaminants*, parameter *a* was estimated much worse, requiring more trials in almost all conditions. The drift rate did not suffer much in the three- and four-parameter models, but it often required many more trials in the six- and seven-

[14] The requisite trial numbers for the drift rate are based on the mean squared deviations of the two drift rates.

[15] That is, correlation = $b_0 + b_1/n$, where n is the number of trials.

parameter models. The nondecision time and starting point often did not suffer from the addition of slow contaminants. Finally, EZ estimated the drift rate quite well, but nondecision time and, especially, threshold separation required much higher trial numbers than in the condition with no contaminants and than the other methods.

In the presence of *fast contaminants,* KS continued to display good parameter recovery. By contrast, the results from both ML and HDDM were affected strongly by the occurrence of fast contaminants. This applied to both the threshold separation and nondecision component in the three- and four-parameter models, and to all parameters in the four-parameter model. Here, ML and HDDM stayed above the critical value of 15. In the more complex models, nondecision time was estimated better (probably due to the intertrial variability of nondecision time; see also the findings for the bias measure). Besides, in the six-parameter model, especially, HDDM turned in a good performance for the other parameters as well. Across the different parameter models, the performance of CS was often better than that of ML and HDDM, but still worse than that of KS. Interestingly, EZ showed a good performance for drift rate and nondecision time despite the presence of fast contaminants. For the data sets in which the smaller response distribution comprised fewer than 4 % of the data, the pattern of results was similar.

Since the added contaminant trials were situated partly outside, partly overlapping with the RT distribution, they could not all be identified and excluded before parameter estimation. Applying the frequently used criterion of 200 ms as the lower limit for the condition with fast contaminants to the one-drift design led to the exclusion of 0.6 % of the trials on average (so only a small part of the 4 % contaminants were identified). We additionally applied the Tukey criterion (Tukey, 1977) to exclude further possible contaminants. In the condition with fast contaminants, this led to a total exclusion of 5.5 % of the trials on average. The average percentage of trials correctly identified as fast contaminants was 98.4 %. However, also 4.7 % of the trials were falsely identified as slow contaminants. In the condition with slow contaminants, 7.1 % of the slow trials were excluded, but only 56.3 % of these were “true” slow contaminants (the percentage of falsely identified fast contaminants was very small).

To see whether the exclusion of trials led to an improvement in parameter recovery, we reestimated the parameters for the adjusted data sets. This procedure led to basically the same results as when all trials were used for parameter estimation. For almost all cases in the condition with fast contaminants, the numbers of trials required were equal or higher than the values with the full data set. For data sets with slow contaminants, an improvement was observed for some conditions (mostly in the six- and seven-parameter models), but a deterioration or equal performance in most conditions. In sum, no systematic overall improvement pattern could be identified from the exclusion of trials according to the standard procedure of identifying outliers.

Another option for dealing with possible contaminant trials would be including in the model a further parameter to explicitly estimate the percentage of contaminant trials. To exemplify the effect of this additional parameter, we implemented the approach proposed by Ratcliff and Tuerlinckx (2002) to the ML criterion. The requisite trial numbers resulting from the inclusion of this further parameter were compared to the trial numbers shown in Tables 4 and 5. For the data sets with slow contaminants, we observed improvements for some conditions (almost all of them in the six- and seven-parameter models), but deteriorations in other conditions (in the three- and four-parameter models). For the conditions with fast contaminants, the criteria for low and high precision were—as for the data without adjustments—mostly not reached.[16] In total, the inclusion of this further parameter, at least for the range of trial numbers analyzed in our study, did not have a clear positive effect. The positive effect possibly resulting from the estimation of the proportion of contaminants might have been undermined by the negative effect of adding a further parameter calling for estimation.

Criterion 5: Computation time

Our final evaluation criterion was the computation time needed for parameter estimation, averaged across individual data sets. In the three- and four-parameter models, no relevant time difference was apparent between the three optimization criteria KS, ML, and CS. On average, parameter estimation took less than 5 s per individual data set for all methods. Only after inclusion of the intertrial variabilities did the three optimization criteria differ substantially in terms of computation time, with ML for large trial numbers taking considerably longer than KS and CS. Even then, however, the computation process took no longer than 30 min per data set in the one-drift design, and 40 min in the two-drift design. Accordingly, as long as the traditional methods are used, computation time will probably not affect a researcher's choice of optimization criterion. The HDDM approach, however, involved longer computation times, requiring anything up to 5 h per data set. As EZ is based on closed-form equations, the computation time was negligibly small.

[16] HDDM also permits estimation of the proportion of contaminants, using an approach similar to the one by Ratcliff and Tuerlinckx (2002). We applied this approach to our data and found results very similar to those observed for the non-Bayesian ML approach. Most importantly, the inclusion of the additional parameter did not have a consistent positive effect on parameter estimation.

Discussion

As demonstrated by the increasing number of research articles applying Ratcliff's diffusion model (Ratcliff, 1978), the interest in diffusion modeling is growing in various fields of psychology (Voss, Nagler, et al., 2013). This development can be attributed to a recognition of the main benefit of the diffusion model, that is, its capacity to disentangle several latent cognitive processes. The recent increase in popularity of the diffusion model is further fostered by the availability of user-friendly software solutions. Due to these programs the growing interest in diffusion modeling is not hampered by any lack of mathematical or programming skills (Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2007; Wagenmakers et al., 2007). However, knowledge is still scarce about the preconditions of diffusion modeling. In any diffusion model study, the probably most important issue that a researcher has to examine is validity of parameters. In recent years, several experimental validation studies (e.g., Arnold, Bröder, & Bayen, 2015; Voss et al., 2004; Wagenmakers, Ratcliff, et al., 2008) and correlational analyses (e.g., Ratcliff, Thapar, & McKoon, 2011; Schubert, Hagemann, Voss, Schankin, & Bergmann, 2015) have supplied promising results regarding parameter validity. However, for any new paradigm, the validity has to be first examined.

A second prerequisite for diffusion modeling is robustness of parameter estimation. One important question here regards the amount of data that are required. Typically, very large numbers of trials (> 1,000) have been used in diffusion model analyses (e.g., Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, et al., 2008). The present article aimed at clarifying whether this convention could be corroborated by data. Only very few studies have systematically analyzed the effects of different numbers of trials on the precision of parameter estimation (e.g., Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009). To fill this gap, we ran a set of simulation studies using different numbers of trials with the aim of deducing guidelines for the necessary trial numbers. In these studies, the precision of parameter estimation was compared for models differing with regard to the number of parameters while using different optimization criteria. In particular, we analyzed parameter recovery for three-parameter (a, v, t_0), four-parameter (a, v, t_0, z_r), six-parameter (a, v, t_0, s_v, s_{t0}, s_{zr}), and seven-parameter (a, v, t_0, z_r, s_v, s_{t0}, s_{zr}) models, with either one drift rate or two different drift rates. Data sets were simulated either without contaminated trials or with 4 % of slow or fast contaminants. Then, parameters were reestimated using the KS, ML, and CS methods, as well as a Bayesian approach (HDDM; Wiecki et al., 2013). Besides, the EZ-diffusion model (Wagenmakers et al., 2007) was applied to the data of the three-parameter model.

Parameter estimation performance was evaluated using different criteria. (1) First, we analyzed correlations between true and reestimated parameters, which is of relevance for researchers interested in relationships between diffusion model parameters and external criteria. (2) Second, biases (i.e., deviations between true and reestimated parameters) were examined. We were also interested in the influence of the true value of the parameter on the bias, as a positive (negative) relationship can lead to overestimation (underestimation) of a difference between conditions. (3) Third, for the design with two drift rates, we additionally performed power analyses to elicit indications on the number of participants required for the detection of a drift rate difference. (4) The precision of estimation was our fourth criterion. Recently, the idea of using diffusion model parameters for individual diagnostics has been introduced (Aschenbrenner et al., 2016; Ratcliff & Childers, 2015). Certainly, with such an aim it is crucial that parameters be estimated precisely for each person. As a measure of precision, we computed squared deviations of the recovered parameter values from the true values. Thereby—in contrast to the bias measure—over- and underestimations would not cancel each other out. In addition, each parameter's squared deviation was standardized, thereby taking into account the different scales of the different parameters. As a standard value for each parameter, best-possible accuracy was used, which was defined from an optimal condition of parameter recovery (5,000 trials, at least 4 % of trials at each threshold, no contaminants, using ML for parameter recovery). On the basis of this measure of parameter recovery, we propose guidelines for how many trials are required for low or high precision in parameter recovery.

Criterion 1: Correlations between true and reestimated parameter values

Regarding the correlations between true and reestimated parameters, all methods turned in a satisfying performance, with the exception of CS performing worse in small samples.

Criterion 2: Parameter estimation biases

In terms of biases, it is noteworthy that biases sometimes decrease with the number of trials. In contrast, for the three- and four-parameter models with fast contaminants, ML and HDDM showed increasing overestimation of the threshold separation and an increasing underestimation of nondecision times. This pattern was not observed for the more complex models. We suppose that the intertrial variability of the nondecision time (present in both the six- and seven-parameter models) helped to capture the negative effects of fast contaminants.

Note that the decreasing and increasing biases are in contradiction with the hypothesis that only the standard deviation, but not the bias, changes with trial numbers (van Ravenzwaaij & Oberauer, 2009). Importantly, the biases get stable at around 200 to 500 trials. Thus, a further increase in trial numbers does not have a notable influence on the size of the bias.

The trial numbers also sometimes had an influence on the relationship between the true parameter value and the bias. For example, for data with slow contaminants, the relationship between the true value of the threshold separation and the bias increased notably with the number of trials (e.g., from r = -.07 for n = 24 up to r = .89 for n = 5,000 for ML estimation in the three-parameter model). While positive relationships between true parameter values and bias lead to an overestimation of the true effect, negative relationships make it more difficult to detect a true difference in parameters. The starting point reveals a consistent pattern of such negative relationships. Thus, the detection of a significant difference in z_r between conditions would be impeded. For the nondecision time, on the other hand, there were no relationships between the size of the true value and the bias.

Criterion 3: Number of participants required for detection of a drift rate difference

The most important finding in terms of our power analyses is that an increase in trial numbers beyond 500 trials does not lead to essential further reductions in the requisite number of participants. Interestingly, EZ-based model fits proved to have a high power to detect differences between drift rates. For small trial numbers, EZ outperformed KS, CS, and HDDM. Only ML performed better than EZ.

Criterion 4: Estimation precision

On the basis of the squared deviations between the true and reestimated values, we defined criteria for requisite trial numbers. The results reveal that in the absence of contaminants, parameters can be accurately recovered even with small trials numbers. Analyses for the separate parameters showed that the required trial number was lowest for nondecision times, whereas a precise estimation of the threshold separation required especially high trial numbers. For the condition with no contaminants, HDDM usually led to the most precise estimates, followed by ML and KS. CS showed the worst results. Again, for the three-parameter model EZ could recover especially the drift rates very precisely. In the three- and six-parameter models, due to the fixed starting point, parameters were estimated also for data sets with fewer than 4 % of trials at one of the two thresholds. However, in this case more trials were required to achieve the same precision.

We now turn to the question of precision of parameter estimation in the presence of contaminants. When contaminants are slow, both ML and HDDM still provide better results than the other criteria. With fast contaminants, however, KS outperforms the other criteria in almost all conditions. In particular, ML and HDDM are generally affected strongly by fast contaminants. Even our criterion for low precision was in many conditions never reached—that is, even very high trial numbers could not compensate for the presence of fast contaminants. Interestingly, similar to KS, EZ was barely affected by fast contaminants.

We also investigated up to which point additional trials appreciably increase the accuracy of the results. As the slope of the relationship of trial number on precision decreases, increasing the trial number becomes less and less advantageous. Therefore, exceeding a certain number of trials is of limited utility, because the costs will probably be greater than the benefits. For example, it is plausible for the number and percentage of contaminants to increase when participants get tired or bored in long experimental sessions. Splitting sessions over a number of days may also cause problems, since performance may vary from one day to another depending on fatigue, motivation, mood, and so forth. A slope of -0.01 was used to define the point at which more trials did not increase precision notably. Most importantly, the results revealed that it is usually not advisable to increase the number of trials to many hundreds or even thousands, as this improves parameter recovery only marginally.

Number of parameters

The results of our study also provide some insights into the role of additional free parameters. From the three- to the seven-parameter models, there was mostly an increase in the trial numbers required. These results are in line with our finding that the inclusion of a parameter modeling the proportion of contaminants did not lead to any consistent improvement in parameter recovery. In the comparison of the design with one drift rate to the design with two drift rates the additional parameter had a negative effect on threshold separation and drift rate. However, nondecision time was estimated very similarly in both designs, and the starting point was estimated even better in the two-drift design.

One topic urgently calling for further exploration is the poor estimation of the intertrial variabilities s_{zr} and s_v. Even in the condition with 5,000 trials, parameter estimates of s_{zr} and s_v displayed correlations with true values lower than .50 (for similar results, see Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009; Vandekerckhove & Tuerlinckx, 2007). Typically, these parameters are included in the model to explain fast (s_{zr}) or slow (s_v) error RTs. One study in which the role of the intertrial variabilities has been explicitly tackled was conducted by Lerche and Voss (2016a). They examined the question of whether fixing these parameters at zero might result in better overall estimation of the remaining parameters, even if there is moderate variability in the true parameter values. To this end, they compared differently complex parameter models analyzing both simulated and empirical data sets. The results showed that the seven-parameter model often provides poorer results than less complex models. In line with these findings is a study by van Ravenzwaaij, Donkin, and Vandekerckhove (in press), who compared the power to detect parameter differences between EZ (Wagenmakers et al., 2007) and a full diffusion model estimation (i.e., inclusive of all three intertrial variabilities). Although the data-generating model included intertrial variabilities, the EZ model (ignoring these variabilities) led to better power than the more complex model for the detection of differences in drift rate and threshold separation. Note that in our analyses, EZ also proved to be very good at estimating drift rates.

Choice of estimation procedure

It is important to note that our results cannot provide one clear-cut answer to the questions of which estimation method should be used and how many trials are required. Several aspects (e.g., type of contamination, presence of intertrial variabilities) have an influence on which method will produce the most reliable results. In the following, we shortly sketch some guidelines that can help researchers to make qualified decisions for their analyses.

Firstly, researchers have to think about an appropriate experimental paradigm to analyze their research question. Several experimental paradigms have already been analyzed in terms of validity (experimentally or by means of correlations with external criteria). Completely new paradigms should first be validated before applying them to analyze new research questions. Note that in our study we only analyzed two rather simple experimental designs (one-drift and two-drift designs). We suppose, however, that the main patterns of results will remain similar (e.g., best performance of HDDM/ML for uncontaminated data and of KS in the presence of fast contaminants).

Second, the number of trials of an experiment has to be defined. This question will often be related to the chosen paradigm. Especially, if material is restricted, it might be difficult to compose high trial numbers. The homogeneity of the material also influences the decision process, with more heterogeneous material resulting in higher intertrial variability of the drift. Besides, the researcher has to consider the fatigue that the type of task might cause. For tasks that are very demanding and that take very long, a higher percentage of contaminants is to be expected.

Third, after collecting the data, the researcher should analyze the data quality before applying a diffusion model. This means, for example, figuring out whether there are supposedly many contaminants. If, for example, the RT distributions include many statistical outliers according to typical outlier detection procedures (e.g., Tukey, 1977), this might indicate a high level of contaminants. Note that the exclusion of outliers does not necessarily lead to better parameter estimates, as our additional analyses showed. The problem is that not all contaminants will be detected (especially not if they are situated overlapping with the true RT distribution), and "real" RTs might be accidentally removed from the distribution (false positives). Thus, an estimation method that is robust to contaminants (like KS) is in such cases more adequate than an overly strict data cleaning. Besides, estimation of the intertrial variability of the nondecision time (but not of the rather poorly estimated other two variability parameters) can help to counteract the influence of fast contaminants.

Furthermore, one can analyze whether there might be a response bias for one of the two stimuli. If, for example, correct responses to stimulus A are faster than errors, whereas for stimulus B the errors are faster than the correct responses, the starting point might be positioned closer to stimulus A than to B. In such a case, the researcher should not collapse over the two stimuli by estimation of a model with correct and error responses at the two thresholds. Rather, he or she should use a model with the two different stimuli at the thresholds and freely estimate the starting point. Besides, an analysis of the mean RTs of correct and error responses can give a hint as to whether there might be high intertrial variability in the data. Finally, on the basis of these analyses, the researcher can decide which parameters to estimate and which estimation method to use.

Thus, one main message of this article is that there is no single type of diffusion model analysis. Several aspects influence the parameter estimation, and thus, the estimation procedure has to be carefully selected. Our work is intended as a first step in the development of general guidelines for diffusion modeling.

Conclusions

Whereas several hundred or even several thousand trials are often used in the application of the Ratcliff diffusion model (Ratcliff, 1978), our simulation studies—executed with the newest version of *fast-dm* (Voss et al., 2015)—indicate that in most cases considerably lower trial numbers are sufficient. Using a lot more than the necessary number of trials can also be more detrimental than useful. It leads to higher costs (e.g., longer preparation and execution time of the experiment, or fatigue of the participants) without clearly improving parameter estimation performance. In this article, we give guidelines for the number of trials required, depending on the optimization criterion applied, the number of parameters estimated, and the presence of contaminants.

Our simulations provide the following stable patterns of results: (1) CS is generally not advisable for small to moderate trial numbers; (2) parameter recovery often does not improve much if more than around 500 trials are used; (3) for less complex models (i.e., exclusive of intertrial variabilities), notably smaller trial numbers are sufficient; (4) ML and HDDM perform best for uncontaminated data; and (5) KS and EZ are the methods least affected by fast contaminants.

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*(5), 882-898. doi: 10.1007/s00426-014-0608-y

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology, 30*(2), 225-238. doi: 10.1037/neu0000222

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition, 40*(1), 70-82. doi: 10.3758/s13421-011-0128-6

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

Champely, S. (2012). pwr: Basic functions for power analysis (R package version 1.1.1). Available at http://CRAN.R-project.org/package=pwr.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16*(6), 1129-1135.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: a synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117*(2), 309.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience, 31*(47), 17242-17249. doi: 10.1523/JNEUROSCI.0309-11.2011

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin, 40*(2), 217-231.

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53*(2), 55-68. doi: 10.1016/j.jmp.2009.01.006

Heathcote, A., & Brown, S. (2004). Reply to Speckman and Rouder: A theoretical basis for QML. *Psychonomic Bulletin & Review, 11*(3), 577-578.

Horn, S. S., Bayen, U. J., & Smith, R. E. (2011). What can the diffusion model tell us about prospective memory? *Canadian Journal of Experimental Psychology, 65*(1), 69-75. doi: 10.1037/a0022808

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*(3), 353-368. doi: 10.1037/0022-3514.93.3.353

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

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*(7), 651-687.

Lerche, V., & Voss, A. (2016a). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Manuscript submitted for publication.*

Lerche, V., & Voss, A. (2016b). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24. doi: 10.1007/s00426-016-0770-5

McKoon, G., & Ratcliff, R. (2013). Aging and predicting inferences: A diffusion model analysis. *Journal of Memory and Language, 68*(3), 240-254. doi: 10.1016/j.jml.2012.11.002

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27*(2), 193-200. doi: 10.1037/a0031533

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion, 13*(4), 739-747. doi: 10.1037/a0031628

R Development Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org/.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

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

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*(4), 237-279. doi: 10.1037/dec0000030

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

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

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278. doi: 10.1037/0882-7974.19.2.278

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*(4), 408-424. doi: 10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*(3), 464-487. doi: 10.1037/a0023810

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481. doi: 10.3758/bf03196302

Ratcliff, R., & Van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review, 16*(4), 742-751. doi: 10.3758/PBR.16.4.742

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. doi: 10.1037/a0026003

Schmitz, F., & Voss, A. (2014). Components of task switching: A closer look at task switching and cue switching. *Acta Psychologica, 151*, 184-196. doi: 10.1016/j.actpsy.2014.06.009

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence, 51*, 28-46. doi: 10.1016/j.intell.2015.05.002

Spaniol, J., Madden, D. J., & Voss, A. (2006). A Diffusion Model Analysis of Adult Age Differences in Episodic and Semantic Long-Term Memory Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 101-117. doi: 10.1037/0278-7393.32.1.101

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin & Review, 11*(3), 574-576.

Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics, 12*(1), 16-32. doi: 10.3758/bf03212837

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (in press). The EZ Diffusion Model Provides a Powerful Test of Simple Empirical Effects. *Psychonomic Bulletin & Review*.

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*(6), 463-473. doi: 10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44*(4), 1048-1056. doi: 10.1016/j.jesp.2007.10.009

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, 142*(2), 536-559. doi: 10.1037/a0029459

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 539-555. doi: 10.1348/000711009x477581

Voss, A., Voss, J., & Lerche, V. (2015). Assessing Cognitive Processes with Diffusion Model Analyses: A Tutorial based on fast-dm-30. *Frontiers in Psychology, 6*, 336. doi: 10.3389/fpsyg.2015.00336

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46*(1), 15-28. doi: 10.3758/s13428-013-0369-3

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*(1), 140-159. doi: 10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15*(6), 1229-1235. doi: 10.3758/pbr.15.6.1229

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*(1), 3-22. doi: 10.3758/bf03194023

White, C. N., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition and Emotion, 23*(1), 181-205. doi: 10.1080/02699930801976770

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion, 10*(5), 662-677. doi: 10.1037/a0019474

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*(1), 39-52. doi: 10.1016/j.jmp.2010.01.004

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics, 7*, 14. doi: 10.3389/fninf.2013.00014

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

Yap, M. J., Balota, D. A., & Tan, S. E. (2013). Additive and interactive effects in semantic priming: Isolating lexical and decision processes in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 140-158. doi: 10.1037/a0028520

Table 1
Comparison of optimization criteria

	Chi-Square	Maximum Likelihood	Kolmogorov-Smirnov
Term to be minimized in the optimization process	$\sum \frac{(o_i - p_i)^2}{p_i}$ Note: o_i /p_i correspond to the numbers of responses observed/predicted in bin *i*	$-\sum ln\left(d\left(RT_{i,}k_i\right)\right)$ Note: $d(RT_{i,} k_i)$ corresponds to the density value of the RT observed in trial *i* with response k_i	$\max_{i=1\ldots n} \left\lvert \begin{matrix} \text{eCDF}(RT_i)- \\ \text{pCDF}(RT_i) \end{matrix} \right\rvert$ Note: *n* is the number of responses observed; eCDF/pCDF are the empirical/predicted cumulative distribution functions; RT_i is the RT in trial *i*
Information utilization	low	high	medium
Computation time	low	high	medium

Table 2

Minimum and maximum values of each diffusion model parameter used for the creation of parameter sets, and "possible accuracy" of each parameter

Parameter	Minimum	Maximum	Possible Accuracy[a]
a	0.5	2.0	0.054
v	-4.0	4.0	0.270
t_0	0.2	0.5	0.032
z_r	0.3	0.7	0.035
s_v	0.0	1.0	0.849
s_{t0}	0.0	0.2	0.031
s_{zr}	0.0	0.5	0.402

Note. The diffusion coefficient in fast-dm is set to 1. To compare parameter ranges and accuracies with parameter values cited in studies using a coefficient of .1, the parameters a, v, z_r, s_v, and s_{zr} need to be multiplied by .1.

[a] 95 % quantile of absolute deviations of true values and reestimated values using the ML criterion for uncontaminated simulated data sets with 5,000 trials and at least 4 % of trials at each threshold.

Table 3

Juxtaposition of the four evaluation criteria of parameter estimation performance

Evaluation Criterion	Aim of Researcher
Correlations between true and reestimated parameter values	Detection of relationships between diffusion model parameters and external criteria (e.g., between drift rate and intelligence)
Parameter estimation biases (i.e., deviations of reestimated from true parameter values)	Detection of parameter differences between conditions; interpretation of effect sizes (over- or underestimation of true effect?)
Number of participants required for detection of drift rate difference	Sample size computation for detection of parameter differences between conditions
Estimation precision—squared deviations of reestimated from true parameter values	Diagnostic use of diffusion model parameters (e.g., drift rate for the measurement of intelligence)

Table 4

Number of trials required in the one-drift design for data sets with at least 4 % of trials at each threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Four-Parameter Model				Six-Parameter Model			Seven-Parameter Model			
		a	*v*	t_0	*a*	*v*	t_0	z_r	*a*	*v*	t_0	*a*	*v*	t_0	z_r
No contaminants	KS	125; 403 436	37; 119 241	26; 83 203	199; 589 546	94; > 5,000 500	< 24; 78 195	163; > 5,000 500	223; 1,122 551	102; 479 402	< 24; 91 199	650; 3,616 824	195; > 5,000 500	< 24; 310 217	> 5,000 500
	ML	55; 158 283	26; 78 197	< 24; < 24 89	57; 158 286	42; 126 251	< 24; < 24 99	38; 109 235	272; 783 641	277; 727 649	56; 223 309	318; 1,012 687	300; 944 668	54; 335 313	132; 548 449
	CS	287; 498 628	105; 219 358	50; 104 229	259; 478 599	149; 352 457	56; 117 246	146; 348 452	759; 1,693 1,156	713; 1,609 1,112	94; 421 388	592; 1,397 992	453; 1,276 833	124; 517 438	278; 816 647
	HDDM	48; 142 267	26; 79 198	< 24; < 24 86	53; 152 277	35; 112 235	< 24; < 24 86	26; 85 206	> 5,000 500	94; > 5,000 500	< 24; 35 134	378; 2,788 626	288; 4,488 486	< 24; 51 150	41; 290 284
	EZ	272; 836 638	68; 201 318	109; 354 408											
Slow contaminants	KS	555; > 5,000 644	28; 128 234	38; 147 200	1,031;>5,000 997	81; > 5,000 200	< 24; 106 188	> 5,000 500	> 5,000 968	94; 605 385	28; 253 100	> 5,000 1,411	> 5,000 200	39; 195 200	> 5,000 1,000
	ML	116; 1,945 372	< 24; 116 201	< 24; < 24 74	135; 1,963 395	39; 508 278	< 24; < 24 75	41; 224 281	> 5,000 1,987	1,493;>5,000 1,047	30; 257 256	> 5,000 1,925	> 5,000 1,019	< 24; 350 215	218; 2,801 465
	CS	192; > 5,000 500	83; 292 362	47; 91 200	719; 1,870 1,081	186; 722 525	57; 96 200	211; 564 562	> 5,000 4,593	1,329; 4,032 1,405	84; 1,310 342	> 5,000 3,202	791; 3,355 973	146; 1,074 446	375; 1,150 748
	HDDM	143; 2,230 398	< 24; 121 204	< 24; < 24 86	146; 1,881 413	24; 427 238	< 24; < 24 78	< 24; 133 211	> 5,000 1,000	68; > 5,000 200	< 24; 30 100	> 5,000 500	> 5,000 200	< 24; 40 100	29; 1,475 234
	EZ	> 5,000 1,208	69; 475 343	1,246;>5,000 587											
Fast contaminants	KS	109; 402 412	26; 133 231	29; 92 200	144; 540 467	89; > 5,000 500	< 24; 99 200	> 5,000 500	232; 1,407 544	69; 302 340	< 24; 110 205	199; > 5,000 500	245; > 5,000 500	29; 149 100	> 5,000 500
	ML	> 5,000 48	45; > 5,000 100	> 5,000 100	> 5,000 100	> 5,000 500	> 5,000 200	> 5,000 100	578; 2,659 831	287; 1,155 634	28; 341 252	> 5,000 1,841	> 5,000 1,814	< 24; > 5,000 100	> 5,000 200
	CS	418; > 5,000 500	146; > 5,000 200	694; > 5,000 100	217; > 5,000 500	100; > 5,000 200	87; > 5,000 100	> 5,000 500	> 5,000 1,000	> 5,000 500	132; 3,040 367	> 5,000 1,000	> 5,000 1,000	97; > 5,000 200	> 5,000 100
	HDDM	> 5,000 24	47; > 5,000 100	> 5,000 100	> 5,000 100	> 5,000 500	> 5,000 200	> 5,000 24	177; > 5,000 500	57; > 5,000 200	< 24; 38 100	> 5,000 500	> 5,000 200	< 24; > 5,000 100	> 5,000 24
	EZ	257; 923 612	46; 269 294	296; > 5,000 409											

Note. The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row).

Table 5

Number of trials required in the one-drift design for data sets with fewer than 4 % of trials at one threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Six-Parameter Model		
		a	*v*	t_0	*a*	*v*	t_0
No contaminants	KS	> 5,000 2,086	101; 473 400	35; 147 253	3,813;>5,000 1,316	213; 1,918 493	< 24; 94 202
	ML	463; 1,003 870	74; 183 314	< 24; 47 148	1,584; 3,491 1,777	760; 1,516 1,182	41; 126 250
	CS	1,703; 3,492 1,916	206; 502 550	102; 221 358	4,988;>5,000 > 5,000	4,949;>5,000 > 5,000	36; 163 259
	HDDM	439; 1,275 816	48; 142 267	< 24; 25 112	> 5,000 1,000	> 5,000 500	< 24; 32 130
	EZ	> 5,000	4,087;>5,000 500	< 24; 4,544 1,000			
Slow contaminants	KS	> 5,000 500	98; > 5,000 200	< 24; 102 192	> 5,000 1,000	316; > 5,000 492	33; 76 100
	ML	603; 2,965 830	44; 420 100	< 24; < 24 91	4,821;>5,000 2,397	505; 1,114 914	< 24; 142 214
	CS	> 5,000 500	298; > 5,000 500	87; 100 200	> 5,000	4,938; 4,997 > 5,000	< 24; 175 229
	HDDM	568; 2,928 799	32; 356 100	< 24; < 24 91	> 5,000 1,000	> 5,000 500	< 24; < 24 77
	EZ	> 5,000 1,000	4,295;>5,000 500	< 24; > 5,000 500			
Fast contaminants	KS	> 5,000 500	124; > 5,000 200	< 24; 175 200	> 5,000 500	180; 1,547 469	< 24; 113 182
	ML	> 5,000 1,000	100; 3,465 200	> 5,000 200	> 5,000	> 5,000	71; 99 200
	CS	3,417;>5,000 2,653	> 5,000 200	> 5,000 200	> 5,000 1,000	> 5,000 1,000	46; 394 100
	HDDM	> 5,000 4,762	80; 1,840 322	> 5,000 283	> 5,000 1,000	> 5,000 200	< 24; 28 98
	EZ	> 5,000 1,000	498; > 5,000 1,000	< 24; > 5,000 500			

Note. The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row).

Table 6

Number of trials required in the two-drift design for data sets with at least 4 % of trials at each threshold, depending on the parameter model, estimated parameter, type of contamination, and estimation method

		Three-Parameter Model			Four-Parameter Model				Six-Parameter Model			Seven-Parameter Model			
		a	*v*	t_0	*a*	*v*	t_0	z_r	*a*	*v*	t_0	*a*	*v*	t_0	z_r
No contaminants	KS	1,056; 3,680 1,189	121; 389 429	39; 122 245	1,010;>5,000 791	152; > 5,000 500	< 24; 52 157	95; > 5,000 200	1,234;>5,000 1,122	341; 1,240 695	41; 156 266	863; > 5,000 848	295; > 5,000 500	< 24; 103 210	175; > 5,000 500
	ML	201; 509 545	59; 161 289	< 24; < 24 100	171; 434 500	83; 226 345	< 24; < 24 93	29; 85 207	536; 1,348 928	498; 1,091 907	43; 135 258	412; 1,149 794	454; 1,043 856	32; 101 223	64; 175 301
	CS	816; 1,535 1,248	183; 398 505	72; 150 286	677; 1,255 1,109	226; 545 579	64; 129 260	53; 180 292	1,790; 3,443 2,034	1,504; 3,260 1,736	82; 280 358	894; 1,833 1,298	761; 1,582 1,174	45; 146 266	103; 289 388
	HDDM	155; 454 481	40; 124 248	< 24; < 24 86	143; 416 462	59; 182 299	< 24; < 24 82	< 24; 64 179	369; > 5,000 500	127; 479 500	< 24; 44 150	> 5,000 500	> 5,000 500	< 24; 34 134	28; 111 227
	EZ	295; 1,012 655	49; 160 278	28; 93 214											
Slow contaminants	KS	1,974;>5,000 1,318	122; 736 426	26; 178 240	> 5,000 200	133; > 5,000 200	< 24; 48 200	96; > 5,000 500	> 5,000 1,872	350; 1,646 671	32; 331 262	> 5,000 500	402; 3,964 596	28; 84 100	276; 2,854 522
	ML	261; 1,452 577	47; 279 200	< 24; < 24 80	227; 1,263 547	90; > 5,000 200	< 24; < 24 70	26; 120 226	> 5,000 2,214	1,533;>5,000 1,306	< 24; 139 223	> 5,000 1,896	1,277;>5,000 1,138	< 24; 91 195	50; 216 296
	CS	1,039; 2,596 1,334	245; 907 597	42; 137 258	798; 1,847 1,179	288; 1,065 642	50; 91 200	52; 188 292	> 5,000 4,357	2,143;>5,000 2,004	79; 396 362	4,358;>5,000 2,815	1,083; 2,473 1,407	31; 155 247	101; 333 394
	HDDM	255; 1,692 554	34; 570 264	< 24; < 24 85	231; 1,320 548	82; 2,301 317	< 24; < 24 78	< 24; 71 174	> 5,000 1,000	> 5,000 200	< 24; < 24 88	> 5,000 1,000	> 5,000 200	< 24; < 24 88	< 24; 135 216
	EZ	> 5,000 1,000	93; 2,595 327	117; > 5,000 195											
Fast contaminants	KS	1,876;>5,000 1,291	126; 667 437	30; 132 239	1,142;>5,000 780	153; > 5,000 500	< 24; 49 152	96; > 5,000 500	1,991;>5,000 1,110	275; 970 632	38; 163 264	1,393;>5,000 829	388; 3,496 602	< 24; 75 179	170; > 5,000 500
	ML	> 5,000 500	86; > 5,000 200	> 5,000 200	> 5,000 1,000	> 5,000 200	> 5,000 200	> 5,000 200	2,254;>5,000 1,839	1,148; 3,301 1,334	< 24; 264 229	1,309; 4,322 1,346	1,329;>5,000 1,256	33; 94 200	378; > 5,000 200
	CS	2,425;>5,000 2,163	> 5,000 200	> 5,000 200	2,045; 4,857 1,982	> 5,000 200	> 5,000 100	> 5,000 100	> 5,000 500	> 5,000 4,906	87; 1,574 340	> 5,000 200	> 5,000 500	40; 136 200	> 5,000 200
	HDDM	> 5,000 1,122	90; > 5,000 200	> 5,000 131	> 5,000 843	> 5,000 200	> 5,000 200	> 5,000 200	> 5,000 500	78; > 5,000 200	< 24; 44 100	294; > 5,000 500	193; > 5,000 200	< 24; 57 100	408; > 5,000 329
	EZ	> 5,000 500	48; 473 200	< 24; 102 194											

Note. The cells comprise the requisite trial numbers for low and high precision (first row) and the limit (i.e., the number of trials not worth exceeding, since performance then improves only marginally; second row).

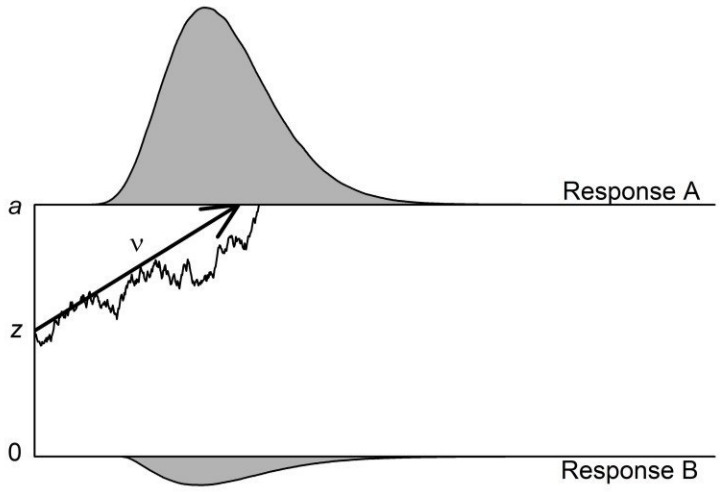


Figure 1. Example illustration of the decision process of the diffusion model. The process starts at *z* (here situated in the middle of threshold distance *a*) and moves with mean drift rate *v* until a threshold is hit (here the upper threshold). In the following, the motoric execution of the associated response (here Response A) is initiated.

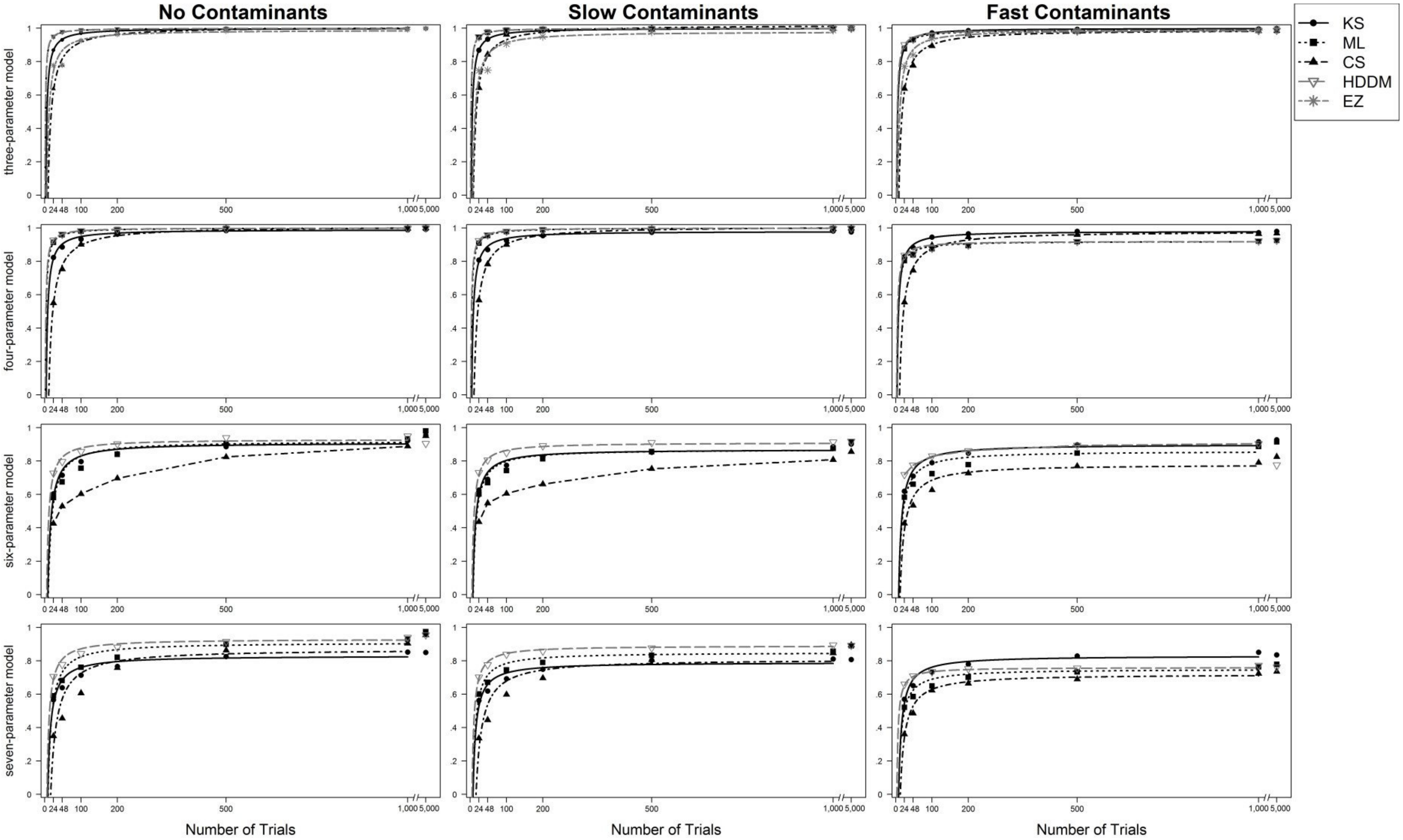


Figure 2. Scatter plot of mean correlation between true and reestimated parameters in the one-drift design, for uncontaminated data sets (left column), data sets with slow contaminants (middle column) and data sets with fast contaminants (right column). On the basis of data sets with at least 4 % of trials at each threshold. Power functions were fitted to the data. Whenever the curve was a poor fit (R^2 < .80), lines were drawn between adjacent trial numbers.

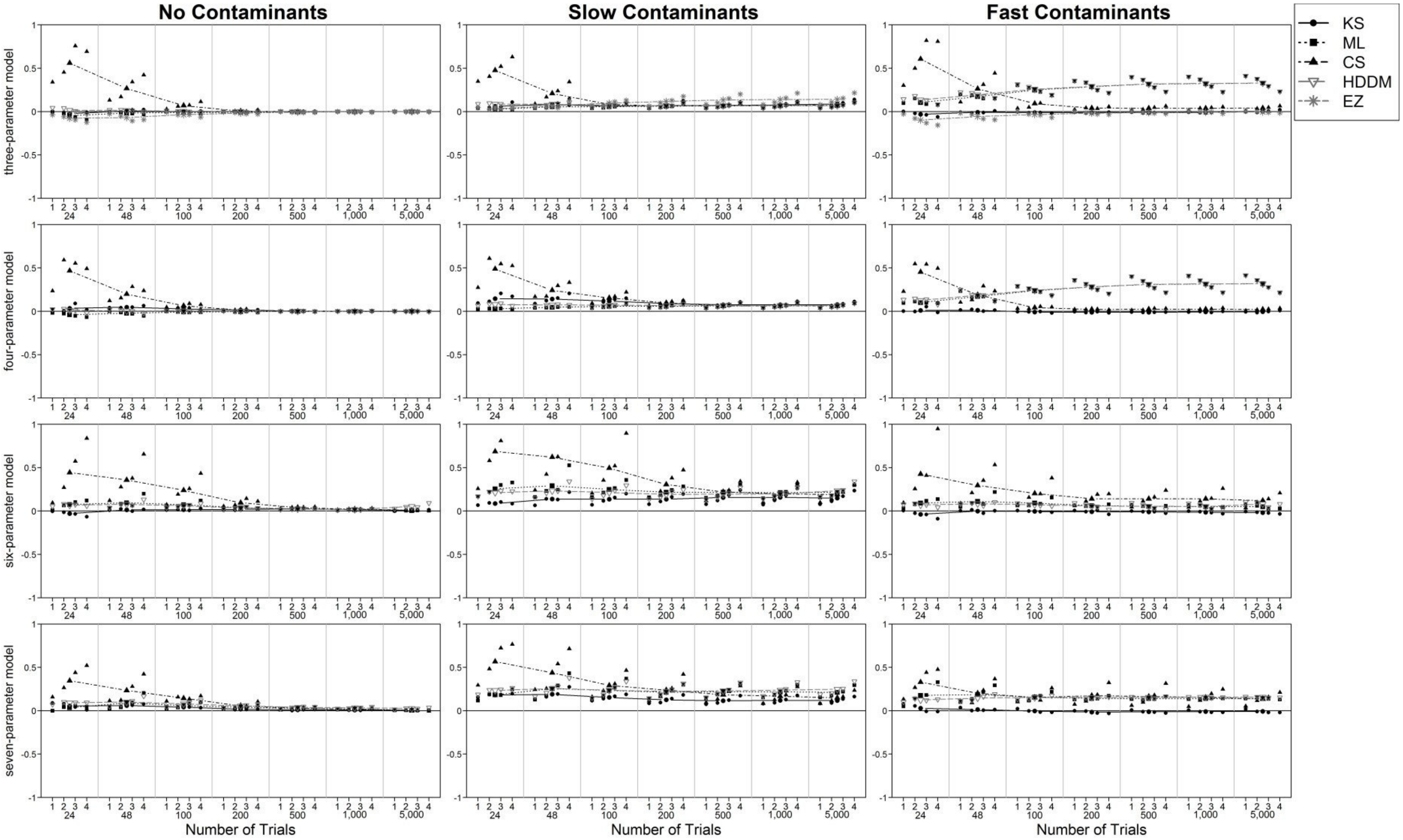


Figure 3. Mean differences between estimated and true values of *parameter a* for each quartile of the true parameter values (numbers 1-4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits.

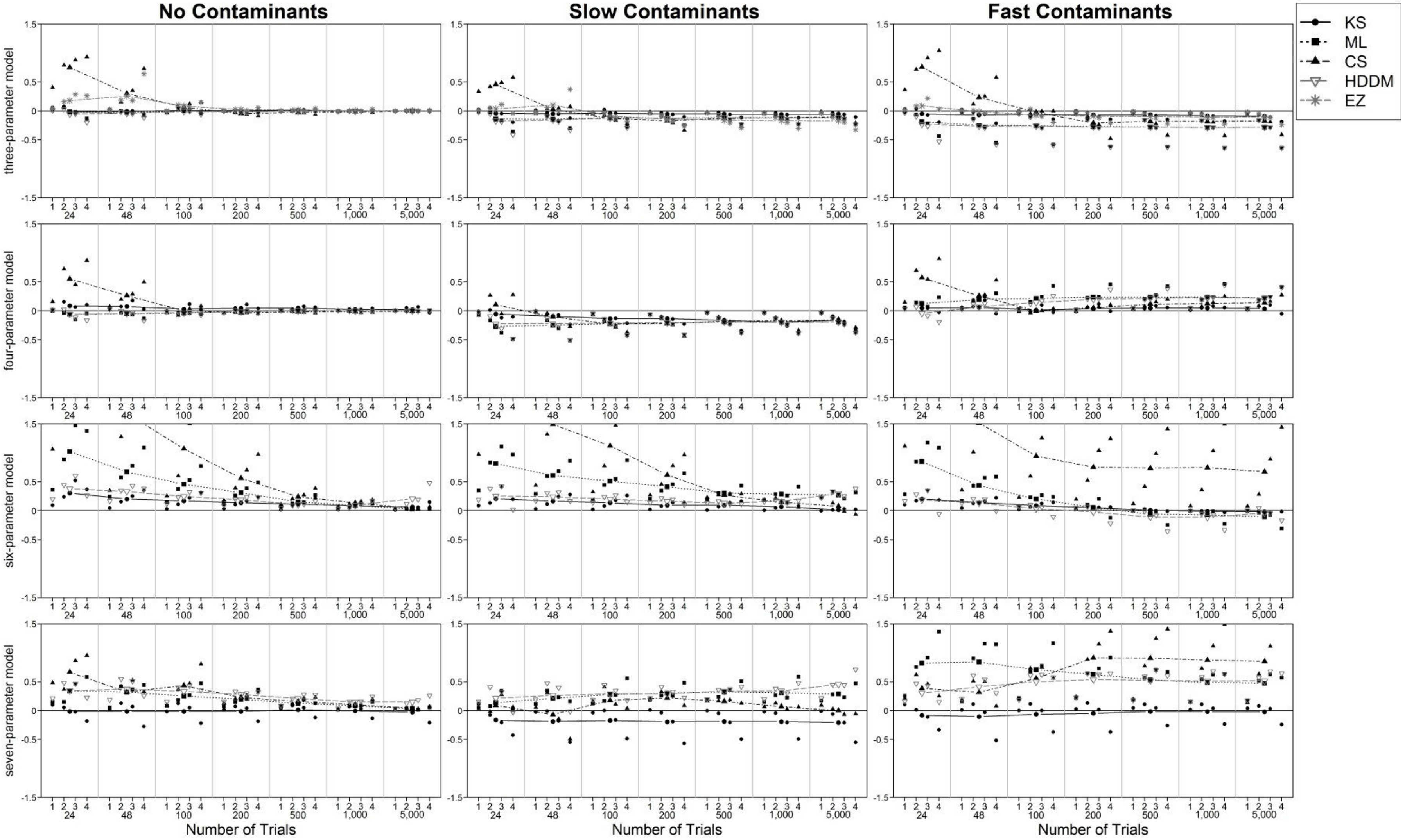


Figure 4. Mean differences between estimated and true values of *parameter v* for each quartile of the true parameter values (numbers 1-4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. All negative drift values were transformed to positive values so that the true values are all located between 0 and 4. On the basis of data sets with at least 4 % of trials at each threshold.

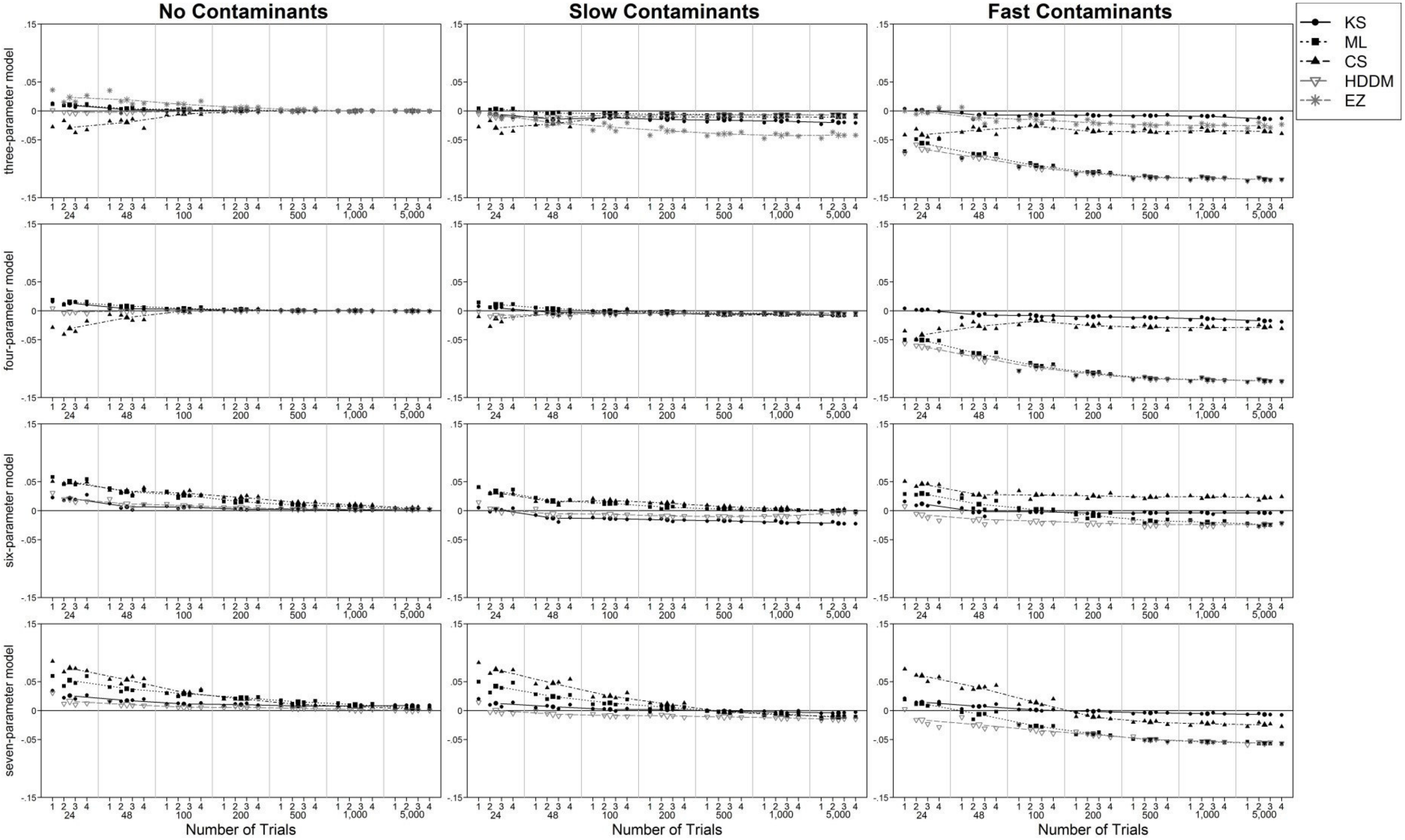


Figure 5. Mean differences between estimated and true values of *parameter t_0* for each quartile of the true parameter values (numbers 1-4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits.

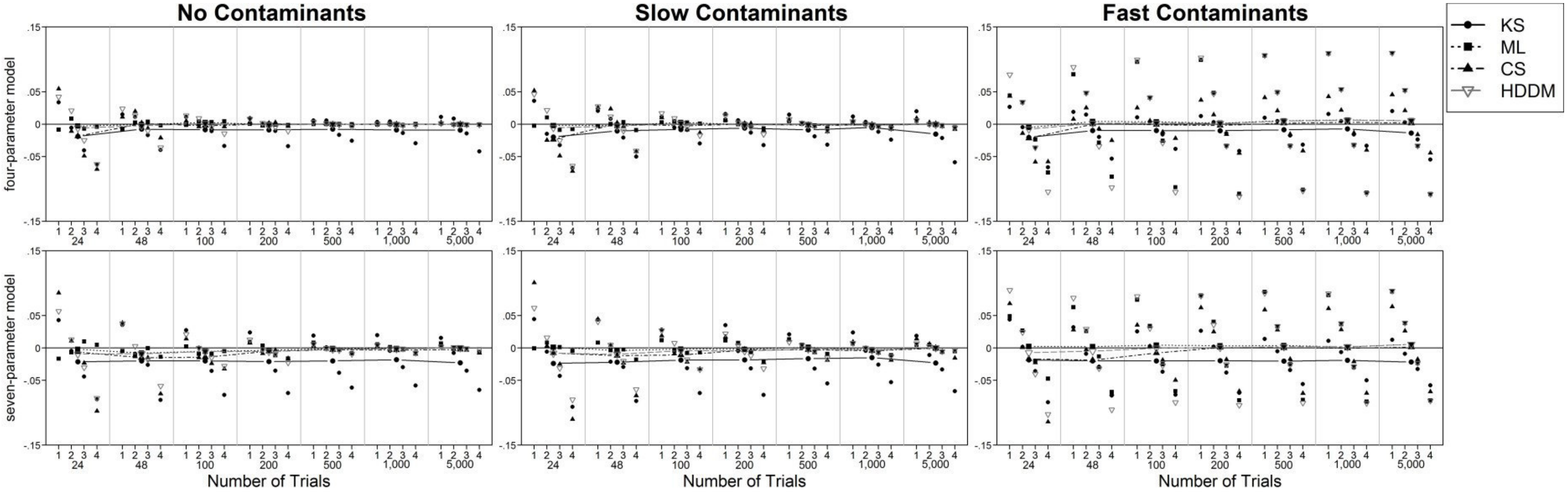


Figure 6. Mean differences between estimated and true values of *parameter* z_r for each quartile of the true parameter values (numbers 1-4; small symbols) and for all datasets (larger symbols connected by lines) depending on the contamination condition, parameter model, estimation method and number of trials. On the basis of data sets with at least 4 % of trials at each threshold. Few values are not depicted as they fall outside the *y*-axis limits.

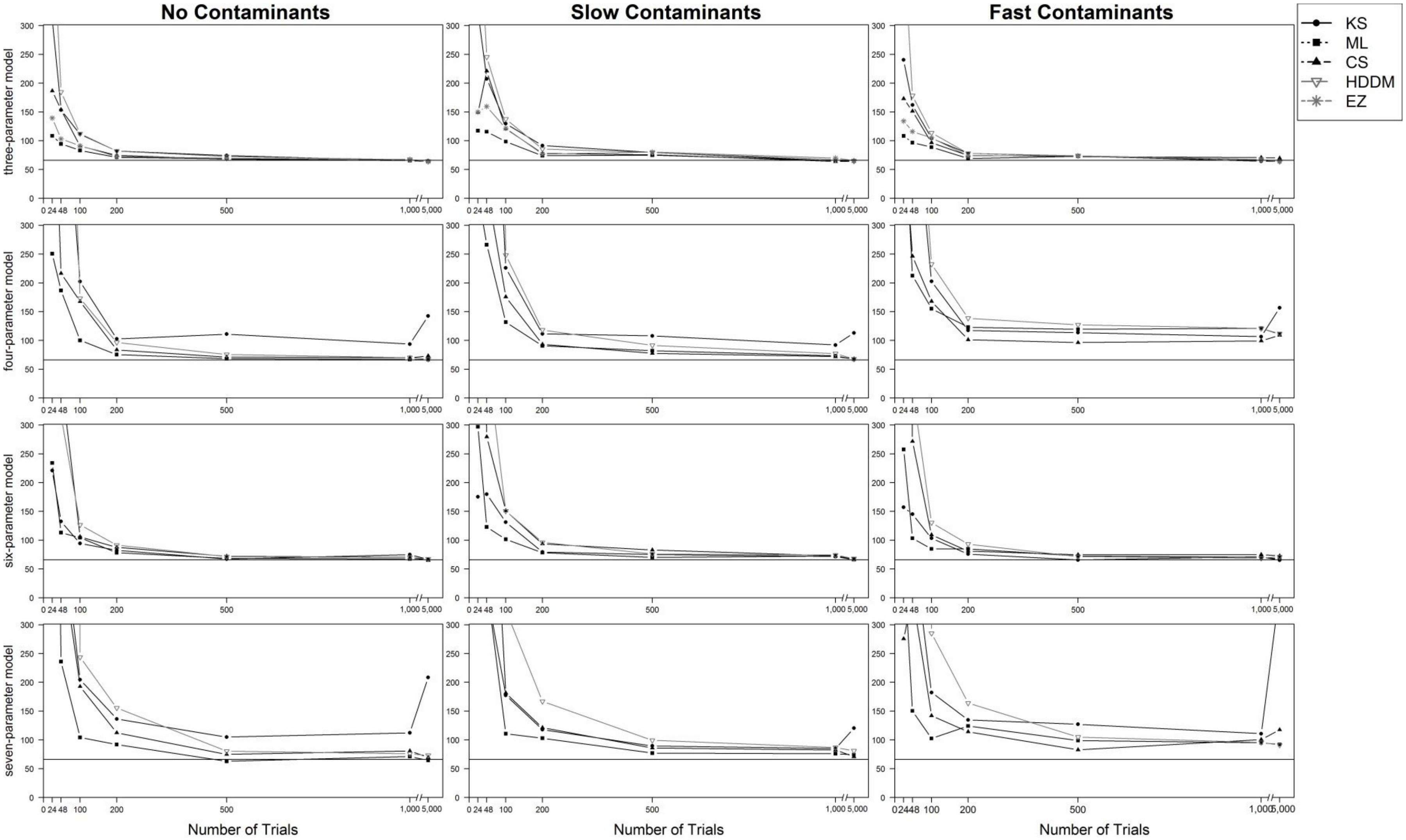


Figure 7. Scatterplot of the number of participants required for the detection of a difference in reestimated drift rates, depending on the number of trials and the estimation method. The horizontal line indicates the number of participants required for the original effect size (n = 66 for d_z = 0.35). On the basis of data sets with at least 4 % of trials at each threshold. Required numbers of participants exceeding 300 are not depicted.

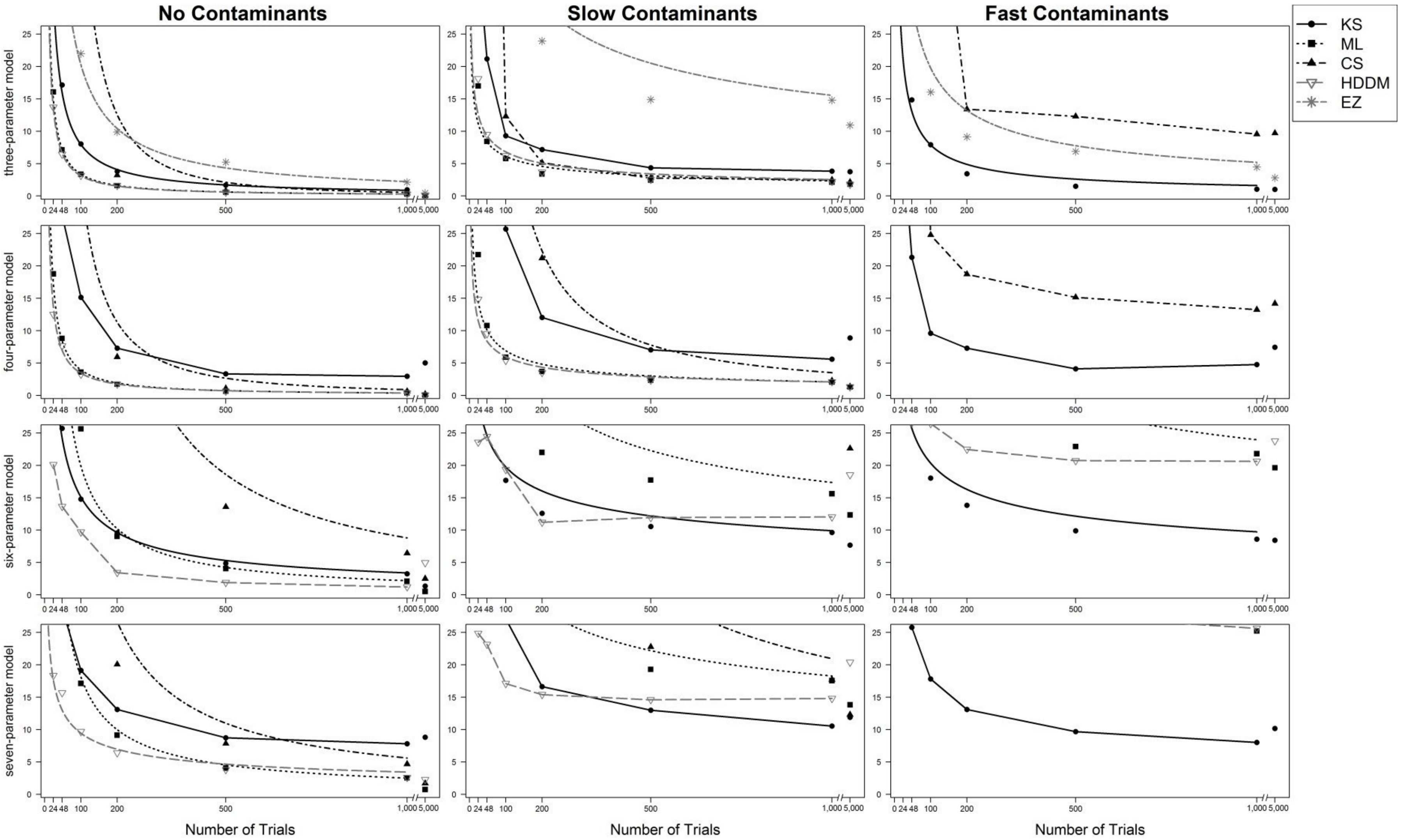


Figure 8. Scatterplot of 95 % quantiles of mean deviation between true and reestimated parameters in the one-drift design, for uncontaminated data sets (left column), data sets with slow contaminants (middle column) and data sets with fast contaminants (right column). On the basis of data sets with at least 4 % of trials at each threshold. Quantiles exceeding a mean deviation of 25 are not depicted. Power functions were fitted to the data. Whenever the curve was a poor fit ($R^2 < .80$), lines were drawn between adjacent trial numbers.

Appendix A 4

Manuscript 4: Lerche, V., & Voss, A. (2016). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24.

doi: 10.1007/s00426-016-0770-5

Retest reliability of the parameters of the Ratcliff diffusion model

Veronika Lerche and Andreas Voss
Ruprecht-Karls-Universität Heidelberg

Author Note

Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany.

This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant Nr. VO1288/2-1).

Correspondence concerning this article should be addressed to Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstrasse 47-51, D-69117 Heidelberg, Germany, email: veronika.lerche@psychologie.uni-heidelberg.de, telephone: +49-6221-54-7322.

ABSTRACT

In the recent years, there is a growing interest to use the Ratcliff Diffusion Model (1978) for diagnostic purposes as the parameters of the model capture interindividual differences in specific cognitive processes. The parameters are estimated using reaction time data from binary classification tasks. For a potential diagnostic application of parameter values sufficient reliability is a necessary precondition. In two studies, each with two sessions separated by one week, the retest reliability of the diffusion model parameters was assessed. In Study 1, 105 participants completed a lexical decision task and a recognition memory task. In Study 2, 128 participants worked on an associative priming task. Results show that the reliability of the main parameters of the Ratcliff Diffusion Model (in particular of the speed of information accumulation and the threshold separation with *r*s > .70 for all three tasks) is satisfying. Besides, we analyzed the influence of the number of trials on the retest reliability using different estimation methods (Kolmogorov-Smirnov, Maximum Likelihood, Chi-Square and EZ) and both empirical and simulated data sets.

Keywords: diffusion model, test-retest reliability, fast-dm, EZ, mathematical models, reaction time methods

Retest reliability of the parameters of the Ratcliff diffusion model

While so far the Ratcliff Diffusion Model (1978) has mostly been employed for the analysis of group differences, recently the aim of employing the model as a diagnostic tool has been expressed (Aschenbrenner, Balota, Gordon, Ratcliff, & Morris, 2015; Ratcliff & Childers, 2015). Regarding the validity for the measurement of specific cognitive functions, several approaches supply promising results. Experimental validation studies (e.g., Voss, Rothermund, & Voss, 2004) indicate that the parameters of the Diffusion Model have clear psychological interpretations. Other studies investigated criterion validity. For example, it has been shown that the drift parameter correlates with general intelligence (e.g., Ratcliff, Thapar, & McKoon, 2010; Ratcliff, Thapar, & McKoon, 2011; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007) and with neurophysiological correlates of information-processing speed (Schubert, Hagemann, Voss, Schankin, & Bergmann, 2015), and that threshold separation is related to response inhibition (Stahl et al., 2014). Besides, differences in parameter values between groups of individuals differing in variables like anxiety (White, Ratcliff, Vasey, & McKoon, 2010a, 2010b), ADHD (e.g., Metin et al., 2013; Weigard & Huang-Pollock, 2014) or depression (e.g., Pe, Vandekerckhove, & Kuppens, 2013; Vallesi, Canalaz, Balestrieri, & Brambilla, 2015) have been observed. This gives rise to the idea that parameters of the Ratcliff Diffusion Model might also be used as diagnostic variables, for example in the clinical field. While there are convincing arguments for the validity of the parameters of the diffusion model, the reliability, however, has rarely been investigated (for an exception, see Yap, Balota, Sibley, & Ratcliff, 2012). Prior to further proponing the use of the diffusion model as diagnostic tool, the reliability and stability of its parameters need to be ascertained.

In the following, we first give a brief introduction to the Ratcliff Diffusion Model including information on the estimation of the model parameters and on their validity (for a more expanded introduction, see for example Voss, Nagler, & Lerche, 2013). Then, we sum up previous findings regarding the retest reliability of the diffusion model parameters. This is followed by Study 1 in which we present retest reliability coefficients for a lexical decision task and a recognition memory task. In Study 2, retest reliability coefficients for an associative priming task are given. Finally, in Study 3, we analyze the influence of the number of trials on retest reliability using both empirical and simulated data sets.

Introduction to the diffusion model

The Diffusion Model (Ratcliff, 1978) is a mathematical model that allows for disentangling different cognitive processes involved in decision making. In particular, the model is applied to response time (RT) data from decision tasks with two possible responses. An example of such a binary decision task is a lexical decision task (LDT) in which participants have to decide whether a presented letter string is a word (response 1) or a non-word (response 2). One important assumption of the Diffusion Model is that decisions are based on continuous information sampling that stops when one of two thresholds (e.g., the word- or the non-word-threshold) is reached. Figure 1 depicts a decision process according to the Diffusion Model as it might occur in an LDT after the presentation of a word-stimulus. In this example, the information accumulation process starts centered (starting point: parameter *z*) between both thresholds (distance of thresholds: parameter *a*) and moves with a certain speed in a certain direction (drift parameter *v*). To the linear trend of the mean drift rate—which is here directed to the upper threshold, i.e., the word-threshold—adds Gaussian noise. When a threshold is hit, the decision has been taken and a corresponding motoric response is initiated (e.g., press of a response key). The duration of pre-decisional encoding processes and post-decisional motoric processes is captured by the model with an additional parameter (non-decision time t_0; not depicted in Figure 1). Three of the four main parameters of the Ratcliff Diffusion Model are assumed to vary from trial to trial, namely, the drift rate *v*, the starting point *z* (Ratcliff & Rouder, 1998) and the non-decision time t_0 (Ratcliff & Tuerlinckx, 2002).

Psychological interpretation of the diffusion model parameters

For most applications, the drift rate, the threshold separation, the starting point or the non-decision time are of interest, because these parameters have straightforward psychological interpretations. Experimental validation studies have been conducted supporting the validity of some or all of these four parameters using a color-discrimination task (Voss et al., 2004), a lexical-decision task (Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), a motion discrimination task (Ratcliff & McKoon, 2008), brightness discrimination tasks (Ratcliff & Rouder, 1998) and—recently—also a recognition memory task (RMT; Arnold, Bröder, & Bayen, 2015). Furthermore, in correlational studies relationships between parameters and external criteria such as intelligence have been observed (e.g., Ratcliff et al., 2011). In the following, we shortly present some important findings regarding these four parameters.

The *drift rate (v)* maps the speed with which information is gathered and is thus a measure of speed of information processing. High absolute values of the drift result in short decision times and high accuracy. Empirically, it has been shown for different experimental paradigms that easier tasks (or easier trials) cause higher drift rates than more difficult tasks or trials (e.g., Ratcliff & McKoon, 2008; Voss et al., 2004). Furthermore, the drift rate correlates with general intelligence, with individuals higher in intelligence manifesting higher (absolute) drift rates (Ratcliff et al., 2010, 2011). This is interesting as it suggests the possibility to assess intelligence via the diffusion model's drift rate. Besides, it has been shown that the drift rate is also related to working memory (Schmiedek et al., 2007).

The *threshold separation* (a) reveals whether more or less information is accumulated before a decision is made. Accordingly, an accuracy- (speed-) instruction causes higher (lower) threshold separations (e.g., Ratcliff & McKoon, 2008; Voss et al., 2004). Correlational findings show that threshold separation is related to age (e.g., Ratcliff, Spieler, & McKoon, 2000; Ratcliff, Thapar, & McKoon, 2001) and to performance in tasks requiring response inhibition (Stahl et al., 2014).[1] Recently, it has also been shown that task-switching costs are based in large parts on high threshold separation in switch trials (Schmitz & Voss, 2012, 2014).

The *starting point* (z) reveals whether there is a bias in favor of one of the two responses. In the example illustrated in Figure 1, the starting point is centrally aligned between the two thresholds (i.e., the relative starting point $z_r = z/a = .5$). In this case, the decision maker manifests no decision bias for words or non-words. A bias (i.e., $z_r < .5$ or $z_r > .5$) can result from unequally rewarded responses (e.g., Voss et al., 2004), valence of stimuli (Voss, Rothermund, & Brandtstädter, 2008), or a manipulation of stimulus frequency (Leite & Ratcliff, 2011).

Finally, the *non-decision time* (t_0) informs about the time required for processes taking place before and after the decision process (specifically, the encoding of information and the motoric execution of the response). It is higher if the required motoric response is prolonged (movement of finger vs. direct key press, see Voss et al., 2004). An example for the manipulation of pre-decision time is provided by Schmitz and Voss (2012): They show that non-decision time is increased in task switch trials for unexpected task switches (cf. also Schmitz & Voss, 2014). Presumably, the reconfiguring of the task set in working memory takes time that is captured in t_0. Besides, non-decision time is generally related to age (e.g., Ratcliff et al., 2010; Spaniol, Madden, & Voss, 2006; Spaniol, Voss, & Grady, 2008), and—in an LDT—to vocabulary knowledge: Individuals higher in vocabulary knowledge show smaller non-decision times (Yap et al., 2012).

[1] Noteworthy, in unpublished studies from our lab, we failed to find correlations of threshold separation with self-reported impulsivity using standard speeded response time tasks (cf. also Stahl et al., 2014). However, when using more difficult tasks that required a long duration of information accumulation (RT > 5 sec) weak to moderate correlations emerged.

Estimation of the diffusion model parameters

For the estimation of the parameters of the Ratcliff Diffusion Model, several software implementations are available, amongst them *DMAT* (e.g., Vandekerckhove & Tuerlinckx, 2007, 2008) and *fast-dm* (Voss & Voss, 2007, 2008). Both programs allow the estimation of all seven parameters of the Ratcliff Diffusion Model. Besides, parameters can be estimated depending on conditions. For example, within the same model separate drift rates for words and non-words can be estimated. In the latest version of *fast-dm* (Voss, Voss, & Lerche, 2015), the user can choose between three different optimization criteria for parameter estimation: Kolmogorov-Smirnov (KS), a Maximum Likelihood (ML) and a Chi-Square based criterion (CS). Lerche, Voss, and Nagler (2015) compared the accuracy of parameter recovery between these three criteria for different trial numbers and different levels of model complexity. In addition, in their simulation study, data were either uncontaminated (i.e., all trials emanated from a diffusion process), or contaminated with 4 % of either fast or slow contaminants. Their main criterion of estimation performance was a bias measure based on deviations between true and estimated parameter values. The analyses revealed that in some conditions even for small trial numbers ($n < 100$) acceptable parameter estimates can be obtained. Specifically, ML outperformed KS and CS for uncontaminated data whereas for data contaminated by fast contaminants KS showed best performance.

Wagenmakers, van der Maas, and Grasman (2007) proposed an easy option for parameter estimation termed *EZ* which is based on closed-form equations. Entering accuracy rate and mean and variance of correct responses, these equations allow the computation of three parameters of the diffusion model: drift rate, threshold-separation and non-decision time (for extensions of EZ, see also Wagenmakers, Grasman, & Molenaar, 2005; Wagenmakers, van der Maas, Dolan, & Grasman, 2008). EZ assumes a fixed starting point ($z_r = 0.5$) and no intertrial variability (i.e., all three intertrial variabilities are fixed to zero). As a simulation study by van Ravenzwaaij and Oberauer (2009) demonstrates, EZ—despite using less information than other estimation procedures—allows a good recovery of the three main diffusion model parameters (see also Arnold et al., 2015).

Test-retest reliability

While the validity of the diffusion model parameters has been tested in several studies, to our knowledge the test-retest reliability has been examined only once, namely by (Yap et al., 2012) (see also Yap, Sibley, Balota, Ratcliff, & Rueckl, 2015). They re-analyzed data from an LDT of the English Lexicon Project (Balota et al., 2007) based on two sessions with at most one week in between. Their number of participants ($N = 819$) was very high as well as their number of trials per person (n = 3,374). All parameters of the Ratcliff Diffusion Model were estimated (with one drift rate for words and one for non-words) and CS was used as optimization criterion. Both between-session and within-session reliability were examined. Within-session reliability was assessed via split-half correlations of estimates for odd- and even-numbered trials. The psychologically most interesting parameters showed a high within-session reliability (threshold separation: .906; drift: .814/.827; starting point: .910; non-decision time: .930) and between-session reliability was acceptable (threshold separation: .708; drift: .692/.645; starting point: .720; non-decision time: .736). The intertrial variabilities performed more poorly in terms of within-session reliability (s_{t0}: .647; s_v: .649; but s_{zr}: .812) and especially between-session reliability (s_{t0}: .497; s_v: .403; s_{zr}: .388).

In the following, we first report the method and results of Study 1, in which participants worked on two classification tasks (a lexical decision task and a recognition memory task) at two different sessions. We display retest correlation coefficients (i.e., between-session reliability) for the parameters of the Ratcliff Diffusion Model, separately for both tasks and for four estimation methods (KS, ML, CS and EZ). In contrast to standard diagnostic procedures (e.g., personality questionnaires), a diffusion model analysis requires a complex parameter estimation process (with the exception of EZ which is based on closed-form equations). Therefore, any inaccuracies in parameter estimation (in the following termed “parameter estimation error”) add to the unsystematic “measurement error” of any test procedure. Thus, we expect test-retest reliability coefficients to be smaller for the diffusion model parameters than for statistical measures that can be directly computed like the mean reaction time. Support for this reasoning supply the findings by Yap et al. (2012). In their study, the reliability of the mean RT was higher (within-session: .997; between-session: .871) than the reliability coefficients of the diffusion model parameters (within-session: .647-.930; between-session: .388-.736).

In Study 2, the test-retest reliability of a further classification task (an associative priming task) was analyzed. Finally, in Study 3, we simulated data sets with different trial numbers assuming perfect parameter stability. These simulations allow to estimate (1) the influence of the number of trials on the “parameter estimation error” and (2) the maximally possible retest reliability under perfect conditions (i.e., assuming that there is no change in cognitive processes across sessions). They also allow (3) to get rough estimates of the trait- and state-proportions of parameters comparing the reliability of the empirical data (most likely, with state influences) with the reliability of the simulated data (without state influences due to the simulation procedure).

Study 1: test-retest reliability of the diffusion model parameters in a lexical decision task and a recognition memory task

An important prerequisite for the interpretation of the diffusion model parameters as measures of trait-like cognitive styles and abilities is the test-retest reliability of these parameters. However, there is little information available on the reliability and stability of the parameters (see Yap et al., 2012, for an exception). In Study 1, at two sessions separated by one week, participants had to work on two experimental tasks, an LDT and an RMT.

The LDT is an experimental paradigm that has often been analyzed with the Diffusion Model (e.g., Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2007; Ratcliff et al., 2010; Wagenmakers, Ratcliff, et al., 2008). The study of Yap et al. (2012) provides first information on the retest reliability of the different cognitive processes involved in an LDT.

The diffusion model has also been frequently applied to data from RMTs, both for words (e.g., Ratcliff, Thapar, & McKoon, 2004; Ratcliff et al., 2007; Spaniol et al., 2006; White, Ratcliff, Vasey, & McKoon, 2009; White et al., 2010b) and pictures (Bowen, Spaniol, Patel, & Voss, 2015; Ratcliff & McKoon, 2015; Spaniol et al., 2008). We opted for an RMT based on picture stimuli. Thus, we examine two tasks that differ in the paradigm and in the material (LDT: words vs. RMT: pictures). Besides, for both tasks, experimental validation studies have been performed (Arnold et al., 2015; Wagenmakers, Ratcliff, et al., 2008).

Method

Participants

One-hundred-and-five native German speakers participated in Study 1. They were recruited from the participants' pool of the Psychological Institute of the University of Heidelberg, Germany using the *hroot* software (Bock, Baetge, & Nicklisch, 2014). From all participants informed consent was obtained. After the second session participants received either course credit or 20€ (approx. 22 US $). The participants were mostly female (80 %) and were on average 22.0 years old (min = 18, max = 32, *SD* = 2.9). Forty percent of the participants studied psychology or worked as psychologists and the proportion of students added up to 97 %.

Stimuli

For the LDT we used 200 German nouns with 1 or 2 syllables and 4-6 letters. All word stimuli had a low frequency in German language.[2] For each word stimulus, a non-word was created by random vocal replacement.

For the RMT we used 200 pictures from the IAPS (International Affective Picture System; Lang, Bradley, & Cuthbert, 2008) and the Emotional Picture Set (EmoPics; Wessa et al., 2010) with neutral valence (range: 4-6 on a scale ranging from 1 to 9) and low arousal (max. 5, scale: 1-9). Furthermore, only pictures showing humans were selected to have a relatively homogeneous set of stimuli, thus making the task more demanding. In the retention phase, items of three questionnaires were presented as filler task, namely the Personality Research Form (PRF, Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1985), the UPPS Impulsive Behavior Scale (Whiteside & Lynam, 2001) and the NEO Five Factor Inventory (Borkenau & Ostendorf, 2008); different items from these tests were presented in the two sessions.

Design and procedure

The experiment consisted of two sessions that were separated exactly by one week for almost all participants (for five participants, the second session had to be postponed, but took place no more than two weeks after the first session). The order of tasks (LDT and RMT) was counterbalanced across participants. For each participant the sequence of tasks was identical for both sessions.

In the LDT, each session consisted of 400 trials which were presented in four blocks of 100 trials. Each block consisted of 50 word trials and 50 non-word trials. Between blocks participants could take short breaks. Stimuli were presented in a random order that was held constant for all participants and for both sessions. The mean duration of the task was 12 min.

The RMT consisted of two blocks. Each block started with a study phase in which 52 pictures were presented sequentially for 3 s each. The first and last presented picture (primacy and recency buffer) did not appear in the test list. The learning phase was followed by a retention phase in which participants had to fill in questionnaire items on the PC for at least 5 min. The test list of each block consisted of 50 pictures from the study list and 50 new pictures and started with one warm-up trial (a new picture) that was ignored for analyses. In

[2] Words had frequencies below 5 per million (CELEX; Baayen, Piepenbrock, & Gulikers, 1995) and – at the same time – a frequency class of 14 or 15 (online dictionary project of the university of Leipzig in November 2014, see http://wortschatz.unileipzig.de), indicating that the word “der” (“the”) is used 2^{14} or 2^{15} times as often in German language as our stimuli.

the learning phase, a random order of the stimuli was used that was identical for all participants and for both sessions. The order of stimuli in the test phase was also created randomly and was equivalent for all participants but varied from session 1 to session 2. Mean duration of the recognition task was 26 min.

The procedure of each trial was equivalent for the two tasks: First, a fixation target (a black dot) was presented at the center of the screen for 500 ms. Then, the stimulus (a letter combination or a picture) was shown at the same position and the participants had to press either a left key (A) for non-words or new pictures or a right key (L) for words or old pictures. Labels at the bottom corners of the screen showed the mapping of stimulus types to keys throughout the experiment. As soon as a response was given the stimulus was removed from the screen and the next trial started after an intertrial interval of 500 ms. Participants were instructed to respond as fast as possible while avoiding errors at the same time.

Parameter estimation

In order to compare the performance of different estimation methods, we carried out parameter estimation using three different optimization criteria—KS, ML and CS as implemented in fast-dm-30 (Voss & Voss, 2007, 2008; Voss et al., 2015)—and using the method EZ (Wagenmakers et al., 2007). In our previous work we found that less complex models often provide more accurate and stable results even if some aspects of the true cognitive processes cannot be mapped (Lerche & Voss, 2016). Specifically, in our simulations the fixation of the intertrial variabilities of v and z_r to zero had a positive influence on the parameter estimation even when there was such variability in the data generating process. Apparently, the inclusion of these parameters can make the parameter estimation instable in the case of moderate or low trial numbers. Thus, for the fast-dm estimates we will present retest reliability coefficients based on a parameter model including the parameters threshold separation (a), starting point (z_r), two drift rates for the two stimulus types (v_0 for non-words and new pictures associated with the lower threshold, v_1 for words and old pictures associated with the upper threshold), non-decision time (t_0) and intertrial variability of non-decision time (s_{t0}). Note that for the KS method an even more restricted parameter model with additional fixation of the intertrial variability of t_0 to zero would lead to slightly higher reliability coefficients. EZ was applied separately to the trials of each target type, resulting in two drift rates, two threshold separations and two non-decision components[3]. For threshold separation and non-decision time, the mean over the two estimates was calculated.

In addition, we computed retest reliabilities of the arcsine-transformed percentage of correct responses (accuracy) and the mean of the logarithmized response times of correct responses (mean RT). These are transformations that are regularly used for analyses on data from response time experiments to get more normally distributed variables (e.g., Greenwald, McGhee, & Schwartz, 1998; Hintzman & Curran, 1997; Klauer & Dittrich, 2010; Neumann & Strack, 2000).

Results

Due to a technical problem data from the RMT task from session 2 were missing for one participant. Data from another participant were excluded because his or her mean accuracy score was at chance level for the RMT (46 %) and was also the lowest in the complete sample for the LDT (75 %). Therefore, data from 104 participants for the LDT and from 103 participants for the RMT were entered into analyses. For each correlational analysis, bivariate outliers were identified via the Mahalanobis distance (D^2) and participants with extreme values ($p < .001$) were excluded from the respective analysis. This led to an exclusion of at most 4 participants. Reaction times faster than 200 ms (on average 0.02 % for the LDT and 0.03 % for the RMT) were excluded as well as reaction times slower than 2,500 ms (on average 0.38 % for the LDT and 1.11 % for the RMT).

The test-retest reliability coefficients (Pearson correlation coefficients) were calculated between the two sessions for the diffusion model parameters, the accuracy rate and mean RT and are depicted in Figure 2. For the drift parameter, reliability coefficients are presented additionally for the difference and the sum of the two drift rates. As v_1 is usually positive and v_0 usually negative, the difference ($v_{total} = v_1 - v_0$) can be seen as a measure of overall speed of information processing (i.e., the ability to discriminate between both stimulus types), while the sum ($v_{bias} = v_1 + v_0$) maps a potential bias in drift rate, that is, a general preference in information accumulation for one type of information.

As can be seen in Figure 2, v_{total} ($r > .70$ for KS, ML and EZ) and the threshold separation a ($r > .70$ for ML and EZ) show an acceptable reliability for both tasks. The

[3] As EZ cannot be applied to data sets with an accuracy rate of 100%, we applied an edge correction method which has also been used by Wagenmakers et al. (2007): accuracy = $1 - \frac{1}{2\times n}$, with n being the number of trials. Similarly, in Study 3 we additionally used a correction for a few data sets due to an observed accuracy of 50% in one condition (accuracy = $0.5 + \frac{1}{2\times n}$).

reliabilities of the single drift rates are smaller than those of v_{total}, but they are still in an acceptable range. A comparison of the diffusion model parameters with the "standard" variables mean RT and accuracy rate reveals that the accuracy rate outperforms the diffusion model parameters' reliability, while the mean RT performs worse than some parameters (primarily, v_{total}). The figure also allows for a comparison of the four methods that we used for parameter estimation. As expected, CS often shows the worst performance; ML and EZ manifest the best performance. One further finding is that parameters estimated from the LDT have higher reliability coefficients than those of the RMT. We will consider possible explanations for this finding in the Discussion.

Besides the retest reliability of the parameters, we also analyzed practice effects from the first to the second session. Our findings indicate that several parameters changed significantly (Table 1). In particular, in both tasks, participants' performance was better at the second session: drift rates (both the single drift rates and v_{total}) increased and non-decision time decreased. Furthermore, the participants showed a more liberal response criterion, indicating that less information was accumulated at the second compared to the first session. In the LDT, there was also a significant change in the starting point. While in the first session the starting point was closer to the upper threshold (indicating a bias in favor of words), in the second session the starting point was unbiased. In the RMT, on the other hand, the starting point was closer to the upper threshold at both sessions (indicating a bias in favor of "old"-responses). Finally, across both tasks participants showed a bias of drift for the stimulus associated with the lower threshold (non-words or new pictures). The absolute values of drift rates for these stimuli were higher compared to the drift rates for words and old pictures. This bias decreased significantly from the first to the second session. We also computed across-task correlations of parameters within one session. These are essentially smaller than the test-retest correlations (see Table 2).

Discussion

The present test-retest study shows consistent findings for retest reliabilities across both tasks: The main diffusion model parameters (specifically, the drift parameter and the threshold separation) have an acceptable retest reliability. These results suggest that these measures may be used for cognitive diagnostics. For example, the drift rate (or the difference of drift rates between stimulus types, i.e., v_{total}) might be used for the assessment of general cognitive speed and—possibly—for intelligence. Threshold separation could be employed as a measure of impulsive decision making. The other main diffusion model parameters (z_r and

t_0) perform worse, but still manifest a robust retest correlation, indicating some trait-like component in these parameters. Besides, also the sum of drift rates manifests a certain reliability. This finding is interesting as it shows that this measure might be used for the diagnosis of a bias (e.g., "green bias", see Allen, Lien, Ruthruff, & Voss, 2014; or memory bias, see Bowen et al., 2015). In the interpretation of the retest reliabilities for the bias measures (z_r and v_{bias}) it is important to note that in our paradigms no interindividual variance was expected. When such biases are more meaningful (e.g., in social-cognitive studies on preferences), we expect retest reliability to be higher.

Generally, we assume that the values of the retest coefficients depend on the paradigm at hand as well as on sample characteristics. Thus, the results reported are limited to the type of task used (lexical decision task and recognition memory task). In Study 2, we report a further test-retest study to challenge the generalizability of our results to a slightly different paradigm (an associative priming paradigm based on a lexical decision task). Here, in addition to the reliability of the single parameters, we also analyze the reliability of a priming effect (in particular, the difference in drift rates between associated and nonassociated prime-target pairs).

Besides the retest reliability of the different parameters, in Study 1 we were also interested in the performance of different optimization criteria. Our results are in line with findings from simulation studies (Lerche & Voss, 2016; Lerche et al., 2015): In these studies, —for small and medium trial numbers as employed in our LDT and RMT—a higher accuracy of parameter recovery for ML, compared to the widely used CS approach, was found. This more precise parameter estimation is assumingly also the reason for higher reliabilities for ML compared to CS in the present study.

While the pattern of the two tasks is very similar (with drift rate and threshold separation manifesting acceptable retest reliability), the reliability coefficients of the RMT are, however, generally smaller than the respective coefficients of the LDT. As the parameter values show larger variances in the RMT compared to the LDT, the smaller retest reliability coefficients cannot be explained by means of limited variance. However, there are several other possible reasons for the differences in reliability coefficients: (1) Differences could be attributed to the higher trial number in the LDT ($n = 400$) compared to the RMT ($n = 200$). (2) They could be a consequence of differences in contamination of the RT distributions. It is possible that for the RMT in a higher percentage of trials performance is not based on a continuous information accumulation as assumed by the Diffusion Model, because the task took longer and might have been experienced as more exhausting. (3) Differences in reliability could be due to differences in stability of the cognitive processes involved in recognition memory compared to lexical decision. In Study 3, we analyze the relationship of reliability and the number of trials of an experiment. This also allows for tackling the question of which of the three possible explanations might hold.

Study 2: test-retest reliability of the diffusion model parameters in an associative priming task

Study 2 is a further empirical test-retest study. In this study, we aimed at testing the generalizability of the results of Study 1 to another experimental paradigm. In particular, we used an associative priming task (APT). This type of task has already been analyzed with diffusion model analyses. In particular, Voss, Rothermund, Gast, and Wentura (2013; Study 1a and Study 3a) used a lexical decision task with the targets (words or non-words) preceded by words that were either associated with the target or not (see also Study 3b for similar findings based on a semantic classification task). Trials with associated primes featured shorter response times for words and a lower error rate than trials with nonassociated primes. To figure out on which cognitive component(s) the priming manifests, Voss, Rothermund, et al. (2013) applied a diffusion model analysis. As expected, in trials with associated primes, drift rates for words were higher than in nonassociated trials. The manipulation did not have an influence on the response-execution bias (parameter *d*; see Voss, Voss, & Klauer, 2010). Less clear-cut were the findings regarding the non-decision time. While in Study 1a it was not affected, in Study 3a significantly shorter non-decision times emerged in the associated compared to the nonassociated condition. In Study 2, we applied a procedure similar to Study 1a in Voss, Rothermund, et al. (2013). We were interested in the reliability of the single diffusion model parameters and in the reliability of a possible priming effect.

Method

Participants

One-hundred-and-twenty-eight participants were recruited using the *hroot* software (Bock et al., 2014). Informed consent was obtained and participants were remunerated with course credit or 20€. Almost all participants were students (97 %), most of them female (77 %) and they had an average age of 22.9 years (min = 17, max = 61, *SD* = 5.0). The percentage of participants that studied psychology or worked as psychologists was 27 %.

Stimuli

Both the primes and targets were taken from the study by Voss, Rothermund, et al. (2013). We used a total of 400 prime-target pairs. In half of these pairs, the prime was highly associated with the target which was either a word (e.g., “Kochtopf” [cooking pot]—“Essen”

[food]) or a non-word (”Banane” [banana]—“Affo” [“Affe” = monkey]). In the other half, the prime was not associated with the target (target is word: e.g., “Möbel” [furniture]—“Füller” [pen]; target is non-word: e.g., “Gürtel” [belt]—“Alzt” [“Arzt” = doctor]).

Design and procedure

For almost all participants, the second session was held exactly one week after the first session (for one participant it had to be preponed and for another postponed by one day). The order of trials (with the order created randomly) and the stimulus-key-mapping (word: key “K”; non-word: key “S”) was identical for all participants and at both sessions. During the experiment, the mapping was shown with labels positioned at the bottom corners of the screen.

The 400 prime-target pairs were divided upon 4 blocks of trials. Each block consisted of 25 associated prime-word pairs, 25 nonassociated prime-word pairs, 25 associated prime-non-word pairs and 25 nonassociated prime-non-word pairs. The task was set in with a block of 30 practice trials without primes and with accuracy feedback. Besides, each block had two additional warm-up trials. The practice and warm-up trials were not part of the test list.

Participants were instructed to respond as fast and accurate as possible. Each trial started with the presentation of a prime for 300 ms, followed by the appearance of the target. The discriminability of the target was hindered by a pixel mask. After the given response the target was removed and following an interval of 500 ms the next trial started. The task took on average 14 min.

Parameter estimation

The parameter estimation procedure was mainly equivalent to the procedure used in Study 1. The only difference was that in Study 1 we estimated two drift rates (one drift rate for each stimulus type: words vs. non-words in the LDT and old vs. new pictures in the RMT) and in Study 2 four drift rates and four non-decision times[4] based on the combinations of prime type (associated vs. nonassociated) and target (word vs. non-word).

Results

The lower response time boundary (200 ms) led to an exclusion of 0.04 % trials on average, the higher boundary (2,500 ms) to an average exclusion of 0.64 %. The use of the

[4] Based on the findings by Voss, Rothermund, et al. (2013), we did not expect the *d*-parameter of the diffusion model (Voss et al., 2010) to be influenced by the prime type. Besides, estimation of this parameter requires very high trial numbers (Voss et al., 2010).

Mahalanobis distance for the bivariate correlation coefficients resulted in an exclusion of at most 3 participants.

For word targets, as expected, in both sessions accuracy rate was higher for associated (e.g., Session 1: *M* = 95.31 %, *SD* = 3.69) than for nonassociated prime-target pairs (*M* = 92.06 %, *SD* = 5.08; *t*[128] = 9.71, *p* <.001, d_z = 0.90). Mean RT was lower (*M* = 789.97, *SD* = 110.85 vs. *M* = 814.84, *SD* = 105.70; *t*[128] = -9.84, *p* <.001, d_z = -0.70) and drift rates were higher (*M* = 2.22, *SD* = 0.55 vs. *M* = 1.84, *SD* = 0.44; *t*[128] = 10.94, *p* <.001, d_z = 0.97) (see also Table 3). Besides, the non-decision time was not significantly different between the two conditions (session 1: *p* = .06; session 2: *p* = .07).[5]

Test-retest coefficients across-sessions are shown in Figure 2 (bottom row). Interestingly, the pattern is similar to the LDT and RMT from Study 1: in all three paradigms v_{total} and *a* have satisfying correlation coefficients. In contrast to Study 1, in the priming paradigm t_0 has a higher reliability. Similar to Study 1 is the performance of the four estimation methods with ML and EZ outperforming the other two methods for most parameters. While in both sessions the expected priming effect on the drift parameter emerged, the reliability of this effect (i.e., of the difference between associated and nonassociated prime-target pairs) is not very high (*r*s < .40).

Information on changes in parameters from Session 1 to Session 2 can be retrieved from Table 4. Like in Study 1, drift rate increased and mean RT and non-decision time decreased from Session 1 to Session 2. Furthermore, participants responded more liberally (lower threshold separation) at Session 2 and the starting point (first closer to the word-threshold) shifted to a more centered position (see also Study 1). Besides, variability in non-decision time decreased from one session to the next.

Discussion

The findings from Study 2 are in line with the results reported by Voss, Rothermund, et al. (2013). For word targets, drift rates were significantly higher for associated compared to nonassociated primes (Cohen's d_zs > .95). For the non-decision component, there was no significant difference between these different prime-word pairings (d_zs < .20). The priming effect on the drift rate did not manifest a high retest reliability coefficient. Note that this is not surprisable given the high correlation of drift rates for associated and nonassociated primes (*r*

[5] Results for non-word targets are also presented in Table 3. Note, that the findings are similar to those reported by Voss, Rothermund, et al. (2013).

= .71 at session 1 and *r* = .77 at session 2 for ML estimation). These correlation coefficients are higher than the retest reliability coefficients of associated (*r* = .59) and nonassociated prime-word pairs (*r* = .62) and thus, make a high retest reliability coefficient of the difference measure less likely (e.g., Guilford, 1954). Besides, our results are in line with findings by Stolz, Besner, and Carr (2005) who, using a semantic priming task, found retest reliability coefficients that were in most experiments smaller than *r* = .30.

One important finding of Study 2 is that both drift rate and threshold separation revealed satisfying reliability coefficients. This is in line with the findings from Study 1. Note that in Study 2 the non-decisional component showed a higher reliability than in Study 1. This might have resulted from greater differences between individuals in encoding of information on the target stimulus due to the preceding prime stimulus. In fact, the participants seem to have used diverse approaches to deal with the prime as the answers to an open-framed question on their use of strategies revealed. Amongst these strategies were simply “ignoring”, counting the words, or even shortly closing the eyes. Others reported having tried to pay attention to the prime thinking that it might help them to identify the target. These different strategies could have had an influence on the subsequent encoding of the target stimulus. They can result in a higher variance and—assuming that the participants used similar strategies at both sessions—higher stability of the encoding process. Note that the closely following presentation of prime and target might also be interpreted as a form of task switching which can have an influence on the non-decision time (Schmitz & Voss, 2012).

As other studies (e.g., Lerche et al., 2015; Wiecki, Sofer, & Frank, 2013) demonstrate, the accuracy of parameter estimation depends on the number of trials. While in Study 1 and Study 2 only one trial number (200 for the RMT, 400 for the LDT and for the APT) was analyzed, in Study 3 we compared several different trial numbers. Besides, in Study 3, a simulation study was executed. One problem associated with empirical data is that we do not know the true parameter values of the participants and thus cannot separate error that results from inaccuracies in parameter estimation from instability of cognitive processes. In a simulation study, however, we know the true parameter values based on which we created the data sets. In sum, in Study 3 we had three main aims: (1) analyzing the influence of the number of trials on parameter estimation, (2) explaining the differences in retest reliability between the LDT and RMT from Study 1, and (3) getting a rough estimate of state proportions of parameters.

Study 3: Influence of the number of trials on test-retest reliability

In Study 3, data from Study 1 and Study 2 were re-analyzed using different subsamples of trials. This allows us to check for the influence of the number of trials on retest coefficients. Besides, we extended the empirical data of Study 1 with simulated data. Retest reliability coefficients from the empirical data sets were compared to coefficients resulting from simulated data sets. Because data were simulated under "optimal" conditions—that is, without any changes in the data-generating process—the simulations provide an upper bound for the possible reliability coefficients for different trial numbers. This allows us to estimate the influence of changes in cognitive states.

Method

First, we estimated diffusion model parameters from subsamples of trials of each participant of Study 1 and Study 2. We used sample sizes of 32 (for LDT and RMT)[6], 48 and 100 trials (for all three tasks) and of 200 trials (for LDT and APT). For the subsamples, a sequence of trials from the beginning of each data set was used, thus mimicking complete data sets from shorter experiments. Like in Study 1 and 2, parameters were estimated with fast-dm-30. Using Fisher's Z-transformation we calculated the mean over the retest reliability coefficients of the EZ diffusion model parameters (specifically, a, v_{total} and t_0) to obtain an overall measure of retest reliability (in the following termed "mean retest reliability").

Additionally, we report reliability coefficients within sessions. In particular, exemplarily, correlations between the first and second block of the RMT are contrasted with the correlation between the first and second session (using the first 100 trials to have equal trial numbers for the computation of all coefficients). Similarly, for the LDT and APT we calculated correlations between the first and second 200 trials (thus, blocks 1-2 vs. blocks 3-4) within sessions and compared findings to across-session reliability (using the first 200 trials).

As a second approach, we simulated data sets under the assumption of perfect parameter stability. Specifically, we first created 1,000 parameter sets based on a multivariate normal distribution of the parameter sets from the real data from Study 1 (Table 5), using *mvrnorm* from the MASS *R* package (R Development Core Team, 2014; Venables & Ripley, 2002). For each parameter set and each trial number (32; 48; 100; 200; 400; 1,000; 5,000),

[6] In each condition, fast-dm requires at least 10 trials (independent of the type of response) for ML and KS estimation and 12 trials (of the same response) for CS estimation (Voss et al., 2015). Thus, for the data of the APT due to the higher number of conditions no retest coefficients could be computed for 32 trials and for CS neither for 48 trials.

two data sets were simulated with *construct-samples*[7]. All differences within the simulated pairs of data sets are due to random noise in the diffusion process. Therefore, the correlation between the parameters re-estimated from the pairs deviates from a perfect reliability coefficient of $r = 1$ only by a “parameter estimation error“ (i.e., the error resulting from the complex multidimensional estimation procedure). In two further conditions, we made the data generation more realistic by adding contaminants, that is, trials in which the diffusion model is not the data generating process. For this purpose, we substituted 4 % of simulated trials by either fast or slow contaminants following the procedure described in Lerche et al. (2015).

Results

Unsurprisingly, for smaller trial numbers the retest reliability is lower than for higher trial numbers. This pattern can be observed for both the empirical (Figure 3 and Figure 4) and the simulated data sets (Figure 4). Interestingly, using more than about 200-400 trials does not have a substantial positive effect on retest reliability. The curve of the simulated data sets then still increases, but only to a very small degree. Figure 4 and Figure 5 show the influence of the number of trials on retest reliability of mean RT and accuracy rate. While for the accuracy rate, the curves clearly increase, for mean RT the number of trials seems to have a very small influence. This is plausible as the accuracy rate is based on data of a lower level of measurement (nominal) than the mean RT (metric).

The data resulting from the simulation study also allows an assessment of the proportion of “parameter estimation error”. Interestingly, there is basically no systematic difference between uncontaminated data and data with fast or slow contaminants. The distance between the reliability of empirical data and the correspondingly simulated data reveals information on state influences on parameters. We want to add the caveat that other factors may also reduce reliability of empirical data (e.g., a higher percentage and/or different type of contamination). However, the distance of reliability graphs in Figure 3 provides an estimation of the *maximal* proportion of state influences.

As evident from the figures, the higher the trial numbers, the smaller is the parameter estimation error and the higher the maximal proportion of state influences. For very small trial numbers (e.g., 48 trials), it is difficult to disentangle these two influences. From around 100 or 200 trials on, however, the distance between the simulated and empirical lines remains

[7] *Construct-samples* is part of fast-dm (Voss et al., 2015). For the simulation of data sets we used a high precision setting of $p = 4$.

roughly stable (only increasing slightly due to further decrease in parameter estimation error). Thus, if, for example, we were interested in the moderation of practice effects by a personality trait, we should better use at least 200 trials.

Figure 3 also allows for a comparison of parameter stability across the two tasks. Remember that in Study 1 we found lower reliability coefficients for the RMT compared to the LDT when using the total number of trials of each task (200 and 400 trials for RMT and LDT, respectively). As the increase in retest reliability with the number of trials indicates, the trial number does have an influence (even if this influence gets smaller with an increase in the trial number). Besides, for the simulated data, higher reliability coefficients are observed for the RMT compared to the LDT (which is probably due to the higher variability of parameters observed in the RMT). This results in slightly greater distances between the simulated data and the empirical data for the RMT than for the LDT. Taken together, these findings suggest that the differences found in Study 1 are not exclusively attributable to the different trial numbers, but also partly to differences in parameter stability. Specifically, the processes involved in the LDT result more stable than those involved in the RMT. This might be because memory processes are more influenced by situational factors or because the assessment is more contaminated (by sources of contamination not assumed for our simulation of data sets). We give more weight to the first proposed explanation as our simulation study reveals that contamination generally does not have a clear negative influence on correlation coefficients.

While Figure 3 is based on the average of the four main parameters, Figure 6 illustrates the patterns for the single parameters, exemplarily using ML as optimization criterion (the pattern is similar for the other two criteria). The comparison of the different parameters shows which cognitive processes are more or less stable. For example, we can see that the cognitive speed factor v_{total}—in particular for the LDT—is very stable. The distance between the simulated and empirical lines is approximately zero meaning that there are basically no state influences on cognitive speed as measured with the drift rate.[8] In line with these findings are also the results from the comparison of within- and across-session reliability coefficients. While for drift and threshold separation no systematic pattern of

[8] Note that, as already mentioned, it is possible that the contamination in the empiric data is different from the type and amount of contamination that we assumed for the simulation of data. Thus, it could be that the estimation of the drift rate suffers less from contamination than for example the estimation of t_0 and that not (only) the stability of the parameters is responsible for the different distances between the lines of simulated and empiric data. Our study, thus, only allows getting an approximate idea of the state proportions. A clear disentangling of state and trait proportions would require larger samples of participants and data points.

differences between within- and across-session reliability coefficients emerges (within-session reliability coefficients are not always higher and if they are higher, there is only a rather small difference), there is a clearer pattern for t_0 and z_r with the within-reliability coefficients being higher than the across-session reliability coefficients (see Table 6). This cautiously gives rise to the idea that interindividual differences in drift and threshold separation might be less influenced by situational characteristics than interindividual differences in non-decision time and starting point.

The analysis of the retest correlation coefficients is one way to examine differences between estimation methods. However, correlation coefficients can mask potential biases in parameter estimation. If a parameter is systematically over- or underestimated (i.e., this happens at both sessions), still high retest correlation coefficients can result. In order to examine the vulnerability of the different methods to estimation biases, we conducted further analyses. In particular, for the data of the simulation study, we compared the true parameter values (i.e., the values, the simulation was based on) with the estimated parameter values. In Figure 7 (for the LDT simulation study) and Figure 8 (for the RMT simulation study), boxplots of the residuals (i.e., the differences between estimated and true values) are presented, exemplarily for Session 1[9]. Positive values indicate an overestimation and negative values an underestimation of the true parameter. The different diffusion model parameters have very different ranges. To enhance comparability of parameter biases, we weighted the differences between estimated and true values against parameter accuracies reported in Lerche et al. (2015). These accuracies are based on parameter estimation under perfect conditions (i.e., a very high trial number, no contaminants, ML estimation).

First, it is apparent that both threshold separation and drift rate are often underestimated. The non-decisional component and the starting point are estimated without a notable, systematic bias by the fast-dm methods. EZ, however, underestimates the non-decisional component (especially, in the RMT simulation study). The four methods also differ in that EZ has higher interquartile ranges in the RMT simulation study and, most importantly, in that CS has a higher number of outliers than the other methods. This is in line with the results from the retest coefficients that were mostly lower for CS than for the other methods. As for the retest coefficients, differences between types of contamination are negligible in size. Finally, Figure 9 exemplifies the influence of the number of trials on

[9] Results are very similar for Session 2.

parameter estimation bias (exemplarily, for data with no contaminants from the LDT simulation study). Importantly, from 400 trials on or even earlier any biases are very stable.

Discussion

We conducted Study 3 with three main objectives: Most importantly, we wanted to derive guidelines on the trial number required for reliable parameter estimation. As our results show, an increase in trial numbers has a positive effect up to about 100 or 200 trials. For higher trial numbers, however, the improvement in reliability is negligible. Additionally, biases in parameter recovery are also rather stable from this number of trials on.

Our second aim was to explain the differences observed in retest reliability between the LDT and RMT from Study 1. In this regard, we could infer that the differences are not exclusively attributable to the different trial numbers. The cognitive processes involved in the RMT seem to be a bit less stable than those involved in the LDT. Finally, our third aim was to make a rough estimate of state proportions of the different parameters. These seem to be smaller for drift rate and threshold separation than for non-decision time and starting point.

General discussion

The retest reliability of the parameters of the Ratcliff Diffusion Model (1978) was evaluated using three different experimental paradigms. The lexical decision task (LDT) and the recognition memory task (RMT; Study 1) are both tasks that have been frequently used for diffusion model analyses. Besides in Study 2, an associative priming task (APT) was analyzed. In both studies, participants (Study 1: 105; Study 2: 128) worked on the tasks at two sessions separated by an intersession interval of one week. The analyses revealed satisfying reliability coefficients. In particular, the parameters with the highest reliability in all three tasks are the drift rate that is a measure of speed of information accumulation and the threshold separation that measures the amount of information required for decision making. Note that the accuracy rate (percentage of correct responses) showed higher retest correlations than drift rate and threshold separation. One might think that this challenges the applicability of the Ratcliff Diffusion Model. In this context, we want to point out that the great strength of the Diffusion Model is the discriminant validity of its parameters, that is, its capacity to separate different cognitive processes. For example, whenever accuracy and mean RT are unrelated (or even negatively related), both measures may lead to different conclusions. This trade-off can often be solved using more direct measures of cognitive processes as those provided by the Diffusion Model (e.g., Ratcliff, Thompson, & McKoon, 2015).

Generalizability and stability

In recent years, the interest in applying the diffusion model to measure interindividual differences is growing rapidly (e.g., Aschenbrenner et al., 2015). At the same time, strikingly little is known about the stability of the underlying processes and the reliability of different parameter estimation procedures. In view of this, the high similarity of the pattern of results (in terms of satisfying reliability coefficients of drift rate and threshold separation in all three paradigms) is in our eyes quite promising for a potential future diagnostic use.

Note, however, that the lower reliability coefficients for the starting point in all three tasks do not necessarily imply that this parameter is much less stable in general. In fact, the comparison of the three paradigms demonstrated that the size of the retest coefficient of a parameter also depends on the type of paradigm. Specifically, we found a higher reliability of t_0 in the APT than in the LDT and RMT. This might be a result of stability in strategy use across sessions (i.e., strategies to deal with the prime stimulus that influenced the encoding of the subsequently presented target). Similarly, there might be tasks in which the starting point has more variance and stability (e.g., in tasks based on emotional stimuli with the stimulus valence depending on characteristics of the individual). Thus, we cannot conclude that z_r is less stable in general, but only in the specific paradigms that we analyzed. In this context, we also want to stress that the results of our studies only supply information on the reliability of diffusion model parameters in lexical decision (with and without priming) and recognition memory. They do not automatically generalize to any other type of task that the diffusion model has been applied to. The stability of parameters and the reliability of the parameter estimation procedure need to be carefully tested for each new paradigm. This also applies to the validity of diffusion model parameters in new paradigms that needs to be examined using experimental validation studies (e.g., Arnold et al., 2015; Voss et al., 2004) and analyses of criterion validity (e.g., Schubert et al., 2015).

In addition to the collection of empirical data, we also conducted a simulation study (Study 3). For the generation of data sets we used the parameter ranges observed in the empirical data of Study 1 and assumed perfect stability of parameters. Thereby, we got the maximally possible retest coefficients (given a sample as used in our study). A comparison of the retest coefficients for the empirical data with the maximally possible retest coefficients allows an estimation of the "parameter estimation error" (i.e., the error resulting from the complex estimation procedure) and it allows a rough estimate of the maximal size of state influences. This analysis revealed that the processes involved in the LDT seem to be more stable than those in the RMT.

In line with these differences between the two tasks are also the across-task correlation coefficients that are essentially smaller than the retest reliability coefficients. Strikingly, the drift parameter has only rather small across-task correlations indicating that the two tasks do not measure a common speed of information accumulation. This contrasts previous studies in which large across-task correlations have been found. For example, Ratcliff et al. (2015) used four number-based tasks (number discrimination, numerosity discrimination and memory for numbers composed of either two or three digits) and report a mean correlation of $r = .52$ for the drift parameter. Ratcliff et al. (2010) even found a correlation of $r = .63$ between the drift in an RMT and an LDT. Their sample, however, was more heterogeneous than our sample and they used word stimuli in the RMT, so that the material was more similar in the two tasks compared to the material in our study. Note that we intentionally selected tasks differing in material (LDT/APT: words; RMT: pictures) to test the generalizability of our results.

Influence of the estimation procedure

One further aim of our study was a comparison of different estimation methods: Kolmogorov-Smirnov (KS), a Maximum Likelihood (ML) and a Chi-square based criterion (CS) and the method EZ (Wagenmakers et al., 2007) were used. Our results are mainly in line with findings from simulation studies by Lerche et al. (2015). Specifically, ML outperformed KS and CS. Of interest is the good performance of EZ (for similar findings see also van Ravenzwaaij, Donkin, & Vandekerckhove, 2016; van Ravenzwaaij & Oberauer, 2009). The retest coefficients achieved by EZ are comparable with those of ML. The fact that EZ performs so well despite its fixations of parameters (centered starting point; no intertrial variabilities) is in line with recent findings by van Ravenzwaaij et al. (2016). Note, however, the higher interquartile ranges in the estimation bias measure (Study 3). Regarding estimation bias, KS and ML show the best performance.

In Study 3, we also varied the number of trials used for the estimation of parameters. As expected, with the number of trials increases the retest reliability. However, interestingly, using more than approximately 200-400 trials does not substantially increase retest reliability. This aligns well with the findings by Lerche et al. (2015) who also noted that from around 500 trials on parameter estimation performance improves only marginally. High retest reliability coefficients do not necessarily imply that parameters are estimated perfectly. In fact, as our analyses on parameter recovery show, especially drift rate and threshold separation are estimated with a certain bias.

One limitation of our study is that we collected only 200 trials (for the RMT) and 400 trials (for the LDT and APT). Note, however, that it is possible that an increase in trials beyond the total number of trials used in the experiments would have led to additional data contamination due to increasing fatigue, unwillingness to participate and distraction from the task. Therefore, retest reliability coefficients might have been even lower than for the 200 or 400 trials used.

To our knowledge, so far only one further published study has assessed the retest reliability of the diffusion model parameters: Yap et al. (2012) used a similar intersession interval and also an LDT (see also Yap et al., 2015). Interestingly, in their study, the four main diffusion model parameters showed a similar performance (across-session coefficients between .645 and .736) while in our LDT study drift rate and threshold separation outperformed the other parameters.

These differences might be attributable to the differently complex models used for parameter estimation. Specifically, Yap et al. (2012) estimated all parameters of the Ratcliff Diffusion Model while we fixed the intertrial variability of v and z_r to zero. More restrained parameter models can—despite their wrong fixations of parameters—lead to better estimates of the main diffusion model parameters (Lerche & Voss, 2016; van Ravenzwaaij et al., 2016). Besides, Yap et al. (2012) had a very high number of trials (n > 3,000) while in our study the LDT was comprised of only 400 trials. As our findings and previous findings (Lerche et al., 2015) demonstrate, the CS criterion—also employed by Yap and colleagues—requires higher numbers of trials in order to supply good parameter estimates than other estimation methods. Thus, the parameter estimation procedure employed by Yap and colleagues might not be adequate for moderately sized data sets as in our studies. Another difference between their study and our study lies in the employed stimulus material. In contrast to Yap et al. (2012), we used the same stimuli at both sessions.

To exemplarily demonstrate the influence of model complexity, in a set of further analyses (see also Lerche & Voss, 2016) we used a more complex parameter model in which we estimated all intertrial variabilities (thus, all seven different diffusion model parameters were estimated instead of only five parameters) using both CS and ML. Besides, we estimated parameters both for the total number of trials of the Lexical Decision Task (i.e., 400) and—in order to have a trial number comparable to the number of trials by Yap et al. (2012)—for the condition with 5,000 trials of the LDT simulation study.

For 400 trials, using the seven-parameter model (7-PM) instead of the five-parameter model (5-PM) leads to mainly worse parameter estimates. Most importantly, the coefficient of the drift parameter (v_{total}) decreases to a large amount for both CS (.69 [5-PM] vs. .39 [7-PM]) and ML (.86 [5-PM] vs. .52 [7-PM]). For the threshold separation also lower values were found for the seven-parameter model for CS (.70 [5-PM] vs. .64 [7-PM]) and ML (.79 [5-PM] vs. .71 [7-PM]). For t_0, only for ML the five-parameter model performed better than the seven-parameter model (.54 [5-PM] vs. .45 [7-PM]) and no difference was observed for CS. Finally, for z_r there were negligible differences, with the seven-parameter model performing slightly better (.49 [5-PM] vs. .53[7-PM]) for CS and slightly worse for ML (.55 [5-PM] vs. .52 [7-PM]). In sum, the five-parameter model performs better for the main diffusion model parameters than the full diffusion model. In particular, if one is interested in the drift rate, it is advisable to use the less complex model.

While for small to moderate trial numbers the complexity of the parameter model is crucial, it is less for high trial numbers (see also Lerche & Voss, 2016). For the condition with 5,000 trials all retest coefficients were very high (min = .90, max = .99) and thus, there were smaller differences between parameters, estimation methods and between the two differently complex parameter models. Yet the biggest difference between models is observed for the drift rate estimated by CS. Here the five-parameter model performs better (.97) than the seven-parameter model (.91).

Accordingly, the fact that Yap and colleagues observed similar retest coefficients for all four main diffusion model parameters might be mostly influenced by two aspects. (1) They used a higher number of trials (3,000 vs. 400). As we exemplarily demonstrated for our condition with 5,000 trials, differences between parameters and estimation procedures are clearly smaller for such high trial numbers. (2) They used a more complex parameter model inclusive of all three intertrial variabilities while we fixed two of them to zero. It might be that in the analysis of their data, drift rate and threshold separation would have had even higher values if they had used the less complex five-parameter model.

As the previous sections demonstrate, there is no single procedure of diffusion modeling and different procedures can produce different results. To sum up, the three most important methodological findings resulting from our studies are the following: First, enhancing the number of trials seems only worthwhile up to around 200-400 trials. Second, ML and EZ showed the best retest reliability in all of our three tasks. Third, model complexity can influence results. Specifically, the fixation of the intertrial variabilities of starting point and drift seems to be a good strategy to obtain reliable parameter estimates.

Directions for future research

In our studies, participants were mostly students and thus, the variance in vocabulary knowledge and memory performance was probably quite restricted. This homogeneity of our sample lowers the retest reliability coefficients we could reach. The coefficients observed must therefore rather be taken as a lower bound and higher correlation coefficients both between sessions and across tasks are to be expected for more heterogeneous samples. We hope that in future studies several other paradigms will be tested using different samples. One more aspect that might lower retest coefficients are differential learning effects. As a comparison of the first with the second session revealed, parameters changed over time (e.g., drift rate increased from the first to the second session). Our results are in line with findings from Dutilh et al. (2009) who also observed practice effects (see also Petrov, Van Horn, & Ratcliff, 2011). In addition to these general changes, interindividual differences in changes over sessions are plausible and they can contribute to lower retest coefficients. In future studies, it would be interesting to analyze learning effects in more detail (e.g., using more than two sessions). Importantly, despite these factors that can weaken the retest coefficients, drift rate and threshold separation still showed satisfying values in all three tasks (*r*s > .70).

Conclusions

The present test-retest studies analyzed reliability of diffusion model estimation and stability of cognitive processes using three experimental paradigms (lexical decision task, recognition memory task, associative priming task). Results show that the main parameters of the diffusion model (specifically, the drift rate and the threshold separation) reflect stable interindividual differences. Accordingly, a potential use of the diffusion model as diagnostic tool is further promoted. Besides, we present information on the applicability of the diffusion model (e.g., regarding optimization criteria and requisite trial numbers).

References

Allen, P. A., Lien, M.-C., Ruthruff, E., & Voss, A. (2014). Multitasking and aging: Do older adults benefit from performing a highly practiced task? *Experimental Aging Research, 40*(3), 280-307. doi: 10.1007/s00426-014-0608-y

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*(5), 882-898. doi: 10.1007/s00426-014-0608-y

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2015). A Diffusion Model Analysis of Episodic Recognition in Preclinical Individuals With a Family History for Alzheimer's Disease: The Adult Children Study. *Neuropsychology*. doi: 10.1037/neu0000222

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM). Linguistic Data Consortium*. Philadelphia, PA: University of Pennsylvania.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459.

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review, 71*, 117-120.

Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren Inventar: nach Costa u. McCrae; NEO-FFI*: Hogrefe, Verlag f. Psychologie.

Bowen, H. J., Spaniol, J., Patel, R., & Voss, A. (2015). A Diffusion Model Analysis of Decision Biases Affecting Delayed Recognition of Emotional Stimuli. *Manuscript submitted for publication.*

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review, 16*(6), 1026-1036. doi: 10.3758/16.6.1026

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General, 126*(3), 228-247. doi: 10.1037/0096-3445.126.3.228

Klauer, K. C., & Dittrich, K. (2010). From sunshine to double arrows: An evaluation window account of negative compatibility effects. *Journal of Experimental Psychology: General, 139*(3), 490.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8*. University of Florida, Gainesville, FL.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*(7), 651-687.

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Manuscript submitted for publication.*

Lerche, V., Voss, A., & Nagler, M. (2015). How Many Trials are Required for Robust Parameter Estimation in Diffusion Modeling? Comparison of Different Estimation Algorithms. *Manuscript submitted for publication.*

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27*(2), 193-200. doi: 10.1037/a0031533

Neumann, R., & Strack, F. (2000). Approach and avoidance: the influence of proprioceptive and exteroceptive cues on encoding of affective information. *Journal of Personality and Social Psychology, 79*(1), 39.

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion, 13*(4), 739-747. doi: 10.1037/a0031628

Petrov, A. A., Van Horn, N. M., & Ratcliff, R. (2011). Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. *Psychonomic Bulletin & Review, 18*(3), 490-497. doi: 10.3758/s13423-011-0079-8

R Development Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org/.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*(4), 237-279. doi: 10.1037/dec0000030

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review, 111*(1), 159-182. doi: 10.1037/0033-295x.111.1.159

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

Ratcliff, R., & McKoon, G. (2015). Aging Effects in Item and Associative Recognition Memory for Pictures and Words. *Psychology and Aging*. doi: 10.1037/pag0000030

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

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin and Review, 7*(1), 1-25.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278. doi: 10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16*(2), 323.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*(4), 408-424. doi: 10.1016/j.jml.2003.11.002

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. doi: 10.1037/0882-7974.22.1.56

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*(3), 464-487. doi: 10.1037/a0023810

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115-136. doi: http://dx.doi.org/10.1016/j.cognition.2014.12.004

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481. doi: 10.3758/bf03196302

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology - General, 136*(3), 414-429. doi: 10.1037/0096-3445.136.3.414

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. doi: 10.1037/a0026003

Schmitz, F., & Voss, A. (2014). Components of task switching: A closer look at task switching and cue switching. *Acta Psychologica, 151*, 184-196. doi: 10.1016/j.actpsy.2014.06.009

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence, 51*, 28-46. doi: 10.1016/j.intell.2015.05.002

Spaniol, J., Madden, D. J., & Voss, A. (2006). A Diffusion Model Analysis of Adult Age Differences in Episodic and Semantic Long-Term Memory Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 101-117. doi: 10.1037/0278-7393.32.1.101

Spaniol, J., Voss, A., & Grady, C. L. (2008). Aging and emotional memory: Cognitive mechanisms underlying the positivity effect. *Psychology and Aging, 23*(4), 859-872. doi: 10.1037/a0014218

Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tüscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity. *Journal of Experimental Psychology: General, 143*(2), 850-886. doi: 10.1037/a0033981

Stolz, J., Besner, D., & Carr, T. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition, 12*(2), 284-336.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D., & Beloch-Till, H. (1985). *Deutsche personality research form*: Hogrefe.

Vallesi, A., Canalaz, F., Balestrieri, M., & Brambilla, P. (2015). Modulating speed-accuracy strategies in major depression. *Journal of Psychiatric Research, 60*, 103-108. doi: 10.1016/j.jpsychires.2014.09.017

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2016). The EZ Diffusion Model Provides a Powerful Test of Simple Empirical Effects. *Manuscript submitted for publication*.

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*(6), 463-473. doi: 10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. A practical introduction. *Experimental Psychology, 60*(6), 385-402.

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology, 44*(4), 1048-1056. doi: 10.1016/j.jesp.2007.10.009

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, 142*(2), 536-559. doi: 10.1037/a0029459

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 539-555. doi: 10.1348/000711009x477581

Voss, A., Voss, J., & Lerche, V. (2015). Assessing Cognitive Processes with Diffusion Model Analyses: A Tutorial based on fast-dm-30. *Frontiers in Psychology, 6*, 336. doi: 10.3389/fpsyg.2015.00336

Wagenmakers, E.-J., Grasman, R. P. P. P., & Molenaar, P. C. M. (2005). On the relation between the mean and the variance of a diffusion model response time distribution. *Journal of Mathematical Psychology, 49*(3), 195-204. doi: 10.1016/j.jmp.2005.02.003

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*(1), 140-159. doi: 10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15*(6), 1229-1235. doi: 10.3758/pbr.15.6.1229

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*(1), 3-22. doi: 10.3758/bf03194023

Weigard, A., & Huang-Pollock, C. (2014). A diffusion modeling approach to understanding contextual cueing effects in children with ADHD. *Journal of Child Psychology & Psychiatry, 55*(12), 1336-1344. doi: 10.1111/jcpp.12250

Wessa, M., Kanske, P., Neumeister, P., Bode, K., Heissler, J., & Schönfelder, S. (2010). EmoPics: Subjektive und psychophysiologische Evaluation neuen Bildmaterials für die klinisch-biopsychologische Forschung. *Zeitschrift für Klinische Psychologie und Psychotherapie, Supplementum 1/11, 77.*

White, C. N., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion-model analysis. *Cognition and Emotion, 23*(1), 181-205. doi: 10.1080/02699930801976770

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010a). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion, 10*(5), 662-677. doi: 10.1037/a0019474

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010b). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*(1), 39-52. doi: 10.1016/j.jmp.2010.01.004

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and individual differences, 30*(4), 669-689.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics, 7*, 14. doi: 10.3389/fninf.2013.00014

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

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 597-613. doi: 10.1037/xlm0000064

Table 1

Comparison of parameters from Session 1 with Session 2 (Study 1)

	Session 1		Session 2				95 % *CI*		
Parameter	*M*	*SD*	*M*	*SD*	t^a	*p*	*LL*	*UL*	Cohen's d_z
Lexical decision task									
a	1.26	0.24	1.16	0.21	7.09	<.001	0.07	0.12	0.69
$v_{non\text{-}word}$	-2.86	0.54	-2.95	0.64	2.14	<.05	0.01	0.17	0.21
v_{word}	2.09	0.53	2.47	0.64	-8.48	<.001	-0.47	-0.29	-0.83
v_{total}	4.94	0.92	5.41	1.11	-8.26	<.001	-0.58	-0.36	-0.81
v_{bias}	-0.77	0.55	-0.48	0.65	-4.47	<.001	-0.42	-0.16	-0.44
t_0	0.459	0.038	0.450	0.044	2.02	<.05	0.000	0.017	0.20
z_r	0.53	0.05	0.50	0.06	6.08	<.001	0.02	0.04	0.60
s_{t0}	0.14	0.06	0.15	0.08	-0.90	.37	-0.02	0.01	-0.09
Mean RT	694.69	87.82	651.27	87.18	7.41	<.001	31.45	55.40	0.71
Accuracy	94.20	3.60	94.17	4.03	-0.51	.61	-0.30	0.35	0.02
Recognition memory task									
a	1.42	0.27	1.30	0.25	6.02	<.001	0.08	0.15	0.59
v_{new}	-2.03	0.59	-2.40	0.83	5.51	<.001	0.23	0.49	0.54
v_{old}	1.73	0.94	2.29	1.08	-6.71	<.001	-0.72	-0.39	-0.66
v_{total}	3.77	1.34	4.69	1.72	-8.03	<.001	-1.15	-0.69	-0.79
v_{bias}	-0.30	0.83	-0.11	0.87	-2.00	<.05	-0.38	0.00	-0.20
t_0	0.588	0.048	0.536	0.044	10.43	<.001	0.042	0.062	1.03
z_r	0.54	0.07	0.54	0.07	-0.23	.82	-0.02	0.01	-0.02
s_{t0}	0.14	0.08	0.13	0.08	0.98	.33	-0.01	0.02	0.10
Mean RT	898.73	103.82	784.68	90.05	16.31	<.001	99.51	128.58	1.53
Accuracy	90.17	7.82	91.88	7.63	-4.23	<.001	-2.66	-0.76	-0.35

Note. Results are based on parameter estimation with ML. In fast-dm the diffusion coefficient is set to 1. Multiply a, v and z_r by 0.1 to compare parameter ranges with those in studies with constant 0.1. *CI* confidence interval; *LL* lower limit; *UL* upper limit. $v_{total} = v_1$ (word/old picture) - v_0 (non-word/new picture) and $v_{bias} = v_1 + v_0$. *t* test for mean RT (accuracy rate) is based on logarithmized (arcsine-transformed) values, but *M*, *SD*, 95 %-*CI* and Cohen's d_z are based on the untransformed values.

[a] *df* = 103 for the lexical decision task and *df* = 102 for the recognition memory task.

Table 2

Across-task correlations of the corresponding diffusion model parameters from the lexical decision and recognition memory task, separated for Session 1 and Session 2

	a	v_0	v_1	v_{total}	v_{bias}	t_0	z_r	s_{t0}
				Session 1				
KS	.48***	-.03	.08	.06	.01	.25*	.00	.25**
ML	.51***	.11	.20*	.17	.04	.27**	.24*	.19
CS	.32***	.06	.11	.07	.14	.27**	.16	.24*
EZ	.51***	.18	.32***	.30**	.07	.20*	–	–
				Session 2				
KS	.44***	.18	.26**	.22**	.21*	.54***	.16	.37***
ML	.49***	.34***	.27**	.35***	.12	.59***	.38***	.38***
CS	.39***	.33***	.16	.27**	.00	.44***	.31**	.11
EZ	.47***	.34***	.39***	.42***	.18	.34***	–	–

Note. N = 104 for the first session and N = 103 for the second session. Non-words and new pictures were associated with the lower (v_0) and words and old pictures with the upper threshold (v_1).

*p < .05; **p < .01; ***p < .001.

Table 3

Priming effects (Study 2)

	associated		nonassociated				95 % *CI*		
Parameter	*M*	*SD*	*M*	*SD*	*t*(127)	*p*	*LL*	*UL*	Cohen's d_z
					Target = word				
mean RT	789.97	110.85	814.84	105.70	-9.84	<.001	-31.07	-18.66	-0.70
	723.18	86.99	745.01	90.14	-9.65	<.001	-26.95	-16.71	-0.75
accuracy	95.31	3.69	92.06	5.08	9.71	<.001	2.62	3.89	0.90
	94.98	4.49	92.02	5.67	9.30	<.001	2.30	3.63	0.78
v	2.22	0.55	1.84	0.44	10.94	<.001	0.31	0.45	0.97
	2.53	0.61	2.14	0.55	10.91	<.001	0.31	0.45	0.96
t_0	0.526	0.041	0.529	0.043	-1.91	.06	-0.007	0.000	-0.17
	0.511	0.042	0.514	0.042	-1.84	.07	-0.007	0.000	-0.16
					Target = non-Word				
mean RT	891.54	113.45	884.37	114.67	3.44	<.001	1.67	12.67	0.23
	793.82	92.00	785.07	90.76	4.87	<.001	4.18	13.31	0.34
accuracy	92.43	6.31	93.84	4.74	-3.73	<.001	-2.04	-0.77	-0.39
	92.82	6.46	93.71	6.07	-3.22	<.01	-1.48	-0.29	-0.26
v	-2.19	0.50	-2.30	0.52	3.62	<.001	0.05	0.16	0.32
	-2.42	0.58	-2.56	0.61	3.97	<.001	0.07	0.22	0.35
t_0	0.559	0.059	0.560	0.059	-0.28	.78	-0.008	0.006	-0.02
	0.536	0.052	0.540	0.048	-1.58	.12	-0.010	0.001	-0.14

Note. The first row always refers to the first, the second row to the second session. Results are based on parameter estimation with ML. In fast-dm the diffusion coefficient is set to 1. Multiply *v* by 0.1 to compare parameter values with those in studies with constant 0.1. *CI* confidence interval; *LL* lower limit; *UL* upper limit. *t* test for mean RT (accuracy rate) is based on logarithmized (arcsine-transformed) values, but *M*, *SD*, 95 %-*CI* and Cohen's d_z are based on the untransformed values.

Table 4

Comparison of parameters from Session 1 with Session 2 (Study 2)

	Session 1		Session 2				95 % *CI*		
Parameter	*M*	*SD*	*M*	*SD*	*t*(127)	*p*	*LL*	*UL*	Cohen's d
a	1.39	0.26	1.24	0.24	10.43	<.001	0.13	0.19	0.92
$v_{non\text{-}word,\ associated}$	-2.19	0.50	-2.42	0.58	5.18	<.001	0.14	0.31	0.46
$v_{non\text{-}word,\ nonassociated}$	-2.30	0.52	-2.56	0.61	5.61	<.001	0.17	0.36	0.50
$v_{word,\ associated}$	2.22	0.55	2.53	0.61	-6.65	<.001	-0.40	-0.22	-0.59
$v_{word,\ nonassociated}$	1.84	0.44	2.14	0.55	-7.64	<.001	-0.38	-0.22	-0.67
v_{total}	4.27	0.82	4.82	0.96	-9.64	<.001	-0.67	-0.44	-0.85
v_{bias}	-0.22	0.46	-0.16	0.53	-1.13	.26	-0.16	0.04	-0.10
$t_{0\ non\text{-}\mathrm{word},\ associated}$	0.559	0.059	0.536	0.052	4.94	<.001	0.014	0.032	0.44
$t_{0\ non\text{-}\mathrm{word},\ nonassociated}$	0.560	0.059	0.540	0.048	4.08	<.001	0.010	0.029	0.36
$t_{0\ \mathrm{word},\ associated}$	0.526	0.041	0.511	0.041	4.35	<.001	0.008	0.021	0.38
$t_{0\ \mathrm{word},\ nonassociated}$	0.529	0.043	0.514	0.042	4.27	<.001	0.008	0.021	0.38
z_r	0.56	0.06	0.54	0.06	3.01	<.01	0.01	0.03	0.27
s_{t0}	0.14	0.08	0.12	0.06	3.26	<.01	0.01	0.03	0.29
mean RT	844.64	107.08	761.41	86.69	15.88	<.001	72.46	94.00	1.35
accuracy	93.41	4.19	93.38	4.99	-0.62	.54	-0.42	0.48	0.01

Note. Results are based on parameter estimation with ML. In fast-dm the diffusion coefficient is set to 1. Multiply a, v and z_r by 0.1 to compare parameter ranges with those in studies with constant 0.1. *CI* confidence interval; *LL* lower limit; *UL* upper limit. $v_{total} = 0.5 \times ((v_{word,\ associated} + v_{word,\ nonassociated}) - (v_{non\text{-}word,\ associated} + v_{non\text{-}word,\ nonassociated}))$ and $v_{bias} = 0.5 \times ((v_{word,\ associated} + v_{word,\ nonassociated}) + (v_{non\text{-}word,\ associated} + v_{non\text{-}word,\ nonassociated}))$. *t* test for mean RT (accuracy rate) is based on logarithmized (arcsine-transformed) values, but *M*, *SD*, 95 %-*CI* and Cohen's d_z are based on the untransformed values.

Table 5

Means, standard deviations and intercorrelations of parameters used for the simulation of data sets

				Lexical decision task							
			M (*SD*)	1	2	3	4	5	6	7	8
	M (*SD*)		–	1.42 (0.32)	-4.01 (1.13)	3.10 (1.11)	0.48 (0.04)	0.53 (0.06)	0.15 (0.05)	1.34 (0.64)	0.37 (0.25)
Recognition memory task	1.	a	1.60 (0.36)	–	-.15	.16	.15	.11	-.14	.33	-.04
	2.	v_0	-3.07 (1.14)	-.17	–	-.85	-.22	.00	.34	-.85	-.48
	3.	v_1	2.44 (1.20)	.39	-.59	–	.08	-.18	-.36	.80	.58
	4.	t_0	0.61 (0.05)	-.13	.04	-.16	–	-.09	.00	.30	.19
	5.	z_r	0.55 (0.08)	.17	-.26	-.01	-.05	–	-.08	-.13	-.28
	6.	s_{t0}	0.17 (0.08)	-.38	-.05	-.21	.64	-.10	–	-.31	-.18
	7.	s_v	1.41 (0.83)	.15	-.72	.32	.17	.17	.25	–	.49
	8.	s_{zr}	0.15 (0.22)	-.08	-.14	.09	.09	-.45	.09	.14	–

Note. The values are based on the first session and an estimation of the 7-parameter model with ML (with exclusion of one participant due to a significant Mahalanobis distance computed based on all parameters). Values for the lexical decision task are presented above and for the recognition memory task below the diagonal. Words and old pictures were associated with the upper (v_1) and non-words and new pictures with the lower threshold (v_0). Multiply *a, v,* z_r, s_v, and s_{zr} by 0.1 to compare parameter ranges with those in studies with constant 0.1.

Table 6

Across-session and within-session reliability coefficients

	a	v_{total}	t_0	z_r
		Lexical decision task		
Across-session	.77	.82	.49	.44
Within-session				
Session 1	.84	.82	.71	.52
Session 2	.74	.78	.77	.61
		Recognition memory task		
Across-session	.53	.69	.44	.38
Within-session				
Session 1	.65	.71	.59	.53
Session 2	.67	.73	.65	.42
		Associative Priming Task		
Across-session	.74	.69	.56	.33
Within-session				
Session 1	.81	.76	.70	.34
Session 2	.82	.80	.79	.48

Note. Results are based on parameter estimation with ML. N = 104 for the first session and N = 103 for the second session in Study 1, N = 128 in Study 2. The number of trials is 200 for the lexical decision task and associative priming task and 100 for the recognition memory task.

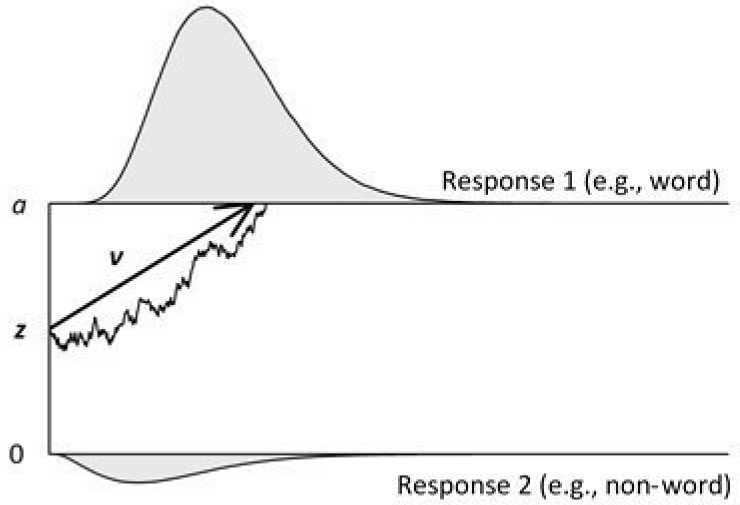


Figure 1. Example of a decision process according to the Ratcliff Diffusion Model. The process starts at *z*, moves with drift *v* and ends at the upper threshold (associated with Response 1).

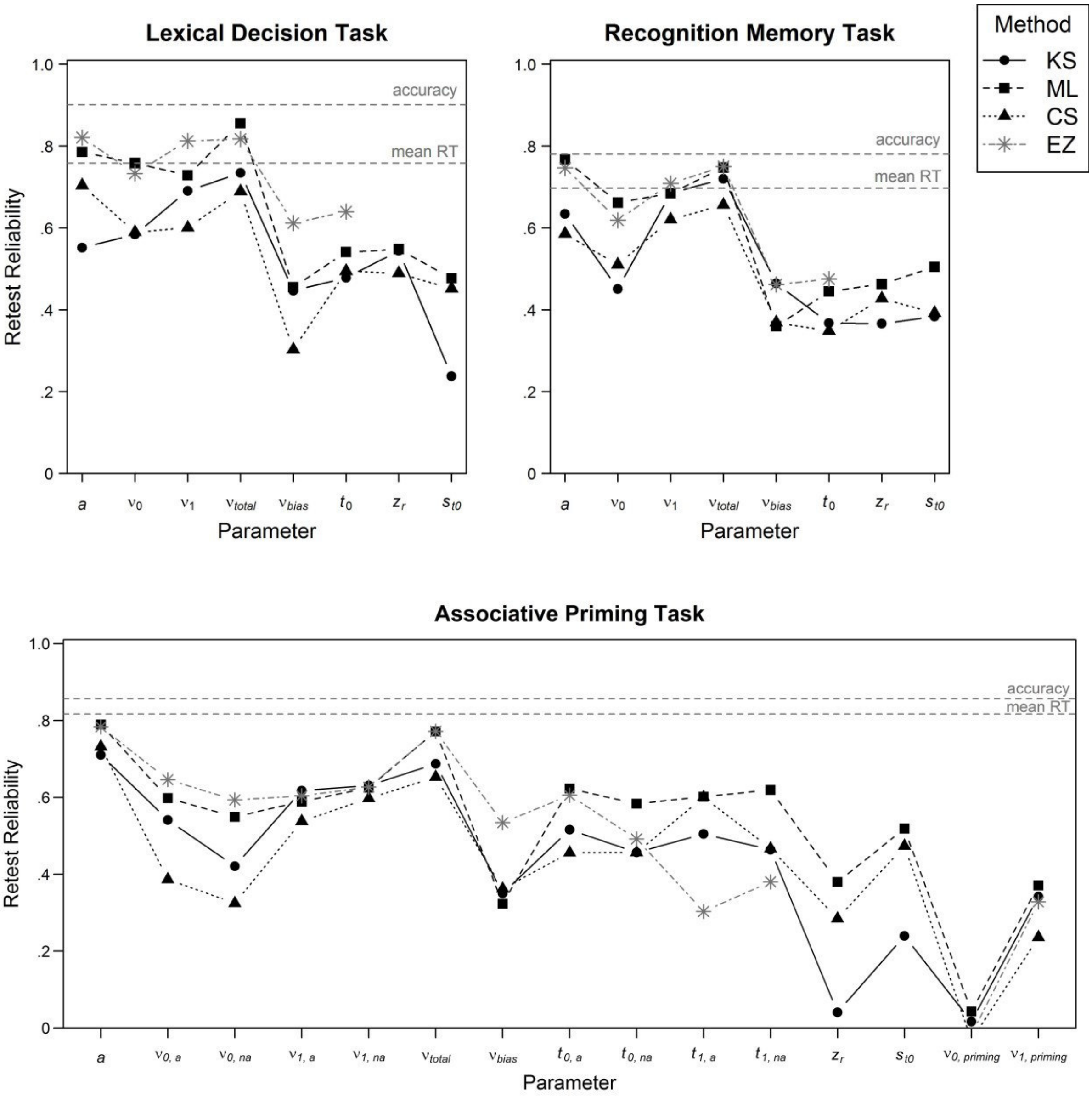


Figure 2. Retest reliability of the diffusion model parameters depending on the task and the method and retest reliability of logarithmized mean RT of correct responses and of arcsine-transformed accuracy rate. For the drift rates the index 0 (1) is used for non-words (words) in the lexical decision task and associative priming task and new (old) pictures in the recognition memory task. In the associative priming task, the index a (na) refers to associated (nonassociated) prime-target pairs and the index "priming" refers to the difference between associated and nonassociated prime-target pairs. Reliability coefficients are based on 400 trials for the lexical decision task and associative priming task and 200 trials for the recognition memory task.

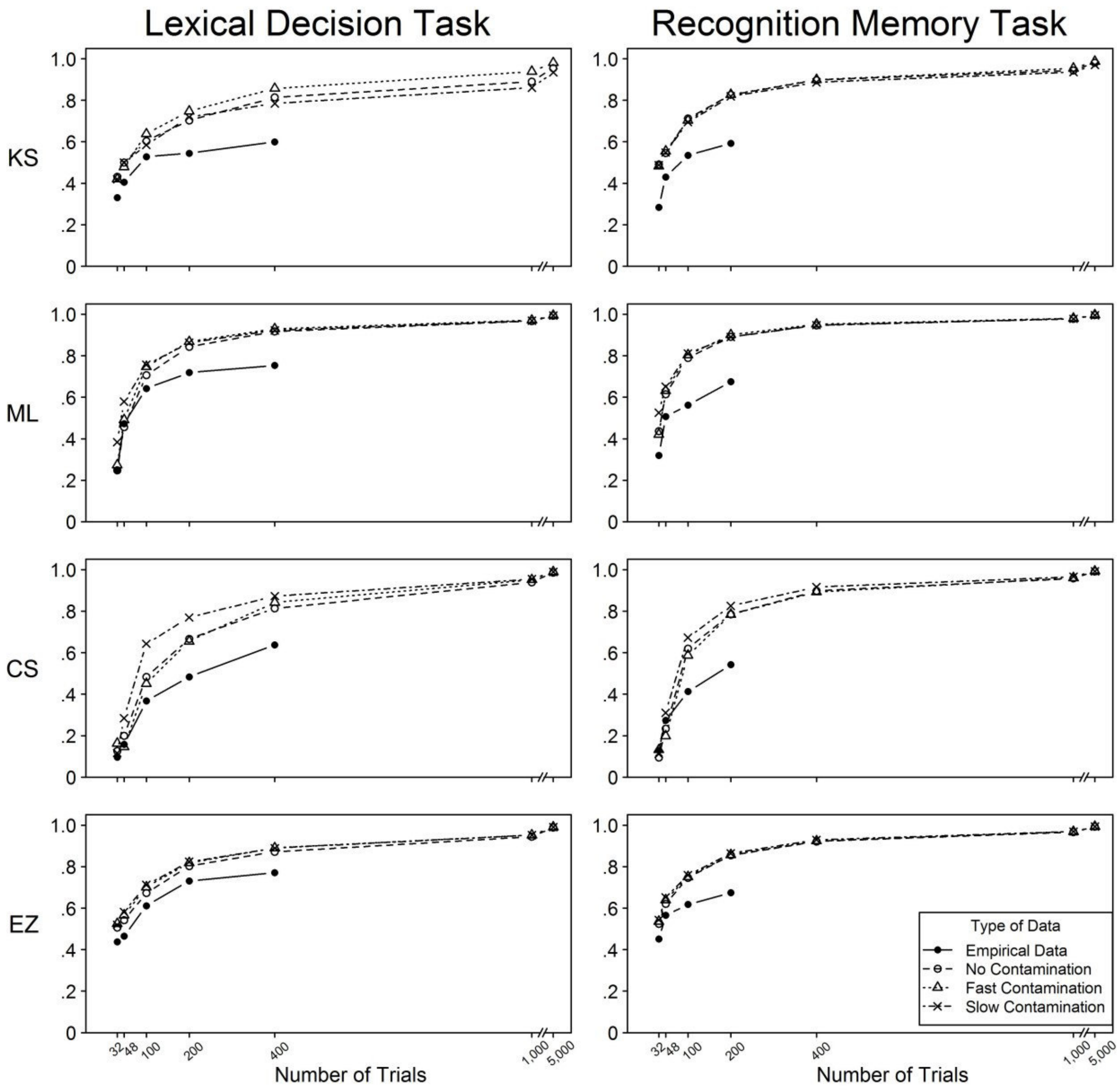


Figure 3. Mean retest reliability (over three of the main parameters of the Ratcliff Diffusion Model: *a*, v_{total}, t_0) for empiric data of Study 1 and simulated data sets, depending on the task, method and number of trials and—for the simulated data sets—the presence and type of contamination (no/fast/slow contamination).

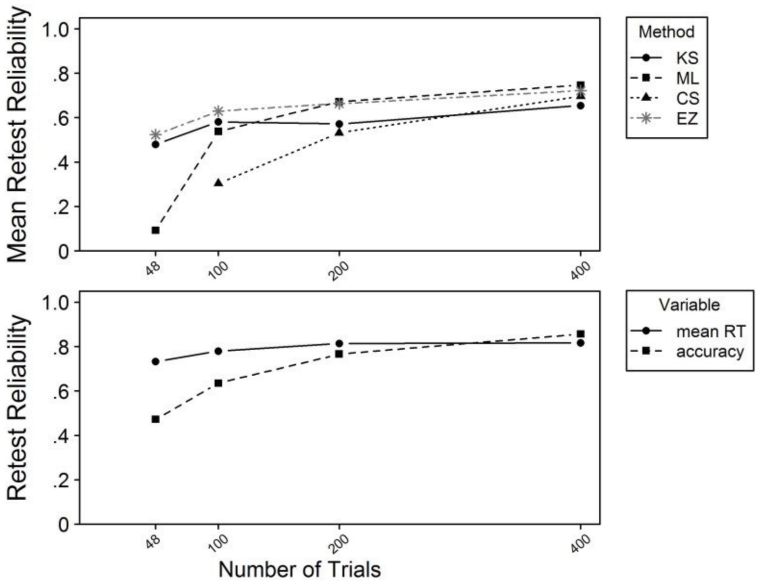


Figure 4. Mean retest reliability (over three of the main parameters of the Ratcliff Diffusion Model: *a*, v_{total} and t_0) for empiric data sets of Study 2, depending on the method and number of trials (*top row*) and retest reliability of mean RT and accuracy rate depending on the number of trials (*bottom row*).

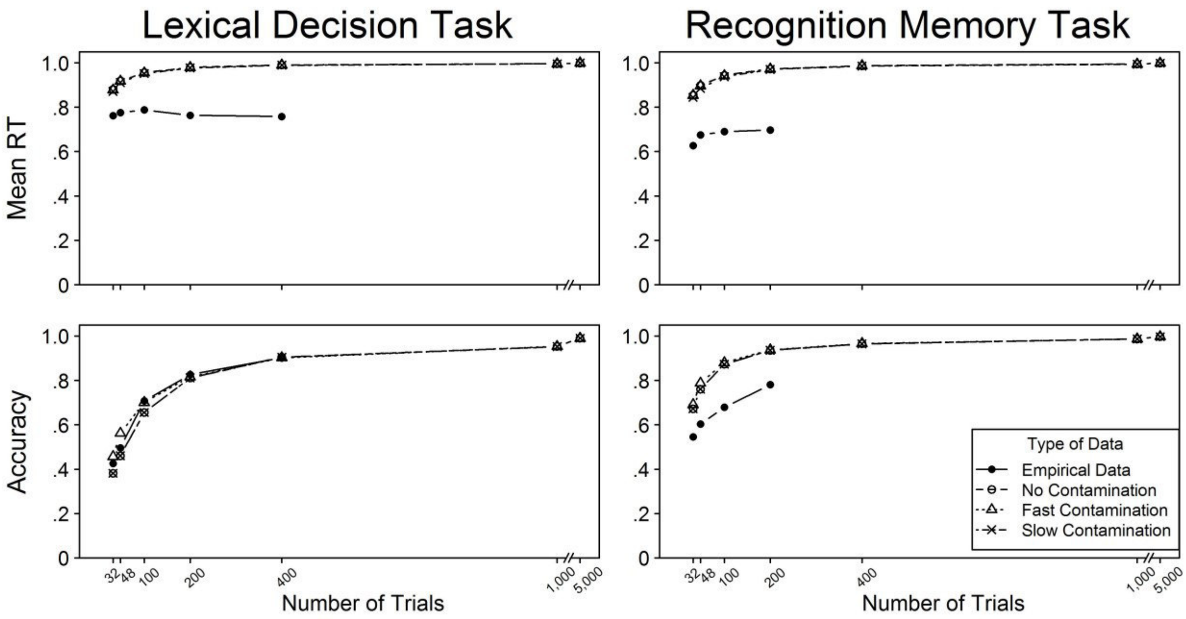


Figure 5. Retest reliability of mean RT and accuracy rate for empiric data of Study 1 and simulated data sets, depending on the task and number of trials and—for the simulated data sets—the presence and type of contamination (no/fast/slow contamination).

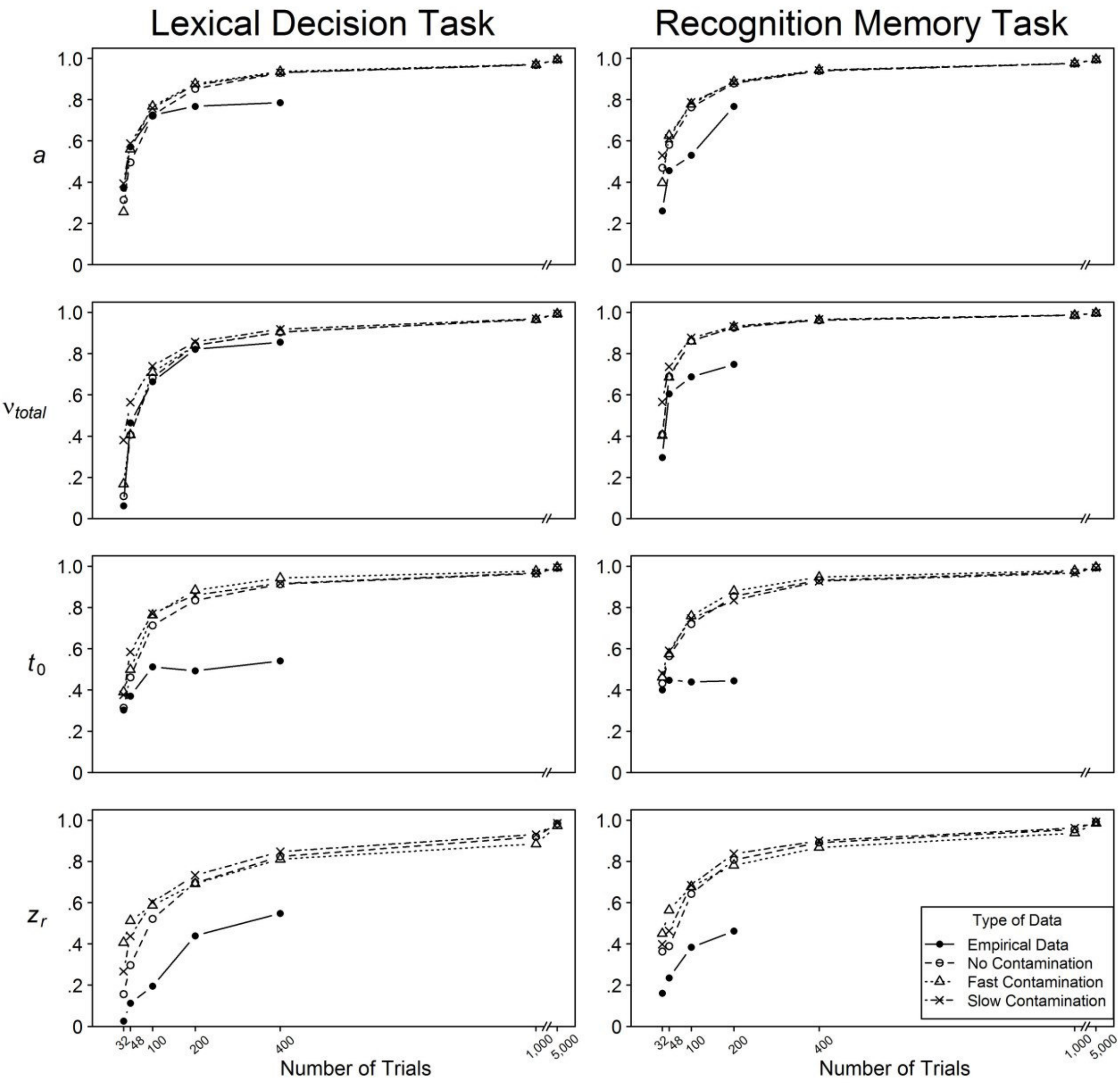


Figure 6. Retest reliability of the four main diffusion model parameters for empiric data sets of Study 1 and simulated data sets, depending on the task and number of trials and—for the simulated data sets—the presence and type of contamination (no/fast/slow contamination). The parameters were estimated with ML.

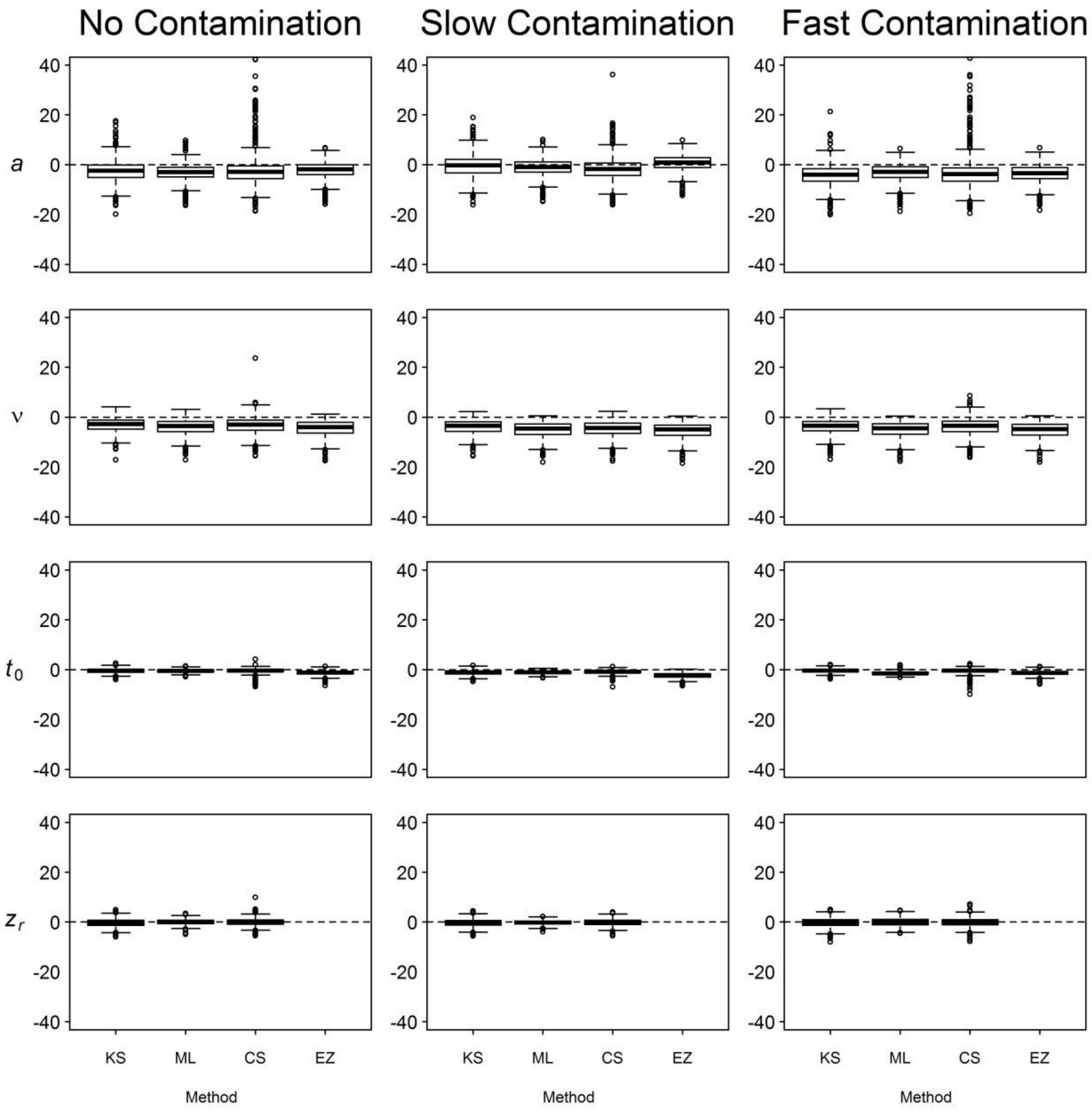


Figure 7. Boxplots of the differences between estimated and true parameters values of the LDT simulation study, depending on the type of contamination, the parameter and method. Based on data from Session 1 and on the first 200 trials of the task. Boxplots show the first, second and third quartile. Outliers are any values greater than 1.5 times the interquartile range from either end of the box. For better comparability of parameters, we report ν (= ν_{total} × 0.5) and we weighted the differences between estimated and true parameter values against parameter accuracies reported in Lerche et al. (2015). Outliers exceeding a weighted difference of ± 40 are not depicted.

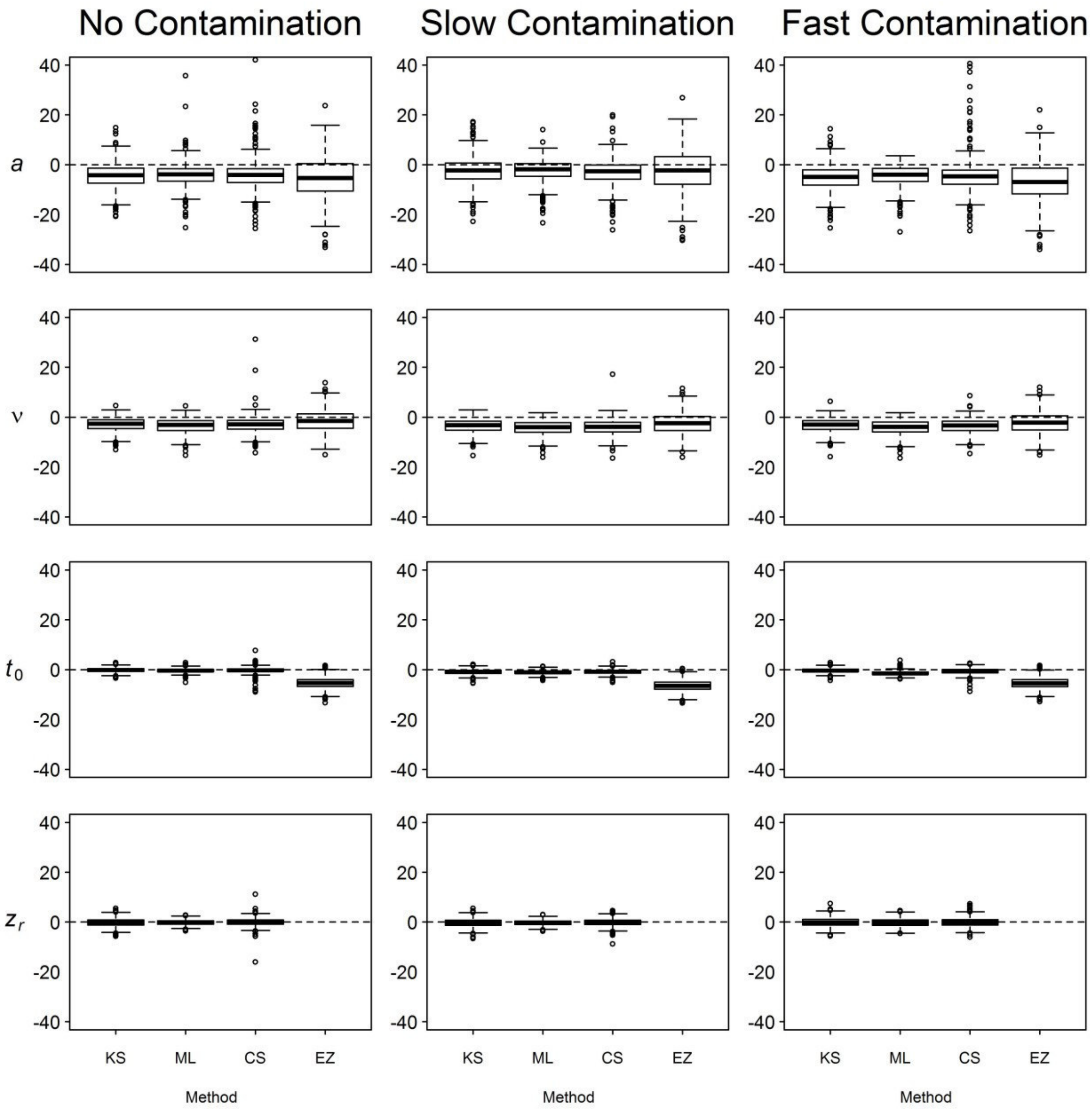


Figure 8. Boxplots of the differences between estimated and true parameters values of the RMT simulation study, depending on the type of contamination, the parameter and method. Based on data from Session 1. Boxplots show the first, second and third quartile. Outliers are any values greater than 1.5 times the interquartile range from either end of the box. For better comparability of parameters, we report ν (= ν_{total} × 0.5) and we weighted the differences between estimated and true parameter values against parameter accuracies reported in Lerche et al. (2015). Outliers exceeding a weighted difference of ± 40 are not depicted.

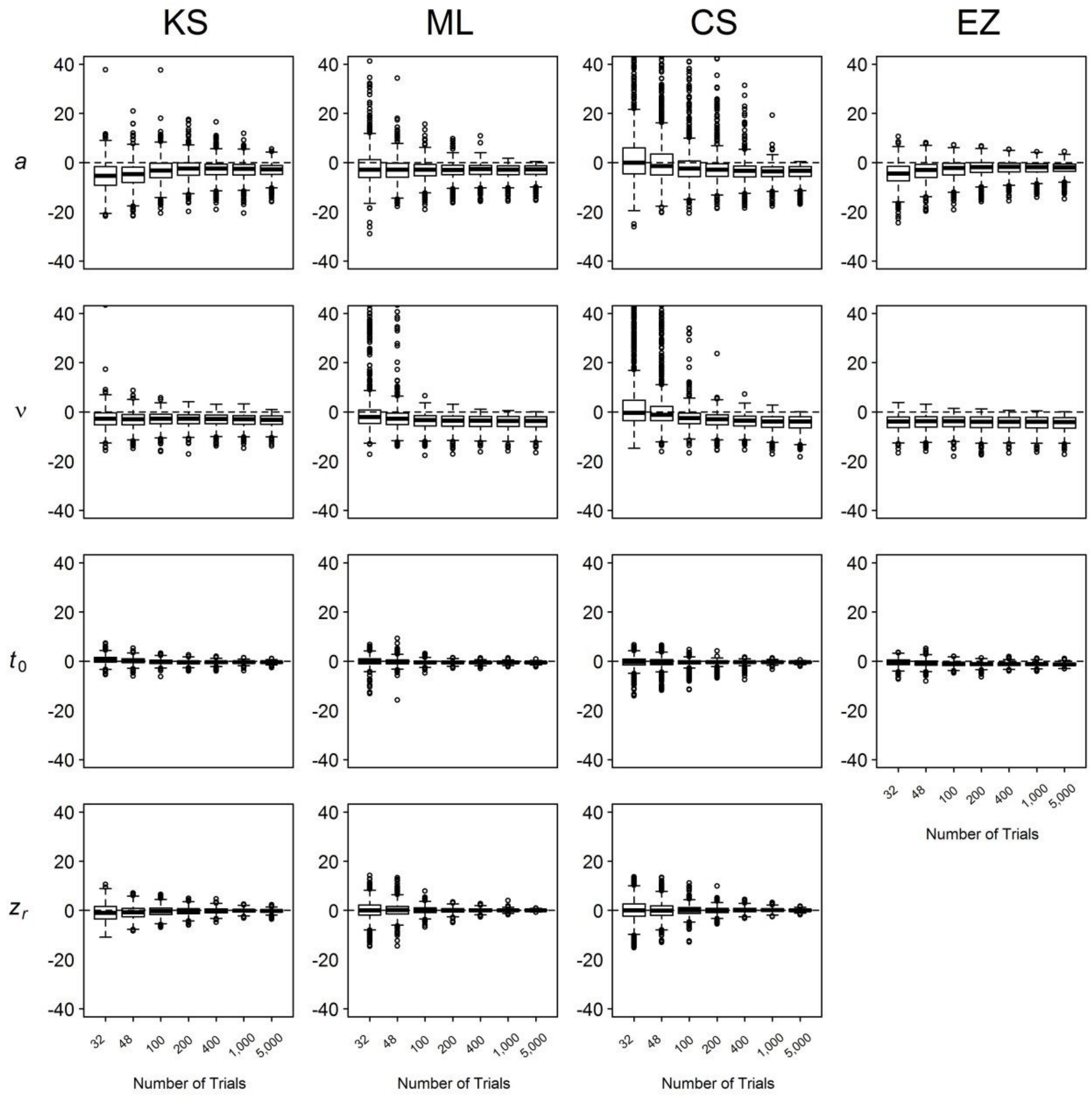


Figure 9. Boxplots of the differences between estimated and true parameters values of the LDT simulation study, depending on the parameter, method and number of trials. Based on data from Session 1 and the condition with no contaminants. Boxplots show the first, second and third quartile. Outliers are any values greater than 1.5 times the interquartile range from either end of the box. For better comparability of parameters, we report v (= v_{total} × 0.5) and we weighted the differences between estimated and true parameter values against parameter accuracies reported in Lerche et al. (2015). Outliers exceeding a weighted difference of ± 40 are not depicted.

Compliance with Ethical Standards

- This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant Nr. VO1288/2-1).
- *Ethical approval*: “All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.“
- *Informed consent*: “Informed consent was obtained from all individual participants included in the study.”

Appendix A 5

Manuscript 5: Lerche, V., & Voss, A. (2016). Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious. *Frontiers in Psychology*, *7*(1324).

doi: 10.3389/fpsyg.2016.01324

Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious

Veronika Lerche and Andreas Voss
Ruprecht-Karls-Universität Heidelberg

Author Note

Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany; Andreas Voss, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany.

This research was supported by a grant from the German Research Foundation to Andreas Voss (Grant Nr. VO1288/2-2).

Correspondence concerning this article should be addressed to Veronika Lerche, Psychologisches Institut, Ruprecht-Karls-Universität Heidelberg, Hauptstrasse 47-51, D-69117 Heidelberg, Germany, email: veronika.lerche@psychologie.uni-heidelberg.de, telephone: +49-6221-54-7322.

ABSTRACT

The diffusion model (Ratcliff, 1978) takes into account the reaction time distributions of both correct and erroneous responses from binary decision tasks. This high degree of information usage allows the estimation of different parameters mapping cognitive components such as speed of information accumulation or decision bias. For three of the four main parameters (drift rate, starting point and non-decision time) trial-to-trial variability is allowed. We investigated the influence of these variability parameters both drawing on simulation studies and on data from an empirical test-retest study using different optimization criteria and different trial numbers. Our results suggest that less complex models (fixing intertrial variabilities of the drift rate and the starting point at zero) can improve the estimation of the psychologically most interesting parameters (drift rate, threshold separation, starting point and non-decision time).

Keywords: diffusion model, *fast-dm*, parameter estimation, mathematical models, reaction time methods

Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious

The diffusion model (Ratcliff, 1978) is a popular mathematical model that recently attracted the attention of researchers of diverse fields of psychology (see Voss, Nagler, & Lerche, 2013, for a recent review; see for example Brown & Heathcote, 2008, for another popular sequential sampling model). The model provides information about the cognitive processes underlying binary decision tasks. This becomes possible because the diffusion model parameters validly map specific latent cognitive processes (e.g., speed of information accumulation, decision bias). Despite the increased popularity of the diffusion model, there is a lack of research investigating how different model specifications influence the quality of the parameter estimation (see Lerche, Voss, & Nagler, 2016, for an exception). In particular, little to no information is available on the costs and benefits of model complexity. While the basic diffusion model (Ratcliff, 1978) comprises only four parameters, Ratcliff and Rouder (1998) and Ratcliff and Tuerlinckx (2002) suggested that it may be necessary to allow for intertrial variability of parameter values, because psychological processes (such as expectations or attention) will shift from trial to trial. This led to the inclusion of three so-called intertrial variability parameters.

Since then, these additional parameters have been estimated in almost all published diffusion model studies (e.g., Allen, Lien, Ruthruff, & Voss, 2014; Ratcliff, Thapar, & McKoon, 2004; Spaniol, Voss, & Grady, 2008; van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014; Yap, Balota, Sibley, & Ratcliff, 2012), even if trial numbers were small to moderate (e.g., Metin et al., 2013). This might be problematic, because in this case the parameter estimation might become unstable.

The aim of the present article is to compare the performance of more parsimonious with more complex models. In doing so, we do not question the theoretic rationale of the intertrial variabilities. We are aware that in all applications there will be fluctuations in psychological processes. Nonetheless, we argue that sometimes the available data might not suffice to get reliable estimates for the full diffusion model. Thus, neglecting these fluctuations might lead to more accurate and stable results.

In the following sections, we first give a short introduction to the diffusion model. Then, we elaborate on necessary choices regarding estimation procedures and model specifications. Finally, we present data from a simulation study (Study 1) and from a test-retest study (Study 2).

Parameters of the diffusion model

The diffusion model can be applied to binary decision tasks (e.g., lexical decision tasks [LDTs], or perceptual tasks such as color discrimination). One central supposition is that information is accumulated continuously and that this accumulation process ends as soon as one of two thresholds is reached. Each threshold is associated with one of the two responses of the binary task (or, alternatively, with correct vs. erroneous responses). Figure 1 shows an example of such a decision process.

The four parameters of the basic diffusion model are the (1) drift rate (v), (2) threshold separation (a), (3) starting point (z), and (4) non-decision time (t_0). The *drift rate* v informs about the speed and direction of information accumulation. Positive (negative) drift rates indicate an average slope of information accumulation toward the upper (lower) threshold. The absolute value of the drift rate is a measure of the speed of information uptake with higher values indicating faster accumulation. The drift rate can be interpreted as a measure of subjective task difficulty: (absolute) drift rates will be higher for easier tasks. The diffusion model assumes that information uptake is a stochastic (i.e., noisy) process. Thus, the process does not necessarily end at the same time or at the same threshold, even if the same information is available.

The *threshold separation* (a) represents the chosen response criterion. Higher distances go along with longer information uptake and fewer erroneous responses. While in Figure 1 the process is assumed to start in the center between the two thresholds, it might also start at a position closer to the upper or lower threshold. If the *starting point* z (or, $z_r = z/a$) is located closer to one of two thresholds, less evidence needs to be accumulated before the participant decides for this option.

Finally, to the time taken by the decision process (illustrated in Figure 1) adds the *non-decision time* t_0. It includes the duration of all processes that take place before (e.g., encoding of information) and after (e.g., motoric response execution) the decisional process. In most diffusion model studies one or more of these four parameters are in the focus of the research questions. Importantly, in several validation studies it was demonstrated that these parameters are sensitive to specific experimental manipulations, which supports the parameters' validity (e.g., Arnold, Bröder, & Bayen, 2015; Voss, Rothermund, & Voss, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

Ratcliff and Rouder (1998) suggest the inclusion of intertrial variabilities for two parameters, namely for the drift rate (s_v) and the starting point (s_{zr}) (see also Laming, 1968, for an earlier account on intertrial variability). An important advantage of including these intertrial variability parameters in the model is that they provide an explanation for differences in speed of correct responses and errors. Specifically, if the *drift rate* varies from trial to trial, the model predicts *slower errors than correct responses*. Imagine trials with a drift rate that is higher than the average drift rate. In this case, all responses (including errors) are fast while the error rate is low. A drift rate that is lower than the average, on the other hand, results in a higher percentage of errors which are slow. Thus, the intertrial variability of the drift causes the majority of errors to be slow. A pattern of *faster errors than correct responses* can be explained by intertrial variability of the *starting point*. A starting point that is close to the lower (error) threshold increases the number of errors and decreases the decision time for those. If, on the other hand, the starting point is closer to the upper threshold (associated with correct responses), errors are slow but rare.

Later, a third variability parameter was included into the model: the intertrial variability of the non-decision time (s_{t0}; Ratcliff & Tuerlinckx, 2002). A high intertrial variability of non-decision time accounts for a higher number of fast responses (i.e., the skew of the predicted RT distribution is reduced). Thereby, the model might also become less susceptible to the impact of fast contaminants. With the three intertrial variabilities, the diffusion model includes seven parameters (for a model with one further parameter, see Voss, Voss, & Klauer, 2010).

In most diffusion model studies intertrial variabilities are included not because they are important to answer a psychological research question, but rather to improve model fit and, possibly, to avoid a bias in the other parameters. In the present article, we test whether excluding the intertrial parameters derogates the estimation of the four main diffusion model parameters.

Necessary choices in estimation procedures and model specifications

In the first decades after the introduction of the diffusion model in 1978, the parameter estimation was restricted to researchers with sound mathematical and programming skills. Now, several user-friendly software solutions exist that enable any researcher to apply a diffusion model to their data. Amongst these programs are *EZ* (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008; Wagenmakers, van der Maas, & Grasman, 2007), *DMAT* (Vandekerckhove & Tuerlinckx, 2007, 2008), *fast-dm* (Voss & Voss, 2007, 2008; Voss, Voss, & Lerche, 2015), and *HDDM* (Wiecki, Sofer, & Frank, 2013). Even if these programs are easy to use, they require the users to make several choices in terms of the parameter estimation procedure (with the exception of *EZ* that works with closed-form equations and offers fewer degrees of freedom in model definition). One such choice regards the optimization criterion, another the complexity of the model (i.e., the number of estimated parameters).

Optimization Criterion

The diffusion model programs allow the choice between different optimization criteria. *Fast-dm-30* (Voss et al., 2015), for example, allows the choice between Kolmogorov-Smirnov (KS), a chi-square (CS) and a maximum likelihood (ML) based criterion. These criteria differ in the degree of usage of information with CS taking account of the least amount of information (RTs are grouped into bins) and ML using data from each single trial. On a continuum of information usage, with CS at the one end and ML at the other, KS can be positioned somewhere in between (see Voss et al., 2015, for a more detailed comparison of these three criteria). Related to information usage is the performance in parameter recovery. As a row of simulation studies by Lerche et al. (2016) shows, ML performs best, followed by KS and CS. The high efficiency of ML, however, comes with a cost: in the presence of fast contaminants (i.e., data not resulting from a diffusion process with the RTs situated at the lower tail of the distribution), the estimates obtained with ML are often severely biased. KS, on the other hand, turned out to be the least influenced by these contaminants.

Model Complexity

Most diffusion model programs allow an estimation of all seven parameters of the diffusion model. Furthermore, they also offer the possibility of fixing one or more of the parameters to a constant value, thereby specifying less complex models. As already mentioned, the intertrial variabilities are usually estimated not due to the theoretical interest in these parameters (see Ratcliff, 2008; Starns & Ratcliff, 2012, for an exception), but to avoid a biased estimation of the basic diffusion model parameters.

However, several simulation studies show that these parameters (especially, the variability of drift rate and starting point) are estimated less accurately than the other parameters (e.g., Lerche et al., 2016; van Ravenzwaaij & Oberauer, 2009; Vandekerckhove & Tuerlinckx, 2007). This raises the question of whether the inclusion of intertrial variability parameters really improves the estimation of the other parameters. Based on such findings, in some recent studies the intertrial variabilities have been deliberately fixed. For example, Germar, Schlemmer, Krug, Voss, and Mojzisch (2014) fixed all three intertrial variabilities at zero (see also Ratcliff & Childers, 2015). Note that also in earlier work the intertrial variabilities have sometimes been fixed at zero, because the application of the *EZ* method does not allow to include these parameters (e.g., Dutilh, Forstmann, Vandekerckhove, &

Wagenmakers, 2013; Grasman et al., 2009; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007; van Ravenzwaaij, Dutilh, & Wagenmakers, 2012; Wagenmakers, van der Maas, et al., 2008; Wagenmakers et al., 2007).

Whereas Ratcliff and Rouder (1998) and Ratcliff and Tuerlinckx (2002), who argued for the inclusion of intertrial variabilities, typically used very high trial numbers (at least 1,000 trials per participant), more recently the model has also been applied to data sets with significantly smaller trial numbers (e.g., with only 100, see Metin et al., 2013). This raises the question of whether small data sets provide enough information to estimate the full (seven-parameter) model. Lerche et al. (2016) systematically investigated the number of trials that allow for a precise estimation of the diffusion model parameters. They simulated data sets both on the basis of a seven-parameter model (i.e., with the assumption of intertrial variabilities) and on the basis of more restricted models. For example, in a four-parameter model the three intertrial variabilities were fixed at zero both for the generation of data and for the reestimation of parameters. The comparison of these models revealed that—as expected—for more complex models higher trial numbers are required. Besides, as Lerche et al. (2016) show, the required number of trials also depends on the used optimization criterion. The authors found that the three optimization criteria KS, ML and CS perform equally well for very high trial numbers. However, for small and moderate trial numbers, accuracy of estimates from CS based parameter search was inacceptable.

The findings by Lerche et al. (2016) raise the issue of whether less complex models (i.e., models with fixations) also perform better when the true (data generating) model is more complex (i.e., includes variabilities). A study by van Ravenzwaaij, Donkin, and Vandekerckhove (in press) speaks in favor of this hypothesis. The authors compared the performance of *EZ* (Wagenmakers et al., 2007) with the performance of a diffusion model estimation including all three intertrial variability parameters (using Quantile Maximum Proportion Estimation, see Heathcote, Brown, & Mewhort, 2002). Interestingly, the power of between-group difference detection for both drift rate and threshold separation was higher for *EZ* than for the more complex model even if there were substantial intertrial variabilities in the data generating models. Thus, it seems that simpler models can outperform more complex models.

We further tackled this question in two studies, a simulation study (Study 1) and a test-retest study (Study 2). In Study 1, the performance of the estimation procedure is measured by deviations and correlations between the true and the recovered parameter values. In Study 2, the estimation performance is assessed by means of the correlations between the parameters of two different sessions.

Study 1: Simulation study

Study 1 is a simulation study in which we reanalyzed data sets of the seven-parameter model from Lerche et al. (2016).

Method

Lerche et al. (2016) simulated data sets with different numbers of trials and reestimated parameters in order to deduce guidelines on requisite trial numbers. In Study 1, we reanalyzed a part of their data sets, namely the data sets that were created on the basis of the seven-parameter model (i.e., the model that includes intertrial variabilities and a bias in the starting point; see also Table 1). Here, we only briefly present their study design with a focus on the differences between the two studies. Please refer to Lerche et al. (2016) for more details on their simulation procedure.

The authors constructed data sets for two different experimental designs: a one-drift design and a two-drift design. Whereas the one-drift design simulates choices between two stimuli with the same absolute drift rate value, in the two-drift design the drift rate for one stimulus is larger than for the other stimulus ($d_z = 0.35$). Accordingly, in the one-drift design, only one drift rate was estimated. In the two-drift design, two drift rates (with opposite signs) were estimated simultaneously. One-thousand different parameter sets with random parameter values were used for each experimental design. For each parameter set seven data sets were created, using *construct-samples*[1], with different trial numbers (24—48—100—200—500—1,000—5,000). Then, 4 % of the simulated trials were randomly selected and substituted for by either fast or slow contaminants, resulting in three contamination conditions (no contaminants—fast contaminants—slow contaminants). More specifically, in the condition with fast contaminants, the responses of the contaminant trials were set by chance to 0 or 1 (simulating guesses) and the simulated RTs from these trials were substituted for by RTs situated at the lower edge of the original distribution (range: t_{min} – 100 msec to t_{min} + 100 msec, with $t_{min} = t_0 - s_{t0}/2$). In the condition with slow contaminants, only the response times were replaced, using values lying 1.5 - 5 interquartile ranges above the third quartile of the original RT distribution.

[1] *Construct-samples* is part of *fast-dm* and offers the possibility of constructing data sets based on a diffusion process.

For each condition (stimulus design × trial number × contamination condition), Lerche et al. (2016) reestimated all seven parameters and compared them with their true values (in the remainder of this article termed “seven-parameter model”). In the present study, we additionally use more parsimonious models for parameter estimation. In particular, in the “five-parameter model”, two of the intertrial variabilities (s_v and s_{zr}) were fixed at zero (i.e., we assumed that these two parameters do not vary from trial to trial). We fixed these two intertrial variabilities, because several studies have shown that they are recovered poorly (e.g., van Ravenzwaaij & Oberauer, 2009). The intertrial variability of the non-decision time, on the other hand, is estimated better and could counteract the negative influence of fast contaminants. Thus, this parameter was kept in the model even if it is psychologically less interesting than the main diffusion model parameters (a, v, t_0, z_r). Furthermore, we used a “four-parameter model” (i.e., the “basic” model) with additional fixation of the intertrial variability of the non-decision time (i.e., $s_{t0} = 0$). Note that these fixations are always false assumptions (“false fixations”), since the data generating model included all three intertrial variabilities. Finally, we estimated a “three-parameter model” in which we additionally fixed the starting point to the center between the two thresholds (i.e., $z_r = .5$). For the parameter estimation, we used *fast-dm-30* (Voss et al., 2015) and estimated the parameters with each of the three implemented optimization criteria (i.e., KS, ML and CS).

Our evaluation criteria are similar to those by Lerche et al. (2016): We analyzed (1) correlations between the true and the reestimated parameter values, (2) biases (i.e., deviations between the true and the reestimated parameter values) and (3) estimation precision (i.e., squared deviations between the true and the reestimated parameter values). For criterion 1 and criterion 3 we additionally computed an average measure across parameters. Specifically, for criterion 1, we calculated the mean correlation over the four main diffusion model parameters using Fisher’s Z-transformation.[2] The mean estimation precision was calculated on the basis of the formula stated below. Most importantly, differences between the estimated and the true parameter values were computed and weighted against the best possible accuracy that can be reached by each parameter. In contrast to Lerche et al. (2016), we computed the mean based on only the four basic diffusion model parameters (i.e., a, v, t_0 and z_r).[2, 3]

[2] In the three-parameter model, the mean was based on a, v and t_0. In the two-drift design, first the mean for the criterion performance of the two drift rates was calculated.

[3] “Best possible accuracies“ of the main diffusion model parameters: a – 0.054; ν – 0.270; t_0 – 0.032; z_r – 0.035. These values are based on an optimal condition of parameter estimation (5,000 trials, no contaminants, ML estimation; for more details, please refer to Lerche et al., 2016).

$$mean\ estimation\ precision = \frac{1}{4} \cdot \sum_{\mathrm{k}=1}^{4} \left[\frac{\mathrm{estimated}_k - \mathrm{true}_k}{\mathrm{best\ possible\ accuracy}_k}\right]^2$$

If the interest of the researcher lies in relationships between the diffusion model parameters and external criteria, the correlation criterion is of most relevance. A disadvantage of correlation coefficients is that they can mask possible biases in parameter estimation (e.g., if a parameter is systematically over- or underestimated, still high correlation coefficients result). The bias criterion tackles such systematical deviations in parameter estimation. Finally, the estimation precision criterion is the strictest criterion, since it takes into account any inaccuracy in parameter estimation. This criterion is of relevance if the diffusion model parameters are to be used as diagnostic measures. Such a potential future use of diffusion model parameters requires very accurate parameter estimates.

Results

In Figure 2, results are presented for the one-drift design for uncontaminated data. Figure 3 and Figure 4 show results for the conditions of slow and fast contaminants, respectively. In the left column, the 95 % quantiles of the mean estimation precision (criterion 3) are shown (thus, for most data sets, the mean estimation precision is smaller than the values from the figure). In the right column, mean correlation coefficients (criterion 1) are depicted. Results are presented as a function of number of trials, optimization criterion and model complexity.[4] Additionally, Table 2 (for the one-drift design) and Table 3 (for the two-drift design) sum up which model (model with 3, 4, 5, or 7 parameters) shows the best performance in terms of the correlations (first value), the mean bias across data sets (second value) and the 95 % quantiles of estimation precision (third value) depending on the optimization criterion (KS/ML/CS), type of contamination (none/fast/slow) and number of trials. Note that in some conditions, several models manifest almost identical performance

[4] Surprisingly, in some conditions, the estimation precision of KS decreased from 1,000 to 5,000 trials. This effect is based on a few models with very bad fit that strongly influence the reported 95 % quantiles. If medians are examined instead of the 95 % quantiles, the estimation precision—as expected—augments from 1,000 to 5,000 trials, or decreases only marginally. The KS-based search is more prone to get stuck in local minima for larger data sets. Artificial local minima can arise when calculation precision is too low. Exemplarily, we selected the ten data sets that showed the worst performance in the condition with 5,000 trials in the one-drift model with no contaminants. We then reestimated parameters for these data sets with the seven-parameter model with increased precision of calculation (the fast-dm precision criterion was increased from 3 to 4). This improved parameter estimation notably for the condition with 5,000 trials. More specifically, the mean across these ten data sets dropped to less than half, whereas there was less improvement for the condition with 1,000 trials. Accordingly, for higher trial numbers, we recommend using higher precision settings in fast-dm.

and that in these tables no information on the size of the differences between the models is given.

One main finding is that in most conditions the seven-parameter model does not provide the most accurate or unbiased estimates, although this is the true model. For ML, the pattern is quite consistent: in most cases, the five-parameter model reveals the best results. For CS, the findings are similar: The five-parameter model shows the best performance. In contrast to the results from ML, the CS procedure more often gets best results from the full seven-parameter model, even for smaller trial numbers. Note, however, that for small trial numbers the performance of CS is generally so poor for all models that results cannot be reasonably interpreted. Therefore, we generally do not recommend using CS for small trial numbers (see also Lerche et al., 2016). For KS more often than for ML and CS, models less complex than the five-parameter model (i.e., the three- or four-parameter models) bring forth the best results. Furthermore, here, more often than for ML and CS, the seven-parameter model performs best. A comparison of the different parameters reveals that for a and t_0 the five-parameter model and for v and z_r the four-parameter model result in the best recovery.

Discussion

Study 1 demonstrates that even if the three parameters a, v and t_0 vary from trial to trial (and the starting point is not situated centrally), the seven-parameter model does not always provide the most accurate results.

For data sets with fast contaminants, Lerche et al. (2016) (focusing on the mean precision criterion) showed that a KS based parameter search generally recovers parameters better than ML and CS. Interestingly, in the present analyses, ML and CS show a good performance for data contaminated by fast contaminants, if the five-parameter model is used (see Figure 4). Thus, the inclusion of the intertrial variability of t_0 seems to help to counteract the negative influence of fast contaminants. For KS, on the other hand, a similarly good performance is found for all applied models.

To test the stability of our results, we conducted additional analyses in which the parameter search started with other initial values for the intertrial variabilities. The default initial values of the intertrial variabilities incorporated in *fast-dm* are the following: $s_v = 0.5$; $s_{zr} = 0.3$; $s_{t0} = 0.2$. In one of the additional estimation series, we set all three intertrial variabilities to zero. In another, we set them to the maximum values used for simulation of data sets (see Table 1). Finally, in a third series of parameter estimation, we set them to half of the maximum values. The main results are very similar for all series of analyses in that the seven-parameter model is mostly outperformed by less complex models.

A caveat of our simulation study is that we made assumptions about the proportion and type of contamination that might not accurately reflect the contamination of real data. We are also not sure about the true range of intertrial variabilities in empirical studies. Another way to analyze the performance of different estimation procedures is provided by a test-retest study.

Study 2: Test-retest study

The main aim of Study 2 was to test whether the conclusions from Study 1 also hold for empirical data. For this purpose, we reanalyzed data from a test-retest study by Lerche and Voss (2016).

Method

In Study 1 of Lerche and Voss (2016), 105 participants worked at two sessions—separated by one week—on an Lexical Decision Task (LDT) and a Recognition Memory Task (with pictures as stimuli; RMT). As in Study 1 we used *fast-dm-30* and fitted the model using KS, ML, and CS procedures. We also compared the four models differing in complexity as introduced in Study 1. One response (“words” in the LDT and “old pictures” in the RMT) was assigned to the upper threshold, the other response (“non-words” and “new pictures”) to the lower threshold. In each model, we estimated two drift rates (for the different stimulus types). Both drift rates were then combined to an overall measure of speed of information accumulation, termed v_{total} by computing the difference between the drift for words (old pictures) and for non-words (new pictures).

For each of the basic diffusion model parameters (a, v_{total}, t_0 and z_r) the Pearson correlation between the two sessions was calculated.[5] To make results more accessible, as in Study 1, the mean over these four coefficients (without z_r in the three-parameter model) was computed using Fisher’s Z-transformation (in the remainder of this article termed “mean retest reliability”). Retest correlation coefficients were computed not only for parameters estimated from the actual data (i.e., 200 trials from the RMT, and 400 trials from the LDT), but also for parameters estimated from subsets of data with smaller trial numbers (specifically, for the first 32, 48, 100 and 200 trials of each participant).

Additionally, we wanted to test whether our main findings from Study 1 hold for a different strategy of data simulation. The parameter sets by Lerche et al. (2016) were created using uniform distributions across value ranges typically observed in previous diffusion

[5] Prior to the correlational analyses, we identified bivariate outliers with the Mahalanobis distance (D^2) and excluded participants with extremely high values ($p < .001$) from the respective analysis (resulting in at most 4 excluded participants).

model studies (only for the drift rates in the two-drift design a multivariate normal distribution was used). Lerche and Voss (2016), on the other hand, based their random parameter sets on multivariate normal distributions defined by the means, standard deviations and correlations of parameter estimates from the data of the LDT and RMT (Table 1). Importantly, as in the simulation study by Lerche et al. (2016), there were substantial intertrial variabilities. Data sets were created using different trial numbers (32—48—100—200—400—1,000—5,000) and assuming equal parameter sets for both sessions (i.e., no state influences). This allows an estimation of the maximum retest reliability coefficients. Again, in contrast to Lerche and Voss (2016), we estimated parameters using models with different complexity.

Results

In Figure 5, the retest reliabilities are presented for the four main diffusion model parameters for both LDT and RMT as a function of model complexity (estimations are based on the complete data, i.e., 400 and 200 trials for LDT and RMT, respectively). Again, applying the full seven-parameter model does not result in the highest correlations; retest-reliability is higher for less complex models. Whereas for non-decision time and starting point retest reliabilities for all models are similar, there are larger differences for drift rate and threshold separation. Notable is the poor estimation of drift rates from the seven-parameter model for estimations based on ML or CS. For ML and CS, the five-parameter shows the best performance, whereas for KS, the even more restricted four-parameter model mostly outperforms the other models. Figure 6 shows the influence of the number of trials on retest reliability. Mean reliability coefficients are shown both for the empirical data sets (depicted in black) and the data sets that were simulated on the basis of the parameter ranges observed in the empirical data (depicted in grey). Most importantly, for neither the empirical nor the simulated data does the seven-parameter model show the highest retest correlations. It is noteworthy that for CS and ML, even in the condition with 1,000 trials, the seven-parameter model be still worse than the other models.[6]

Discussion

The main findings from Study 2 are in line with those from Study 1 in that the seven-parameter model does not always show the best performance (here, in terms of the test-retest correlation coefficients). In fact, it is mostly outperformed by less complex models such as

[6] Note that we also analyzed the Associative Priming Task presented in Lerche and Voss (2016; Study 2) using models with different complexity. We found very similar results in that the seven-parameter model did not show the highest retest reliabilities.

the five-parameter model. In the simulation study—which was based on the multivariate distributions of estimated parameters—a similar pattern emerged. This suggests that the main findings do not depend on the particular simulation strategy of Study 1.

Interestingly, using the CS or ML criterion, only at 5,000 trials does the seven-parameter model catch up with the more restricted models. Note that sometimes CS has been used for data sets with such high trial numbers. In these studies, the use of a seven-parameter model is justified. Our results, however, suggest that it would be equally effective to use a more restricted model. In addition, it would be more efficient, since the time needed for parameter estimation is prolonged when models with intertrial variabilities are estimated. For smaller trial numbers, on the other hand, the use of the seven-parameter model can lead to worse parameter estimates than the use of more restricted models.

General discussion

In recent years, an increase in the number of researchers interested in the diffusion model and a higher variability regarding the addressed research topics and experimental designs is evident. For example, while in the past the diffusion model has almost exclusively been used for data sets with very large trial numbers (even > 1,000; e.g., Leite & Ratcliff, 2011; Ratcliff, Thapar, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, et al., 2008), more recently, it has also often been employed for studies with small to moderate trial numbers (e.g., Arnold et al., 2015; Boywitt & Rummel, 2012; Karalunas & Huang-Pollock, 2013; Karalunas, Huang-Pollock, & Nigg, 2012; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Metin et al., 2013; Pe, Vandekerckhove, & Kuppens, 2013).

Usually, complex models (i.e., with all seven distinct diffusion model parameters and, additionally, parameters varying between several conditions) are used. This has been done even if the number of trials is essentially smaller (e.g., 100 trials, see Metin et al., 2013) than in the studies that originally argued for the inclusion of intertrial variabilities (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). Especially for small to moderate trial numbers, the choices of model complexity and of optimization criteria for parameter estimation are crucial. Therefore a systematic comparison of different estimation procedures and a spreading of this knowledge is important in order to support a reasonable use of the diffusion model. With the studies reported here we make a step in this direction.

With two diverse approaches, we analyzed the influence of the model complexity on the accuracy of parameter estimation. We were particularly interested in the influence of the intertrial variabilities (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002) that have proven to be more difficult to estimate than the other diffusion model parameters (e.g., van Ravenzwaaij & Oberauer, 2009). In Study 1, we reanalyzed data sets from a simulation study by Lerche et al. (2016). The data sets were created assuming the presence of intertrial variabilities and a starting point of the diffusion process that was allowed to differ from the center between the thresholds. In Study 2, data from a test-retest study and a further simulation study by Lerche and Voss (2016) were analyzed. While in Study 1 deviations and correlations between the true and the recovered parameter values served as the performance measures, in Study 2 we examined the retest reliability coefficients. In both studies, the parameters were estimated using differently complex models.

Our results for both the simulated and the empirical data sets indicate that the most complex model (the "full" model comprising all seven parameters) is often not the best choice. A five-parameter model (with fixation of s_v and s_{zr} to zero) generally provides accurate estimates, especially when the maximum likelihood (ML) or the chi-square (CS) criterion is applied. For ML and CS, an additional fixation of s_{t0} is not advisable, since these two criteria are sensitive to the presence of fast contaminants (see also Lerche et al., 2016) and s_{t0} helps to counteract the negative influence of this type of contamination. Thus, keeping s_{t0} in the model can help to reach better estimation of the psychologically most interesting parameters (a, v, t_0 and z_r). For Kolmogorov-Smirnov (KS)—a criterion that is generally less sensitive to fast contaminants—the even less complex four-parameter model (i.e., the basic diffusion model with all intertrial variabilities fixed at zero) often provides the most accurate results.

Note that our results are in line with recent findings by van Ravenzwaaij et al. (in press). In their study, a model with fixed intertrial variabilities had a higher power to detect differences between conditions than a model including intertrial variabilities. Specifically, results from the EZ approach (Wagenmakers et al., 2007)—which fixes the starting point at the center between the two thresholds and the intertrial variabilities at zero—were compared to the application of a full diffusion model analysis. Even if the data were generated based on a full diffusion model, EZ outperformed the full diffusion model both for detection of drift rate and threshold separation differences. For non-decision time, the efficiency of both procedures was similar.

For future research, it would be interesting to analyze further experimental paradigms using test-retest studies. Besides, one could use different fixation strategies (e.g., instead of fixation at zero, the intertrial variabilities could be fixed at values typically observed in previous studies). To sum up, our results generally speak in favor of the use of less complex models. Thus, if the diffusion model is applied to get accurate estimates of cognitive processes (mapped by a, v, t_0, or z_r), a less complex model will often supply more reliable estimates. In particular, it is helpful to fix the intertrial variabilities of starting point and drift rate (s_{zr} and s_v) at zero.

References

Allen, P. A., Lien, M.-C., Ruthruff, E., & Voss, A. (2014). Multitasking and aging: Do older adults benefit from performing a highly practiced task? *Experimental Aging Research, 40*(3), 280-307. doi: 10.1007/s00426-014-0608-y

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*(5), 882-898. doi: 10.1007/s00426-014-0608-y

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition, 40*(1), 70-82. doi: 10.3758/s13421-011-0128-6

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

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging, 28*(1), 64.

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin, 40*(2), 217-231.

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53*(2), 55-68. doi: 10.1016/j.jmp.2009.01.006

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review, 9*(2), 394-401. doi: 10.3758/BF03196299

Karalunas, S. L., & Huang-Pollock, C. L. (2013). Integrating impairments in reaction time and executive function using a diffusion model framework. *Journal of Abnormal Child Psychology, 41*(5), 837-850. doi: 10.1007/s10802-013-9715-2

Karalunas, S. L., Huang-Pollock, C. L., & Nigg, J. T. (2012). Decomposing attention-deficit/hyperactivity disorder (ADHD)-related effects in response speed and variability. *Neuropsychology, 26*(6), 684-694. doi: 10.1037/a0029936

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*(3), 353-368. doi: 10.1037/0022-3514.93.3.353

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

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making, 6*(7), 651-687.

Lerche, V., & Voss, A. (2016). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, 1-24. doi: 10.1007/s00426-016-0770-5

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. doi: 10.3758/s13428-016-0740-2

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology, 27*(2), 193-200. doi: 10.1037/a0031533

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion, 13*(4), 739-747. doi: 10.1037/a0031628

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi: 10.1037/0033-295x.85.2.59

Ratcliff, R. (2008). Modeling aging effects on two-choice tasks: Response signal and response time data. *Psychology and Aging, 23*(4), 900-916. doi: 10.1037/a0013930

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*(4), 237-279. doi: 10.1037/dec0000030

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

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278. doi: 10.1037/0882-7974.19.2.278

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*(4), 408-424. doi: 10.1016/j.jml.2003.11.002

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481. doi: 10.3758/bf03196302

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology - General, 136*(3), 414-429. doi: 10.1037/0096-3445.136.3.414

Spaniol, J., Voss, A., & Grady, C. L. (2008). Aging and emotional memory: Cognitive mechanisms underlying the positivity effect. *Psychology and Aging, 23*(4), 859-872. doi: 10.1037/a0014218

Starns, J. J., & Ratcliff, R. (2012). Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. *Psychonomic Bulletin & Review, 19*(1), 139-145. doi: 10.3758/s13423-011-0189-3

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General, 143*(5), 1794-1805.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (in press). The EZ Diffusion Model Provides a Powerful Test of Simple Empirical Effects. *Psychonomic Bulletin & Review*.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology, 219*(4), 1017-1025. doi: 10.1007/s00213-011-2435-9

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*(6), 463-473. doi: 10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi: 10.3758/bf03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi: 10.3758/brm.40.1.61

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. A practical introduction. *Experimental Psychology, 60*(6), 385-402. doi: 10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi: 10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767-775. doi: 10.3758/bf03192967

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52*(1), 1-9. doi: 10.1016/j.jmp.2007.09.005

Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 539-555. doi: 10.1348/000711009x477581

Voss, A., Voss, J., & Lerche, V. (2015). Assessing Cognitive Processes with Diffusion Model Analyses: A Tutorial based on fast-dm-30. *Frontiers in Psychology, 6*, 336. doi: 10.3389/fpsyg.2015.00336

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*(1), 140-159. doi: 10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review, 15*(6), 1229-1235. doi: 10.3758/pbr.15.6.1229

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*(1), 3-22. doi: 10.3758/bf03194023

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics, 7*, 14. doi: 10.3389/fninf.2013.00014

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

Table 1

Parameter ranges (Study 1) and means and standard deviations (Study 2) used for generation of parameter sets

	Study 1: ranges		Study 2: *M* (*SD*)	
parameter	minimum	maximum	Lexical Decision Task	Recognition Memory Task
a	0.5	2.0	1.42 (0.32)	1.60 (0.36)
v	-4.0	4.0	-	-
v_0	-2.35 (1.0)[a]		-4.01 (1.13)	-3.07 (1.14)
v_1	2.00 (1.0)[a]		3.10 (1.11)	2.44 (1.20)
t_0	0.2	0.5	0.48 (0.04)	0.61 (0.05)
z_r	0.3	0.7	0.53 (0.06)	0.55 (0.08)
s_v	0.0	1.0	1.34 (0.64)	1.41 (0.83)
s_{t0}	0.0	0.2	0.15 (0.05)	0.17 (0.08)
s_{zr}	0.0	0.5	0.37 (0.25)	0.15 (0.22)

Note. Parameter sets of Study 1/Study 2 were created on the basis of a uniform distribution/multivariate normal distribution, respectively. *Fast-dm* uses a diffusion coefficient of 1. For comparison with parameters used in studies with diffusion coefficient .1 multiply a, v, z_r, s_v, and s_{zr} by .1.

[a] The drift rates in the two-drift design were created on the basis of a multivariate normal distribution with the given means and standard deviations.

Table 2

Model Superiority for the One-Drift Design, depending on the Type of Contamination, the Method, the Parameter and the Number of Trials

			Number of Trials						
	Method	Parameter	24	48	100	200	500	1,000	5,000
none	KS	a	4/4/3	7/3/3	7/3/3	3/4/4	7/4/7	7/4/7	4/4/7
		ν	3/7/3	4/7/3	4/7/4	4/7/4	4/7/4	4/7/4	4/7/4
		t_0	4/4/4	7/5/3	7/5/5	4/5/7	4/5/7	4/5/5	4/5/5
		z_r	7/4/4	4/4/4	4/4/4	7/4/4	7/4/7	7/4/7	4/4/4
	ML	a	3/3/3	3/3/3	5/5/5	5/5/5	5/7/5	5/7/5	7/7/7
		ν	4/4/3	4/5/4	4/4/4	5/4/5	5/4/5	5/4/5	7/7/7
		t_0	3/4/4	3/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/7/7
		z_r	4/7/4	5/4/4	5/5/5	5/4/5	5/4/5	5/5/5	5/7/7
	CS	a	7/7/7	7/5/7	7/5/5	5/5/5	5/3/5	5/3/5	7/7/7
		ν	3/5/5	4/5/5	5/3/5	5/4/5	5/5/5	5/7/5	7/7/7
		t_0	7/5/7	7/5/5	7/5/5	5/5/5	5/5/5	5/5/5	5/7/5
		z_r	7/7/7	7/4/7	7/4/7	5/4/5	5/4/7	5/4/5	5/4/7
slow	KS	a	7/5/3	7/7/3	7/4/3	5/4/7	7/4/7	7/3/7	4/3/7
		ν	3/7/3	4/4/4	4/7/4	4/4/4	7/7/7	4/7/4	4/7/4
		t_0	7/5/7	7/5/5	4/7/7	4/7/5	7/7/7	7/7/7	5/7/7
		z_r	7/4/4	5/4/7	5/4/7	7/4/7	7/4/7	7/4/7	7/4/7
	ML	a	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5
		ν	4/4/4	4/4/4	4/4/5	5/4/5	5/4/5	5/4/5	5/4/5
		t_0	7/7/7	5/7/7	5/7/7	5/7/7	5/7/5	5/7/5	5/7/5
		z_r	5/4/4	5/4/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5
	CS	a	7/7/7	7/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5
		ν	3/5/7	4/5/5	5/4/4	4/5/4	5/5/5	5/4/5	5/4/5
		t_0	7/5/7	7/5/7	5/5/5	5/7/5	5/7/7	5/7/5	5/7/5
		z_r	7/4/7	7/4/7	5/4/7	5/5/5	5/7/5	5/5/5	5/7/5
fast	KS	a	4/3/3	4/3/3	4/3/3	3/3/3	4/3/3	3/3/5	4/3/5
		ν	3/7/3	4/4/4	4/7/4	4/7/4	7/7/4	4/7/4	4/7/4
		t_0	7/5/7	3/5/4	7/7/7	4/7/5	4/7/7	4/7/7	4/7/7
		z_r	4/4/4	4/4/4	4/4/4	4/4/4	4/4/4	4/4/4	4/4/4
	ML	a	3/5/3	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5
		ν	4/7/3	4/7/4	4/7/4	5/4/4	5/4/4	5/4/4	5/4/4
		t_0	4/5/4	4/5/5	5/7/5	5/7/5	5/7/5	5/7/5	5/7/5
		z_r	4/7/4	5/4/4	5/5/5	5/4/5	5/5/5	5/4/5	5/5/5
	CS	a	7/5/7	7/5/5	7/5/5	5/5/3	5/5/5	5/5/5	5/5/5
		ν	3/7/5	3/4/5	5/7/4	5/7/4	5/7/5	5/7/5	5/7/7
		t_0	7/5/7	7/5/7	5/5/5	4/5/5	5/7/5	5/7/5	5/7/5
		z_r	4/4/4	4/4/4	4/4/4	4/4/4	5/5/5	5/4/5	5/7/5

Note. The first value is based on the correlation criterion, the second on the bias criterion, and the third on the estimation precision criterion. In the five-parameter model, the intertrial variabilities s_v and s_{zr} are fixed at zero, in the four-parameter model additionally the intertrial variability s_{t0} is fixed at zero and in the three-parameter model also the starting point z_r is fixed ($z_r = 0.5$). For conditions with 95 % quantiles of parameter estimation precision (weighted against the best possible accuracy) exceeding 25, values are depicted in grey. On the basis of data sets with at least 4 % of trials at each threshold.

Table 3

Model Superiority for the Two-Drift Design, depending on the Type of Contamination, the Method, the Parameter and the Number of Trials

	Method	Parameter	Number of Trials						
			24	48	100	200	500	1,000	5,000
none	KS	a	4/4/4	7/4/4	7/7/7	7/4/5	7/4/7	7/5/7	7/7/7
		ν	4/3/4	4/3/4	4/5/4	7/3/4	7/3/5	5/3/5	7/7/4
		t_0	4/4/4	7/5/4	7/5/4	5/7/7	5/7/7	7/7/7	7/7/7
		z_r	7/4/4	7/4/4	7/4/4	7/7/7	7/7/7	7/7/7	7/4/7
	ML	a	7/5/5	7/5/5	5/5/5	5/5/5	5/5/5	5/7/7	7/7/7
		ν	4/4/4	4/4/4	4/5/4	5/5/5	5/5/5	5/7/5	7/7/7
		t_0	4/5/4	5/5/5	5/5/5	5/5/5	5/7/5	5/7/7	7/7/7
		z_r	4/4/4	4/7/4	5/7/4	5/5/5	5/7/5	5/5/7	7/7/7
	CS	a	5/5/7	7/5/5	5/5/5	5/5/5	7/5/5	5/7/7	7/7/7
		ν	3/4/3	3/5/3	4/5/5	5/3/5	5/5/5	5/7/5	7/7/7
		t_0	7/5/5	7/5/7	5/5/5	5/5/5	5/5/5	5/7/5	7/7/7
		z_r	4/4/4	7/7/4	7/7/4	7/7/5	7/7/7	7/4/7	7/5/7
slow	KS	a	4/4/4	7/7/4	7/7/4	7/4/7	5/7/7	5/7/7	5/4/7
		ν	4/3/4	4/5/4	4/5/4	7/7/5	5/7/7	7/7/7	7/7/7
		t_0	4/4/4	7/5/5	4/7/7	7/7/7	7/7/7	7/7/7	7/7/7
		z_r	7/4/4	7/4/4	7/4/7	7/4/7	7/4/7	7/4/7	7/4/7
	ML	a	4/5/5	5/5/5	7/5/5	5/5/5	5/5/5	5/5/7	5/5/7
		ν	3/4/3	3/5/3	5/5/5	5/7/5	5/7/5	5/7/5	5/7/7
		t_0	7/5/7	7/7/7	5/7/7	5/7/7	5/7/7	5/7/7	5/7/7
		z_r	7/7/7	5/5/7	5/7/5	5/7/7	5/5/5	5/7/5	5/7/5
	CS	a	7/5/4	7/5/5	5/5/5	5/5/5	5/5/5	5/5/5	5/5/4
		ν	4/3/3	3/5/3	4/5/5	5/3/5	5/5/5	5/5/5	5/5/5
		t_0	7/5/7	7/5/7	7/5/7	5/7/7	5/7/7	7/7/7	7/7/7
		z_r	4/7/4	7/7/4	7/5/7	5/5/7	7/5/7	7/7/7	7/7/7
fast	KS	a	4/5/4	7/5/5	7/5/5	7/5/4	4/5/4	5/4/5	5/4/4
		ν	7/5/4	7/7/4	5/7/5	7/7/7	5/7/7	5/7/7	5/7/7
		t_0	4/4/4	7/7/7	4/7/7	4/7/7	5/5/5	5/5/5	5/5/5
		z_r	4/4/4	7/7/4	5/4/4	5/7/4	5/7/5	5/7/5	7/4/7
	ML	a	5/5/5	5/5/5	5/5/5	5/5/5	5/5/5	7/5/5	7/5/5
		ν	3/5/3	4/5/4	4/5/4	4/7/5	5/7/5	5/7/7	5/7/7
		t_0	4/5/4	5/5/5	5/7/5	5/7/5	5/7/7	5/7/7	7/7/7
		z_r	4/7/4	5/4/4	7/7/4	7/4/5	5/4/5	7/4/7	7/4/7
	CS	a	7/5/5	7/5/5	5/5/5	5/5/5	5/5/5	5/5/3	7/4/7
		ν	3/3/3	3/5/3	4/4/5	5/3/5	4/7/5	7/7/7	7/7/7
		t_0	7/5/5	7/7/7	5/7/5	5/7/5	5/7/7	5/7/7	7/7/7
		z_r	4/5/4	7/5/4	7/7/7	7/4/7	7/7/7	7/7/7	7/7/7

Note. The first value is based on the correlation criterion, the second on the bias criterion, and the third on the estimation precision criterion. In the five-parameter model, the intertrial variabilities s_v and s_{zr} are fixed at zero, in the four-parameter model additionally the intertrial variability s_{t0} is fixed at zero and in the three-parameter model also the starting point z_r is fixed (z_r = 0.5). For conditions with 95 % quantiles of parameter estimation precision (weighted against the best possible accuracy) exceeding 25, values are depicted in grey. On the basis of data sets with at least 4 % of trials at each threshold.

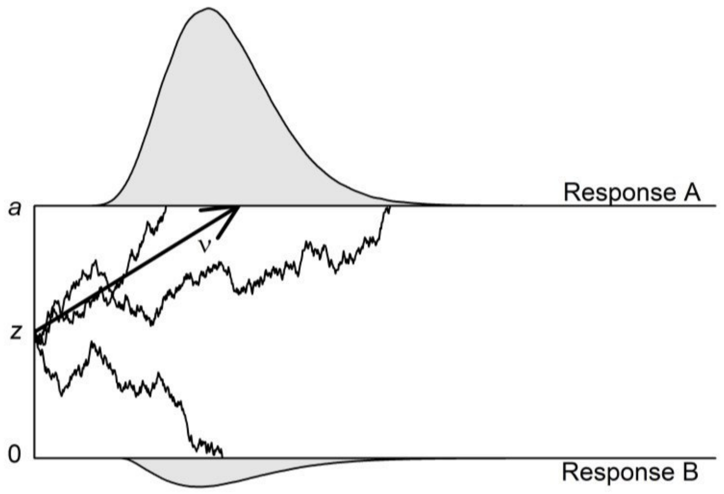


Figure 1. Illustration of the diffusion model with three of its four main parameters. The two thresholds that are associated with Response A (upper threshold; correct response in this illustration) and Response B (lower threshold; erroneous response) are separated by the distance *a.* The accumulation of information starts at starting point *z,* which is here centered between the thresholds. The mean drift rate (*v*) is positive so that the upper threshold is reached more often than the lower threshold. In two of the three exemplary trials, the processes reach the upper threshold—resulting in one fast and one very slow correct response—and in one trial, the process reaches the lower threshold. The non-decisional component (t_0) as well as the intertrial variabilities (s_{t0}, s_v and s_{zr}) are not depicted.

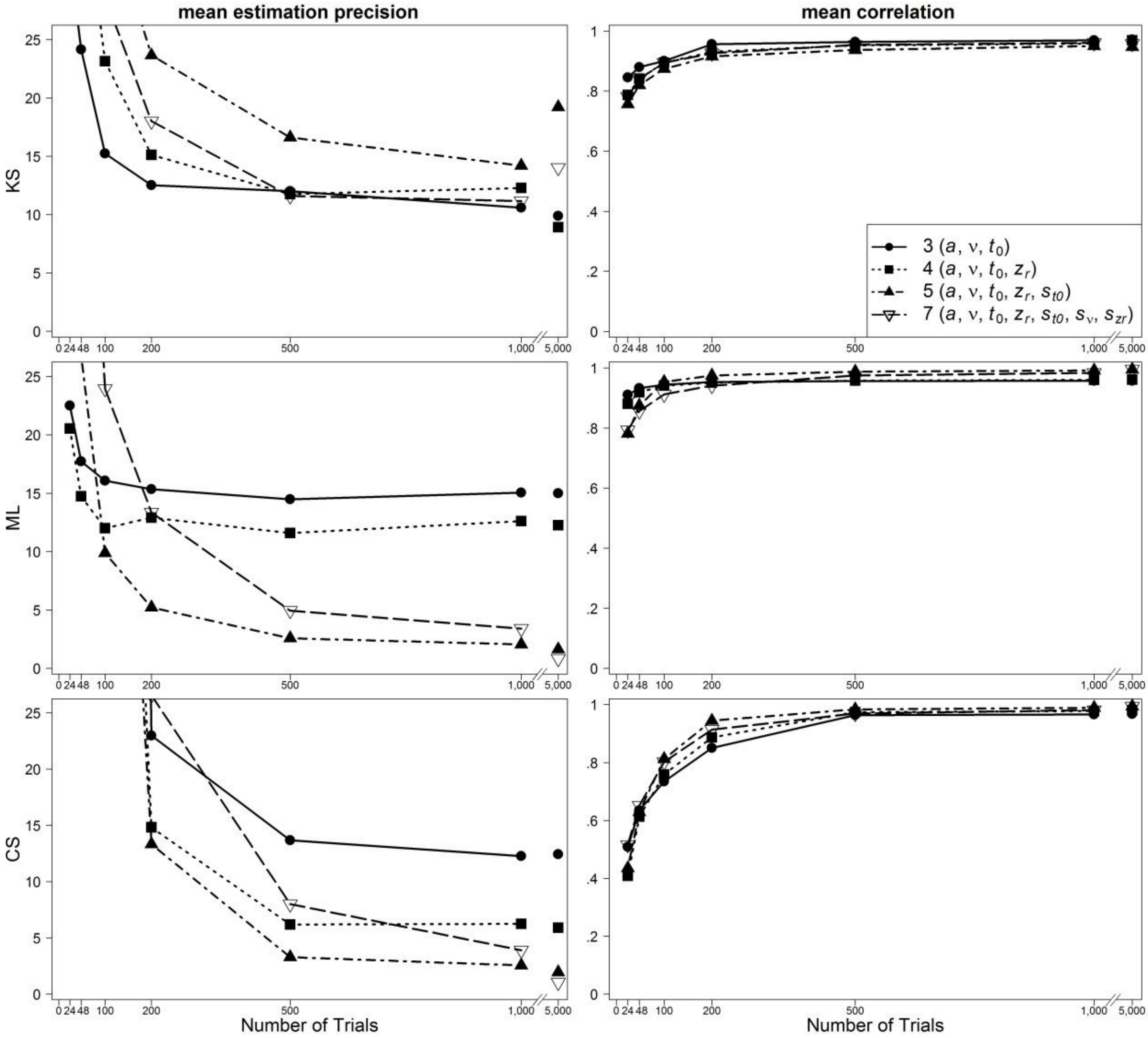


Figure 2. Scatter plot of 95 % quantiles of mean estimation precision (left column) and mean correlation between true and reestimated parameters (right column) for *uncontaminated data sets* in the one-drift design. On the basis of data sets with at least 4 % of trials at each threshold. Quantiles exceeding the mean estimation precision of 25 are not depicted.

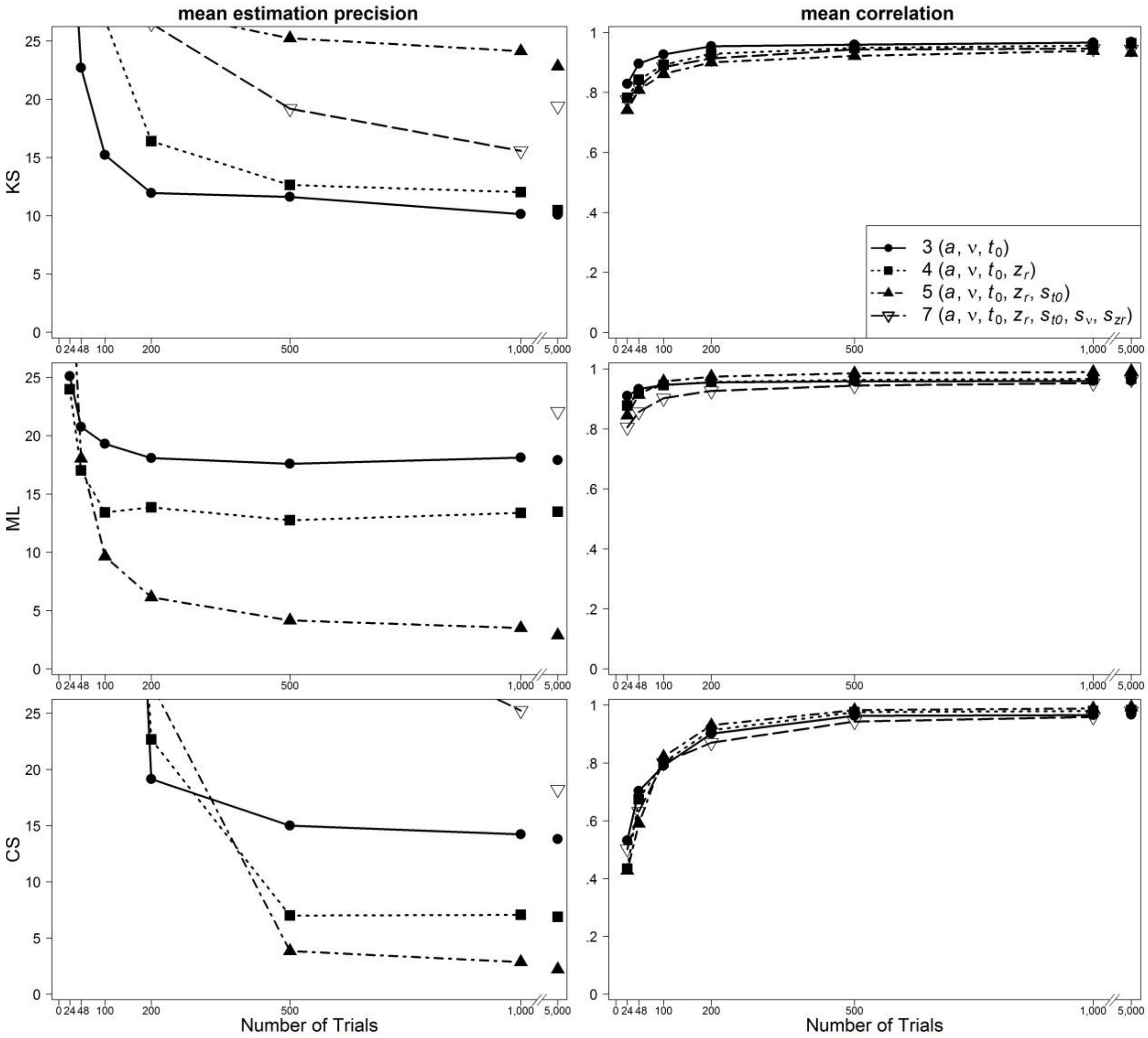


Figure 3. Scatter plot of 95 % quantiles of mean estimation precision (left column) and mean correlation between true and reestimated parameters (right column) for *data sets with slow contaminants* in the one-drift design. On the basis of data sets with at least 4 % of trials at each threshold. Quantiles exceeding the mean estimation precision of 25 are not depicted.

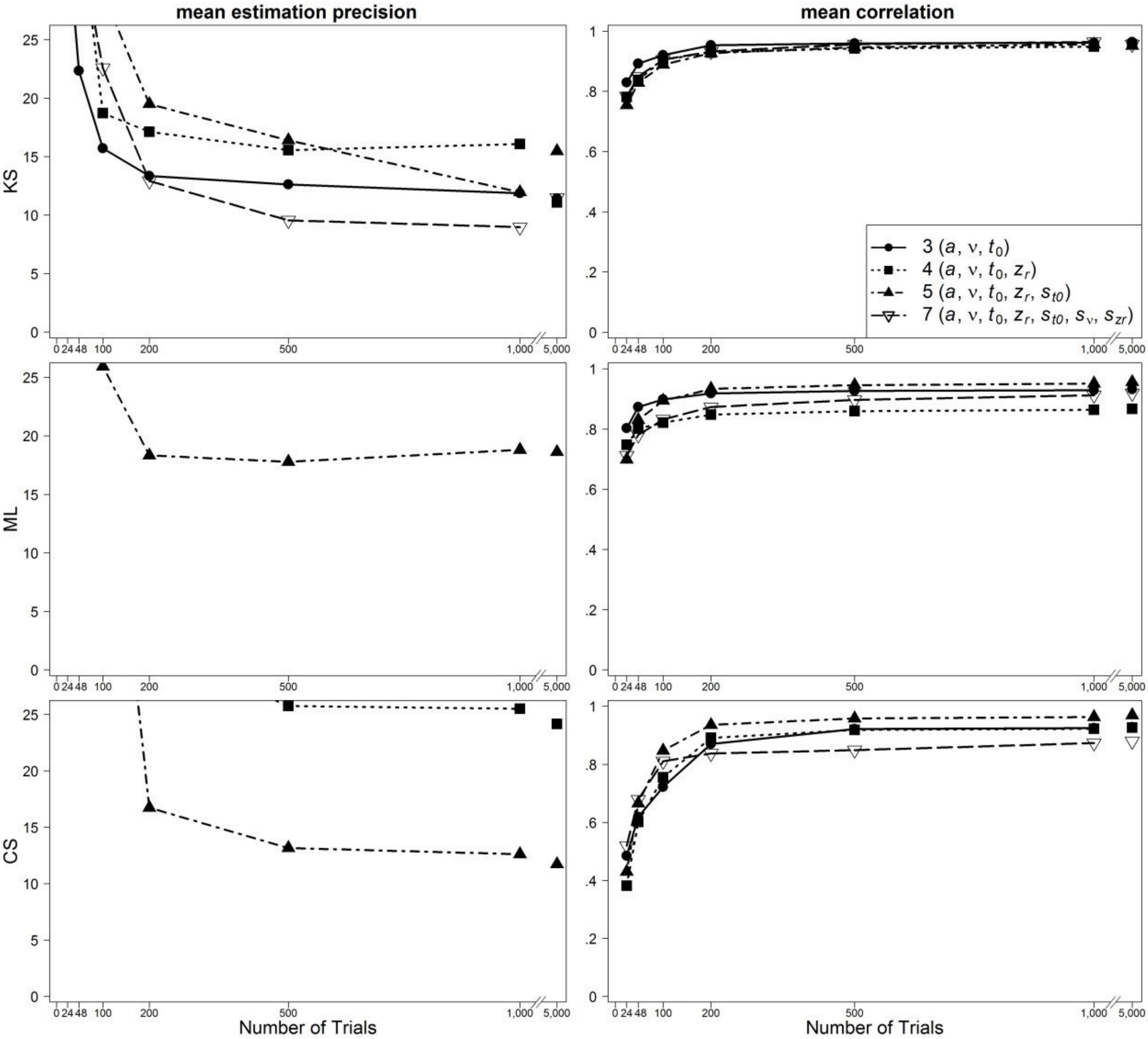


Figure 4. Scatter plot of 95 % quantiles of mean estimation precision (left column) and mean correlation between true and reestimated parameters (right column) for *data sets with fast contaminants* in the one-drift design. On the basis of data sets with at least 4 % of trials at each threshold. Quantiles exceeding the mean estimation precision of 25 are not depicted.

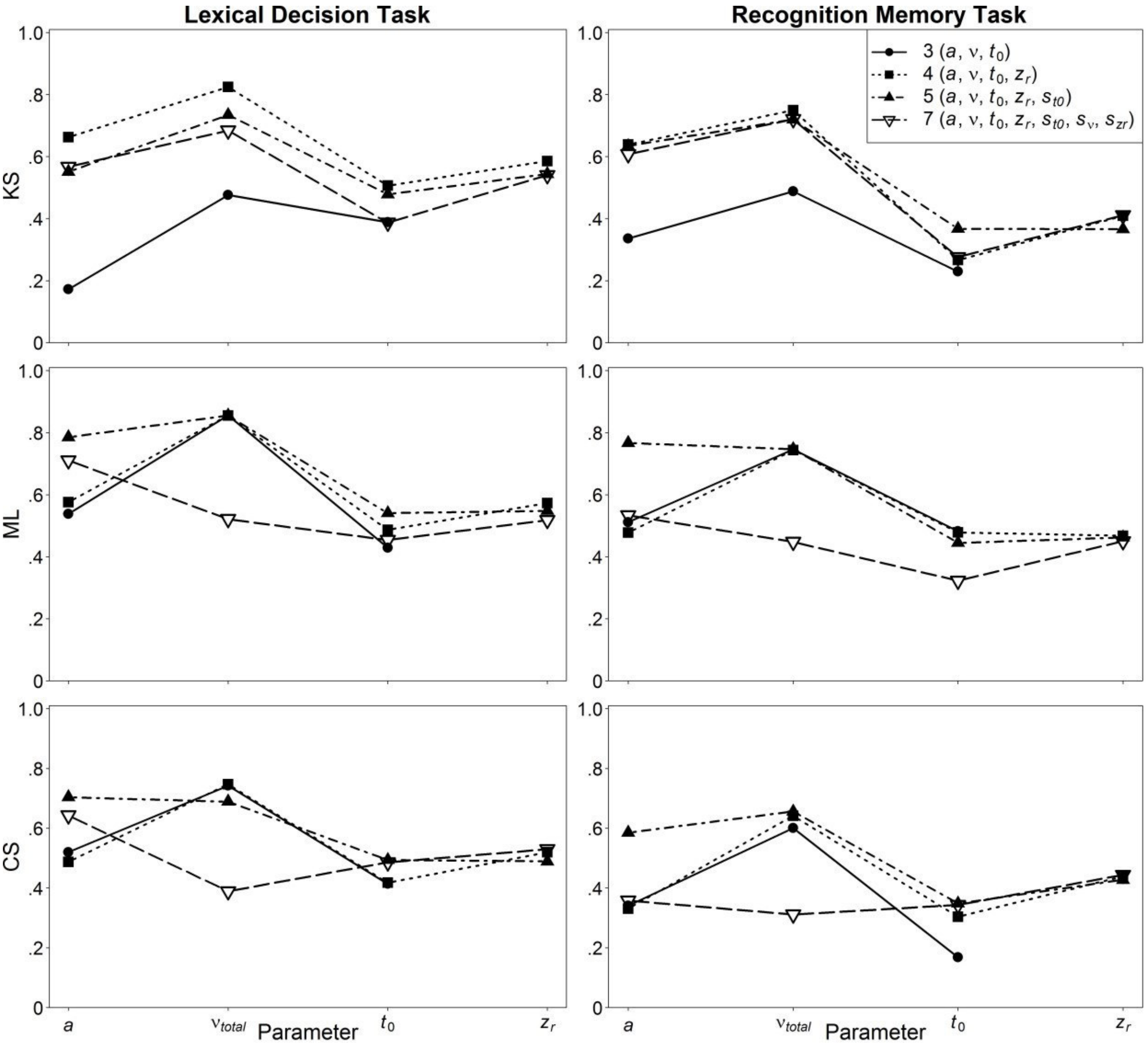


Figure 5. Retest reliability depending on model complexity and method.

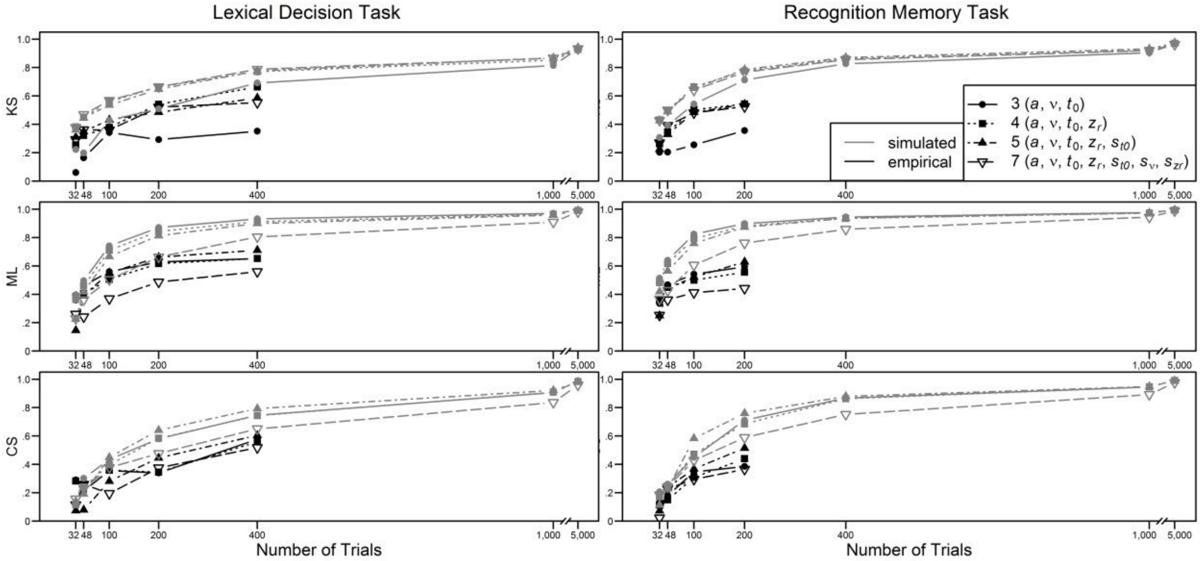


Figure 6. Mean retest reliability depending on model complexity, method, type of data (empirical vs. simulated) and number of trials.

Erklärung gemäß § 8 Abs. (1) c) und d) der Promotionsordnung der Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg

Doctoral Committee of the Faculty of Behavioural and Cultural Studies, of Heidelberg University

Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Declaration in accordance to § 8 (1) b) and § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.

I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.

I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname Veronika Lerche

Datum, Unterschrift ______________________________