Chapter 2
An Introduction to Good Practices in Cognitive Modeling

Andrew Heathcote, Scott D. Brown and Eric-Jan Wagenmakers

Abstract Cognitive modeling can provide important insights into the underlying causes of behavior, but the validity of those insights rests on careful model development and checking. We provide guidelines on five important aspects of the practice of cognitive modeling: parameter recovery, testing selective influence of experimental manipulations on model parameters, quantifying uncertainty in parameter estimates, testing and displaying model fit, and selecting among different model parameterizations and types of models. Each aspect is illustrated with examples.

2.1 Introduction

One of the central challenges for the study of the human mind is that cognitive processes cannot be directly observed. For example, most cognitive scientists feel confident that people can shift their attention, retrieve episodes from memory, and accumulate sensory information over time; unfortunately, these processes are latent and can only be measured indirectly, through their impact on overt behavior, such as task performance.

Another challenge, one that exacerbates the first, is that task performance is often the end result of an unknown combination of several different cognitive processes. Consider the task of deciding quickly whether an almost vertical line tilts slightly to the right or to the left. Even in this rather elementary task it is likely that at least four different factors interact to determine performance: (1) the speed with which perceptual processes encode the relevant attributes of the stimulus; (2) the

A. Heathcote (✉) · S. D. Brown
School of Psychology, University of Newcastle, University Avenue,
Callaghan, NSW 2308, Australia
e-mail: Andrew.Heathcote@newcastle.edu.au

S. D. Brown
e-mail: Scott.Brown@newcastle.edu.au

E. J. Wagenmakers
Department of Psychological Methods, University of Amsterdam,
Weesperplein 4, 1018 XA, Amsterdam, The Netherlands
e-mail: EJ.Wagenmakers@gmail.com

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_2

efficiency with which the perceptual evidence is accumulated; (3) the threshold level of perceptual evidence that an individual deems sufficient for making a decision; and (4) the speed with which a motor response can be executed after a decision has been made. Hence, observed behavior (i.e., response speed and percentage correct) cannot be used blindly to draw conclusions about one specific process of interest, such as the efficiency of perceptual information accumulation. Instead, one needs to untangle the different cognitive processes and estimate both the process of interest and the nuisance processes. In other words, observed task performance needs to be decomposed in terms of the separate contributions of relevant cognitive processes. Such decomposition almost always requires the use of a cognitive process model.

Cognitive process models describe how particular combinations of cognitive processes and mechanisms give rise to observed behavior. For example, the linear ballistic accumulator model (LBA; [1]) assumes that in the line-tilt task there exist two accumulators—one for each response—that each race towards an evidence threshold. The psychological processes in the LBA model are quantified by parameters; for instance, the threshold parameter reflects response caution. Given the model assumptions, the observed data can be used to estimate model parameters, and so draw conclusions about the latent psychological processes that drive task performance. This procedure is called cognitive modeling (see Chap. 1 for details).

Cognitive modeling is perhaps the only way to isolate and identify the contribution of specific cognitive processes. Nevertheless, the validity of the conclusions hinges on the plausibility of the model. If the model does not provide an adequate account of the data, or if the model parameters do not correspond to the psychological processes of interest, then conclusions can be meaningless or even misleading. There are several guidelines and sanity checks that can guard against these problems. These guidelines are often implicit, unspoken, and passed on privately from advisor to student. The purpose of this chapter is to be explicit about the kinds of checks that are required before one can trust the conclusions from the model parameters. In each of five sections we provide a specific guideline and demonstrate its use with a concrete application.

2.2 Conduct Parameter Recovery Simulations

One of the most common goals when fitting a cognitive model to data is to estimate the parameters so that they can be compared across conditions, or across groups of people, illuminating the underlying causes of differences in behavior. For example, when Ratcliff and colleagues compared diffusion-model parameter estimates from older and younger participants, they found that the elderly were slower mainly due to greater caution rather than reduced information processing speed as had previously been assumed [2].

A basic assumption of investigations like these is adequate parameter recovery—that a given cognitive model and associated estimation procedure produces accurate and consistent parameter estimates given the available number of data points. For

standard statistical models there is a wealth of information about how accurately parameters can be recovered from data. This information lets researchers know when parameters estimated from data can, and cannot, be trusted. Models of this sort include standard statistical models (such as general linear models) and some of the simplest cognitive models (e.g., multinomial processing trees [3]).

However, many interesting cognitive models do not have well-understood estimation properties. Often the models are newly developed, or are new modifications of existing models, or sometimes they are just existing models whose parameter estimation properties have not been studied. In these cases it can be useful to conduct a parameter recovery simulation study. An extra advantage of running one's own parameter recovery simulation study is that the settings of the study (sample sizes, effect sizes, etc.) can be matched to the data set at hand, eliminating the need to extrapolate from past investigations. When implementing estimation of a model for the first time, parameter recovery with a large simulated sample size also provides an essential bug check.

The basic approach of a parameter recovery simulation study is to generate synthetic data from the model, which of course means that the true model parameters are known. The synthetic data can then be analysed using the same techniques applied to real data, and the recovered parameter estimates can be compared against the true values. This gives a sense of both the bias in the parameter estimation methods (accuracy), and the uncertainty that might be present in the estimates (reliability). If the researcher's goal is not just to estimate parameters, but in addition to discriminate between two or more competing theoretical accounts, a similar approach can be used to determine the accuracy of discrimination, called a "model recovery simulation". Synthetic data are generated from each model, fit using both models, and the results of the fits used to decide which model generated each synthetic data set. The accuracy of these decisions shows the reliability with which the models can be discriminated.

When conducting a parameter recovery simulation, it is important that the analysis methods (the model fitting or parameter estimation methods) are the same as those used in the analysis of real data. For example, both synthetic data and real data analyses should use the same settings for optimisation algorithms, sample sizes, and so on. Even the model parameters used to generate synthetic data should mirror those estimated from real data, to ensure effect sizes etc. are realistic. An exception to this rule is when parameter recovery simulations are used to investigate methodological questions, such as what sample size might be necessary in order to identify an effect of interest. If the researcher has in mind an effect of interest, parameter recovery simulations can be conducted with varying sizes of synthetic samples (both varying numbers of participants, and of data points per participant) to identify settings that will lead to reliable identification of the effect.

2.2.1 Examples of Parameter Recovery Simulations

Evidence accumulation models are frequently used to understand simple decisions, in paradigms from perception to reading, and short term memory to alcohol intoxication [4, 5, 6, 7, 8, 9]. The most frequently-used evidence accumulation models for analyses such as these are the diffusion model, the EZ-diffusion model, and the linear ballistic accumulator (LBA) model [10, 11, 1]. As the models have become more widely used in parameter estimation analyses, the need for parameter recovery simulations has grown. As part of addressing this problem, in previous work, Donkin and colleagues ran extensive parameter recovery simulations for the diffusion and LBA models [12]. A similar exercise was carried out just for the EZ diffusion model when it was proposed, showing how parameter estimates from that model vary when estimated from known data of varying sample sizes [11].

Donkin and colleagues also went one step further, and examined the nature of parameters estimated from wrongly-specified models [12]. They generated synthetic data from the diffusion model and the LBA model, and examined parameter estimates resulting from fitting those data with the other model (i.e., the wrong model). This showed that most of the core parameters of the two models were comparable—for example, if the non-decision parameter was changed in the data-generating model, the estimated non-decision parameter in the other model faithfully recovered that effect. There were, however, parameters for which such relationships did not hold, primarily the response-caution parameters. These results can help researchers understand when the results they conclude from analysing parameters of one model might translate to the parameters of the other model. They can also indicate when model-based inferences are and are not dependent on assumptions not shared by all models.

To appreciate the importance of parameter recovery studies, consider the work by van Ravenzwaaij and colleagues on the Balloon Analogue Risk Task (BART, [13]). On every trial of the BART, the participant is presented with a balloon that represents a specific monetary value. The participant has to decide whether to transfer the money to a virtual bank account or to pump the balloon, an action that increases the balloon's size and value. After the balloon has been pumped the participant is faced with the same choice again: transfer the money or pump the balloon. There is some probability, however, that pumping the balloon will make it burst and all the money associated with that balloon is lost. A trial finishes whenever the participant has transferred the money or the balloon has burst. The BART task was designed to measure propensity for risk-taking. However, as pointed out by Wallsten and colleagues, performance on the BART task can be influenced by multiple psychological processes [14]. To decompose observed behavior into psychological processes and obtain a separate estimate for the propensity to take risk, Wallsten and colleagues proposed a series of process models.

One of the Wallsten models for the BART task (i.e., "Model 3" from [14], their Table 2) has four parameters: α, β, γ^{+}, and μ. For the present purposes, the precise specification of the model and the meaning of the parameters is irrelevant (for a detailed description see [15, 14]). What is important here is that van Ravenzwaaij and colleagues conducted a series of studies to examine the parameter recovery

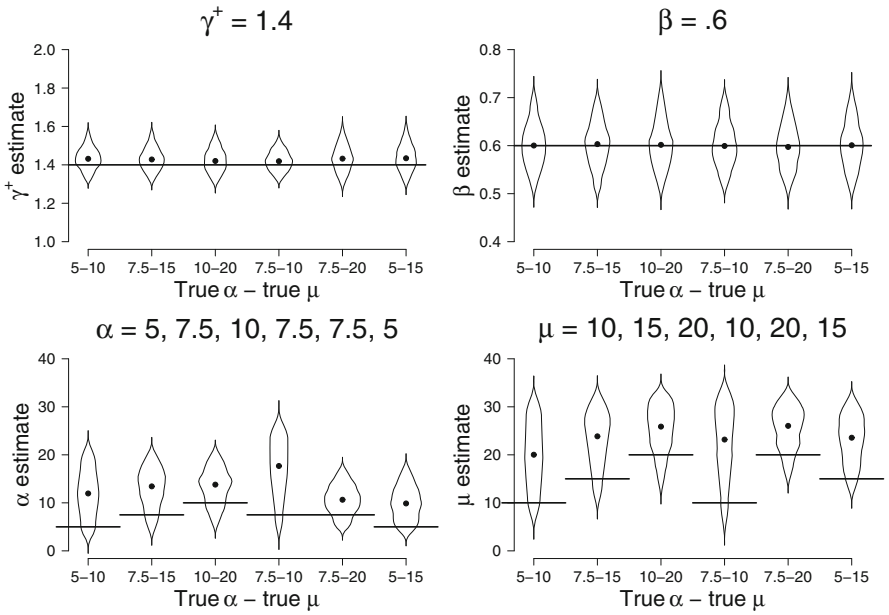

Fig. 2.1 The 4-parameter BART model recovers parameters γ^+ and β, but fails to recover parameters α and μ (results based on a 300-trial BART). The *dots* represent the median of 1000 point estimates from 1000 different BARTs performed by a single synthetic agent. The *violin shapes* around the *dots* are density estimates for the entire distribution of point estimates, with the extreme 5 % truncated [16]. The *horizontal lines* represent the true parameter values

for this model [15].[1] The results of one of those recovery studies are presented in Fig. 2.1. This figure shows the results of 1000 simulations of a single synthetic participant completing 300 BART trials[2], for each of six sets of data-generating parameter values. For each of the 1000 simulations, van Ravenzwaaij et al. obtained a point estimate for each parameter. In Fig. 2.1, the dots represent the median of the 1000 point estimates, and the "violins" that surround the dots represent density estimates that represent the entire distribution of point estimates, with the extreme 5 % truncated. The horizontal lines show the true parameter values that were used to generate the synthetic data (also indicated on top of each panel).

Figure 2.1 shows good parameter recovery for γ^+ and β, with only a slight overestimation of γ^+. The α and μ parameters are systematically overestimated. The overestimation of α increases when the true value of μ becomes smaller (in the bottom left panel, compare the fourth, second, and fifth violin from the left or compare the leftmost and rightmost violins). The overestimation of μ increases when the true value of α becomes larger (in the bottom right panel, compare the first and

[1] Extensive details are reported here: http://www.donvanravenzwaaij.com/Papers_files/BART_Appendix.pdf.

[2] With only 90 trials—the standard number—parameter recovery was very poor.

the fourth violin from the left). Both phenomena suggest that parameter recovery suffers when the true value of α is close to the true value of μ. For the six sets of data-generating parameter values shown on the x-axis from Fig. 2.1, the correlations between the point estimates of α and μ were all high: 0.97, 0.95, 0.93, 0.99, 0.83, 0.89, respectively.

The important lesson here is that, even though a model may have parameters that are conceptually distinct, the way in which they interact given the mathematical form of a model may mean that they are not distinct in practice. In such circumstances it is best to study the nature of the interaction and either modify the model or develop new paradigms that produce data capable of discriminating these parameters. The complete set of model recovery studies led van Ravenzwaaij and colleagues to propose a two-parameter BART model ([15]; but see [17]).

2.3 Carry Out Tests of Selective Influence

Cognitive models can be useful tools for understanding and predicting behavior, and for reasoning about psychological processes, but—as with all theories—utility hinges on validity. Establishing the validity of a model is a difficult problem. One method is to demonstrate that the model predicts data that are both previously unobserved, and ecologically valid. For example, a model of decision making, developed for laboratory tasks, might be validated by comparison against the decisions of consumers in real shopping situations. External data of this sort are not always available; even when they are, their ecological validity is not always clear. For example, it is increasingly common to collect neural data such as electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) measurements simultaneously with behavioral data. Although it is easy to agree that the neural data should have some relationship to the cognitive model, it is not often clear what that relationship should be—which aspects of the neural data should be compared with which elements of the cognitive model.

An alternative way to establish model validity is via tests of selective influence. Rather than using external data as the benchmark of validity, this method uses experimental manipulations. Selective influence testing is based on the idea that a valid model can titrate complex effects in raw data into separate and simpler accounts in terms of latent variables. From this perspective, a model is valid to the extent that it make sense of otherwise confusing data. For example, signal detection models can explain simultaneous changes in false alarms and hit rates—and maybe confidence too—as simpler effects on underlying parameters (i.e., sensitivity and bias). Similarly, models of speeded decision-making can convert complex changes in the mean, variance, and accuracy of response time data into a single effect of just one latent variable.

Testing for selective influence begins with *a priori* hypotheses about experimental manipulations that ought to influence particular latent variables. For instance, from the structure of signal detection theory, one expects payoff manipulations to influence

bias, but not sensitivity. Empirically testing this prediction of selective influence becomes a test of the model structure itself.

2.3.1 Examples of Selective Influence Tests

Signal detection theory has a long history of checking selective influence. Nearly half a century ago, Parks [18] demonstrated that participants tended to match the probability of their responses to the relative frequency of the different stimulus classes. This behavior is called probability matching, and it is statistically optimal in some situations. Probability matching requires decision makers to adjust their decision threshold (in SDT terms: bias) in response to changes in relative stimulus frequencies. Parks—and many since—have demonstrated that decision-makers, from people to pigeons and rats, do indeed change their bias parameters appropriately (for a review, see [19]). This demonstrates selective influence, because the predicted manipulation influences the predicted model parameter, and only that parameter. Similar demonstrations have been made for changes in signal detection bias due to other manipulations (e.g., the strength of memories: [20]).

Models of simple perceptual decision making, particularly Ratcliff's diffusion model ([5, 21, 10]), have around six basic parameters. Their apparent complexity can be justified, however, through tests of selective influence. In seminal work, Ratcliff and Rouder orthogonally manipulated the difficulty of decisions and instructions about cautious vs. speedy decision-making, and demonstrated that manipulations of difficulty selectively influenced a stimulus-related model parameter (drift rate) while changes to instructions influenced a caution-related model parameter (decision boundaries). Voss, Rothermund and Voss [22] took this approach further and separately tested selective influences on the diffusion model's most fundamental parameters. For example, one experiment manipulated relative payoffs for different kinds of responses, and found selective influence on the model parameter representing bias (the "start point" parameter). These kinds of tests can alleviate concerns about model complexity by supporting the idea that particular model parameters are necessary, and by establishing direct relationships between the parameters and particular objective changes or manipulations.

Deciding whether one parameter is or is not influenced by some experimental manipulation is an exercise in model selection (i.e., selection between models that do and do not impose the selective influence assumption). Both Voss et al. and Ratcliff and Rouder approached this problem by estimating parameters freely and examining changes in the estimates between conditions; a significant effect on one parameter and non-significant effects on other parameters was taken as evidence of selective influence. Ho, Brown and Serences [23] used model selection based on BIC [24] and confirmed that changes in the response production procedure—from eye movements to button presses—influenced only a "non-decision time" parameter which captures the response-execution process. However, a number of recent studies have rejected the selective influence of cautious vs. speedy decision-making on decision boundaries [25, 26, 27]. In a later section we show how model-selection was used in this context.

2.4 Quantify Uncertainty in Parameter Estimates

In many modeling approaches, the focus is on model prediction and model fit for a single "best" set of parameter estimates. For example, suppose we wish to estimate the probability θ that Don correctly discriminates regular beer from alcohol-free beer. Don is repeatedly presented with two cups (one with regular beer, the other with non-alcoholic beer) and has to indicate which cup holds the regular beer. Now assume that Don answers correctly in 3 out of 10 cases. The maximum likelihood estimate $\hat{\theta}$ equals $3/10 = 0.3$, but it is evident that this estimate is not very precise. Focusing on only a single point estimate brings with it the danger of overconfidence: predictions will be less variable than they should be.

In general, when we wish to use a model to learn about the cognitive processes that drive task performance, it is appropriate to present the precision with which these processes have been estimated. The precision of the estimates can be obtained in several ways. Classical or frequentist modelers can use the bootstrap [28], a convenient procedure that samples with replacement from the original data and then estimates parameters based on the newly acquired bootstrap data set; the distribution of point estimates across the bootstrap data sets provides a close approximation to the classical measures of uncertainty such as the standard error and the confidence interval. Bayesian modelers can represent uncertainty in the parameter estimates by plotting the posterior distribution or a summary measure such as a credible interval.

2.4.1 Example of Quantifying Uncertainty in Parameter Estimates

In an elegant experiment, Wagenaar and Boer assessed the impact of misleading information on earlier memories [29]. They showed 562 participants a sequence of events in the form of a pictorial story involving a pedestrian-car collision at an intersection with a traffic light. In some conditions of the experiment, participants were later asked whether they remembered a pedestrian crossing the road when the car approached the "stop sign". This question is misleading (the intersection featured a traffic light, not a stop sign), and the key question centers on the impact that the misleading information about the stop sign has on the earlier memory for the traffic light.[3]

Wagenaar and Boer constructed several models to formalize their predictions. One of these models is the "destructive updating model", and its critical parameter d indicates the probability that the misleading information about the stop sign (when properly encoded) destroys the earlier memory about the traffic light. When $d = 0$, the misleading information does not affect the earlier memory and the destructive updating model reduces to the "no-conflict model". Wagenaar and Boer fit the destructive updating model to the data and found that the single best parameter estimate was $\hat{d} = 0$.

[3] The memory for the traffic light was later assessed by reminding participants that there was a traffic light at the intersection, and asking them to indicate its color.

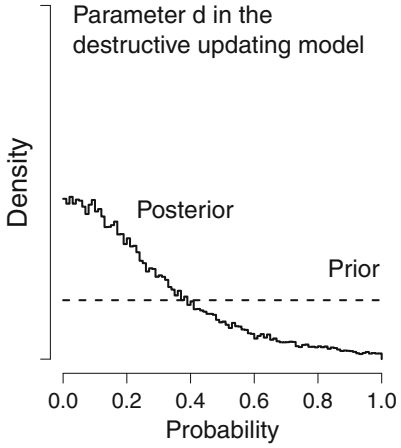

Fig. 2.2 Prior and posterior distributions for the d parameter in the destructive updating model from Wagenaar and Boer (1987), based on data from 562 participants. When $d = 0$, the destructive updating model reduces to the no-conflict model in which earlier memory is unaffected by misleading information presented at a later stage. The posterior distribution was approximated using 60,000 Markov chain Monte Carlo samples. (Figure downloaded from Flickr, courtesy of Eric-Jan Wagenmakers)

Superficial consideration may suggest that the result of Wagenaar and Boer refutes the destructive updating model, or at least makes this model highly implausible. However, a more balanced perspective arises once the uncertainty in the estimate of $\hat{d}$ is considered. Figure 2.2 shows the prior and posterior distributions for the d parameter (for details see [30]). The prior distribution is uninformative, reflecting the belief that all values of d are equally likely before seeing the data. The observed data then update this prior distribution to a posterior distribution; this posterior distribution quantifies our knowledge about d [31]. It is clear from Fig. 2.2 that the most plausible posterior value is $d = 0$, in line with the point estimate from Wagenaar and Boer, but it is also clear that this point estimate is a poor summary of the posterior distribution. The posterior distribution is quite wide and has changed relatively little compared to the prior, despite the fact that 562 people participated in the experiment. Values of $d < 0.4$ are more likely under the posterior than under the prior, but not by much; in addition, the posterior ordinate at $d = 0$ is only 2.8 times higher than the prior ordinate at value $d = 0$. This constitutes evidence against the destructive updating model that is is merely anecdotal or "not worth more than a bare mention" [32].[4]

In sum, a proper assessment of parameter uncertainty avoids conclusions that are overconfident. In the example of Wagenaar and Boer, even 562 participants were not sufficient to yield strong support for or against the models under consideration.

[4] Wagenaar and Boer put forward a similar conclusion, albeit not formalized within a Bayesian framework.

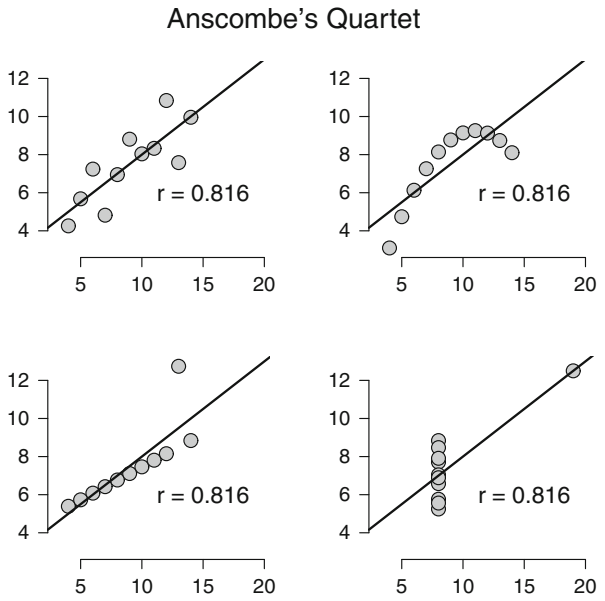

Fig. 2.3 Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, $r = 0.816$. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. Despite the equivalence of the four data patterns in terms of popular summary measures, the graphical displays reveal that the patterns are very different from one another, and that the Pearson correlation (a linear measure of association) is only valid for the data set from the top left panel. (Figure downloaded from Flickr, courtesy of Eric-Jan Wagenmakers)

2.5 Show Model Fit

When a model is unable to provide an adequate account of the observed data, conclusions based on the model's parameters are questionable. It is, therefore, important to always show the fit of the model to the data. A compelling demonstration of this general recommendation is known as Anscombe's quartet [33] replotted here as Fig. 2.3. The figure shows four data sets that have been equated on a number of measures: the Pearson correlation between the x and y values, the mean of the x and y values, and the variance of the x and y values. From the graphical display of the data, however, it is immediately obvious that the data sets are very different in terms of the relation between the x values and the y values. Only for the data set shown in the top left panel does it make sense to report the Pearson correlation (a linear measure of association). In general, we do not recommend relying on a test of whether a single global measure of model misfit is "significant". The latter practice is not even suitable for linear models [34], let alone non-linear cognitive process models, and is subject to the problem that with sufficient power rejection is guaranteed, and therefore meaningless [35]. Rather we recommend that a variety of

http://www.springer.com/978-1-4939-2235-2

An Introduction to Model-Based Cognitive Neuroscience

Forstmann, B.U.; Wagenmakers, E.-J. (Eds.)
2015, XI, 354 p. 81 illus., 55 illus. in color., Hardcover
ISBN: 978-1-4939-2235-2