Drawing conclusions from choice response time models: a tutorial

Chris Donkin,
Indiana University

Scott Brown and Andrew Heathcote
The University of Newcastle

Abstract

Cognitive models of choice and response times can lead to deeper insights into the processes underlying decisions than standard analyses of accuracy and response time data. The application of these models, however, has historically been reserved for the authors of the models, and their associates. Recently, choice response time models have become more accessible through the release of user-friendly software for estimating their parameters. The aim of this tutorial is to provide guidance about the process of using these parameter estimates and model fits to make conclusions about experimental data. In particular, we discuss the steps required to select an appropriate characterization of a given data set in terms of the parameters of a choice model. We also discuss how to evaluate the quality of the agreement between model and data, including some guidelines for presenting model predictions for group-level data.

Introduction

Evidence accumulation models of choice response time (RT) are increasingly used to examine the psychological processes underlying rapid decisions. The central assumption of these models is that the decision maker accumulates evidence for potential choices and makes a decision once the evidence reaches a threshold amount. The predicted time to make a response is the time taken to accumulate evidence, plus "non-decision time", which is the time for other necessary processes, such as stimulus encoding and response execution. The parameters of evidence accumulation models quantify different aspects of the decision process, such as the rate of evidence accumulation, response caution (the amount of evidence required for a response) and response bias (different caution for different responses). Variations among experimental conditions in estimates of these parameters, and in associated estimates of non-decision time, can provide insights into latent psychological processes beyond those available from traditional measures (i.e., independent analyses of accuracy and mean RT).

The evidence accumulation modeling approach has been successfully applied to many different paradigms, including: simple perceptual decisions (Usher & McClelland, 2001), visual short-term memory (Smith & Ratcliff, 2009), absolute identification (Brown, Marley, Donkin, & Heathcote, 2008), lexical decision (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), the link between depression and anxiety (White, Ratcliff, Vasey, & McKoon, 2009, in press), and the neural correlates of behavioral measures (Farrell, Ratcliff, Cherian, & Segraves, 2006; Forstmann et al., 2008; Ho, Brown, & Serences, 2009).

Many evidence accumulation models have been proposed as explanations of a variety of rapid choice tasks, including Ratcliff's diffusion model (Ratcliff, 1978), the Poisson counter model (Pike, 1966; Van Zandt, Colonius, & Proctor, 2000), the accumulator model (Smith & Vickers, 1988), the leaky competing accumulator model (Usher & McClelland, 2001), the linear and non-linear ballistic accumulator models (Brown & Heathcote, 2005, 2008). We will focus on the recently proposed linear ballistic accumulator (LBA) model because it is mathematically simple, and because it was the model used by the authors of the data set we use as an example in this tutorial (Forstmann et al., 2008). Although our focus here is on the LBA model, the techniques we illustrate for model selection and evaluation are applicable to all evidence accumulation models.

Applying an RT model to data involves – at minimum – estimating parameters from data. Brown and Heathcote (2008) and Donkin, Averell, Brown, and Heathcote (2009) provide simple computational routines to make such estimates possible for the LBA. Similarly, Vandekerckhove and Tuerlinckx (2007) provide methods and advice for estimating the parameters of Ratcliff's (1978) diffusion model (see also Tuerlinckx, Maris, Ratcliff, & De Boeck, 2001; Tuerlinckx, 2004; Vandekerckhove & Tuerlinckx, 2008). More generally, Myung (2003) and Van Zandt (2000) provide excellent tutorials on how to estimate parameters for psychological models in general. However, when using a choice RT model, it is not a trivial step to go from estimating free parameters to psychologically meaningful conclusions. The aim of the current tutorial is to bridge the gap between parameter estimation and interpretation. We present a step-by-step analysis of data from a simple perceptual two-choice task (Forstmann et al., 2008) to illustrate this process.

The Linear Ballistic Accumulator

Figure 1 illustrates decision processing in a pair of LBA units. Suppose that the figure represents a single trial in Forstmann et al.'s (2008) experiment, in which participants must choose whether a cloud of dots appears to be moving to the left or to the right, requiring a "left" or "right" response, respectively. Presentation of the stimulus causes evidence to accumulate for both the "left" or "right" responses separately, as indicated by the two lines (one solid and one dotted) in Figure 1. The vertical axis of the figure represents the amount of evidence that has been accumulated, and the horizontal axis shows how much decision time has passed. The amount of evidence in each accumulator increases linearly until one reaches the response threshold, and the decision time is the time taken for the first accumulator to reach threshold. The predicted RT is made up of the decision time plus a

non-decision time, quantified by parameter t_0[1].

The slopes of the lines in Figure 1 indicate the rates at which evidence is accumulated for each response, and are usually referred to as the drift rates. If the physical stimulus favors a "left" response, the drift rate for the "left" response accumulator will usually be larger than for the "right" response accumulator. Drift rates are assumed to be set by physical stimulus properties and by the demands of the task. For example, in Forstmann et al.'s (2008) task, decisions might be made easier by making the displayed dots drift more steadily in one direction. This would provide more evidence that "left" was the correct response, and the drift rate for the left response would increase. Drift rates are also assumed to be modulated by sensory and attentional processing, and the overall efficiency of the cognitive system. For example, Schmiedek, Oberauer, Wilhelm, Süß, and Wittmann (2007) found larger drift rates for participants with higher working memory capacity and fluid intelligence. In the LBA, there are two different drift rates: one for the correct response and another for the incorrect response (the two sloping lines in Figure 1). The relative size of drift rate parameters describes differences in task performance between different conditions or groups. Importantly, although not explicitly illustrated in Figure 1, drift rates in the LBA are assumed to vary randomly from trial-to-trial according to a normal distribution with mean v and standard deviation s, reflecting trial-to-trial fluctuations in factors such as attention and arousal.

The amount of evidence in each accumulator before the beginning of the decision process also varies from trial-to-trial. The starting evidence for each accumulator is assumed to follow a uniform distribution whose minimum value is set (without loss of generality) at zero evidence for all accumulators, and whose upper value is determined by a parameter A. Hence, the average amount of evidence in each accumulator before accumulation begins is $\frac{A}{2}$. The height of the response threshold that must be reached is called b, and is represented by the horizontal dotted line in Figure 1[2]. The value of b relative to the average starting activation ($\frac{A}{2}$), provides a measure of average response caution, because the difference ($b-\frac{A}{2}$) is the average amount of evidence that must be gathered before a response will be triggered. In Figure 1, the same response threshold (b) is used for both accumulators; this indicates that the same amount of evidence is required, on average, before either response is made. If participants choose to favor one particular response (i.e., a response bias), b and/or A might be smaller for the preferred response. Response bias leads to a speed-accuracy trade-off, as the preferred response is made more quickly, but it is also made more often when incorrect, reducing accuracy.

The time taken for each accumulator in the LBA to reach threshold on any given trial is the distance between the response threshold and the start point of activation, divided by the rate of evidence accumulation. The observed decision time on any given trial, however, is the time for the fastest accumulator to reach threshold. The formula for the distribution across trials of the time taken for the fastest accumulator to reach threshold is given by Brown and Heathcote (2008). This formula makes it possible to estimate the model's parameters from data.

[1]Sometimes, the amount of non-decision time is assumed to vary from trial-to-trial with a uniform distribution, especially for Ratcliff's diffusion model (e.g., Ratcliff & Tuerlinckx, 2002).

[2]When estimating parameters for the LBA b is always constrained to be greater than A

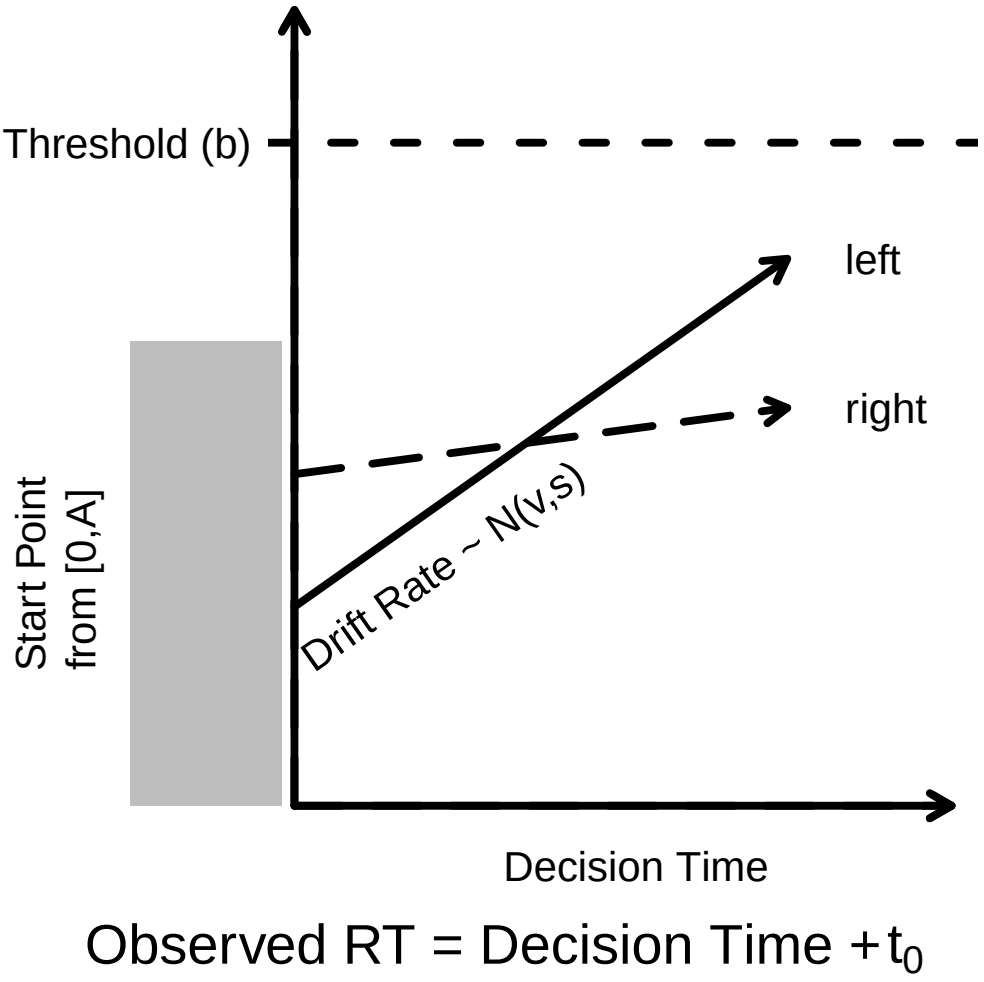

Figure 1. A typical LBA decision for the task in Forstmann et al. (2008). In the current trial a left stimulus has been presented and so drift rates for the left and right accumulators have been sampled normal distributions with means v and $1 - v$,respectively, and a common standard deviation s.

Example LBA Application

Choice RT models are most appropriate for paradigms requiring simple and rapid decisions[3]. Forstmann et al.'s (2008) participants made simple decisions with average RTs around one second, so the paradigm is appropriate. Their experiment investigated the neural correlates of the trade-off between speed and accuracy, by testing predictions from a neurophysiological theory of how response caution is implemented by sub-cortical decision circuits. They presented participants with a cloud of 120 moving dots, of which 60% moved coherently to either the left or right of the screen, while the remaining 40% moved in random directions. Participants were asked in which direction (either "left" or "right") the cloud appeared to move. Before each trial, participants were given one of three cues, indicating whether they should try to make a very accurate response (accuracy emphasis), or a very fast response (speed emphasis), or try to balance accuracy and speed (neutral emphasis). Twenty participants each completed 280 trials per emphasis condition; other methodological details can be found in the original article.

The manipulation of response caution using cues had the expected effect. On average, participants were faster under speed emphasis ($\bar{RT} = 429$ms) than under neutral emphasis ($\bar{RT} = 515$ms) or accuracy emphasis ($\bar{RT} = 555$ms). The faster responses came at the cost of lower accuracy: in the speed condition, 77% of responses were correct, in the neutral condition this was 86% and 87% in the accuracy condition. This pattern of data - trading accuracy for speed - is consistent with the effects of manipulating response caution in a choice response model (i.e., moving the response threshold higher and lower). However, it is also possible that participants were doing something more complicated. For example, non-decision processes (t_0) might also have been faster under speed emphasis, or the quality of information (drift rate, v) might have been greater under accuracy emphasis. Forstmann et al. (2008) examined these possibilities by comparing the fit of the LBA model using a range of different parameter constraints. This analysis allowed them to infer which cognitive processes were influenced by the experimental manipulation. In the next section we address this detailed problem of selecting the best set of parameter constraints. First, however, we briefly review parameter estimation for choice RT models and some other assumed knowledge.

Fitting the Model

Parameter estimation

A choice RT model, like any quantitative theory, is defined by numerical parameters, and changing these parameters alter the model's predictions about RT and accuracy. For example, increasing the response threshold parameter increases accuracy and both slows and increases the variability of predicted RTs. Increasing the drift rate also increases accuracy, but it has the opposite effect on RT, reducing both the mean and variability. Non-decision time affects mean RT, but has no effect on RT variability or accuracy.

The aim of fitting a model to data is to find parameter values which yield model predictions that best match observed data. The degree of match between the observed data

[3] Although similar models have been extended to more complicated judgments (e.g., Busemeyer & Townsend, 1992, 1993).

and model predictions can be quantified by an objective function, which takes into account both decision accuracy and the distribution of RTs for each type of response. Automated search algorithms are used to find the best fitting parameter values – those that optimize the objective function.

There is a great deal of literature, including several tutorials, dealing with the choice of objective function and optimization algorithm (e.g., Heathcote, Brown, & Mewhort, 2002; Myung, 2003; Ratcliff & Tuerlinckx, 2002; Van Zandt, 2000). For the purpose of this tutorial we assume that the reader has a reasonable grasp of the issues associated with parameter estimation. Our starting assumption is that the reader is capable of finding the best-fitting values for some given set of free parameters (and some data, and a model). The appendix contains a general review of issues related to fitting and parameter estimation for choice RT models, with a particular emphasis on the model and fitting methods used herein.

The purpose of this tutorial is go beyond finding the best *estimates* for certain free parameters, and instead to find the best *set* of free parameters to estimate. This issue is critical since choice RT models are often used to determine which decisional processes are influenced by specific experimental manipulations. This is the same as asking which parameters of the model systematically vary across a set of conditions produced by an experimental manipulation. To illustrate, we analyze Forstmann et al.'s (2008) data highlighting several issues regarding the selection of which parameters *can* and *should* change across experimental conditions.

Which parameters change across conditions?

Forstmann et al.'s (2008) experiment had three emphasis conditions (speed, neutral and accuracy) and in each of these conditions there were two types of stimuli (coherent motion to the left or to the right). One of the central tasks of cognitive modeling for these data is to investigate which aspects of cognitive processing were influenced by the experimental factors. In model terms, we want to know which parameters changed across conditions.

A priori assumptions.

To begin, we first decide which parameters potentially *could* change. The LBA model has five parameters that determine the behavior of an evidence accumulator in any condition (b, A, s, t_0, v), and there are always two accumulators for each decision (one for the response "left" and one for the response "right"). This means that there could be up to 10 parameters which vary for each particular combination of stimulus and emphasis conditions, for a total of 60 parameters. Thankfully, this type of freedom, though possible, is not usually required, because sensible *a priori* constraints can be placed on parameters across conditions. We elaborate these constraints by considering three factors in succession: "left" vs. "right" responses; left-moving vs. right-moving stimuli; and decision caution conditions (speed, neutral or accuracy emphasis).

The two possible responses ("left" and "right", corresponding to the two accumulators) should share many parameters. Usually, t_0 can be fixed at the same value for both because in most cases it is reasonable to assume that the time to execute each response is the same. This assumption is plausible for the present data, but may break down in unusual paradigms, such as when one response is harder to produce than another. In contrast, the

evidence threshold parameter (b) and the parameter for the level of between-trail variability in starting points (A) might reasonably vary between responses – because participants might be biased toward one response over the other. For example, if participants are biased to respond "left" rather than "right", this can be reflected in smaller values of b and/or A for the "left" accumulator than the "right" accumulator.

Response biases may or may not occur, depending on individual differences between participants. However, when choice accuracy is above chance, every participant ought to demonstrate a difference in drift rates between "left" and "right" accumulators as a function of which response is correct for a given stimulus. That is, on trials where the stimulus drifts to the right, the mean drift rate should be higher for the accumulator corresponding to the "right" response than for the "left" response. Often, greater simplification can be obtained by fixing the mean drift rate for the incorrect response at one minus the mean drift rate for the correct response. This restriction has commonly been applied because it also satisfies a scaling property applying to all choice RT models, which requires at least one parameter to be fixed in order to obtain unique estimates of the remaining parameters. However, applying this restriction to drift rates across *all* conditions provides greater constraint than necessary, and can result in poor fits. We advise careful consideration about whether this restriction is justifiable on theoretical grounds, before applying it (for further discussion see Donkin, Brown, & Heathcote, 2009).

Finally, though it is possible that between-trial variability in the drift rate, s, can differ across responses, it has been fixed to the same value in all applications of the LBA to date, and the same is true, to our knowledge, of analogous parameters in other evidence accumulation models. We follow this convention here, but note that this is an additional, and untested, assumption.

To sum up, the only parameters that we might allow to take on different values for "left" than "right" response options are b and A. We will further assume that v is constrained to sum to one across "left" and "right" responses, within any condition, as this seems a reasonable assumption for Forstmann et al.'s (2008) experiment, in which increased evidence for one response (e.g., more dots moving left) necessarily implies decreased evidence for the other response (e.g., fewer dots moving right).

Next, consider which parameters could vary between left- and right-moving stimuli. In Forstmann et al.'s (2008) experiment, left- and right-moving stimuli were randomly ordered over trials. It is typically assumed that changing response threshold settings is a relatively slow process, so threshold parameters (b and A) cannot be adjusted in response to which particular stimulus has been presented for the current decision (Ratcliff, 1978). More generally, b and A are usually fixed across conditions whenever the participant is unable to predict which of those conditions will occur next. Since the other parameters (v, s and t_0) are assumed to be determined largely by the properties of the presented stimuli, they should be free to vary across stimuli. For example, left-moving stimuli may, for some reason, provide more salient motion cues than right-moving stimuli, which should be reflected in a higher drift rate.

Finally, consider which parameters might vary between speed, neutral and accuracy emphasis conditions. Following convention, between-trial variability in drift rate (s) is usually fixed across experimental conditions, particularly those not stimulus-based – although this assumption is not strictly necessary. All other parameters (b, A, v and t_0) could feasibly

be influenced by response emphasis.

Together, this relatively liberal set of constraints, based on theoretical plausibility, and some conventions, reduces the number of free parameters from 60 to 26. To keep notation compact, we subscript the parameter names with "left" or "right", to indicate left-moving vs. right-moving stimuli, and with "L" or "R" to indicate evidence accumulators corresponding to "left" vs. "right" responses. The 26 possible free parameters which remain are: s_{left}, s_{right} and three sets of b_L, b_R, A_L, A_R, v_{left}, v_{right}, $t_{0_{left}}$ and $t_{0_{right}}$, one for each emphasis condition.

Which parameters need to change to fit the data?.

We next examine the data to identify which of the 26 free parameters are actually needed. There are two ways this has been approached in the literature. The first is to fit the model to each participant's data with all 26 free parameters, and then examine how parameter estimates differ across conditions. To demonstrate this approach, we fit each individual's data from Forstmann et al.'s (2008) experiment using maximum likelihood estimation (MLE) and a SIMPLEX search algorithm (see the appendix for computational details). As discussed in the appendix, obtaining good parameter estimates for such a complex model (26 free parameters) is not easy – a drawback of this particular approach.

Figure 2 suggests some general ideas about which parameters might vary across conditions. For example, both the drift rate (v) and its standard deviation across trials (s) were larger for left-moving than right-moving stimuli. However, the much smaller difference (relative to the standard error bars) for the non-decision time plot suggests that t_0 probably did not change between stimuli. Similarly, the evidence threshold (b) and start point variability (A) parameters did not change much between "left" and "right" responses. These two parameters, however, changed substantially between the three response-caution conditions (left to right across the plots). Non-decision time and drift rate showed much smaller changes between emphasis conditions.

Differences between parameter estimates can be tested for statistical reliability using a repeated measures analysis of variance (ANOVA). The results of these tests help to decide which parameters are significantly affected by which manipulations. However, such tests bear only on the question of whether the population means for the parameters vary between conditions. It is quite possible that there is no difference in population means between conditions and yet there is substantial and systematic variation among conditions due to individual differences. If this is the case fixing parameters to be the same over conditions for fits to individual participants may introduce substantial misfit that will distort the model selection process. A full solution to this dilemma requires a hierarchical approach, which produces explicit estimates of population means and variability for each parameter type by fitting the model to all data from a group of participants simultaneously. Vandekerckhove, Tuerlinckx, and Lee (2008) develop this approach for Ratcliff's diffusion and we are presently doing the same for the LBA (see Donkin, Averell, et al., 2009). However, hierarchical versions of the models impose greater computational burdens, so in this tutorial we focus on methods based on fitting each participant's data separately.

Given the limitations of this initial free-fitting method, we recommend it be augmented with a second method based on sequential model building. This method also uses individual analysis (without between-subject hierarchies), but it is still sensitive to the need

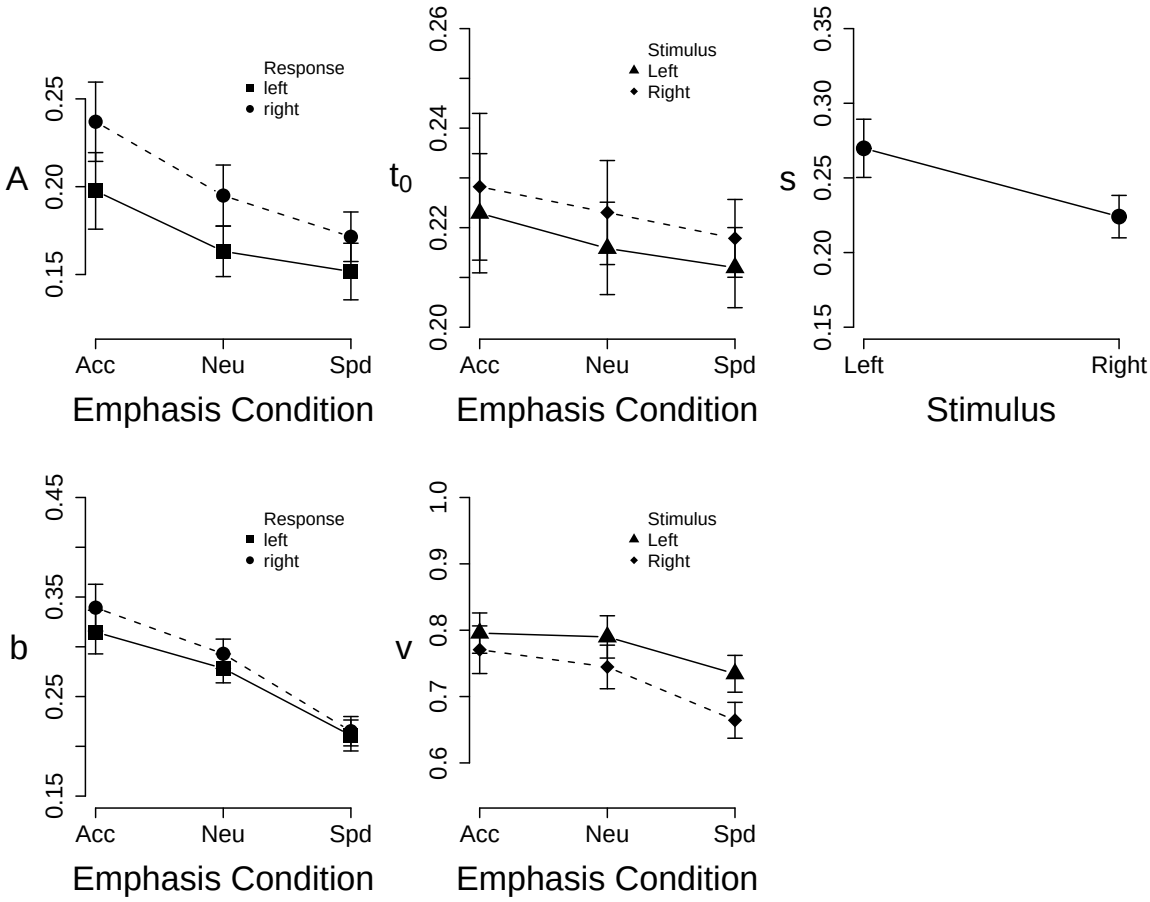

Figure 2. Parameter estimates averaged over participants across emphasis conditions, responses and stimuli. Error bars are +/-1 standard error.

to allow for individual differences. The key to this approach is to fit many different versions of the model, beginning with the simplest version (identical parameters for all conditions; only five free parameters in total) and ending with the most complex (with 26 free parameters, in our example). This approach can be computationally demanding because there might be very many intermediate models to analyze. The intermediate models are formed by considering all factorial combinations of parameter constraints. For example, after estimating the simplest model, one might next estimate a model where drift rate was free to vary between left-moving and right-moving stimuli. After that, both of those first two models would be used to start parameter searches for even more complex models, perhaps with the boundary separation parameter free to vary across speed/accuracy emphasis conditions. This process continues through to the most complex model.

After estimating the parameters of all the intermediate models, one model is selected that best satisfies the trade-off between simplicity and goodness-of-fit. Since each intermediate model is nested within a more complex version, each model's goodness-of-fit must

necessarily improve with the number of estimated parameters. However, when the improvement is small it may be due to "over-fitting", where the extra parameters serve only to account for unsystematic noise in the data. Such over-fitting is undesirable because it leads to a model that poorly predicts new data, and may produce theoretically misleading patterns of parameter estimates.

One method of identifying over-fitting is to choose the model with the smallest AIC (Akaike Information Criterion, Akaike, 1974) or BIC (the Bayesian Information Criterion, Schwarz, 1978). Parameter estimation using maximum likelihood is most appropriate for this purpose as both information criteria are calculated by adding a penalty to -2 times the log-likelihood of the model. The penalty quantifies model complexity based on the number of estimated model parameters, k ($2k$ for AIC and $\log(\mathrm{N})\times k$ for BIC, where N is the number of data points).

We use BIC in our application, as the AIC is known to choose overly complex models in large samples, although we acknowledge that there are other grounds on which to prefer some other approaches. Both methods are also limited because they do not take account of differences in functional form complexity (Pitt & Myung, 2002) between models: that is, both statistics treat all parameters as equal in terms of model complexity, but this may not be true. For example, even when two models have the same number of parameters one model may have more flexibility in fitting data due to differences in the way that the models restrict interactions amongst parameters. Model selection methods that address this issue, such as the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002) require Bayesian estimation (see also Donkin, Averell, et al., 2009).

The ideal data driven version of this second, nested-model, approach requires fitting of all factorial combinations of restrictions on the 26 parameters. However, that is not always feasible because it can require estimating parameters for many thousands of models. If the computational load is too great, one can make an initial simplification by fixing parameters whenever the free estimates (from Figure 2) suggest that those parameters do not change across conditions. For example, non-decision time does not appear to be influenced much by either different stimuli or emphasis conditions, suggesting that just one t_0 estimate will do for all six conditions. Similarly, response threshold and start point variability appear not to vary across different responses, but only for different emphasis conditions. Based on these observations we might narrow down the options to a model with five different ways that parameters could vary over conditions, and 15 free parameters: b, A, v_{left} and v_{right} for the three emphasis conditions, and one each of s_{right}, s_{left} and t_0. It is important at this point to remember the limitations, discussed above, of making inferences about the population parameters based on average parameter estimates from the minimally-constrained model. These limitations mean that the short-cut method we used to move from the 26-parameter model to the 15-parameter model should only be employed when the computational burden associated with exhaustively estimating the intermediate models is too great.

We also note that the interpretation of Figure 2 and the subsequent choices about which parameters should be fixed have a subjective element. For example, we chose to fix A across responses but let it vary across emphasis conditions. It is certainly arguable from the upper left plot in Figure 2 that either both or neither form of parameter variability may be necessary. A more formal method would be to perform ANOVA on the parameter estimates, but we are cautious about recommending the blind application of this approach given its

inherent limitations in terms of providing positive evidence in favor of a null difference. Instead, when it is not clear whether a parameter might vary across conditions, it is best to use the methods outlined in the next section to further investigate. Although we do not allow A to vary across responses in the following analysis, further investigation suggested that A should be fixed across both responses and emphasis conditions – we demonstrate this for emphasis conditions, but not responses, below.

Model Selection Example

The 15 free parameters that may account for our data are generated from five ways that parameters might vary across experimental conditions; b, A, and v might vary over emphasis condition, and v and s might vary across stimuli; for convenience, we will call these the five "features" of the model. One way to determine which of these features are required by the data is to fit all 31 possible combinations of models made up of these features, i.e. one model with all five features, the five models with four features, the 10 models with three features, etc., and select the best model using BIC values. This method is by far the most comprehensive, and indeed could be extended out to all nine features of parameter variations which were earlier deemed plausible. Such an approach using all nine features would require 511 models to be fit separately to each participant's data. Forstmann et al.'s (2008) model selection analysis was based on an exhaustive evaluation of set of models of this size, although with slightly different features[4].

We estimated the parameters for all 31 possible combinations of models made up of the five features. Models were fit individually to each of the 20 participants and for each participant a set of parameters and a BIC was calculated. We use BIC summed over participants, which we will call "total BIC", to describe group-level results. The results of the total BIC analysis for the most complex model and the intermediate model which yielded the smallest BIC value are reported in Table 1. The most complex model, with all five features, has a much larger BIC value (-15226) than the best intermediate model (-15653) indicating that this intermediate model provides a much better compromise between goodness-of-fit and model complexity[5]. The best intermediate model has only two features, and eight parameters. Averaged across participants, those parameters were: $v_{left} = 0.72sec^{-1}$ and $v_{right} = 0.67sec^{-1}$, $b_{acc} = 0.29$, $b_{neu} = 0.27$, $b_{speed} = 0.17$, $s = 0.22sec^{-1}$, $A = 0.15$ and $t_0 = 0.11sec$.

It is interesting to note that the average parameters estimated for our eight parameter model, with the exception of t_0, roughly correspond with the values freely estimated for the speed condition in Figure 2. For example, in Figure 2 average drift rate for left stimuli, 0.73, is close to the average value of 0.72 estimated in our eight parameter model. Similarly, drift rate for right stimuli is 0.66 compared with our estimate of 0.67 in our reduced model. Because RTs are less variable in the speed condition we might expect to see the model under predict variability in accuracy and neutral conditions. The model compensates, however,

[4]Sequential model selection techniques, such as the forward, backward and stepwise methods commonly used in linear regression model selection, provide an alternative method of reducing computational cost. Although not discussed further here we note that such techniques could be applied to selection amongst choice RT models, either based on likelihood ratio tests, as is common practice with linear regression models, or based on BIC (Hoeting, Madigan, Raftery, & Volinsky, 1999).

[5]See Wagenmakers and Farrell (2004) for formal methods of comparing BIC differences that can be employed when results are not so clear cut

Table 1: Which parameters were allowed to vary over which factors (Responses: "left" or "right", Stimuli: Left or Right, and Emphasis: Accuracy, Neutral or Speed), number of parameters (k), and total BIC of the best fitting model from each step of the model selection procedure we used.

	Factor							
Model	Responses	Stimuli		Emphasis			k	Σ BIC
Most Complex	-	v	s	b	A	v	15	-15226
Best Intermediate	-	v	-	b	-	-	8	-15653

by setting t_0 to be smaller in the eight parameter model (0.11 compared to around 0.22 in Figure 2) because this increases RT variance when response thresholds are large.

These results suggest that the participants reported by Forstmann et al. (2008) were able to extract information from left-moving stimuli around 8% faster than for right-moving stimuli. Also, manipulation of response emphasis affected only one cognitive process: the amount of evidence required before responding. Relative to the neutral condition, participants set evidence thresholds 7% higher under accuracy emphasis and 37% lower under speed emphasis. Note that the final model chosen in the present analysis is very similar to that selected in the original paper, except that the model selected here constrained between-trial drift rate variability for left- and right-moving stimuli (this possibility was not examined in the original analysis). However, our results are consistent with the major conclusion of the original paper, that the emphasis instructions selectively affect the response threshold.

Evaluating and Presenting Model Fit

Another important selection criterion is the descriptive adequacy of the model, which can assessed graphically. A model is inadequate if it fails to describe important patterns in the data. Similarly, if the parameter estimates vary across conditions in ways that make no sense, the model is suspect in terms of its theoretical adequacy. The average parameter estimates for the best intermediate model given in the last section appear to be adequate on the latter grounds, as did the corresponding parameter estimates for all individuals. In this section we describe how to graphically check model adequacy.

The match between model and data should be assessed for each individual participant. However, the final communication of results almost always requires a summary of the grouped data. Such averages can fail to represent the individual participants, depending on how they are constructed. As an extreme example, suppose that an experiment had just two participants, one who responded very quickly and another who responded very slowly. In this case, an "average" histogram formed by simply pooling data could be bimodal, and so not be representative of either individual[6]. It is often better to first calculate statistics which summarize the RT distribution and then average those. Regardless of the method used, one should always check how well averaged data matches the individual participants.

[6]Even though data grouped this way will not necessarily look like any individual's data, a similarly-grouped graph of the model predictions still provides a valid assessment of model adequacy.

The agreement between model and data is usually assessed by plotting together predicted and observed statistics that summarize RT distributions and response probabilities. Histograms depicting the observed RT distribution are often overlaid with the predicted PDF from the model, to assess model fit. Such plots are simple to interpret, but do not always highlight the shortcomings of the model. Cumulative probability plots (e.g. Forstmann et al., 2008), or quantile-probability (QP) plots (e.g. Ratcliff & Smith, 2004), are a little more complicated to produce and read, but can better reflect differences between the model and data. Group QP or cumulative probability plots, which are obtained by averaging quantiles for each individual, also have the advantage that they tend to be more representative of individual results (e.g., such averages do not suffer from the bi-modality problem that occurs with histograms). To represent the model predictions on these plots at the group level, one calculates the model's predicted quantiles for each individual and averages these together in the same way as the data. This means that we apply the same averaging process to create summary information for model predictions as for the data, and so both summaries are subjected equally to any distorting effects of averaging.

Figure 3 summarizes Forstmann et al.'s (2008) data and the corresponding LBA model fits using QP plots. QP plots are an efficient way of displaying the important information from a set of choice RT data – the horizontal axis contains response probability (accuracy) information and the vertical axis contains information about the RT distribution. There are six QP plots in Figure 3, with each plot representing the RT distributions from a single experimental condition. Each plot contains two sets of vertically aligned points, illustrating the RT distributions for correct and incorrect responses from one experimental condition. The horizontal position of a set of vertically aligned points represents the proportion of responses making up that RT distribution. For example, in the top left panel of Figure 3 the observed quantiles (solid squares) sit above 0.89 on the horizontal axis, indicating that 89% of responses were correct in that condition (left-moving stimuli under accuracy emphasis). Note also that this implies that 11% (i.e., 100%-89%) of responses were incorrect, and that the quantiles for these errors do indeed sit at 0.11 on the horizontal axis. In general, points to the left and right of 0.5 on a QP plot indicate incorrect and correct responses, respectively. The position of points on the vertical axis are determined by a set of five quantile estimates (.1, .3, .5, .7 and .9). The .1 quantile estimate corresponds to the value below which .1, or 10%, of the RT values in the distribution fall. The five quantile values together, therefore, summarize the RT distribution. For example, the first filled square above 0.89 in the top left panel shows the .1 quantile for correct responses in that condition. The next square above this shows the .3 quantile, and so on. The unfilled squares provide the same information, but for the distributions predicted by the LBA model, rather than for the observed data.

Note that QP plots can be constructed with any desired set of quantiles, such as with deciles or semi-deciles. Using more than five quantile estimates will provide a more detailed description of the RT distributions, but can also make the plots difficult to read. Similarly, results for more than one condition can be given in the same graph. This often works well when the conditions differ sufficiently in accuracy. For example, we could have given results for accuracy and speed conditions in the same panel. In contrast, the neutral and accuracy conditions are quite similar in accuracy, so providing results for these two conditions in the one plot can make the QP plot hard to interpret.

Figure 3 reveals the following general patterns: Responses in the speed emphasis con-

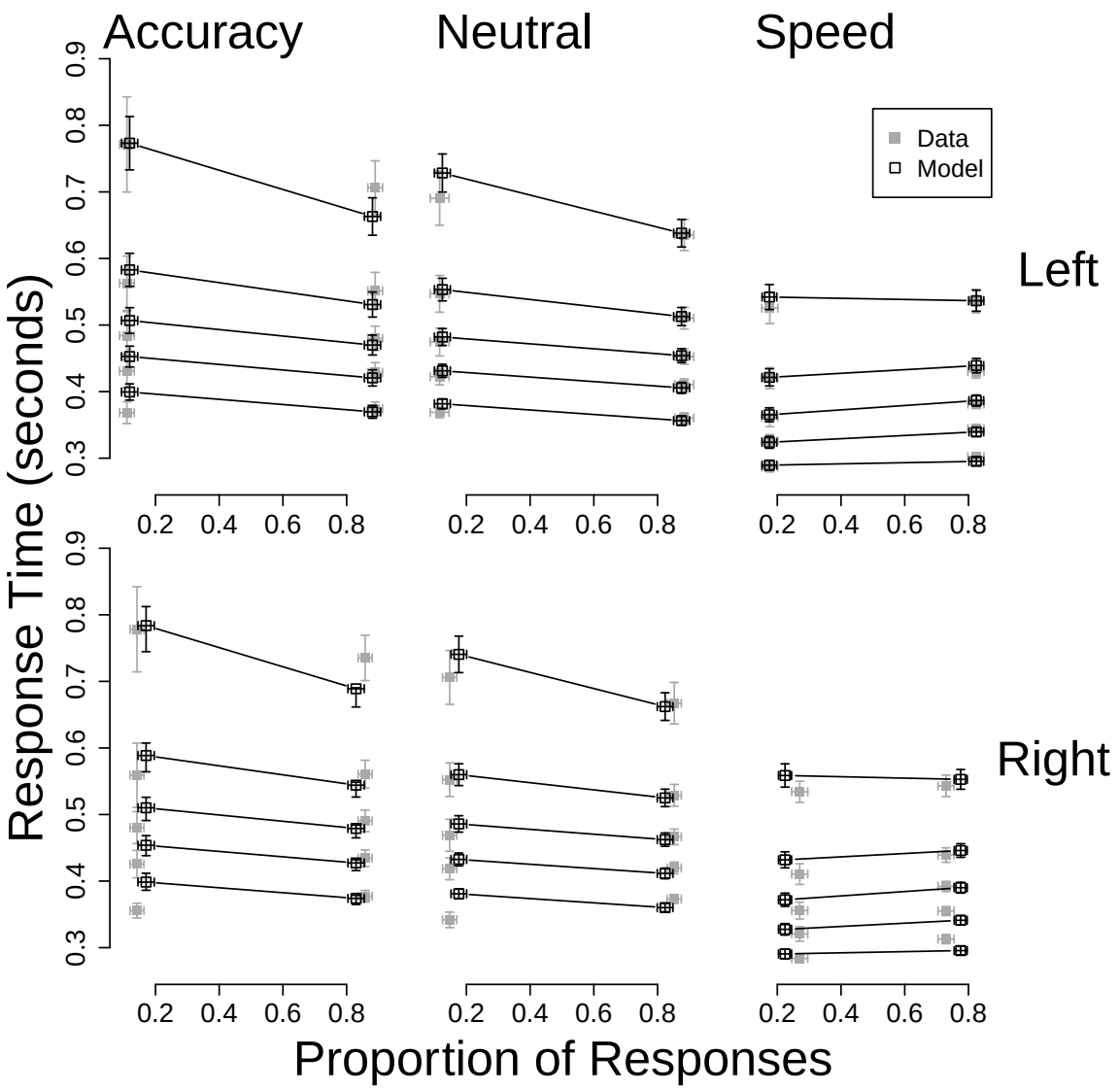

Figure 3. A quantile probability plot for the data from Forstmann et al. (2008). Observed and predicted quantiles are represented by solid and open symbols, respectively. Responses to left-moving and right-moving stimuli are represented in the top and bottom rows, respectively. Accuracy, neutral and speed emphasis conditions are shown in the left, center and right columns, respectively. Error bars show standard errors across participants for both data and model predictions.

ditions (right column of Figure 3) are faster, as indicated by their lower position on the vertical axis, than in accuracy and neutral conditions (left and center columns, respectively). Responses for left-moving stimuli (top row) are more accurate than for right-moving stimuli (bottom row). This shows up in the figure because quantiles in the top row sit at more extreme horizontal positions than those in the bottom row. For example, quantiles for left stimuli in the speed condition are positioned at 0.18 and 0.82 on the horizontal axis, while the same quantiles for right stimuli sit above 0.27 and 0.73. The addition of lines joining corresponding quantiles for correct and incorrect responses in Figure 3 highlights a theoretically important issue, the relative speed of correct and incorrect responses for different emphasis conditions. In these data, incorrect responses are faster than correct responses in the speed-emphasis condition, but this difference is reversed in the accuracy-emphasis condition. The model does a good job of accounting for these patterns. However, the QP plot also reveals shortcomings of the model, with the most evident being a tendency to predict too many incorrect responses for right-moving stimuli. If this failing were considered to be practically or theoretically important selection of a more complex model that addresses this issue might be warranted.

Producing a QP plot requires calculation of the .1, .3, .5, .7, and .9 quantiles for observed and predicted RT distributions. Quantile estimates from the observed data can be calculated using functions available in most statistical software (for more details see Heathcote et al., 2002; Van Zandt, 2000). Quantiles were calculated for each individual participant and then averaged together to create the observed quantiles in Figure 3. It is always important to check that the summary information to be presented in a QP plot is representative of individuals. In the current data set over eighty percent of individual quantile estimates were within 50 ms of their respective average values, suggesting that our averages were representative of individual RT distributions.

Calculating the quantile values predicted by the model is a little more difficult. There are two standard approaches: either using a search algorithm to invert the cumulative distribution function (CDF) of the model, or via simulation. To generate the predictions shown in Figure 3, we used the conceptually simpler, but computationally more expensive, simulation method (see our appendix for details on the search method). To calculate predicted quantiles via simulation, we took each individual's best fitting parameters and used them to simulate a large amount of data (one million data points in each condition). The simulated data followed the exact same design as the empirical data – i.e., three emphasis conditions and two stimulus conditions, where only drift rate changes over stimulus conditions and only response threshold changes over emphasis conditions. For details on simulating data from the LBA see Donkin, Averell, et al. (2009), and for other choice RT models see Brown, Ratcliff, and Smith (2006). Finally, we calculated quantiles from these simulated data, and averaged across participants, in the same way as for the observed data.

Rather than plotting the predicted quantiles averaged over individuals, Ratcliff and colleagues suggest fitting the model to the average observed quantiles to create model predictions for QP plots (e.g., Ratcliff, 2002; Ratcliff, Gomez, & McKoon, 2004). This approach can appear to indicate a better fit than the method we describe here, since model predictions will be based on parameters which optimize that fit. However, one risk with this approach is that the newly estimated parameters may not be representative of the parameters of any individual. For this reason, Ratcliff and colleagues always assess how closely

these new parameters match the average individual participant parameters. Further, a quantile-based objective function must be used for estimating parameters from the average observed quantiles; MLE cannot be used (see the appendix).

Discussion

In recent years, the once-difficult process of estimating parameters for choice RT models has been made much easier by the provision of software that automates the search process. Our aim was to build on these developments by providing advice about, and a detailed example of, the many extra steps involved in moving from simple parameter estimation to a more meaningful analysis. We focused on an application of the LBA model (Brown & Heathcote, 2008) to data reported by Forstmann et al. (2008). We illustrated the canonical problem in such modeling, by first describing how there are – potentially – 60 free parameters even for Forstmann et al.'s quite simple experiment. We then illustrated how this number can be reduced to 26 free parameters in our example by *a priori* considerations. Exploratory analysis of the 26 parameter model identified where parameter estimates changed substantially across experimental conditions. This resulted in an even simpler model with 15 parameters. Finally, we exhaustively fit all 31 simplifications of the 15 parameter model and selected a final model with 8 free parameters, providing the best trade-off between goodness-of-fit and model complexity.

We also showed how to check the descriptive adequacy of the final model using QP plots. The selected model provided a good fit that captured theoretically important features of the data, and is consistent with results from applications of alternative evidence accumulation models to similar paradigms (e.g., Ratcliff & Rouder, 1998), and with Forstmann et al.'s (2008) conclusion that a manipulation of response emphasis exclusively influenced the amount of evidence required for a decision.

The particular model selection process we described relies on models formed by fixing some parameters across different conditions. In a between-subjects manipulation, different conditions are populated by different people, meaning that certain parameters would have to be fixed across participants – this requires modeling random effects. Most often for choice RT analyses, this problem is handled by estimating model parameters separately for individual subjects, then using standard null-hypothesis significance testing (NHST) to determine which parameters vary across the between-subjects conditions. As an example, imagine that Forstmann et al. (2008) had tested both an older and a younger group of participants and thus had an additional between-subject factor. Standard practice would then be to fit each individual from both the younger and older groups with the model we previously selected, giving observed distributions of each parameter for younger and older participants. Standard inferential tests could then be used to determine whether the average of certain parameters differ between younger and older groups. For example, an independent-samples t-test could be used to determine whether the average non-decision time parameter is different for older and younger participants. This approach has been used to identify the effects of aging on decisional processes (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2007).

The method of using NHST to determine differences between subjects carries with it all of the usual drawbacks. These may be particularly relevant for the application of choice RT models because the typical question is whether some parameter does *not* change across

conditions. For example, Ratcliff and colleagues often find that old and young participants do not differ significantly in their drift rate parameters. It is difficult to know if this lack of significant finding is due to power or whether drift rate is truly equal for old and young participants. Bayesian hypothesis tests such as the Savage-Dickey test (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) allow for direct assessment of the truth of the null hypothesis but require the posterior distribution of model parameters. Donkin, Averell, et al. (2009) and Vandekerckhove et al. (2008) offer software for producing these distributions, but at present the approach is limited by its computational cost.

Hierarchical modeling, in which parameters are estimated for the distribution of parameters at the population level, provides another way to determine differences across both between-subject and within-subject conditions. For example, hyper-parameters describing the distribution over younger and older groups for particular types of choice RT model parameters can be compared using Bayesian model selection techniques. Despite the theoretical merits associated with hierarchical Bayesian parameter estimation, most researchers still use the approaches presented in the current tutorial because they are many orders of magnitude faster than Markov-Chain Monte Carlo estimation methods. Given how much slower these methods tend to be, we expect the individual analysis approach of this tutorial to remain relevant for some time to come. Further, the central issue we discuss – choosing a set of model constraints from among many possible sets – apply equally to hierarchical models and Bayesian estimation.

References

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

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

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

Brown, S., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*, 396–425.

Brown, S., Ratcliff, R., & Smith, P. (2006). Evaluating methods for approximating stochastic differential equations. *Journal of Mathematical Psychology*, *50*, 402-410.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*, 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making. *Psychological Review*, *100*, 432–459.

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*, *41*, 1095–1110.

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

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E. J. (2009). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Manuscript submitted for publication*.

Farrell, S., Ratcliff, R., Cherian, A., & Segraves, M. (2006). Modeling unidimensional categorization in monkeys. *Learning and Behavior*, *34*, 86–101.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). The striatum facilitates decision-making under time pressure. *Proceedings of the National Academy of Science*, *105*, 17538–17542.

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

Heathcote, A., Brown, S. D., & Cousineau, D. (2004). Qmpe: Estimating lognormal, wald and weibull rt distributions with a parameter dependent lower bound. *Behavior Research Methods, Instruments, & Computers*, *36*, 277–290.

Heathcote, A., Brown, S. D., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*, 394–401.

Ho, T., Brown, S., & Serences, J. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29*, 8675–8687.

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

Ludwig, C. J., Farell, S., Ellis, L. A., & Gilchrist, I. D. (in press). The mechanism underlying inhibition of saccadic return. *Cognitive Psychology*.

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of exgaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

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

Pike, A. R. (1966). Stochastic models of choice behaviour: Response probabilities and latencies of finite Markov chain systems. *British Journal of Mathematical and Statistical Psychology*, *21*, 161–182.

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

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

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

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two–choice reaction time. *Psychological Review*, *111*, 333–367.

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*, 278–289.

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*, 56–66.

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*, 438–481.

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*, 414–429.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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

Smith, P. L., & Vickers, D. (1988). The accumulator model of two–choice discrimination. *Journal of Mathematical Psychology*, *32*, 135–168.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*, 583–639.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*, 702-716.

Tuerlinckx, F., Maris, E., Ratcliff, R., & De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior Research Methods, Instruments, & Computers*, *33*, 443–456.

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

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

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

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). Hierarchical Bayesian diffusion models for two–choice response times. *Manuscript submitted for publication*.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424-465.

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

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

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

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192-196.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage-dickey method. *Cognitive Psychology*, *60*, 158–189.

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

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion*, *23*, 181–205.

White, C., Ratcliff, R., Vasey, M., & McKoon, G. (in press). Sequential sampling models and psychopathology: Anxiety and reaction to errors. *Journal of Mathematical Psychology*.

Appendix – Additional Details

The objective function

A widely used objective function is the likelihood of data $\mathbf{x}$ given a model with parameter $\boldsymbol{\theta}$: $L(\mathbf{x}|\boldsymbol{\theta})$. Finding parameter estimates by optimizing this objective function is called maximum likelihood estimation (MLE, see Myung, 2003 for a tutorial on MLE methods). Choice RT models predict RT distributions with associated likelihood functions that can be used for MLE. The likelihood function corresponds to the model's joint density over the latency of the response (i.e., RT) and the identity of the response (i.e., the choice made). Thus, MLE based on these joint densities naturally takes into account both accuracy and RT information (see Brown & Heathcote, 2008, for details of the LBA model's joint density functions).

MLE is a default choice in many areas of statistics because it is unbiased for large samples, and because no other method is more efficient, as long as certain regularity conditions on the model are satisfied. However, choice RT models do not usually satisfy these conditions because they predict distributions whose support is determined by an estimated parameter, t_0 (Heathcote, Brown, & Cousineau, 2004). This can cause maximum likelihood methods to spuriously estimate t_0 as equal to the minimum RT in a data sample, with concomitant mis-estimation of the other parameters. Although it is important to be aware of this problem, it can usually be avoided by censoring implausibly fast RT data (e.g., responses faster than 200ms, which are likely the result of anticipation). Slow outliers, due to processes such as distraction, are more problematic as they are harder to detect than fast outliers. Heathcote et al. (2002) showed that – even when estimating simple and regular RT models – an estimation method based on data quantiles (quantile maximum probability estimation, QMPE, see also Heathcote & Brown, 2004) could be more efficient and less biased than MLE in small samples.

Appropriately selected quantiles can summarize an RT distribution, and more quantiles lead to a more accurate summary. There are several objective functions that use quantiles to summarize the observed RT distributions, and compare these against model predictions. Besides QMPE, these functions include the Kolmogorov-Smirnov statistic (Voss, Rothermund, & Voss, 2004; Voss & Voss, 2007), χ^2 (Ratcliff, 2002), and weighted least squares error (Ratcliff & Tuerlinckx, 2002). In practice, the quantile-based objective functions require evaluation of the model's cumulative distribution function (CDF); this is different from MLE which requires evaluation of the model's probability density function (PDF). The difference means that quantile-based methods are especially useful for models that have easy-to-use algorithms to calculate their CDF, but not PDF (such as Ratcliff's diffusion model).

With the exception of the Kolmogorov-Smirnov based approach, choice RT modelers have mostly used a coarse set of five quantiles: .1, .3, .5, .7, and .9 (Ratcliff & Tuerlinckx, 2002), which we will describe as the "standard" quantile set. Summarizing the observed RT distributions with such a coarse set has the advantage that fitting is only weakly influenced by fast and slow outliers. For example, even if 2% of the data were from a fast- guessing contaminant process, these would all fall below the 10% quantile estimate, which would thus be only mildly affected. Note, however, that there is no necessity for a coarse set of quantiles between the smallest and largest values to gain this advantage in robustness, and

that Heathcote and Brown (2004) found that the advantages of QMPE in small samples only emerged with fine-grained quantile sets.

Forstmann et al. (2008) analyzed their data using QMPE. Above, we reported analyses of the same data using MLE, and we found that the two approaches agreed closely in this data set. In general, we have found that, as long as sensible precautions are enforced pertaining to outliers, and good starting points are used, MLE performs very well for the LBA model. MLE also enjoys a substantial advantage over quantile based methods in terms of computational speed. A further advantage is that MLE produces true maximized likelihood values, which can be useful for performing the model selection analyses discussed in the main body of the paper.

Finding optimal parameters

A variety of optimization methods are available, but in our experience these algorithms differ mostly in speed rather than their ability to find the best set of parameters. Here we use the SIMPLEX algorithm (Nelder & Mead, 1965), which is the most commonly used search algorithm for choice RT fitting because of its ease of use. More computationally efficient optimization algorithms, which require analytic derivatives of the objective function, are not generally used because the required derivatives are not easily available. Other algorithms operate by numerically estimating derivatives, which can improve efficiency, but also decrease numerical stability.

All parameter search algorithms, such as SIMPLEX, need a set of parameters to begin their search, and the consequences of a poor set of starting parameters can be dire – the parameter search can get stuck on an estimate that matches the data better than all nearby parameter sets but which is much worse than estimates further away (a "local optimum"). The search should always begin with parameters that produce model predictions that are reasonably close to the data. Identifying such starting values is a large problem in itself, with no general solution. Typically, it is best to have two or three different ways to generate start values, and then evaluate the goodness-of-fit (i.e., the objective function) at each, and choose the best. One easy way to generate start points is to use parameters which have been reliably shown to produce reasonable predictions for RT data. For example, Matzke and Wagenmakers (2009) provide the average parameter values for the diffusion model based on a large number of fits of the model to data, and Donkin, Brown, Heathcote, and Wagenmakers (2009) provide equivalent values for the LBA. Another approach to address estimation failures is to run a number of searches from a range of start points obtained by randomly perturbing the initial start point.

Alternatively, a good set of start points for parameter search might be generated by heuristics, which can be automatically applied to any data to be fit. We have found the following set of heuristics useful for the LBA model. We first set the drift rate distribution parameters: standard deviation, $s = 0.3$, and mean, $v = \frac{1}{2} + \frac{1}{2}\Phi(\frac{p}{s\sqrt{2}})$, where p is the probability of making the response for this accumulator, and Φ is the standard normal CDF. For t_0 we use 9/10 of the value of the minimum RT from the data. For the maximum of the uniform start point distribution, A, we take twice the inter-quartile range of the RT distribution, and finally we set the response threshold, b, at $1.25 \times A$. In general, we calculate these heuristic values separately for each experimental condition, and then average start points for parameters that are constrained across conditions.

Even when a good start point is obtained it is also possible that the search algorithm may terminate in a local optimum. When using the SIMPLEX algorithm, this can sometimes be avoided by performing a sequence of searches, with each new search using the best estimate found by the last fit as its starting point. This method can work because each new search typically starts with a large simplex that explores parameter values relatively distant from those explored in the final stages of the previous search.

As experimental designs increase in complexity so too can the number of parameters that must be estimated, making the search more difficult. For example, if an experiment has different levels of task difficulty, it is likely that different drift rates should be estimated for those conditions, increasing the total number of free parameters. Nested model fitting is a useful technique for overcoming this problem. This involves first fitting simple models (such as a model with a single drift rate parameter for all difficulty conditions), and later using the resulting parameter estimates as start points when fitting increasingly complicated models. For example, consider data from a 2x3 factorial design. Suppose the two factors, A and B, are both assumed to affect drift rate. We might first fit a simple model with just one drift rate for all six conditions, starting from parameters obtained by averaging heuristic estimates based on data from each condition: say that the best-fitting drift rate value was 1. We could then fit a model in which drift rate varied across the two levels of factor A, say A1 and A2, using 1 as the start point for both levels. Suppose the best fitting parameters turned out to be 0.5 and 1.5 for A1 and A2, respectively. The two best fitting parameters can then serve to create start points for the full factorial model in which drift rate varies over both factors A and B (i.e., the start points for the three levels of factor B in A1 would be 0.5, and the start points for the three levels of A2 would be 1.5). However, even under this approach there are no guarantees that a search algorithm will find the best parameter estimates, and this becomes increasingly true as the number of estimated parameters increases[7]. Hence, it is important to check the quality of fits graphically, as we described in the main body of this paper, and to try different starting values in order to see if improvements can be obtained for any poor quality fits.

As well as checking goodness-of-fit graphically, the parameter estimates themselves should also be checked for *a priori* plausibility. Plausibility can be judged relative to typical parameter ranges (e.g., for the LBA see Donkin, Brown, Heathcote, & Wagenmakers, 2009). When fitting data from a number of participants performing under identical conditions consistency among corresponding parameters estimates from different participants can be assessed. It sometimes happens that these checks reveal wildly large or small estimates of a particular parameter. This usually occurs when the value of that parameter receives little constraint from the data (i.e., its value has little influence on the objective function). This lack of constraint can be seen by making a graph of the objective function for a range of values of the suspect parameter around its estimated value while keeping the remaining parameter values fixed at their estimated values. A flat graph indicates a poorly constrained parameter.

For the LBA, poor constraint is most commonly associated with the drift rate param-

[7]The performance of the SIMPLEX algorithm degrades markedly when the number of parameters being optimized is large. We have generally found adequate performance with up to 20-30 parameters, at least when repeated fits are employed. Beyond this range specialized search algorithms designed for dealing with high dimensional search spaces may be required.

eter for incorrect responses. This is particularly the case where observed accuracy is high, as there are then very few incorrect responses to constrain the estimate of this parameter. This problem did not occur with our example data, as Forstmann et al.'s (2008) participants made incorrect responses on 13% of trials even in the most accurate condition, and a large number of trials were performed, so this equates to about 36 error RT observations per participant in this condition. However, in paradigms where accuracy is very high, prohibitively many trials may be required to obtain enough error responses. Ludwig, Farell, Ellis, and Gilchrist (in press) describe a method of circumventing this problem, using the LBA, that relies only on an estimate of the proportion of error responses rather than using error RT.

An alternative approach, useful for addressing under-constraint for any type of parameter, is to constrain parameters to be the same across conditions based on theoretical considerations. For example, the manipulation of response caution used in Forstmann et al.'s (2008) experiment is commonly assumed to not effect drift rate parameters. If this constraint is enforced by assuming the same value of error drift rate across a range of conditions estimates of this parameter will be constrained as long as there are sufficiently many incorrect responses in total across all conditions.

Maximum likelihood estimation

In this section, we describe in detail how maximum likelihood estimation might be carried out for the example data above. First, consider data just from one emphasis condition in Forstmann et al. (2008). Let us assume that only drift rate differs between left and right responses. We could fit an LBA model for these data with five parameters to estimate: $(b, A, v_{left}, v_{right}, s, t_0)$. The v_{left} and v_{right} parameters represent mean drift rates for correct responses to left and right stimuli, respectively, and in the following we refer to them generically as v_c. We fix the mean drift rates for error responses at $v_e = 1 - v_c$ for both left and right stimuli.

The likelihood function for the LBA model (see Brown & Heathcote, 2008) is relatively easy to compute as it is specified in terms of basic functions and the integral of a normal distribution, which has fast and accurate numerical approximations. Brown and Heathcote also provide computer code that evaluates the likelihood function. These routines take in a set of parameter values, and a response time, say t, and return the probability that the accumulator corresponding to the first response has reached threshold before any other, and at time t. To get the likelihood for a *correct* response then the drift rate for the first accumulator is set at v_c and for the second accumulator at v_e (to get the likelihood for an *incorrect* response, one simply swaps the values of the drift rates to be v_e and v_c, respectively).

The LBA likelihood function for a single response corresponds to a *defective* (or joint, over accuracy and RT) probability density function (PDF), because it does not integrate to one, but instead to the probability that the corresponding accumulator was the first to reach threshold. For example, if the defective density for the accumulator for correct responses in the easiest condition integrates to .95 given a set of parameters, then the model predicts 95% correct responses in the easy condition. This also means that the same set of parameters will produce a defective likelihood function for the error accumulator that integrates to .05.

To construct the maximum likelihood objective function, we evaluate the likelihood

function at each and every observed RT value, and multiply them together – this gives the likelihood of parameter set θ given the entire data set. We use every RT value because we want the set of parameters $\theta = (b, A, v_{left}, v_{right}, s, t_0)$ that is most likely given the four RT distributions under consideration: correct and error responses for left and right stimuli. For every RT value in each of these distributions, we take the following steps:

1. Identify the appropriate drift rate (v_left or v_right) depending on the stimulus presented on the given trial (left or right).
2. If the response associated with this RT was correct, set v_1 to the drift rate identified from Step #1. If the response was incorrect, set v_1 to one minus the drift rate from Step #1.
3. Set v_2=1-v_1.
4. Subtract the parameter t_0 from the observed RT, as the likelihood equations given by Brown and Heathcote (2008) provide the likelihood for the *decision* time which is RT-t_0.
5. Using the equation for the defective PDF (Equation 3 in Brown & Heathcote, 2008), and the drift rates from Steps #2 and #3, and the parameters from above, evaluate the likelihood function for this observation.

Once this operation is performed for every observation, the likelihood function can be obtained by simply multiplying together all the likelihood values (from Step #5) for all the data. However, it is usual to instead take the logarithm of all likelihoods, and then add these log-likelihoods together, to improve numerical stability.

A fast method for producing predicted quantiles

Predicted quantiles can be obtained more quickly than through the simulation method described in the main text by evaluation of the inverse of the CDF of a model. The CDF, $F(t|\theta)$, of the model gives the proportion of responses made before time t, given parameters θ. To find predicted quantile values, we require the inverse of the CDF – that proportion of responses made before time t. For choice RT models, however, we are interested in the *defective* CDF, which does not necessarily reach a probability of 1 as t increases. For example, if a model predicted that only 65% of responses were accurate in a given task, the defective CDF for correct responses would only reach a probability of .65. To evaluate the predicted .1, .3, .5, .7 and .9 quantiles for the *correct* RT distribution, we must identify those values of t for which the defective CDF $F(t|\theta)$ equals $.1p, .3p, .5p, .7p$ and $.9p$, where p is the predicted response accuracy ($p = .65$ in our example). When calculating quantiles for the incorrect response distribution, the value p is replaced by $1-p$, which is the probability of an incorrect response ($1 - p = .35$ for errors in our example).

For example, consider the .1 quantile of a correct RT distribution. The LBA predicts that .1 of this RT distribution is reached at $F^{-1}(.1p)$, where p is the predicted proportion of correct responses. In other words the predicted .1 quantile value, say t, is the inverse defective CDF for correct responses evaluated at $.1p$. To get the predicted quantile value we first need to calculate the predicted value of p, which is done by evaluating the CDF

at ∞: $p = F(\infty)$. The inverse of the CDF does not have a closed-form expression that can be easily evaluated. Instead, we employ a numerical solution. We are attempting to solve $F^{-1}(0.1p) = t$, which is equivalent to $F(t) = .1p$. We are left, therefore, with an expression we will call Equation 1, $F(t) - .1p = 0$, which we can now solve using a root finding algorithm. The value of t returned by this algorithm, plus t_0, is exactly the .1 quantile prediction for correct responses. We can repeat this process for error responses by replacing all instances of p with $(1 - p)$.

The following steps summarize the calculation of a given quantile corresponding to probability q for correct and incorrect responses:

1. Do steps #1-#3 specified above for maximum likelihood estimation to get the appropriate sets of parameters for correct (θ_c) and incorrect (θ_e) responses.
2. Obtain the predicted proportion of correct responses (p) by evaluating the CDF at infinity given the parameters for say the correct response chosen in Step #1.
3. Use a root finding algorithm to get the correct and incorrect quantiles. These correspond to the value of t for which $F(t|\theta_c) = qp$ and $F(t|\theta_c) = q(1 - p)$, respectively.

A QP plot then involves plotting both data and model quantiles for correct and incorrect responses in each experimental condition, as explained in the main body of the manuscript.