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.

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

Andrew Heathcote
University of Newcastle, School of Psychology, University Avenue, Callaghan, 2308, NSW, Australia e-mail: Andrew.Heathcote@newcastle.edu.au

Scott D. Brown
University of Newcastle, School of Psychology, University Avenue, Callaghan, 2308, NSW, Australia e-mail: Scott.Brown@newcastle.edu.au

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

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 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 Chapter 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 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.

2.1 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 for this model [15].[1] The results of one of those recovery studies are presented in Figure 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 Figure 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).

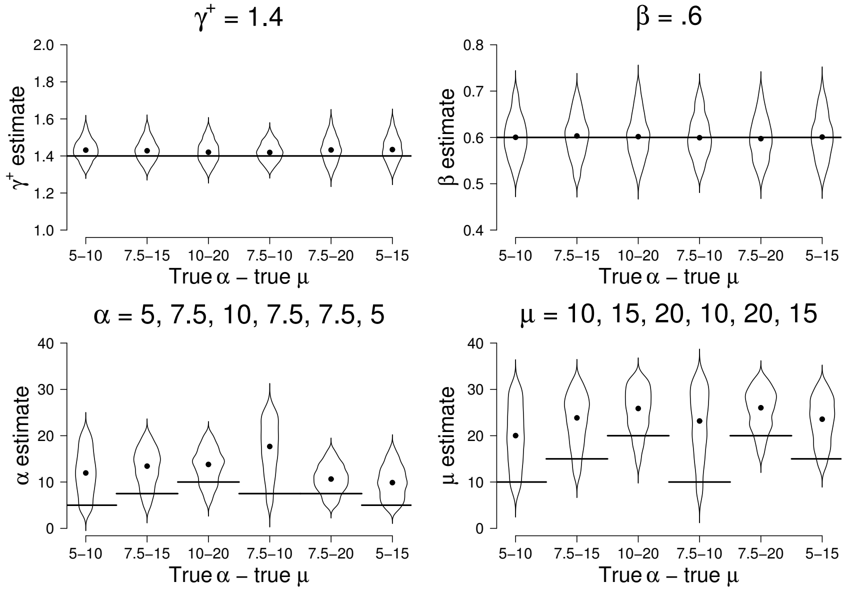

Fig. 1 The 4-parameter BART model recovers parameters γ^+ and β, but fails to recover parameters α and μ (results based on a 300-trial BART). The dots represent the median of 1000 point estimates from 1000 different BARTs performed by a single synthetic agent. The violin shapes around the dots are density estimates for the entire distribution of point estimates, with the extreme 5% truncated [16]. The horizontal lines represent the true parameter values.

Figure 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

[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.

value of α becomes larger (in the bottom right panel, compare the first and 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 Figure 1, the correlations between the point estimates of α and μ were all high: $.97, .95, .93, .99, .83, .89$, 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]).

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.

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.

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 $3/10 = .3$, 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.

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

[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.

(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 $\hat{d} = 0$.

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 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 Figure 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.

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 Figure 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,

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

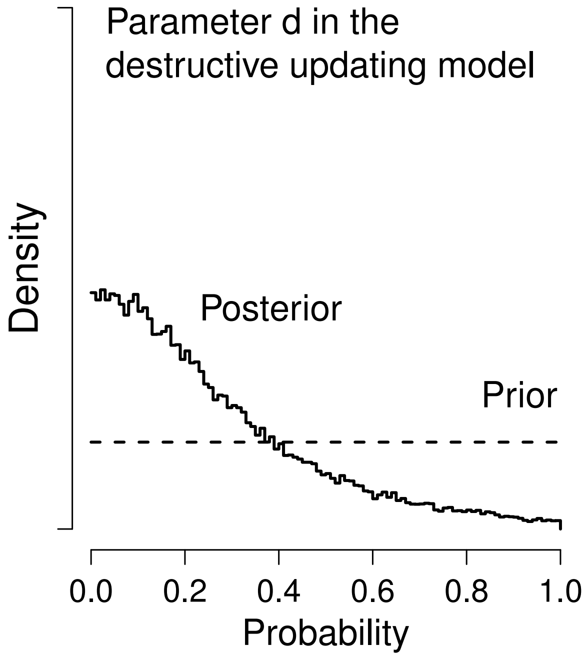

Fig. 2 Prior and posterior distributions for the *d* parameter in the destructive updating model from Wagenaar and Boer (1987), based on data from 562 participants. When $d = 0$, the destructive updating model reduces to the no-conflict model in which earlier memory is unaffected by misleading information 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.

and is subject to the problem that with sufficient power rejection is guaranteed, and therefore meaningless [35]. Rather we recommend that a variety of graphical checks be made and a graphical summary of the relevant aspects of model fit be reported.

Displaying and checking model fit can be difficult when data come from many participants in a complicated multiple-factor design. When the model is nonlinear, as is almost always the case with cognitive process models, fitting to data averaged over participants should be avoided, as even a mild nonlinearity can introduce systematic distortions (e.g., forgetting and practice curves [36, 37, 38]). For the purpose of displaying overall model fit it is fine to overlay a plot of the average data with the average of each participant's model fit, as both averages are subject to the same distortions. However, analogous plots should also be checked for each indi-

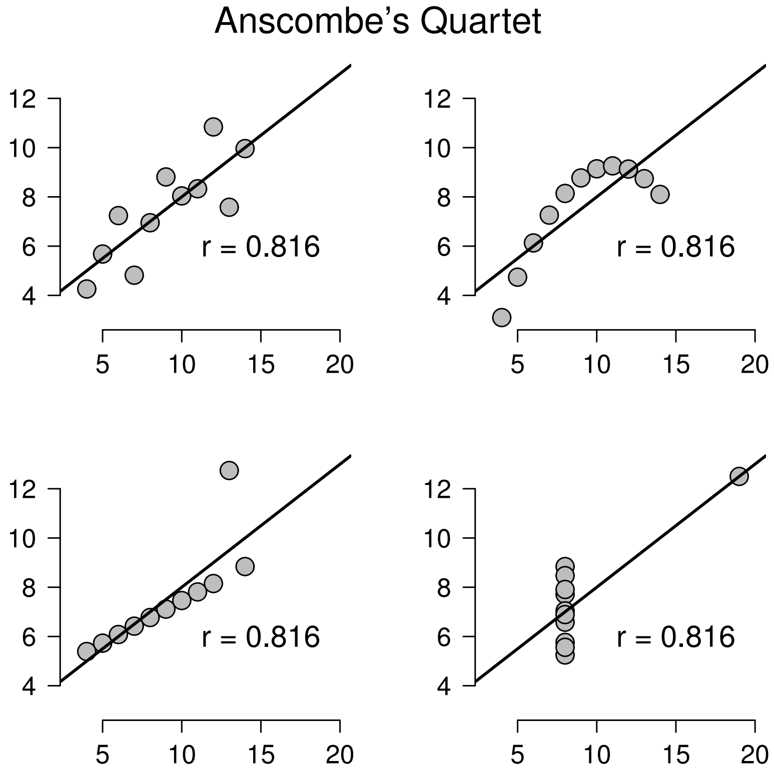

Fig. 3 Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the *x* and *y* values is the same, $r = .816$. In fact, the four different data sets are also equal in terms of the mean and variance of the *x* 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.

vidual, both to detect atypical participants, and because it is common for initial fit attempts to fail with some participants. In some cases individual plots can reveal that an apparently good fit in an average plot is due to “cancelling out” of under- and over-estimation for different groups of participants. Similarly, it is important to check plots of fit broken down by all of the influential factors in the experimental design. Even when interest focuses on the effects of a subset of factors, and so it is appropriate to average over other (effectively “nuisance”) factors when reporting results, such averages can hide tradeoffs that mask systematic misfit. Hence, in the first instance it is important to carry out a thorough check graphical check of fit broken down by all factors that produce non-negligible effects on data.

In the case of continuous data it is practically difficult to display large numbers of data points from many participants in complex designs. An approach often used with evidence accumulation model fit to continuous response time (RT) data is to summarize the distribution of RT using quantiles (e.g., the median and other percentiles). A common choice is the 10^{th}, 30^{th}, 50^{th}, 70^{th} and 90^{th} percentiles (also called the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles). This five-quantile summary may omit some information, but it can compactly capture that are usually considered key features of the data. Of particular importance are the 10^{th} percentile, which summarises the fastest RTs, the 50^{th} percentile or median, which summarises the central tendency, and the 90th percentile, which summaries the slowest RTs. The spread between the 90^{th} and 10^{th} percentiles summarises variability in RT and a larger difference between the 90^{th} and 50^{th} percentiles compared to the 50^{th} to 10^{th} percentile summarises the typically positive skew in RT distribution.

Further complication arises when data are multivariate. For example, cognitive process models are usually fit to data from choice tasks. Where one of two choices is classified as correct, the rate of accurate responding provides a sufficient summary. However, participants can trade accuracy for speed [39], so in many cases it is important to also take RT into account. That is, the data are bivariate, consisting of an RT distribution for correct responses, an RT distribution for error responses, and an accuracy value specifying the rate at which correct and error responses occur. Latency-probability (LP) plots [40, 41] deal with the bivariate nature of choice data by plotting mean RT on the y-axis against response probability on the x-axis. As error responses commonly occur with low probability, error data appear on the left of the plot and correct response data on the right of the plot. In the two-choice case the x-values occur in pairs. For example, if the error rate is 0.1 then the corresponding correct-response data must be located at 0.9. Quantile-probability (QP) plots [42] generalize this idea to also display a summary of RT distribution by plotting quantiles on the y-axis (usually the five-quantile summary) instead of the mean. Although the QP plot provides a very compact representation of choice RT data that can be appropriate in some circumstances, we do not recommend it as a general method of investigating model fit for reasons we illustrate in the following example. Rather, we recommend looking at separate plots of accuracy and correct and error RT distributions (or in the $n>2$ alternative case, RT distributions for each type of choice).

5.1 Examples of Showing Model Fit

Wagenmakers and colleagues [43] had participants perform a lexical decision task – deciding if a letter string constituted a word or nonword, using high, low and very-low frequency word stimuli and nonword stimuli. In their first experiment participants were given instructions that emphasised either the accuracy or speed of responding. They fit a relatively simple 12-parameter diffusion model to these data, assuming that instructions selectively influenced response caution and bias, whereas stimulus type selectively influenced the mean drift rate. Rae and colleagues [44] refit these data, including two extra participants not included in the originally reported data set (17 in total), in order to investigate the selective influence assumption about emphasis. They fit a more flexible 19-parameter model allowing (1) emphasis to affect the trial-to-trial standard deviation of bias as well as the mean and standard deviation of non-decision time; (2) stimulus type to affect the trial-to-trial standard deviation of the drift rate; and (3) allowing for response contamination. Their interest was in whether instruction emphasis could affect drift rate parameters, so they contrasted this 19 parameter "selective influence" model with a 27-parameter ("least constrained") model allowing speed emphasis to affect the mean and standard deviation of drift rates. We discuss this contrast in a following section but for now we focus on the fit of the selective-influence model.

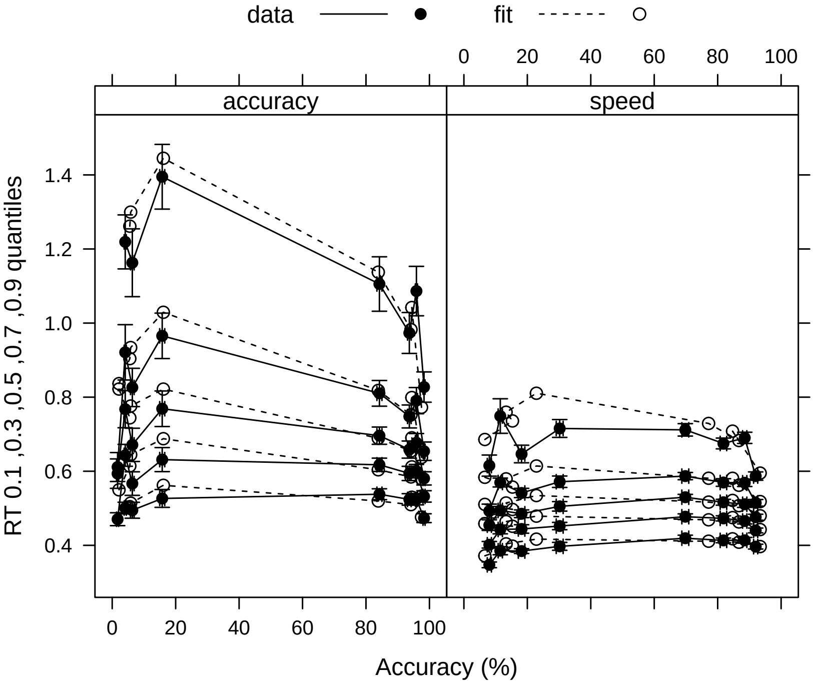

Fig. 4 Quantile-probability plot of the average over 17 participants from Wagenmakers and colleagues' [43] Experiment 1 and fits of the "selective influence" model. In both emphasis conditions accuracy was ordered from greatest to least: high-frequency (hf) words, nonwords (nw), low-frequency words (lf) and very-low-frequency (vlf) words. Each data point is accompanied by a 95% confidence interval assuming a Student t distribution and based between-subject standard errors calculated as $SD(x)/\sqrt{n}$, where $SD(x)$ is the standard deviation over participants and n is the number of participants.

Figure 4 is a quantile-probability plot of the selective-influence model. Data points are plotted with 95% confidence intervals based on conventional standard errors assuming a normal distribution in order to convey an idea of the likely measurement error, and model fit is indicated by points joined by lines. Uncertainty can also be conveyed by other means, such as bootstrap methods [28] applied to the data [43] or model fits [45]. In any case, it is important to plot points for the model as well as the data; plotting only lines for either can hide mis-fit because the eye can be fooled by intersections that do not reflect an accurate fit. The figure demonstrates the utility of QP plots in illustrating an important regularity in choice RT data [46]; the overall decreasing lines from left to right in the accuracy condition show that errors are slower than corresponding correct responses, whereas the symmetric lines around $accuracy = 50\%$ in the speed condition indicate approximately equal correct and error speed.

Overall, Figure 4 demonstrates that the model captures the majority of trends in the data, with many of the fitted points falling within 95% data confidence intervals. However there is also some evidence of misfit, especially in regard to accuracy in the speed condition. Rae and colleagues [44] focused on this failure of the selective-influence model to account for the effect of emphasis instructions, motivated by similar findings for the LBA model [48], and the same pattern of under-estimation in experiments they reported using perceptual stimuli and in recognition memory (see also [49]). Figure 5 more clearly illustrates the speed-accuracy tradeoff induced by emphasis instructions –with accuracy displayed in the upper panels and speed of correct responses in the lower panels– and the under-estimation of the accuracy effect in the lexical decision data. For clarity, data from each stimulus type is plotted in a separate panel with emphasis condition plotted on the x-axis and joined by lines in order to emphasise the difference of interest. Each row of panels is tailored to examine a different aspect of the data. The upper panels show accuracy and the middle panels the distribution of correct RT. The lower panels plot both the central tendency (median) of correct RT (circle symbols) and error RT (triangle symbols) in order to highlight the relative speed of incorrect responses.

Figure 5 also uses within-subject standard errors appropriate to the focus on the (within-subject) difference between speed and accuracy emphasis. These standard errors reflect the reliability of the difference between speed and accuracy conditions, making it clear that the selective-influence model is unable to account for the effect of emphasis, particularly for very-low frequency and low-frequency words. The middle panels make it clear that, in contrast, this model provides a highly accurate account of its effect on RT distributions for correct responses, even for the 90^{th} percentile, which has much greater measurement error than other quantiles. Finally, the lower panel clearly shows that the model predicts slow errors in the accuracy condition for all stimuli, whereas slow errors occur in the data only for the least word-like (very-low frequency and nonword) stimuli. In general, we recommend that a variety of multi-panel plots such as those in Figure 5 be examined, each tailored to particular aspects of the data, with standard errors appropriate to the questions of interest. We also highly recommend plotting versions of these plots for individual participants, which is made easy using trellis graphics [50].

6 Engage in Model Selection

Cognitive models typically have several parameters that can sometimes interact in complex ways. How does a modeler decide which experimental design factors affect each parameter? Initial guidance is provided by conventions based on *a priori* assumptions and past research. In the realm of evidence accumulation models, for example, it has been widely assumed that stimulus-related factors selectively affect parameters related to the evidence flowing into accumulators (e.g., evidence accumulation rates) whereas instruction-related factors (e.g., an emphasis on speed vs. accuracy) affect accumulator-related parameters (e.g., bias and the amount of evidence required to make a decision) [46]. However, such conventional settings may not always hold, and in new paradigms they may not be available. Hence, it is prudent, and sometimes necessary, to engage in model selection: comparing a variety of different model parameterisations (variants) so that one or more can be selected and differences in parameter estimates among experimental conditions interpreted.

Even in relatively simple designs the number of model variants can rapidly become unmanageable. Suppose there are two experimental design factors, A and B. Each parameter might have a single estimated value (i.e., an intercept only model, often denoted ˜1), a main effect of or A or B (˜A or ˜B), or both main effects and their interaction (denoted ˜A*B = A + B + A:B, where "+" indicates an additive effect and ":" an interaction effect). If only parameterisations of this sort are considered there are 2^f models to select amongst, where f is the number of factors. If each of m types of model parameter are allowed this freedom then the total number of variants is $2^{f \times m}$. One might also consider additive models (e.g., ˜A + B), in which case it is important to note that fit of an additive model depends on parameter scale (e.g., the model˜A + B can fit differently for a linear vs. log-scaled parameter, whereas A*B will fit identically). Further, so far we have only considered hierarchical models, where the higher terms can only occur accompanied by their constituents. If all possible non-hierarchical models are allowed (e.g., ˜A + A:B and ˜B + A:B) the increase in the number of variants with f is much faster.

Once an appropriate set of model variants is selected its members must be compared in some way. Ideally misfit is quantified by a function of the likelihood of the data under each model variant, such as in Bayesian or maximum likelihood estimation methods, or some approximation, such as in quantile-based methods like maximum probability [51, 52, 53] or minimum likelihood-ratio χ^2 (i.e., G^2) estimation [49]. As the number of estimated parameters increases a model becomes more flexible and so is able to better fit data. In particular, a nested model –a model variant that is a special case obtained by fixing one or more parameters of a more complex model– necessarily has greater misfit than models that nest it. The best model variant cannot be selected based on goodness-of-fit alone, as the least-constrained model would always be selected, and over-fitting (i.e., capturing unsystematic variation specific to a particular data set) would be rife.

One approach –similar to sequential regression methods such as stepwise selection– is to choose a model based on the significance of changes in fit as parameters are added or deleted. The maximised log-likelihood (L) is convenient for this purpose as the deviance misfit measure (D = -2×L) has a χ^2 distribution.[5] However, this approach is limited to selection amongst nested models. Preferable approaches use a model selection criterion that includes a penalty for model complexity; the model with the lowest criterion (i.e., least penalised misfit) is selected. Such criteria can be used not only to compare non-nested variants of the same model but also to compare different cognitive process models. In a Bayesian framework the Bayes factor is the ideal criterion [54], but this is rarely easy to directly compute for cognitive process models (although see [55, 56]). When estimation is achieved by posterior sampling, DIC [57], or its bias-corrected variant BPIC [58], are easy-to-compute alternatives. With maximum likelihood estimation it is convenient to use BIC = D + k$\times \log(n)$, a Bayes factor approximation, or AIC = D + 2×k (*n* is the number of data points and *k* is the number of parameters) [59]. As is evident from these formulae, BIC applies a harsher penalty for complexity for typical sample sizes ($n \geq 8$).

Although we recommend using penalised-misfit criteria to guide model selection we do not recommend rigid adherence. Different criteria are optimal under different assumptions and they are often based on approximations that can be sensitive to the size of effects and samples (see, for example, the comparison of BIC and AIC in [60]). Further, it is seldom possible to check all possible models, and when the data-generating model is omitted model selection may err. Suppose, for example, the data were generated by an additive model, A + B but only the A, B, and A*B models are fit. Depending on the size of the A, B, and A:B effects relative to the size of the complexity penalty any of these three models may be selected. Even if the A + B model is fit, things can still go wrong if the data-generating model is additive on, for example, a logarithmic scale. Given the appropriate scale is usually not known, it is apparent that it is difficult to be absolutely sure that data-generating model is included even in quite exhaustive sets of variants (although clearly the chance of problems reduces when selection is among a large set of variants!).

In short, some judgement –albeit aided by a number of sources of evidence– must be used in model selection. For example, in smaller samples BIC can often select a model that is so simple that plots of fit reveal that the model is unable to account for practically or theoretically important trends in the data. On the other end of the spectrum over-fitting is indicated when model parameters are unstable (i.e., take on implausibly large or small values and/or values that can vary widely with little effect on fit) or take on patterns that appear nonsensical in relation to the way the parameter is interpreted psychologically. In both cases it is prudent to consider alternative model selection methods or possibly to seek further evidence. It is also worth reconsidering whether selection of a single model is required for the purpose at hand. For example, predictions averaged over models weighted by the evidence for each model are often better than predictions made by a single model [35]. Similarly, different criteria may select models that differ in some ways but are consistent with respect to the theoretical issue under investigation. We illustrate the process in the next section.

[5] The absolute value of the deviance depends on the measurement units for time and so only relative values of deviance are meaningful. Deviance is on an exponential scale, and as a rule of thumb a difference less than 3 is negligible and and difference greater than 10 indicates a strong difference.

6.1 Examples of Model Selection

Our examples again use the lexical decision data from Wagenmakers and colleagues [43], focusing on variant selection assuming a diffusion model [44]. The example is based on maximum-likelihood fits to individual participant data, with 2^9=512 diffusion variants fit with the methods described in [61]. Diffusion fits were based on quantile data, so the likelihood is only approximate [52, 53]. The least-constrained variant allowed rate parameters to vary with emphasis, that is, it did not make the conventional assumption that accumulation rate cannot be affected by instructions.

Before examining the model selection process in these examples it is important to address issues that can arise due to individual differences. When each participant's data are fit separately, different models are often selected for each participant. Considering participants as random effects provides a useful perspective on this issue. Even if there is no effect of a factor on the population mean of a parameter, when the population standard deviation is sufficiently large individual participants will display reliable effects of the factor. That is, selecting models that include the effect of a factor for some individuals does not imply that factor affects the corresponding population mean. Hierarchical models –which make assumptions about the form of population distributions– enable simultaneous fitting of all participants and direct estimation of population means. Even in this approach, however, individual participant estimates must be examined to check assumptions made by the hierarchical model about the form of the population distribution. For example, it is possible that some individual variation results from participants being drawn from different populations (e.g., a mixture model where in one population a factor has an effect and in another it does not), in which case assuming a single unimodal population distribution is problematic. Caution must also be exercised in case the random effects model is incorrectly specified and the shrinkage (i.e., the averaging effect exerted by the assumed population distribution) masks or distorts genuine individual differences. Hierarchical modeling is best applied to relatively large samples of participants and usually requires Bayesian methods. These methods can sometimes be difficult in practice with cognitive process models where strong interactions among parameter make posterior sampling very inefficient, as was the case for the LBA model until recent advances in Markov chain Monte Carlo methods [62].

With maximum-likelihood fits to individuals it is possible to select an overall model based on and aggregate BIC or AIC.[6] The selected model, which can be thought of as treating participants as fixed effects, is usually sufficiently complex to accommodate every individual. However, it is important to be clear that selecting a model that contains a particular factor does not necessarily imply an effect of that factor on the random effect population mean. In the individual-fitting approach such questions can be addressed by testing for differences over a factor in the means of individual participant parameter estimates. These approaches to individual-participant

[6] Note that deviance can be summed over participants, as can AIC, but BIC cannot, due to the nonlinear $\log(n)$ term in its complexity penalty. Instead the aggregate BIC is calculated from the deviance, number of parameters and sample size summed over participants

fitting were taken by Rae and colleagues [44], with results summarised in Table 1. The table reports aggregate misfit measures minus the minimum deviance. Hence the best-fitting model (necessarily the least-constrained model with the largest number of parameters, k) has a zero entry in the deviance column.

Table 1 Diffusion model variants specified by design factors that effect each parameter, number of parameters per participant for each variant (k), and misfit measures (D = deviance, AIC = Akaike Information Criterion and BIC = Bayesian Information Criterion) minus the minimum value in each column. Factors are emphasis (E: speed vs. accuracy) and stimulus (S: high/low/very-low frequency words and non words). Diffusion parameters: a = response caution parameter, distance between response boundaries; accumulation rate parameters, v = mean, sv = standard deviation; start-point (bias) parameters, z = mean relative to lower boundary, sz = uniform distribution width; non-decision time parameters: t_0 = minimum time for stimulus encoding and response production, st = width of uniform distribution of non-decision time. Note that, for example, the notation v˜E*S means that the v parameter can be affected by the main effects of the E and S factors as well as their interaction.

Model Type	Model Definition	k	D	AIC	BIC
Least Constrained	a˜E, v˜E*S, sv˜E*S, z˜E, sz˜E, t_0˜E, st˜E	27	0	70	984
AIC Selected	a˜E, v˜E*S, sv˜S, z˜E, sz˜E, t_0˜E, st˜E	23	66	0	552
BIC Selected	a˜E, v˜S, sv˜1, z˜1, sz˜E, t_0˜E, st˜1	14	635	263	0
Selective Influence	a˜E, v˜S, sv˜S, z˜E, sz˜E, t_0˜E, st˜E	19	237	35	225
AIC Selected	a˜E, v˜S, sv˜S, z˜E, sz˜1, t_0˜E, st˜E	19	237	35	225
BIC Selected	a˜E, v˜S, sv˜1, z˜E, sz˜1, t_0˜E, st˜1	14	635	263	0

The top three rows in Table 1 report results for the least constrained diffusion model and the models selected from the full set of 512 by aggregate AIC and BIC. The bottom three rows report results for the variant within the full set that imposes a minimal selective effect assumption –that the emphasis manipulation cannot affect rate parameters (v and sv)– and the AIC and BIC selected models among the subset of 2^7 = 128 variants nested by the selective influence variant. Among the full set of 512 variants, AIC selected a variant where emphasis did affect the mean rate (i.e., violating selective influence), whereas BIC selected a variant that had no influence of emphasis on rate parameters. This pattern of results nicely exemplifies that fact that selection criteria can lead to theoretically important differences in conclusions, requiring researchers to seek other sources of evidence.

Rae and colleagues pointed out that the penalty for complexity imposed by BIC was likely too harsh. As shown in the upper panels of the Figure 5 even the full 25-parameter selective influence LBA variant (which necessarily fits better than the 17 parameter variant selected from the overall set by BIC) fails to accommodate the effect of emphasis on accuracy. In agreement, there is a highly significant decrease in fit from the least-constrained to the BIC model as illustrated by the difference of deviances in Table 1 (i.e., df = 17×(27-19) = 136, $\chi^2(136) = 237$, $p < .001$). In contrast, the 19-parameter variant selected from the overall set by AIC that allows emphasis to affect mean rate does much better in accommodating the effect of emphasis on accuracy. Further, the reduction in fit relative to the least constrained model does not approach significance (i.e., df = 17×(27-23) = 68, $\chi^2(68) = 66$, p=.55). Finally, parameter estimates for the least constrained model were stable and there was a significant effect of emphasis on mean rate. This overall pattern of results confirmed a failure of traditional selective influence assumption and was consistent with findings from perceptual an mnemonic paradigms reported by Rae and colleagues and others [26].

In closing we return to what is perhaps one of the most difficult questions in cognitive process modeling, absolute fit, which might be couched as a model-selection question: “when should all model variants be rejected”? A model may provide a very accurate account of some conditions or some aspects of data but systematically misfit in other cases. For example, Rae and colleagues found even the least-constrained diffusion model misfit error RT distribution in some conditions. Heathcote and Love [48] showed the same was true in those data, although to a lesser degree, for the LBA model. Should both the diffusion and LBA models be rejected? Clearly some judgement is required to decide, since some misfit can be forgiven. A case in point for evidence-accumulation models is sequential effects. Sequential effects can be quite strong in the sorts of paradigms to which such models are fit [63, 64], but that fitting almost invariably assumes data are independently distributed over trials.

On this issue we think Box’s [65] famous dictum that “all models are false but some are useful” is salutary. That is, some misfit can be tolerated, especially when no better alternatives are available, as long as the model captures theoretically important features of a data set. To the degree this happens, parameter estimates are likely to provide an accurate distillation of the data and a more meaningful characterisation than simple summary statistics (e.g., mean RT or accuracy alone can be confided by speed-accuracy tradeoff). Further, if that distillation is sensible in terms of the underlying rationale of the model, and consistent with conclusions based on alternative analyses that do not reply on the model, then it is likely that the cognitive process model has served a useful role.

7 Concluding Remarks

Good standards are important in all areas of science. In cognitive modeling, good standards include not only careful model development and checking but also transparency of method and, ideally, sharing of model code and data. Transparency is obviously of key importance, as it allows interested colleagues to implement the model at hand, whether for teaching, for testing, or for extending the approach to other contexts. Even relatively simple models can sometimes be surprisingly difficult to replicate because crucial information is missing. It is therefore common practice to make available the model code, either on a personal website, archived together with the journal publication as supplementary material, or in a public repository (e.g., the OSU Cognitive modeling Repository, `http://cmr.osu.edu/`).

It is useful to make the model code available with several example data sets so that the new user can confirm that the code works as it should. Ideally, the entire data set that is modeled is made freely available online as well. This benefits the new user (who may be able to use the data set for a different purpose, or for a test of alternative explanations), but it also benefits the author, as the online data set may easily become a modeling benchmark.

In this introductory chapter we have discussed a suite of standard sanity checks that every modeler should turn to on a regular basis. This holds especially true for cognitive process models that are relatively new and untested. By applying the suite of sanity checks the modeler can gain confidence in the validity of a model, and consequently make a more compelling case that the model yields a reliable connection from observed behavior to unobserved psychological process.

8 Exercises

1. You fit a model of task switching to some data. The model includes a parameter which reflects how often people actively prepare for the upcoming task, and you find that the best estimate of this parameter is 50%. What should you also consider, before concluding that task preparation occurs half of the time?
2. If you fit a complex model and a simpler model to some data, and found that the simple model had best BIC but the complex model had the best AIC, which would you expect to give the closest fit to data? And how could you resolve the tension between the two criteria?
3. You examine data plots with panels showing the probability of making different choices and for each choice the median, 10th, and 90th percentiles of the time to make each choice. What characteristics of which plot would tell you about the the average time to make a response and the variability in response times? What measurement derived from the plot tells you about the skew of the response time distributions? What relationship between the plots would be indicative of a speed-accuracy tradeoff?
4. An easy approach to model selection is to construct strong prior assumptions about which experimental effects will influence which parameters, effectively ruling out all other model variants. For example, one might make the prior assumption that a manipulation of stimulus strength can influence only sensitivity in a signal detection model, and not criterion placement. Name one danger of this method.
5. If you conducted a parameter recovery simulation for your new cognitive model, and found that there was unacceptably large variance in the recovered parameters (i.e., large inaccuracies that vary randomly with each new synthetic data set), what might you do?

9 Solutions (These Go in a Separate Book of Answers)

1. Uncertainty in the parameter estimate. A point estimate of 50% could mean that preparation occurs half of the time, but it does not mean this if a confidence interval for the point estimate runs from 10% to 90%.
2. The complex model, because AIC penalizes less harshly for complexity than BIC (also because, all else being equal, complex models tend to fit better due to flexibility). To resolve the tension, consider other model properties, such as how interpretable the parameter estimates are.
3. One danger is the inability to observe violations of the assumptions. Another is the possibility of missing some effects that are present in the data, but not the model, which can cause errors in parameter estimates.
4. Central tendency is indicated by the median (50th percentile). Variability is indicated by the distance between the 90th and 10th percentile. Skew is indicated by 90th minus 50th percentile subtracted from the 50th minus 10th percentile (positive values indicated positive skew, i.e., with a longer slow than fast tail). Speed-accuracy tradeoff is indicated by higher accuracy in condition A vs. condition B but also by a slower RT in condition A than B (or vice versa).
5. You could consider larger data samples – further parameter recovery simulations with larger simulated sample sizes could inform this decision. You might also consider adjusting your model to improve its estimation properties. For example, re-parameterization to reduce parameter correlation, or simplifying the model by fixing or removing some parameters.

10 Further Reading

1. Here are four course books on cognitive modeling, take your pick: Lewandowsky and Farrell [66], Busemeyer and Diederich [67], Hunt [68], and Polk and Seifert [69].
2. A hands-on, Bayesian approach to cognitive modeling is presented in Lee & Wagenmakers [31]; see also `www.bayesmodels.com`.
3. The series of tutorial articles in the *Journal of Mathematical Psychology* are a good source of information that is relatively easy to digest.

References

1. S.D. Brown, A.J. Heathcote, Cognitive Psychology **57**, 153 (2008)
2. R. Ratcliff, A. Thapar, G. McKoon, Psychology and Aging **21**, 353 (2006)

3. D.M. Riefer, B.R. Knapp, W.H. Batchelder, D. Bamber, V. Manifold, Psychological Assessment **14**, 184 (2002)
4. M. Mulder, E.J. Wagenmakers, R. Ratcliff, W. Boekel, B.U. Forstmann, Journal of Neuroscience **32**, 2335 (2012)
5. R. Ratcliff, Psychological Review **85**, 59 (1978)
6. R. Ratcliff, P. Gomez, G. McKoon, Psychological Review **111**, 159 (2004)
7. R. Ratcliff, J.J. Starns, Psychological Review **116**, 59 (2009)
8. R. Ratcliff, H.P.A. van Dongen, Psychonomic Bulletin & Review **16**, 742 (2009)
9. P.L. Smith, R. Ratcliff, Trends in Neurosciences **27**, 161 (2004)
10. R. Ratcliff, F. Tuerlinckx, Psychonomic Bulletin & Review **9**, 438 (2002)
11. E.J. Wagenmakers, H.J.L. van der Maas, R.P.P.P. Grasman, Psychonomic Bulletin & Review **14**, 3 (2007)
12. C. Donkin, S. Brown, A. Heathcote, E.J. Wagenmakers, Psychonomic Bulletin & Review **18**, 61 (2011)
13. C.W. Lejuez, J.P. Read, C.W. Kahler, J.B. Richards, S.E. Ramsey, G.L. Stuart, D.R. Strong, R.A. Brown, Journal of Experimental Psychology: Applied **8**, 75 (2002)
14. T.S. Wallsten, T.J. Pleskac, C.W. Lejuez, Psychological Review **112**, 862 (2005)
15. D. van Ravenzwaaij, G. Dutilh, E.J. Wagenmakers, Journal of Mathematical Psychology **55**, 94 (2011)
16. J.L. Hintze, R.D. Nelson, The American Statistician **52**, 181 (1998)
17. J.J. Rolison, Y. Hanoch, S. Wood, Psychology and Aging **27**, 129 (2012)
18. T.E. Parks, Psychol Rev **73**(1), 44 (1966)
19. N.A. Macmillan, C.D. Creelman, *Detection Theory: A User's Guide*, 2nd edn. (Erlbaum, Mahwah, NJ, 2005)
20. J.T. Wixted, V. Stretch, Psychological Review **107**, 368 (2000)
21. R. Ratcliff, J.N. Rouder, Psychological Science **9**, 347 (1998)
22. A. Voss, K. Rothermund, J. Voss, Memory & Cognition **32**, 1206 (2004)
23. T.C. Ho, S. Brown, J.T. Serences, The Journal of Neuroscience **29**, 8675 (2009)
24. G. Schwarz, Annals of Statistics **6**, 461 (1978)
25. M. Lee, J. Vandekerckhove, D.J. Navarro, F. Tuerlinckx, Presentation at the 40th Annual Meeting of the Society for Mathematical Psychology, Irvine, USA, July 2007 (2007)
26. J.J. Starns, R. Ratcliff, G. McKoon, Cognitive Psychology **64**, 1 (2012)
27. B. Rae, C. Heathcote, A. Donkin, L. Averell, S.D. Brown, Journal of Experimental Psychology: Learning, Memory, and Cognition (in press)
28. B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993)
29. W.A. Wagenaar, J.P.A. Boer, Acta Psychologica **66**, 291 (1987)
30. J. Vandekerckhove, D. Matzke, E.J. Wagenmakers, in *Oxford Handbook of Computational and Mathematical Psychology*, ed. by J. Busemeyer, J. Townsend, Z.J. Wang, A. Eidels (Oxford University Press, 2013)
31. M.D. Lee, E.J. Wagenmakers, *Bayesian Modeling for Cognitive Science: A Practical Course* (Cambridge University Press, in press)
32. H. Jeffreys, *Theory of Probability*, 3rd edn. (Oxford University Press, Oxford, UK, 1961)
33. F.J. Anscombe, The American Statistician **27**, 17 (1973)
34. P.N. McCullagh, J.A. Nelder, *Generalized Linear Models* (Chapman & Hall, London, 1983)
35. A.E. Raftery, Sociological methodology **25**, 111 (1995)
36. L. Averell, A. Heathcote, Journal of Mathematical Psychology **55**, 25 (2011)
37. S. Brown, A. Heathcote, Behavior Research Methods, Instruments, & Computers **35**, 11 (2003)
38. A. Heathcote, S. Brown, D.J.K. Mewhort, Psychonomic bulletin & review **7**, 185 (2000)
39. R.G. Pachella, in *Human Information Processing: Tutorials in Performance and Cognition*, ed. by B.H. Kantowitz (Lawrence Erlbaum Associates, Hillsdale (NJ), 1974), pp. 41–82
40. R.J. Audley, A.R. Pike, The British Journal of Mathematical and Statistical Psychology pp. 207–225 (1965)

41. D. Vickers, D. Caudrey, R. Willson, ACTPSY **35**, 151 (1971)
42. R. Ratcliff, A. Thapar, G. McKoon, Psychology and Aging **16**, 323 (2001)
43. E.J. Wagenmakers, R. Ratcliff, P. Gómez, G. McKoon, Journal of Memory and Language **58**, 140 (2008)
44. B. Rae, A. Heathcote, C. Donkin, L. Averell, S.D. Brown, Journal of Experimental Psychology: Learning, Memory, and Cognition pp. 1–45 (2013)
45. E.J. Wagenmakers, R. Ratcliff, P. Gómez, G.J. Iverson, Journal of Mathematical Psychology **48**, 28 (2004)
46. R. Ratcliff, J.N. Rouder, Psychological Science **9**, 347 (1998)
47. R.D. Morey, Tutorial in Quantitative Methods for Psychology **4**, 61 (2008)
48. A. Heathcote, J. Love, Frontiers in psychology **3** (2012)
49. J.J. Starns, R. Ratcliff, G. McKoon, Cognitive Psychology **64**, 1 (2012)
50. W.S. Cleveland, *Visualizing Data* (Hobart Press, New Jersey, 1993)
51. A. Heathcote, S. Brown, D.J.K. Mewhort, Psychonomic bulletin & review **9**(2), 394 (2002)
52. A. Heathcote, S. Brown, Psychonomic bulletin & review **11**, 577 (2004)
53. P.L. Speckman, J.N. Rouder, Psychonomic bulletin & review **11**, 574 (2004)
54. R.E. Kass, A.E. Raftery, Journal of the American Statistical Association **90**, 773 (1995)
55. T. Lodewyckx, W. Kim, M.D. Lee, F. Tuerlinckx, P. Kuppens, E.J. Wagenmakers, Journal of Mathematical Psychology **55**, 331 (2011)
56. R. Shiffrin, M. Lee, W. Kim, E.J. Wagenmakers, Cognitive Science: A Multidisciplinary Journal **32**, 1248 (2008)
57. D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. Van Der Linde, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**, 583 (2002)
58. T. Ando, Biometrika **94**, 443 (2007)
59. I.J. Myung, M.A. Pitt, Psychonomic bulletin & review **4**, 79 (1997)
60. K.P. Burnham, D.R. Anderson, Sociological Methods & Research **33**, 261 (2004)
61. C. Donkin, S. Brown, A. Heathcote, Journal of Mathematical Psychology **55**, 140 (2011)
62. B.M. Turner, P.B. Sederberg, S.D. Brown, M. Steyvers, Psychological Methods **18**, 368 (2013)
63. S. Farrell, E.J. Wagenmakers, R. Ratcliff, Psychonomic bulletin & review **13**, 737 (2006)
64. D.L. Gilden, Psychological Science **8**, 296 (1997)
65. G.E.P. Box, in *Robustness in statistics: Proceedings of a workshop*, ed. by R.L.L..G.N. Wilkinson (Academic Press, New York, 1979), pp. 201–236
66. S. Lewandowsky, S. Farrell, *Computational Modeling in Cognition: Principles and Practice* (Sage, Thousand Oaks, CA, 2010)
67. J.R. Busemeyer, A. Diederich, *Cognitive Modeling* (Sage, Thousand Oaks, CA, 2010)
68. E. Hunt, *The Mathematics of Behavior* (Cambridge University Press, Cambridge, 2006)
69. T.A. Polk, C.M. Seifert (eds.), *Cognitive Modeling* (MIT Press, Cambridge, MA, 2002)

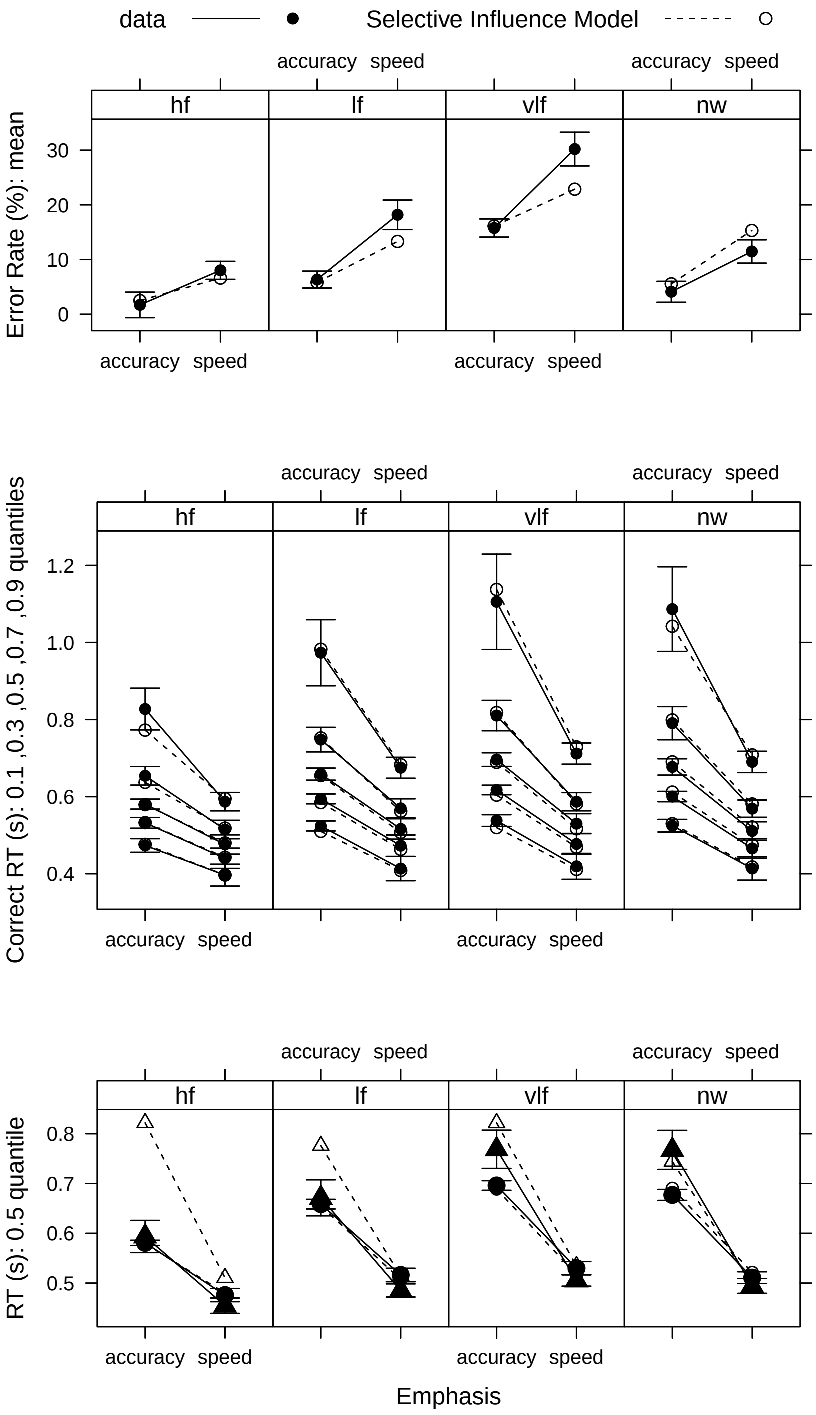

Fig. 5 Accuracy (upper panels), correct RT distribution (middle panels), median correct (circle symbols) and error (triangle symbol) RT (lower panels) plots of the average over 17 participants from Wagenmakers and colleagues' [43] Experiment 1 for high-frequency (hf), low-frequency (lf) and very-low-frequency (vlf) words and nonwords (nw). Each data point is accompanied by a 95% confidence interval assuming a Student *t* distribution and based on within-subject standard errors calculated as using the bias-corrected method described in [47].