CHAPTER 2

Bayesian Methods in Cognitive Modeling

MICHAEL D. LEE

INTRODUCTION

Proponents of Bayesian statistical methods argue that these methods provide a complete and coherent framework for the basic challenge of relating scientific models to data (Jaynes, 2003; Jeffreys, 1961). The defining feature of the Bayesian statistical approach is its use of probability distributions to represent uncertainty (Lindley, 1972). Prior probabilities over models and their parameters are transformed by Bayes' rule to posterior probabilities based on the evidence provided by data. Bayesian methods allow probability theory to be applied to draw inferences about parameters and models and to describe and predict data.

This chapter is about the application of Bayesian methods to cognitive psychology, but deals with only one of the three ways in which Bayesian methods have been used (Lee, 2011). To make the scope of the chapter clear, we first describe the three types of applications.

1. **Bayesian models of the mind:** Since Bayesian statistics provides a rational solution to the problem of making inferences about structured hypotheses based on sparse and noisy data, it provides an interesting and potentially compelling metaphor for the mind. Treating the mind as solving the problems it faces according to the principles of Bayesian inference has proved productive in areas ranging from vision to language to development to decision making. For a broad range of cognitive phenomena, the Bayesian metaphor complements other useful metaphors, such as information processing and connectionism. Models of cognition based on the Bayesian metaphor are often pitched at the computational level in Marr's (1982) hierarchy, although there are models of cognitive processes inspired by Bayesian sampling techniques. This "Bayes in the head" application of Bayesian statistics is controversial and nuanced (see Jones & Love, 2011, and associated commentaries) and is not the focus of this chapter.
2. **Data analysis:** A cornerstone of psychology as an empirical science is the statistical analysis of data using standard

I thank Joram van Driel for providing the raw data, and for his help in motivating the analyses presented in the case study. I have benefited from Bayesian discussions with many excellent researchers over the past 10 years. I am particularly grateful to Bill Batchelder, Simon Dennis, Geoff Iverson, Richard Morey, Dan Navarro, Jeff Rouder, Rich Shiffrin, Mark Steyvers, Joachim Vandekerckhove, Wolf Vanpaemel, and Eric-Jan Wagenmakers. I especially want to thank E.-J. and Wolf, who have been the most rewarding collaborators one could wish for and have forced me to take seriously the need for Bayesian model evaluation and informative priors, respectively. Finally, I thank John Wixted, E.-J., and Adriana Felisa Chávez De la Peña for their careful reading of an earlier version of this chapter.

Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition, edited by John T. Wixted.

DOI: 10.1002/9781119170174.epcn502

statistical models, typically based on generalized linear models. It has long been appreciated—with various degrees of consternation and urgency for reform—that classical statistical methods for parameter estimation and hypothesis testing have serious conceptual problems and practical limitations (Edwards, Lindman, & Savage, 1963; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015; Wagenmakers, 2007). Perhaps the most prominent application of Bayesian statistics to psychology is as a replacement for classical *t*-tests, *F*-tests, *p*-values, and so on. This can be routinely achieved by considering the same statistical model as the classical test, but applying Bayesian methods for inference (Kruschke, 2010). This application of Bayesian statistics is relatively uncontroversial, although, as with any major change requiring new methodological training and thinking, there are teething problems, and it is a work in progress. This sort of Bayesian data analysis in cognitive psychology is also not the focus of this chapter.

3. **Cognitive models:** As empirical sciences mature, theoretical and empirical progress often lead to the development of models. Cognitive psychology has a rich set of models for phenomena ranging from low-level vision to high-order problem solving. To a statistician, these cognitive models remain naturally interpretable as statistical models, and in this sense modeling can be considered an elaborate form of data analysis. The difference is that the models usually are very different from default statistical models like generalized linear models, but instead formalize processes and parameters that have stronger claims to psychological interpretability. There is no clear dividing line between a statistical and a cognitive model. Indeed, it is often possible for the same statistical model to have valid interpretations as a method of data analysis and a psychological model. Signal detection theory is a good example (e.g., Green & Swets, 1966). Originally developed as a method for analyzing binary decisions for noisy signals, in potentially entirely nonpsychological contexts, it nonetheless has a natural interpretation as a model of cognitive phenomena like recognition memory. Despite this duality, the distinction between data analysis and psychological modeling is a useful one. The use of Bayesian methods to implement, apply, and evaluate cognitive models *is* the focus of this chapter.

Advantages of Bayesian Methods

The usefulness of Bayesian methods in cognitive modeling stems from the combination of two important complementary strengths. Bayesian methods offer a *principled foundation for statistical inference* while simultaneously affording the *creative freedom and modeling flexibility* to develop, test, and use a wide range of cognitive models. Both of these trademarks contrast favorably with classical approaches, which provide a framework for statistical inference that is limited and inefficient at best, and unworkable and pathological at worst, and which consequently constrain the ways in which models of cognition can be expressed and applied to data.

The principled statistical framework afforded by the Bayesian approach stems from its foundations in probability theory, which provides a carefully axiomatized system for scientific inference (Cox, 1961; Jaynes, 2003, Chapter 2). At all stages in analyzing a model and data, the Bayesian approach represents everything that is and is not known about the uncertain parts of a model, such as model parameters, and uncertain parts of the data, such as missing data. It uses probability distributions in a simple, consistent, and interpretable way to

represent this information, and automatically updates what is known as new information, especially in the form of new data, becomes available.

In addition, Bayesian methods make it straightforward to focus on the inferences that are the important ones for the scientific questions being asked. It is possible to examine parameters in isolation, with their uncertainty averaged (or marginalized) over other parameters, it is possible to examine combinations of parameters, or condition what is known about one or more parameters on assumptions about other parameters, and so on. The flexibility and generality of the Bayesian framework make it natural and easy to translate substantive research questions into specific statistical inferences.

The top-left panel of Figure 2.1 shows the structure of a standard cognitive model, involving cognitive variables θ controlling cognitive processes f that generate behavior y. Many, and perhaps most, cognitive models can validly be conceived as mappings of this form $y = f(\theta)$. In many cases, the function f is complicated, and involves many processes, but nonetheless constitutes a single mapping from parameters to data. Given this mapping, Bayesian inference allows for prior knowledge about parameters to be updated to posterior knowledge that incorporates the information provided by the data, and for prior predictions about data based on the model to be updated to posterior predictions. It also allows different models to be compared, based on the evidence provided by data. The modeling freedom afforded by the Bayesian approach stems from its ability to make these inferences and evaluations for more complicated model structures in exactly the same way as for a single mapping from parameters to data. The introduction of more complicated modeling assumptions does not require a shift in the principles by which the model is analyzed and applied.

The remaining panels of Figure 2.1 highlight three important classes of extended modeling approaches made feasible by using Bayesian methods. The top-right panel shows a hierarchical structure. The key assumption is that the basic model parameters θ are themselves generated by a psychological process. Hierarchical models drive theorizing to deeper, more abstract, and more fundamental levels by including models of how the basic psychological variables that control behavior are generated, rather than just assuming they exist. For example, $y = f(\theta)$ might describe how an individual with a memory capacity captured by θ performs on a recall task, while $\theta = g(\psi)$ might describe the developmental, learning, or neural processes by which the individual came to have that memory capacity. In the context of studying the impact of Alzheimer's disease and related disorders on memory-task performance, Pooley, Lee, and Shankle (2011) modeled the psychological parameters for a simple two-parameter model of recall as depending on a clinical measure of impairment known as the functional assessment staging (FAST) stage (Reisberg, 1988). In this application of hierarchical methods, the hyper-parameter ψ is the FAST stage for an individual, the parameters θ are their two recall-model parameters, and the process g represents the modeling assumptions that map the FAST stage to the recall parameters.

The bottom-left panel of Figure 2.1 shows a latent-mixture model structure. The key assumption is that observed behavioral data y do not come from a single source, but instead arise as a combination of outcomes from different cognitive processes $f_1, f_2, \ldots, f_n$ controlled by potentially different cognitive parameters $\theta_1, \theta_2, \ldots, \theta_n$. How the behaviors that are produced by these different processes are combined is controlled by a mixing process h that itself is indexed by parameters ϕ. The ability to make these indicators latent, so that the combinations

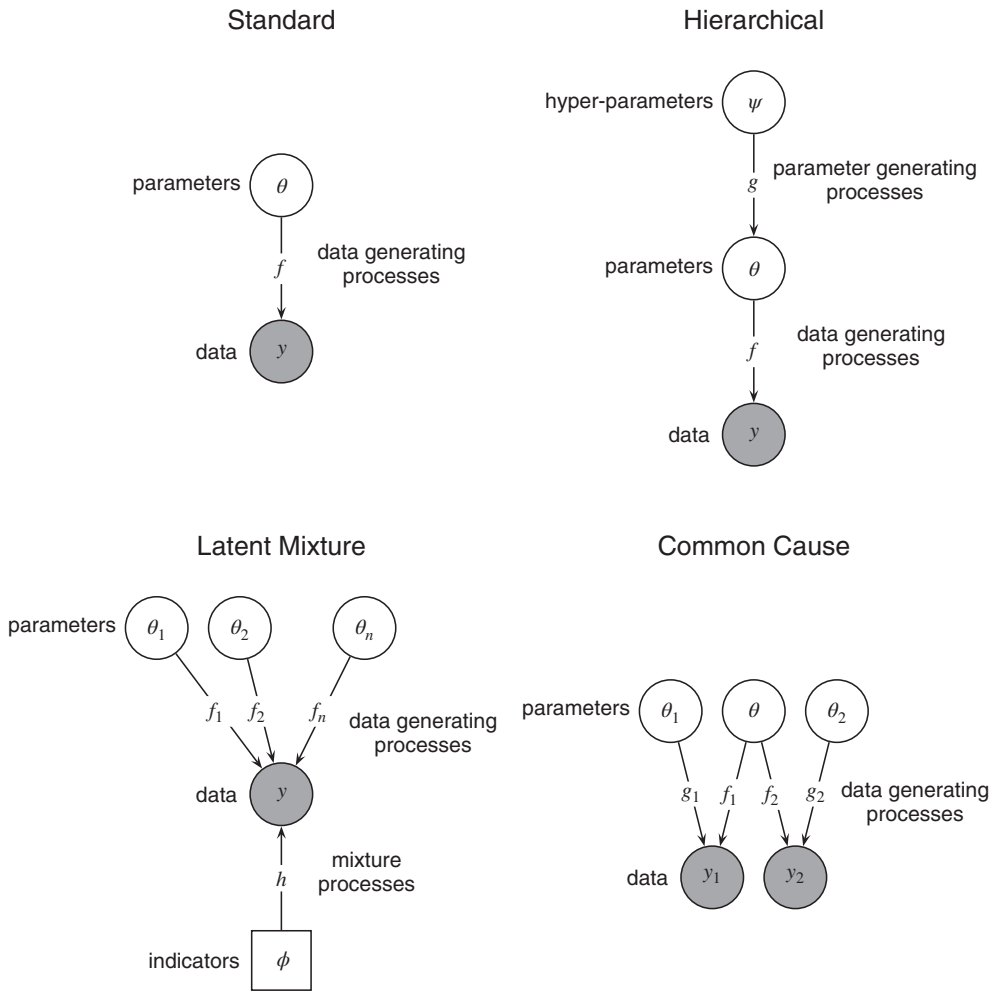


Figure 2.1 Overview of general cognitive modeling structures made possible by using the Bayesian approach to relate models to data. The standard model defines a process f controlled by parameters θ for generating behavioral data y. A hierarchical model structure extends the standard model by including a process g controlled by parameters ψ that generates the original parameters θ. The latent mixture structure allows for different data-generating processes $f_1, f_2, \ldots, f_n$ controlled by different parameters $\theta_1, \theta_2, \ldots, \theta_n$ to combine to generate the data, according to some mixing process h controlled by parameters ϕ. The common cause structure allows for different data y_1 and y_2 to be in part generated by the same parameters θ.

present in the data are inferred from the data themselves, is a powerful tool in building models of cognition, especially in accounting for qualitative or discrete differences. For example, the cognitive processes might be different strategies used by people to make decisions (Hilbig & Moshagen, 2014; Lee, 2016), so that the overall observed behavior comes from multiple sources that are best modeled separately. In this case, the indicator variables ϕ would correspond to which strategy each person used.

Finally, the bottom-right panel of Figure 2.1 shows a common-cause model structure. The key assumption is that some psychological variables influence multiple

sorts of cognitive capabilities. This means that two different data sets y_1 and y_2 relating to potentially different tasks and stimuli are both influenced by the same psychological variables represented by θ. The processes that generate data for each task f_1 and f_2 may be different, but are controlled by the same θ. Each data set may also depend on task-specific parameters and processes, as shown, but there is some level of common cause across the two tasks. This is a powerful theoretical assumption familiar, for example, from empirical sciences like physics, in which the same physical variables, like mass, influence observations for multiple phenomena, like momentum, gravity, and charge. In cognitive modeling, for example, θ might represent memory acuity or discriminability, the data y_1 and y_2 might be recall and recognition tasks, and f_1 and f_2 would be models of the recall and recognition processes involved.

Hierarchical, latent-mixture, and common-cause modeling structures all extend the standard structure significantly and allow for the formalization of much more elaborate accounts of cognition. There is nothing inherently Bayesian about any of these structures. The key point is that Bayesian methods work in exactly the same way for any of these structures, or any others that could similarly be developed. For example, there is nothing stopping latent-mixture models from being extended hierarchically and connecting with multiple data sets, which would combine all three of the novel structures in Figure 2.1. This freedom in model formulations allows a theorist to propose detailed, ambitious, and speculative accounts of cognition, safe in the knowledge that it can be applied to data and evaluated against data in the same way as a simple model. The principled nature of Bayesian inference, which draws only the inferences justified by the data and values simplicity, will rein in theoretical excess. Bayesian inference will diagnose a model that is not useful or is too complicated to be justified by available evidence. In this way, Bayesian inference allows a modeler to chart new theoretical territory, leading to more complete and useful models, and better inferences and predictions.

Overview

The advantages of Bayesian methods just described and the model structures presented in Figure 2.1 are generally applicable to all areas of cognitive modeling. There are worked examples of hierarchical, latent-mixture, and common-cause modeling in areas including perception, memory, learning, development, categorization, decision making, and problem solving. Trying to cover all of these areas in a single chapter is impossible. Instead, the remainder of this chapter considers a single extended case study. The case study is designed to give tutorial examples of the Bayesian approach to cognitive modeling being applied to realistic research questions, models, and data. Where the specific tutorial examples in the case study raise more general issues—especially those that highlight misconceptions in the field or important directions for future development—they are discussed in separate subsections. The hope is that this approach demonstrates the feasibility, applicability, and intuitiveness of the Bayesian approach in a concrete way, while still reviewing the strengths and weaknesses of its current application in the field as a whole and discussing relevant conceptual and theoretical issues.

A CASE STUDY

Our case study comes from the domain of psychophysics (Kuss, Jäkel, & Wichmann, 2005). It involves a relatively standard

experimental design and relatively simple models, but nonetheless raises a rich set of theoretical and empirical questions. The case study works through a set of these questions, using them to demonstrate how they can be addressed by the Bayesian approach to cognitive modeling. In the course of addressing the questions, many of the basic properties of Bayesian inference and all of the extended model structures in Figure 2.1 are demonstrated in concrete ways. All of the code for all of the analyses presented in the case study is available, together with raw data, on the Open Science Framework project page at https://osf.io/zur8m.

Experimental Data

The data come from experiments associated with the research reported by van Driel, Knapen, van Es, and Cohen (2014) and involve two basic psychophysical duration discrimination tasks. In the auditory task, subjects judged the duration of auditory beeps (500 Hz sine waves with 5 ms ramp-up/down envelopes, played by speakers left and right from the screen). In the visual task, they judged the duration of a red LED light located at the center of a computer screen. In both tasks, each trial consisted of a 500 ms standard, followed by a 1,000 ms interstimulus interval (ISI), followed by a target stimulus of variable duration. Subjects indicated with a key press whether they perceived the target stimulus to be longer or shorter than the standard. They were required to respond within 1,500 ms of the target offset and were provided with feedback after each trial. The same 19 subjects completed three blocks of 80 trials for both the auditory and visual tasks.

We focus on just six subjects, chosen to allow some of the most important features and methods in Bayesian analysis to be demonstrated. Figure 2.2 summarizes the behavioral data for these six subjects. Each panel corresponds to a subject, the *x*-axis shows the target duration, and the *y*-axis shows the proportion of times that duration was perceived to be longer than the standard. All of the subjects show behavioral patterns consistent with standard psychophysical theory. Targets much shorter than the standard are perceived as such, and targets much longer than the standard are perceived as such. For targets near the standard, there is greater uncertainty, with a rise in the proportion of longer responses as the duration of the target increases. The data in Figure 2.2 suggest similarities and differences between the subjects, and between how the subjects perceive auditory versus visual stimulus durations. For example, the behavior of subject F appears far less regular than that of the other subjects, and subject A appears more similar in their response patterns to auditory and visual stimuli than subject D.

Research Questions

The main research motivation for van Driel et al. (2014) was to examine how the brain integrates temporal information across modalities. Addressing this general question raises a series of more specific research questions, many of which are naturally treated as cognitive modeling challenges. The following is a list of research questions for which the behavioral data in Figure 2.2 should be useful.

- What is the form of the psychophysical function that maps the physical measure of target stimulus duration to the psychological measure of the probability the target is perceived to be longer or shorter than the standard? The literature is full of established possibilities, including the logistic, Cauchy, Gumbel, Weibull, and others (Kuss et al., 2005).
- For any particular form of psychophysical function, what parameterization best

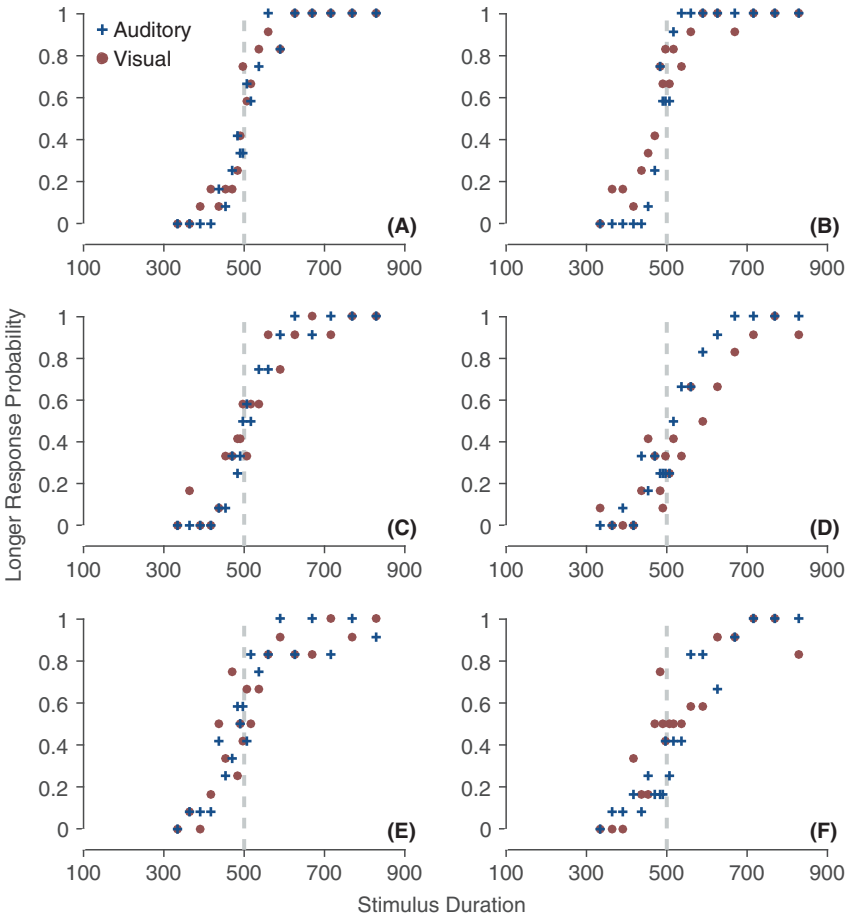


Figure 2.2 Behavioral data for six subjects, labeled A to F, in two psychophysical duration discrimination tasks. Each subject is in a different panel, with the *x*-axis giving the duration of the target auditory tones and visual lights, and the the *y*-axis giving the proportion of responses for which the target stimulus was judged by the subject to have a longer duration than the standard. The broken vertical line shows the duration of the standard.

describes the psychophysical mapping, and do the values of those parameters have psychologically meaningful interpretations?

- Are there individual differences in psychophysical function or parameterizations different people use? These individual differences could be qualitative, in the sense that different people use different psychophysical functions, or they could be more quantitative, in the sense that different people use different parameter values for the same psychophysical function.
- Are there modality-specific differences in the perception of duration? The behavioral data are especially well suited to addressing this question, since each subject completed both the auditory task and the visual task.
- Are there sequential dependencies in the responses for a subject doing many consecutive trials? In other words, is the perception of the duration of the current

target stimulus independent of previous stimuli and responses, or is there some carryover effect for perception or behavior on the current trial?

- Do subjects produce contaminant responses on some trials that suggest a lack of motivation or attention? If so, how do these trials affect inferences about the cognitive processes they use when they are attending to the task?

This list is necessarily incomplete, and each individual question could be tackled in many ways. Thus, our goal is not to answer each in detail. Rather, we use these research questions to motivate the specific modeling analyses that follow, and to highlight the generality of Bayesian methods to enable a wide range of cognitive modeling questions to be addressed.

Model Development

Psychophysics has proposed and examined many possible psychophysical relationships between the stimulus and response (e.g., Kuss et al., 2005, Figure 2b). We consider just two possibilities. The first is a logistic function of the form

$$\theta = 1 \Big/ \left(1 + \exp\left(-\frac{x - s - \alpha}{\beta}\right)\right). \quad (1)$$

The second is a Cauchy function of the form

$$\theta = \arctan\left(\frac{x - s - \alpha}{\beta}\right) \Big/ \pi + \frac{1}{2}. \quad (2)$$

In both cases, θ is the probability of responding "longer" for a target stimulus of length x compared to a standard of length s, and α and β are parameters.

Figure 2.3 shows both the logistic and Cauchy psychophysical functions at different parameterizations. It is visually clear that the two functions are very similar when they have the same parameterization. The important difference is that the Cauchy has fatter tails, in the sense that target stimuli that are very different from the standard correspond to response probabilities a little further from 0 and 1 than the logistic function. Figure 2.3 also makes clear the effects that the two

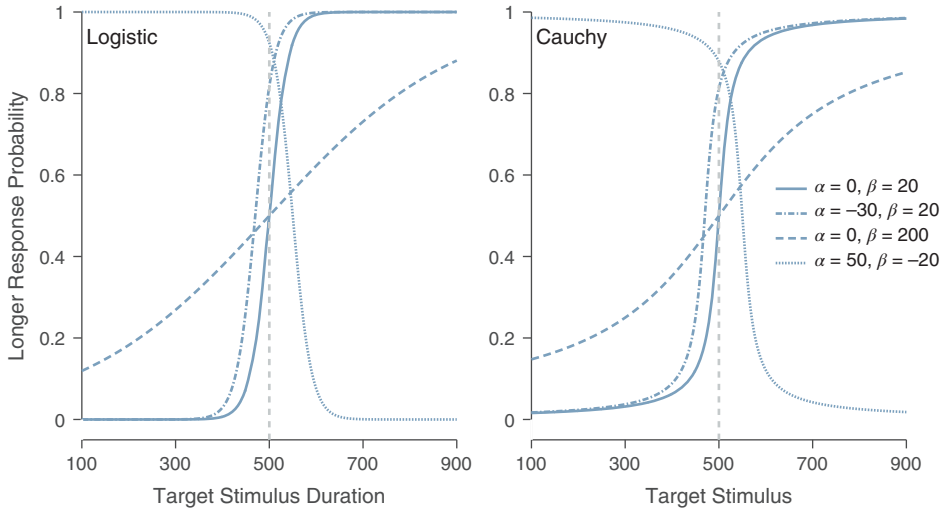


Figure 2.3 Logistic (*left*) and Cauchy (*right*) psychophysical functions, both shown at five different parameterizations.

model parameters have on the psychophysical function, and how the parameters can be interpreted. These effects and interpretations are very similar for both the logistic and Cauchy functions.

The model parameter α corresponds to a shift in the psychophysical function, serving to bias responses toward "longer" decisions when $\alpha > 0$, and "shorter" decisions when $\alpha < 0$. Geometrically, when $\alpha = 0$ the psychophysical function produces the maximally uncertain response probability of 0.5 when the presented stimulus equals the standard. The psychophysical function shifts to the left as α decreases, and to the right as α increases.

The model parameter β corresponds to the scale of the psychophysical function, measuring how quickly and in what direction the function changes as the target stimulus duration changes. Positive values of β correspond to functions that increase as the stimulus becomes longer, whereas negative values correspond to functions that decrease as the stimulus becomes longer. The smaller the absolute value of β, the sharper the change in the psychophysical function, with $\beta = 0$ corresponding to a step function. The larger the absolute value of β, the shallower or more gradual the change in the psychophysical function.

To specify a complete model capable of making predictions, a prior distribution is required for the joint distribution of the parameters α and β. This prior distribution expresses modeling assumptions about plausible shifts and scales for the psychophysical function. These priors are developed in the same way that likelihoods—in this case, the logistic form of the psychophysical function—are traditionally developed in cognitive psychology: through some creative application of applicable theory, knowledge of previous data, and relevant logical constraints (Lee & Vanpaemel, in press).

In general, Bayesian methods require the specification of joint prior distributions, giving the prior probability of each possible combination of model parameters. For the current model, we make the simplifying assumption that the prior for each parameter can be specified separately, and that the prior probability of any combination of parameters is just the product of their individual probabilities. Technically, the assumption is that the joint prior distribution is the product of the marginal prior distributions for each parameter. This seems plausible, given the separate psychological interpretation of the two parameters, and is consistent with the guiding principle of selective influence in cognitive modeling. Selective influence is the idea that experimental manipulations influence just one model parameter, with the conception that each parameter represents a different psychological variable (Batchelder & Alexander, 2013; Voss, Rothermund, & Voss, 2004).

For the shift parameter α, the theoretically optimal value is 0, corresponding to no bias toward either the "longer" or the "shorter" response. This suggests the prior should be centered at zero, and the symmetry of the task, with "longer" and "shorter" answers being logical complements, suggests a symmetric prior. The remaining modeling assumption to be made involves how plausible shifts of different durations might be. Our model assumes a standard deviation of 50 ms, so that

$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

where the Gaussian distribution is parameterized in terms of its mean and precision.

For the scale parameter β, one clear theoretical assumption is that the psychophysical function should increase as the target stimulus duration lengthens. This assumption requires that the psychophysical function should increase from left to right, which

corresponds to the constraint $\beta > 0$. The remaining modeling assumption to be made involves how plausible scales of different magnitudes, corresponding to the steepness of the rise in the function, might be. Based on Figure 2.3 a scale of $\beta = 200$ seems as shallow as is plausible, so our model commits to a standard deviation of 100. Using a truncated Gaussian distribution then allows a peak at 0, corresponding to an optimal step function, resulting in

$$\beta \sim \text{TruncatedGaussian}_{+}(0, 1/100^2).$$

It would, of course, be possible to choose other priors, just as it would be possible to choose a different likelihood function. Changing either will clearly affect the inferences made, which is as it should be. It would be strange if making different assumptions did not affect the results. Nevertheless, there is a widely expressed resistance to Bayesian methods because specifying priors affects the conclusions. The process of choosing priors is somehow seen as arbitrary, despite the same processes being used to choose likelihoods being standard, comfortable, and unquestioned. Sprenger (2015) puts it nicely, saying "The bottom line . . . is that the choice of the prior is, just like any other modeling assumption in science, open to criticism."

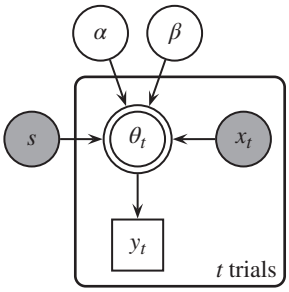
Models Require a Likelihood and a Prior to Make Predictions

A defining property of a scientific model is that it makes predictions. This allows the model to be validated, falsified, and, ultimately, applied (Feynman, 1994, Chapter 7). The defining feature of Bayesian statistics is that it represents the uncertainty about parameters using a prior distribution. Together, the likelihood function and the prior combine to form the predictions of the model. This means that, in the Bayesian approach, likelihood functions—like the logistic and Cauchy psychophysical functions—are not themselves models. They are not complete as models until a prior distribution is placed on the parameters α and β. In a sense, it is the predictions about data that *are* the model, so both the likelihood and the prior should be conceived as having equal status as components of a model. It is often the case that the likelihood is a natural way to formalize assumptions about the cognitive processes that generate behavioral data, while the prior distribution is a natural way to formalize assumptions about the cognitive variables that control these processes. A key feature of the Bayesian approach is that the prior distribution over parameters has the same status as the likelihood as a vehicle to formalize theory and assumptions (Lee & Vanpaemel, in press; Vanpaemel & Lee, 2012). This Bayesian feature has yet to see full fruition in cognitive modeling, and most often developing the prior is not given the same theoretical attention that is given to developing the likelihood, although there are some promising exceptions (e.g., Donkin, Taylor, & Le Pelley, 2017; Gershman, 2016; Lee & Danileiko, 2014).

Graphical Model Representation

The model just developed is shown as a graphical model in Figure 2.4. The graphical model formalism provides a convenient approach for expressing many probabilistic models of cognition, and has the advantage of being especially well suited to the application of computational methods for Bayesian inference. Graphical models were developed and remain widely used in artificial intelligence and machine learning (e.g., Jordan, 2004; Koller, Friedman, Getoor, & Taskar, 2007; Pearl, 1998) and are progressively being adopted in cognitive psychology (Lee & Wagenmakers, 2014).

In a graphical model, nodes in a graph represent parameters and data, and the graph structure indicates how the parameters generate the data. In Figure 2.4, the parameters α and β are shown as circular nodes, because they are continuous-valued,



$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0, 1/100^2)$$
$$\theta_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Figure 2.4 Graphical model representation of the logistic psychophysical model, with informative priors.

and unshaded, because they are latent or unobserved. The probability that the subject will respond "longer" on the tth trial, θ_t, has a double-bordered node, because it is a deterministic function, following Equation (1). It depends both on the unknown parameters α and β, and on the known values for the standard s and the duration x_t of the target stimulus presented on the tth trial. The nodes for s and x_t are shaded to indicate that they are known values, given, in this case, by the design of the experiment. The response probability θ_t then simply generates the predicted behavioral data

$$y_t \sim \text{Bernoulli}(\theta_t)$$

so that $y_t = 1$ if the subject chooses the "longer" response on the ith trial, which happens with probability θ_t, and $y_t = 0$ if the subject chooses the "shorter" response, which happens with probability $1 - \theta_t$.

Graphical Models Have Their Limits

While graphical models provide a flexible, modular, and interpretable language for formalizing cognitive models, they are far from entirely general. They are poorly suited for expressing some major classes of cognitive models. One example involves nonparametric models in which the parameters of a model are not fixed, but depend on the available data (Navarro, Griffiths, Steyvers, & Lee, 2006). Nonparametric models have been developed in a number of areas of cognitive psychology, including language (Goldwater, Griffiths, & Johnson, 2009), categorization (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011), and stimulus representation (Navarro & Griffiths, 2008). One intuitive application of nonparametric Bayesian modeling is to individual differences, with the insight that if there are groups of people who differ from one another, and a parameter is needed to quantify the psychological differences of each group, the number of parameters needed grows as data from more people are observed, and the number of different groups encountered increases. Another example of the limits of the graphical modeling formalism involves doubly stochastic models, in which (roughly) inferences need to be made about parameters that are themselves inferences. This arises naturally in many cognitive modeling situations, ranging from noble pursuits like modeling people's theory of mind (Baker, Saxe, & Tenenbaum, 2011) to less noble ones like trying to combine the knowledge of game show contestants (Lee, Zhang, & Shi, 2011). It also arises in making inferences about Bayesian models of cognition, since the challenge is for the scientist to make inferences, based on behavioral data, about how a person makes inferences, based on the stimuli the person is presented (Hemmer, Tauber, & Steyvers, 2014; Tauber, Navarro, Perfors, & Steyvers, 2017). A promising alternative probabilistic

programming approach to graphical models, especially well suited for these sorts of cognitive models, is described and implemented by Goodman and Stuhlmüller (2014).

Prior Prediction

A key component—perhaps *the* key component—of a scientific model is the set of predictions it makes. These come from the assumptions about variables and processes, and how they combine to produce data. In the graphical model in Figure 2.4, the responses y_t are shown as unobserved. This allows the model predictions to be examined before the actual data summarized in Figure 2.2 are considered.

Figure 2.5 shows the prior distribution for the parameters α and β, and the prior on the logistic psychophysical function that is implied by the prior distribution. The inset panel shows samples from the joint prior distribution (α, β) in a two-dimensional space as a set of points, and marginal prior distributions for each parameter as histograms to the top and the right. The main panel shows a set of specific logistic functions, each corresponding to a single sample from the prior (i.e., a specific combination of α and β). These samples from the prior of the psychophysical function correspond to the prediction of the model as to how subjects will translate target stimulus durations to response probabilities.

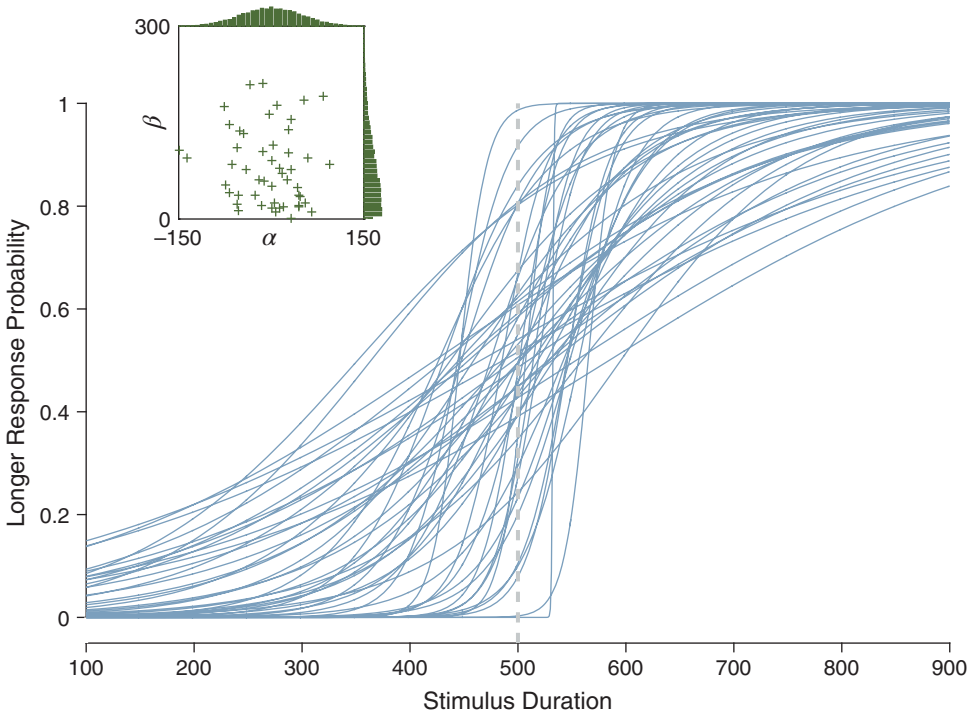


Figure 2.5 The prior distribution for the logistic psychophysical model with informative priors. The inset panel shows samples from the joint prior distribution and the marginal prior distribution for the model parameters α and β. The main panel shows samples from the corresponding prior distribution for the psychophysical function.

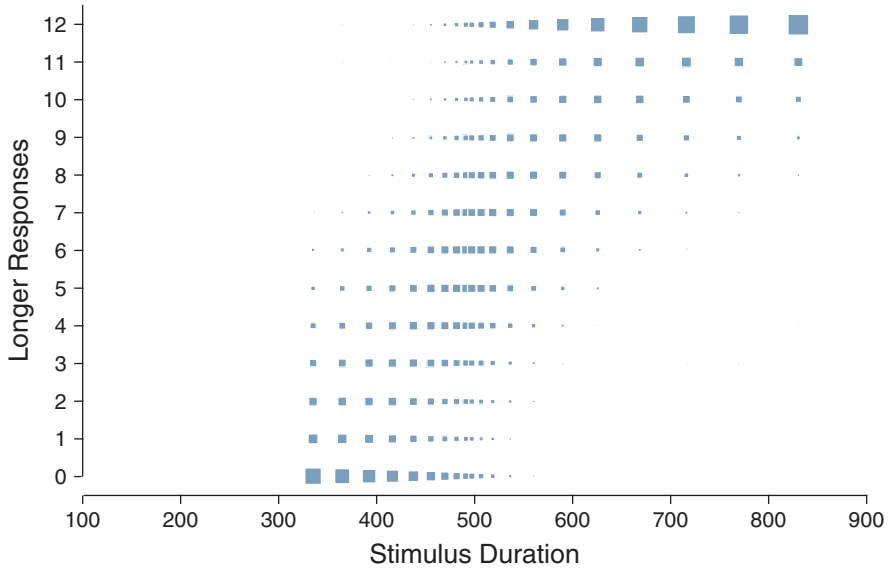


Figure 2.6 The prior predictive distribution for the logistic psychophysical model with informative priors. The x-axis corresponds to the unique target stimulus durations, and the y-axis corresponds to counts of the number of times each target duration is perceived as longer than the standard. Squares show the prior predictive distribution, with the area of each square being proportional to its prior predictive probability.

Figure 2.6 shows a prior predictive analysis. Recall that each of the 20 unique target stimulus durations was presented 12 times during the task. The marginal prior predictive distribution for each unique target duration is a set of probabilities for the counts from 0 to 12, corresponding to the prior probability that the target will be classified as "longer" on that number of trials. These marginal distributions are shown by the squares, with the area of each square being proportional to its mass in the sampled prior predictive distribution (i.e., how frequently that count was sampled as a prior predictive count).

The prior predictive distribution for any specific stimulus duration x is a combination, or average, of the predictions made by each possible parameterization of the model, weighted by the prior probability of that parameterization. Formally, it combines the distribution of response probabilities for a stimulus duration given by the model, $p(\theta \mid x, M)$ and the distribution of predicted behavioral data under the model for each response probability $p(y \mid \theta, M)$. Integrating over these probabilities gives

$$p(y \mid x, M) = \int p(y \mid \theta, M) p(\theta \mid x, M) \quad \mathrm{d}\theta,$$

which corresponds to the overall prediction of the probability of a "longer" response for a target stimulus with duration x. Intuitively, the prior predicted data comes by considering the data that would be seen at every possible response probability θ, weighted by how likely those response probabilities θ are under the model, where the model consists of both the assumed logistic psychophysical function and the joint prior distribution over the shift and scale parameters.

The prior predictive distribution shown in Figure 2.6 seems to be a reasonable one. For target stimuli with durations much shorter than the standard, the prediction is that they

will almost always be perceived as "shorter." For target stimuli with durations much longer than the standard, the prediction is that they will almost always be perceived as "longer." For target stimuli with durations near the standard, wider ranges of possible counts in the data are expected, corresponding to more inconsistency or uncertainty in the perceived duration. The goal of the prior predictive analysis is to verify that the model is making predictions about the outcome of the experiment that match the goals of the model and its constraining theory. The prior predictive distribution shown in Figure 2.6 fares well in this regard.

Alternative Models With Vague Priors

The model developed to this point differs from many applications of Bayesian methods to cognitive psychology, because it uses informative priors. That is, the priors were specified as modeling assumptions about plausible values of the shift and scale model parameters. Often, research in this area has shown an implicit or explicit discomfort with priors (Kievit, 2011), presumably because they are modeling devices unfamiliar from traditional model-fitting approaches. A common reaction to this discomfort is to use priors variously labeled "vague," "weakly informative," "flat," or "diffuse." We think this is a conceptual mistake and—before proceeding to use and extend the model just developed—it is instructive to consider the consequences and problems coming from taking the more standard path of using vague priors.

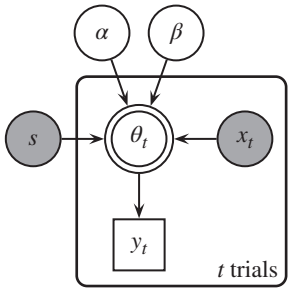
Figure 2.7 shows, as a graphical model, an alternative model that continues to use a logistic psychophysical function, but assumes vague priors on the model parameters. These take the form of very low-precision Gaussian distributions

$$\alpha \sim \text{Gaussian}(0, 0.000001)$$
$$\beta \sim \text{Gaussian}(0, 0.000001)$$

that give approximately equal probability to a very wide range of numbers for both α and β.

Figure 2.8 shows the prior distributions for this alternative model. The vague nature of the prior is clear from the axis limits of the inset panel showing the joint and marginal prior parameter distributions. The effect on the prior for the assumed psychophysical relationship between target stimuli and response probabilities is clear from the main panel. The model now predicts a wide variety of psychophysical functions, most of which run counter to reasonable theoretical and empirical expectations. For example, because the scale parameter β is no longer constrained to be positive, half of the psychophysical functions decrease the probability of a "longer" response as the target stimulus duration increases.

The effect of the vague priors on prior prediction is shown in Figure 2.9, which repeats the analysis shown in Figure 2.6 for the



$$\alpha \sim \text{Gaussian}(0, 0.000001)$$
$$\beta \sim \text{Gaussian}(0, 0.000001)$$
$$\theta_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Figure 2.7 Graphical model representation of the logistic psychophysical model with vague priors.

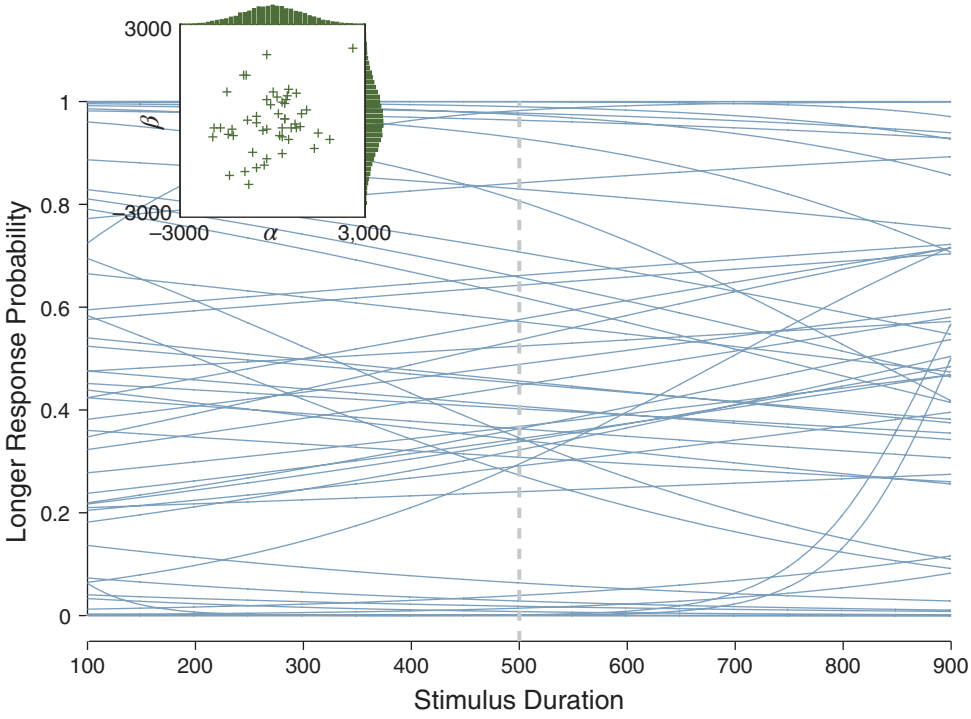


Figure 2.8 The prior distribution for the logistic psychophysical model with vague priors. The inset panel shows samples from the joint prior distribution and the marginal prior distribution for the model parameters α and β. The main panel shows samples from the corresponding prior distribution for the psychophysical function.

alternative model. All possible response patterns are now given significant prior predictive mass. That is, regardless of the duration of the target stimulus, counts of "longer" responses from 0 to 12 are predicted. It is clear that the alternatively parameterized model makes less theoretically and empirically sensible predictions about the underlying psychophysical function and behavioral data. The predictions are also less constrained, and so are less amenable to scientific evaluation or practical application. In a sense, the use of a vague prior neuters the theoretical substance of the logistic psychophysical function, destroying the relationship between the duration of the target stimulus and the expected task behavior.

As if this were not bad enough, the case against the unthinking use of vague priors can be strengthened by considering their effect on alternative parameterizations of the same model. Following the insightful demonstration of Kuss et al. (2005), we now consider the logistic psychophysical function with the scale parameterized differently, as

$$\theta_t = 1/(1 + \exp(-\beta(x_t - s - \alpha))). \qquad (3)$$

This is clearly the same model as Equation (1) if the prior distribution for the scale parameter β is adjusted to convey the same information and produce the same prior predictions. The change in parameterization has a major impact, however, if the same vague priors continue to be applied, because the same prior now corresponds to different information. Figure 2.10 shows the prior for the logistic psychophysical function in Equation (3) for the alternatively parameterized model,

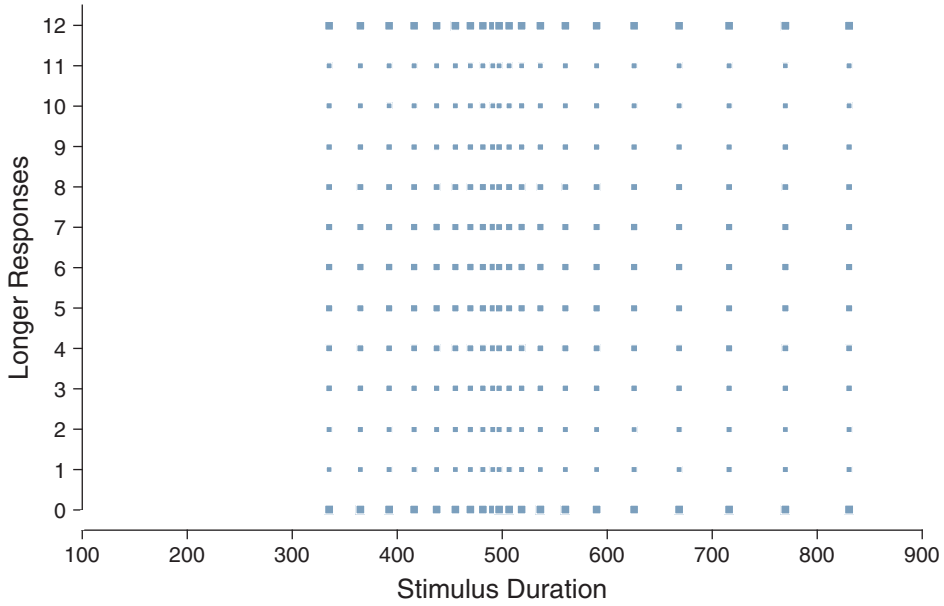


Figure 2.9 The prior predictive distribution for the logistic psychophysical model with vague priors. The *x*-axis corresponds to the unique target stimulus durations, and the *y*-axis corresponds to counts of the number of times each target duration is perceived as longer than the standard. Squares show the prior predictive distribution, with the area of a square being proportional to its prior predictive probability.

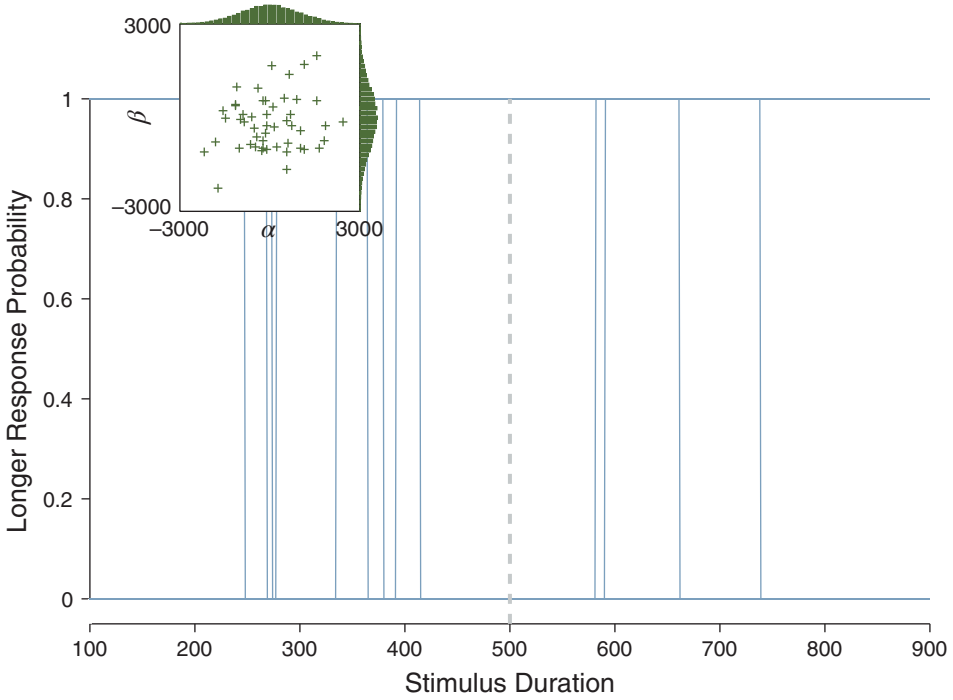


Figure 2.10 The prior distribution for the alternatively parameterized logistic psychophysical function with vague priors. The inset panel shows samples from the joint prior distribution and the marginal prior distribution for the model parameters α and β. The main panel shows samples from the corresponding prior distribution for the psychophysical function.

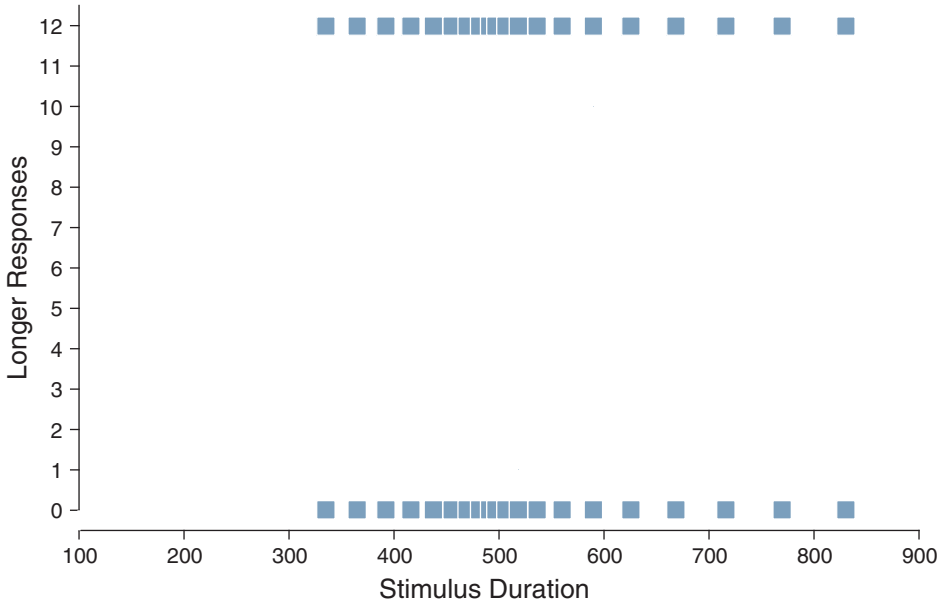


Figure 2.11 The prior predictive distribution for the alternatively parameterized logistic psychophysical model with vague priors. The x-axis corresponds to the unique target stimulus durations, and the y-axis corresponds to counts of the number of times each target duration is perceived as longer than the standard. Squares show the prior predictive distribution, with the area of a square proportional to its prior predictive probability.

using the same vague priors. Now almost all of the predicted psychophysical functions are near step functions. This is because the prior on the (now) inverse-scale parameter $\beta \sim \text{Gaussian}(0, 0.000001)$ has most of its density at large absolute values, and these correspond to very steep changes in response probability.

The consequences of this strange prior for the psychophysical function are emphasized by the resulting prior predictive distribution shown in Figure 2.11. Because the prior almost always gives response probabilities near 0 or 1, the prior prediction is that subjects always classify the same stimulus the same way. That is, each stimulus is always perceived as longer or shorter as the standard on each trial it is presented. While the prior predictive distribution for the flat prior in the original parameterization, shown in Figure 2.9, is unhelpfully vague, the prior predictive in Figure 2.11 is even worse. Vague mispredictions can be overcome by enough data, but specific mispredictions prevent a model from being useful until an overwhelming amount of contradicting evidence is obtained.

Flat and Uninformed Priors Are Not the Same Thing

The analysis summarized in Figure 2.10 is a concrete example of a common conceptual error. It is not the case that "[t]ypically, a non-informative prior would be represented by a distribution with a relatively flat density, where the different values the parameter can take on have approximately equal likelihood under the distribution" (Depaoli & van de Schoot, 2017). The effect of using a relatively flat prior density for the scale parameter β

in Figure 2.10 is to make strong and highly implausible psychophysical assumptions. Relatively flat densities can be vaguely informative for location parameters, but not for the other types of parameters frequently found in statistical and cognitive models of psychological data. Foundational Bayesian texts such as Gelman, Carlin, Stern, and Rubin (2004) make this point, often by considering the effect of uniform priors on precision and other scale parameters.

Parameter Inference

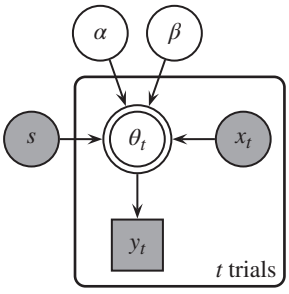
Having observed the problems with using vague priors, we now return to the original model with informative priors. The graphical model in Figure 2.12 uses the informative model again. It differs from Figure 2.4 in that the y_t node is shaded, indicating that the behavioral data are now observed. In this form, the graphical model is immediately amenable to making inferences about the model and its parameters from the available data. Bayes' rule defines the posterior distribution of the model parameters α and β, conditional on the data $y = (y_1, \ldots, y_T)$ and model M, in terms of the prior and likelihood,

$$p(\alpha, \beta \mid y, M) = \frac{p(y \mid \alpha, \beta, M)p(\alpha, \beta \mid M)}{p(y)}. \tag{4}$$

For posterior inference about parameters, the denominator of Equation (4) is constant, so the posterior is simply proportional to the product of the likelihood and the prior.

In practice, the posterior distribution $p(\alpha, \beta \mid y, M)$ can be approximated by drawing samples using computational Bayesian methods, such as Markov chain Monte Carlo (MCMC: Gilks, Richardson, & Spiegelhalter, 1996; MacKay, 2003). We implement all of the graphical models in this chapter using JAGS (Plummer, 2003), which has a simple scripting language for defining graphical models and applies MCMC methods to automate sampling from the joint posterior distribution. The JAGS script implementing the graphical model in Figure 2.12 is shown below. The inputs are the data `y[trial]` giving the observed behavior on each trial, `stimulus[trial]` giving the target stimulus duration on each trial, `standard` giving the duration of the standard stimulus, and `nTrial` giving the total number of trials.

```
# Logistic psychophysical function with
     informative prior
model{
  # Likelihood
  for (trial in 1:nTrials){
    theta[trial] = 1/(1+exp(-(stimulus
[trial]-standard-alpha)/beta))
    y[trial] ~ dbern(theta[trial])
  }
  # Priors
  alpha ~ dnorm(0,1/50^2)
  beta ~ dnorm(0,1/100^2)T(0,)
}
```



$$\alpha \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0,\ 1/100^2)$$
$$\theta_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Figure 2.12 Graphical model representation of the logistic psychophysical model with informative priors, with the behavioral data observed.

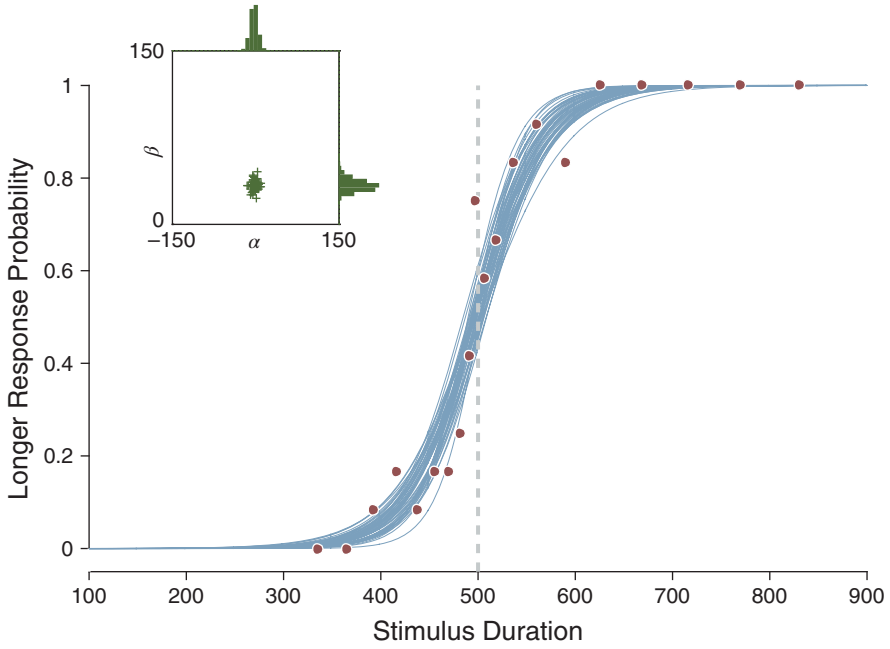


Figure 2.13 The posterior distribution for the logistic psychophysical model with informative priors, based on the visual task data from subject A. The lines show sample psychophysical functions from the posterior, and the circular markers summarize the behavioral response data. The inset panel shows samples from the joint prior distribution and the marginal prior distributions for the shift α and scale β parameters.

The result of using JAGS to apply the model to behavioral data is a set of samples from the joint posterior distribution $p(\alpha, \beta \mid y, M)$. That is, a sequence of `(alpha,beta)` pairs is returned, each pair of which is a draw from the posterior distribution.[1]

The results of this inference for the visual task data for subject A are shown in Figure 2.13. The inset panel summarizes the inferred joint posterior distribution of the parameters α and β by showing a small set of the posterior samples. The histograms summarize the marginal distributions, and are based on all the samples. The main panel summarizes the posterior psychometric function, showing the curves based on the same set of posterior samples used to summarize the joint posterior. The behavioral data on which these posterior inferences are based are presented in the same way as in Figure 2.2, as circular markers showing the proportion of times each unique target stimulus duration was perceived to be longer than the standard.

Posterior distributions represent what is and is not known about parameter values. The interpretation of the joint posterior distribution in Figure 2.13 is that there is a single combination of α and β that is assumed to generate the data, and the probability that any specific combination is this combination

[1]Technical details: The inferences reported are based on three independent chains of 5,000 samples, collected after 1,000 initial burn-in samples were discarded, but without any thinning. Convergence of the chains was checked by visual inspection and by the standard $\hat{R}$ statistic (Brooks & Gelman, 1997), which is a measure of within-to-between chain variance.

is proportional to the density at that point (i.e., proportional to the probability that combination is sampled). In this way, the joint and marginal distributions for α and β represent the uncertainty about their values, quantifying how likely or plausible each possibility can be considered to be, based on the modeling assumptions and the available data. This representation of uncertainty naturally carries forward to any other aspect of the model that depends on the parameters, such as the psychophysical function. Thus, the probability that any specific psychophysical function is the single one assumed to generate the data is proportional to the posterior density of the parameter values that correspond to that function.

Fitting Data Is Not the Same as Predicting Data

The inferences shown in Figure 2.13 correspond to what is often called, in both the Bayesian and non-Bayesian modeling literature, the "goodness of fit" or just "fit" of the model to the data. The word *fit* is often an unhelpful one, at least in the coherent Bayesian context for statistical inference. It can be interpreted as implying that the model is being transformed to match the data, even though nothing of the sort is happening. Once a model—consisting of both a likelihood and prior that together make predictions—is defined and the data are observed, there are no degrees of freedom left for inference. The joint posterior distribution follows logically from probability theory. All that the Bayesian machinery does is calculate or approximate this posterior distribution, so the inferential process is metaphorically more like "reading off" a fixed answer defined by the given model and data than it is "fitting" a model to data. Better terminology might be that what is known is being "updated" from the prior to the posterior, using the additional information provided by the data. This point might seem like semantic nitpicking, but it has some serious carryover consequences. One is the emphasis on model fit over predictive accuracy as a means of model evaluation, evident in observations like "[t]o formally test their theory, mathematical psychologists rely on their model's ability to fit behavioral data" (Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016). A charitable interpretation of this language is that it means the fitness of the model based on data, but that should be achieved using the prior predictive distribution and not through modeling inferences already conditioned on the to-be-predicted data (Roberts & Pashler, 2000, 2002). As Gigerenzer (2016, p. ix) emphasizes, "a model should be evaluated on the basis of its ability to make accurate predictions, not to fit past data."

The prior distributions for the parameters and psychophysical function shown in Figure 2.5 have exactly the same interpretation, except that they are conditional on only the modeling assumptions and represent what is known and unknown without recourse to data. This relationship between the prior and posterior makes clear the conceptual simplicity of Bayesian inference. Inference starts with assumptions about psychological parameters and processes, and inferences are updated as relevant information—typically, but not necessarily, in the form of behavioral data—become available.

Posterior Prediction

This logical progress of analysis as data become available extends to predictions. Figure 2.14 shows a posterior predictive that follows the prior predictive analysis presented in Figure 2.6. To allow these posterior expectations to be compared to the data, the squares corresponding to the actual observed counts are now filled and connected by a line. The analysis in Figure 2.14 suggests that there is reasonable agreement between the posterior predictive distribution and the data, in the sense that the observed counts

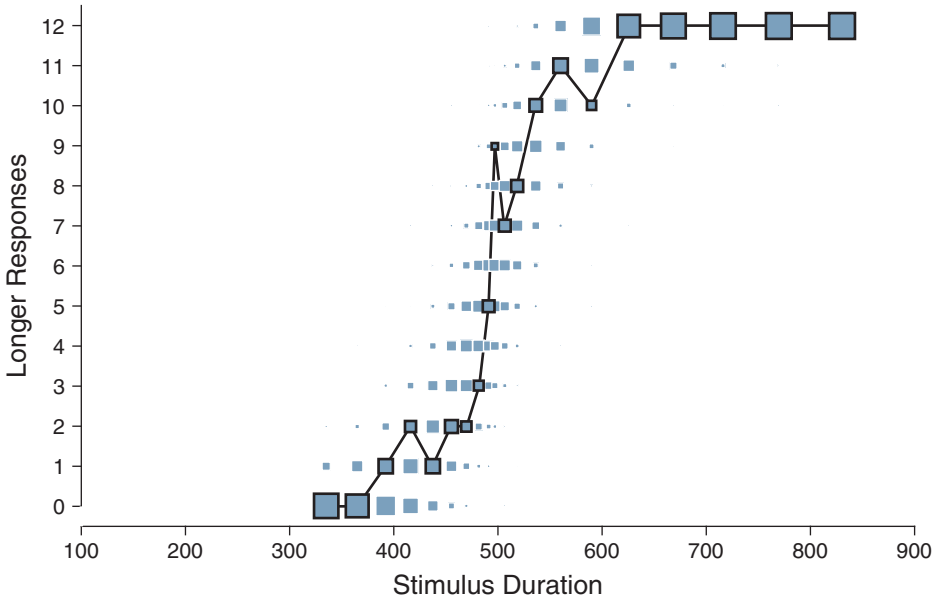


Figure 2.14 The posterior predictive distribution for the logistic psychophysical function with informative priors, applied to the data from subject A in the visual task. The *x*-axis corresponds to the unique target stimulus durations, and the *y*-axis corresponds to counts of the number of times each target duration is perceived as longer than the standard. Squares show the posterior predictive distribution, with the area of a square proportional to its posterior predictive probability. Observed counts are shown by squares with solid borders and connected by a line.

are given significant mass for each target stimulus duration. The representation of the posterior predictive distribution makes clear that there is relative certainty associated with the data expected for the target durations most different from the standard, but much greater uncertainty for those near the standard. In particular, the target durations closest to the standard still give some probability to all possible counts from 0 to 12.

Describing Data Is Not the Same as Predicting Data

The field of cognitive modeling has an unfortunate habit of labeling as "predictions" quantities that are not predictions at all. Often the output of a model for a set of parameters that have been optimized with respect to data is plotted together with the data, and the model values are called "predictions." This is seriously misleading, because predictions logically occur before data are available. In exactly the same unfortunate way, Bayesian posterior predictive distributions do not represent genuine predictions, because they rely on having observed the data. The terminology "posterior predictive distribution" comes from statistics, where "predictive" really means "over the data space." Better terminology might distinguish between prior and posterior in terms of whether or not inferences are conditioned on data, and between parameter and data in terms of whether distributions express uncertainty about latent parameters (or functions of parameters) or data. Instead of thinking of the posterior predictive distribution as being a prediction, it should be conceived as measuring the "descriptive adequacy" of the model. Agreement between observed data and the posterior predictive

distribution assesses whether the model is able to redescribe the data it has observed. Passing a test of descriptive adequacy is not strong evidence in favor of a model, but a major failure in descriptive adequacy can be interpreted as strong evidence against a model (Shiffrin, Lee, Kim, & Wagenmakers, 2008). The genuine predictions in the Bayesian framework are the prior predictive distributions. These are completely determined by the likelihood and the prior, before data are observed. The prior predictive distribution quantifies the relative likelihood of each possible data pattern, according to what the model expects to occur. Prior predictive distributions are rarely presented in the cognitive modeling literature, probably because so little thought goes into priors that they would look more like Figure 2.9 than Figure 2.6, but this should change as the field matures.

Interpreting and Summarizing the Posterior Distribution

Given a model—which includes a likelihood function for generating data, and priors over the parameters that index that function—and data, Bayes' rule immediately defines a joint posterior distribution over the parameters. This joint posterior represents everything that is known and not known, conditional on the model and data. In a sense, that is the end point of a pure fully Bayesian analysis. Often, however, it is convenient or necessary to summarize posterior distributions. By definition, an act of summarizing sacrifices accuracy and completeness for interpretability with respect to specific goals. This is true of summarizing the joint posterior, and it follows that there is no general, correct method for simplifying and reporting posterior inferences.

There are, however, several commonly used approaches that are often effective. Perhaps the most common summary or approximation is to consider only marginal posterior distributions for each parameter, thus losing any information present in the joint distribution that is different from the independent product of the marginals (see Lee & Wagenmakers, 2014, Section 3.6). If this is a reasonable simplification, marginal distributions have the advantage of being easy to display and interpret. Figure 2.15 provides an example showing the marginal posterior distribution of the shift parameter α. It is possible, of course, to summarize this marginal distribution even further. One possibility is to report just the

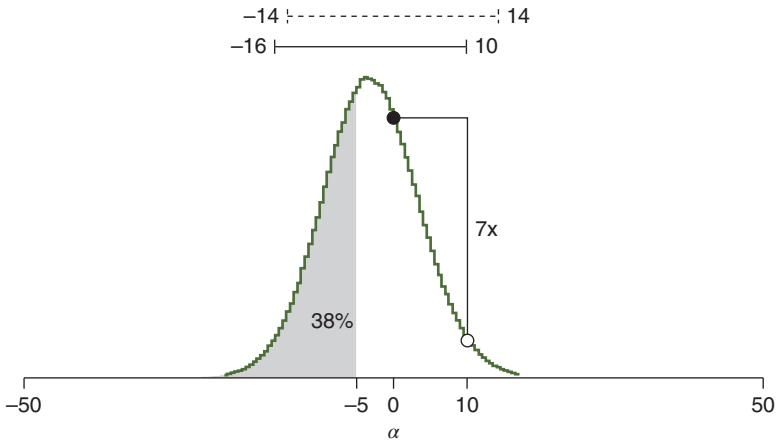


Figure 2.15 The marginal posterior distribution for the shift parameter α, and some example summaries and inferences based on the distribution.

mean, or the mean and the standard deviation. Another is to report a credible interval, which provides a range with some specified proportion of the marginal posterior density. Figure 2.15 shows two different 95% credible intervals. The interval ranging from −14 to +14 is the one that is symmetric about 0 and covers 95% of the posterior density. The interval ranging from −16 to +10 is the one that ranges from the 2.5% percentile to the 97.5% percentile of the distribution. Other credible intervals are obviously possible. An appealing property of credible intervals is that they have the intuitive interpretation—often mistakenly applied to classical confidence intervals (Morey et al., 2015)—of being intervals that are 95% likely to contain the true value of the parameter.

Independent of whether and how a posterior distribution is summarized, it supports a variety of possible inferences. Two of the most important are demonstrated using the marginal posterior distribution for the shift parameter α in Figure 2.15. First, areas under distributions can be interpreted as probabilities, so, for example, the probability that the shift of the psychophysical function is negative—that is, to the left—and more extreme than 5 ms is 0.38. Second, relative densities can be interpreted as likelihood ratios, so, for example, it is about 7 times more likely that the shift is 0 rather than +10.

Model Testing Using Prior and Posterior Distributions

The joint posterior for subject A in the visual task in Figure 2.13 allows for simplifications of the model to be evaluated. For example, the theoretical possibility that the subject is calibrated, in the sense of not having a bias toward "longer" or "shorter" responses, corresponds to the assumption that the subject's behavior is better captured by a model without a shift parameter. Formally, this model can be expressed as special case of the current model, with the restriction that $\alpha = 0$.

The standard Bayesian approach to comparing models is the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2014, Chapter 7). This quantifies the relative evidence that data provide for two models, M_a and M_b, as the ratio

$$BF_{ab} = \frac{p(y \mid M_a)}{p(y \mid M_b)}, \tag{5}$$

which can be conceived as a likelihood ratio, extended the case where one or both models may have parameters. Thus, for example, a Bayes factor of 10 means that the data are 10 times more likely (or provide 10 times more evidence for) M_a than M_b. Whether this level is "significant" is then naturally calibrated by betting, and can be determined in the context of the scientific problem. A number of suggested interpretive scales, with verbal labels for various ranges, exist, although there is a compelling argument that it is better to rely on the inherently meaningful scale itself (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Schönbrodt, 2015).

Applied to the question of whether subject A has a shift from optimality, the Bayes factor compares the full model in Figure 2.4 with $\alpha \sim \text{Gaussian}(0, 1/50^2)$ to the model that restricts $\alpha = 0$. Because the second model is nested within the first—that is, it corresponds to a special case of the first model—it is possible to estimate the Bayes factor using what is known as the Savage-Dickey method (Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Grasman, & Wagenmakers, 2010). This method uses the statistical fact that the Bayes factor is the ratio of the prior and posterior at the point in the parameter space that reduces the full model to the nested model.

The left panel of Figure 2.16 shows how the Savage-Dickey method estimates Bayes factors for testing whether subject A has a

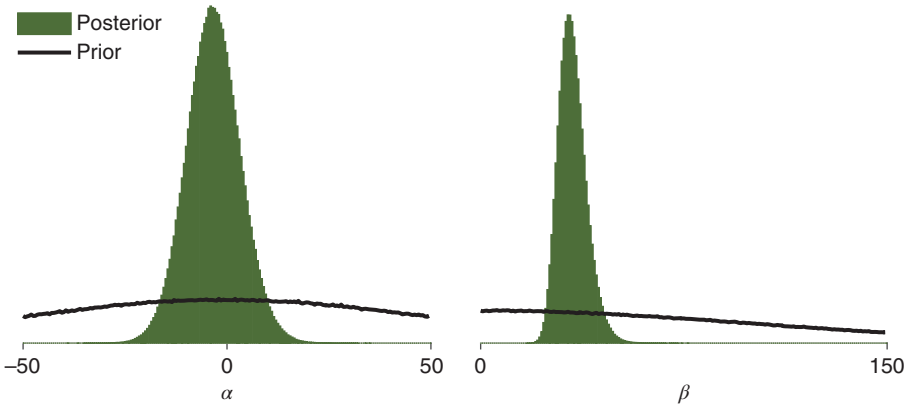


Figure 2.16 The Savage-Dickey method for estimating Bayes factors, applied to the marginal distributions for the shift α (*left*) and scale β (*right*) for subject A in the visual task. The marginal prior distribution is shown by a solid line. The marginal posterior distribution is shown by a histogram. The ratio of the posterior to prior density at the critical points $\alpha = 0$ and $\beta = 0$ approximate the Bayes factors comparing the general model to the nested ones that assume perfect calibration with no shift and perfect discriminability with a step psychophysical function, respectively.

shift. The prior and posterior are shown, and the posterior is about 7 times higher than the prior at the critical point $\alpha = 0$. This means that the Bayes factor in favor of the nested model that says the subject has no shift is about 7. The right panel of Figure 2.16 shows another application of the Savage-Dickey method. This analysis involves the scale parameter and the critical value $\beta = 0$ that corresponds to an optimal step function relating stimulus duration to response probability. The posterior of the scale parameter has almost no density near zero, corresponding to a very large Bayes factor in favor of the original more general model.

Although both of the examples in Figure 2.16 involve a nested model corresponding to setting a single parameter to zero, the Savage-Dickey method is more general. It applies to critical values other than zero, such as, for example, testing whether data provide evidence for a specific response probability of $\frac{1}{2}$. It also applies in the case of nested models that require more than one parameter to take specific values, and to the case of interval nested models, in which the bounds on a parameter or parameters for one model fall within those of another model (Hoijtink, Klugkist, & Boelen, 2008).

Model Selection Inferences Based on Parameter Posteriors Is Perilous

One idea is that statistical inference in psychology should rely heavily (or exclusively) on expressions of uncertainty about parameters, whether classical confidence intervals (Cumming, 2013) or Bayesian credible intervals (Kruschke, 2013). This reliance usually comes at the explicit or implicit expense of hypothesis tests or model selection measures like the Bayes factor. The alternative view is that both model selection and parameter estimation have important complementary roles in analyzing data and models in cognitive psychology (Morey, Rouder, Verhagen, & Wagenmakers, 2014). Choosing a useful model logically precedes making inferences based on that model, which makes it conceptually clear that model selection is important. Basing model selection decisions on ad hoc procedures that rely on posterior distributions is incoherent,

and can lead quickly to arbitrary and unjustified conclusions. A simple example, presented by Wagenmakers, Lee, Rouder, and Morey (2017), makes the point. Suppose the two models being evaluated are that a coin always produces heads or always produces tails (but it is not known which), or that a coin is fair. A single datum (e.g., a head) is completely uninformative with respect to these models, and the Bayes factor correspondingly is 1. But the datum will affect the posterior distribution of the rate at which the coin produces heads or tails, for standard choices like a uniform prior. Inferences about the models based on the posterior distribution will thus, illogically, be impacted by the datum. Given this insight, it does not make sense to infer from the right panel of Figure 2.16 that $\beta \neq 0$, because a 95% credible interval summary of the posterior does not include 0. Intuitively, the posterior is already conditioned on an assumption about the appropriate model and does not allow that assumption to be revisited. Technically, the prior for β on which the posterior depends is not the one needed to choose between models with $\beta = 0$ and $\beta \neq 0$. The appropriate prior would be a so-called spike-and-slab prior, with the density at $\beta = 0$ corresponding to the spike, consistent with its prior possibility as a nested model (Mitchell & Beauchamp, 1988). Even if this prior were used—it could be constructed in practice, for example, by using a latent-mixture model—the full posterior distribution needs to be considered for inference. As Figure 2.15 indicates, there are many possible 95% credible intervals that summarize the posterior, some of which will include any given point or region of interest, and some of which will not. There is no principled basis for choosing which summary to prefer in the context of model selection. The bottom line is that it is possible to use posterior distributions to choose between models only when—as in the case of the Savage-Dickey procedure, or a latent-mixture spike-and-slab prior—the analysis is formally equivalent to a model selection method justified by probability theory, such as the Bayes factor (Morey & Rouder, 2011; Rouder, Haaf, & Vandekerckhove, 2017; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016).

Sensitivity Analysis

Constructing models of human cognition is a creative scientific act, and it is rare that guiding theory is strong, complete, or precise enough to determine all aspects of a model. This means that some parts of most models are based on reasonable assumptions, required to create formal models, but not corresponding to strong theoretical commitments. A natural consequence of this state of affairs is that it is good practice to conduct sensitivity analyses, in which noncore assumptions of the model are varied to other plausible choices, and the results of these changes on the important modeling inferences and conclusions are assessed.

One important target for sensitivity analysis is priors on parameters that are not completely specified by theory. The priors on both the scale and shift parameters in our model are of this type. They were both acknowledged to have some level of arbitrariness in their development. Accordingly, Figure 2.17 summarizes the results of a basic sensitivity analysis. The left-hand panel shows samples from the joint posterior distribution of the original model, with the priors $\alpha \sim \text{Gaussian}(0, 1/50^2)$ and $\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$. The middle panel shows samples from the joint posterior in a modified model with a prior on the shift parameter of $\alpha \sim \text{Gaussian}(0, 1/100^2)$, but with the original prior on the scale parameter left in place. The right-hand panel shows samples from the joint posterior for a model with scale parameter prior $\beta \sim \text{TruncatedGaussian}_+(0, 1/200^2)$, but with the original prior on the shift parameter left in place. All three models result in extremely similar joint posterior distributions, suggesting that inferences about the parameters are not sensitive to the exact form of the priors.

Of course, these two modifications represent a very limited sensitivity analysis, but

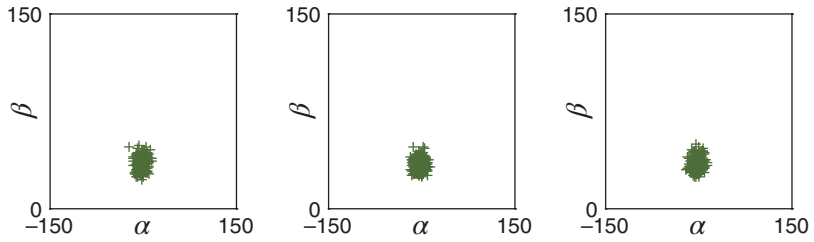


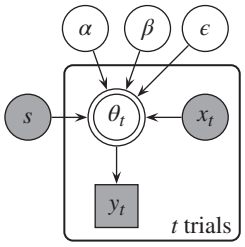
Figure 2.17 Samples from the joint posterior of the original psychophysical model (*left*), from a modified model with shift parameter prior $\alpha \sim \text{Gaussian}(0, 1/100^2)$ (*middle*), and from a modified model with scale parameter prior $\beta \sim \text{TruncatedGaussian}_{+}(0, 1/200^2)$ (*right*).

demonstrate the basic approach. If it is not possible to specify the prior distribution for a model exactly, different reasonable priors should be considered. In the comforting case where the important inferences made using the model—posterior distributions, posterior predictive distributions, Bayes factors, or whatever is important for the problem at hand—do not change enough to affect the substantive conclusions as the priors are varied, those conclusions can be considered robust to the vagaries of model specification. If the inferences are sensitive to the priors, Bayesian methods are highlighting an important deficiency in theory, or the limitations of the available data, or both. The message is that the results depend on aspects of the model that are not well enough understood, and developing a better understanding should become a priority.

Logically, sensitivity analyses are just as important for the likelihood component of a model as they are for the prior (Vanpaemel, 2016). It is rare that the likelihood of a cognitive model does not make simplifying or somewhat arbitrary assumptions, and the sensitivity of inferences to these assumptions should be examined. The current psychophysical model, for example, assumes a complete lack of sequential effects. The probability of a longer response depends only on the target stimulus for the current trial, with no effect from immediately preceding trials. This may or may not be a good assumption. It seems possible that subjects are able to treat each trial in isolation, but there is also substantial evidence for sequential effects in basic psychophysical tasks like absolute identification (e.g., Lockhead, 2004).

A modified model that changes the likelihood to allow for one type of sequential effect is defined by the graphical model in Figure 2.18. The basic idea is that the response the subject made on the previous trial affects the response probability on the current trial. This is formalized by a change ϵ in the response probability in one direction if the previous response was "longer," and in the other direction if the previous response was "shorter." A prior $\epsilon \sim \text{Gaussian}(0, 1/0.1^2)$ is placed on the change, so that it is assumed to be small, but it is possible that a "longer" response either increases or decreases the probability of another "longer" response on the next trial.

Figure 2.19 summarizes the inferences from applying the sequential model to the visual task data from subject A. The joint posterior distribution over the scale and shift parameters and the posterior distribution this implies over the psychophysical functions are shown. These are very similar to those inferred by the original model that did not allow for sequential dependencies.



$$\alpha \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0,\ 1/100^2)$$
$$\epsilon \sim \text{Gaussian}(0,\ 100)$$
$$\theta_t = \begin{cases} 1/\left(1+\exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right) + \epsilon & \text{if } y_{t-1} = 0, t > 1 \\ 1/\left(1+\exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right) - \epsilon & \text{if } y_{t-1} = 1, t > 1 \end{cases}$$
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Figure 2.18 Graphical model representation of the logistic psychophysical model with informative priors and sequential effects between trials.

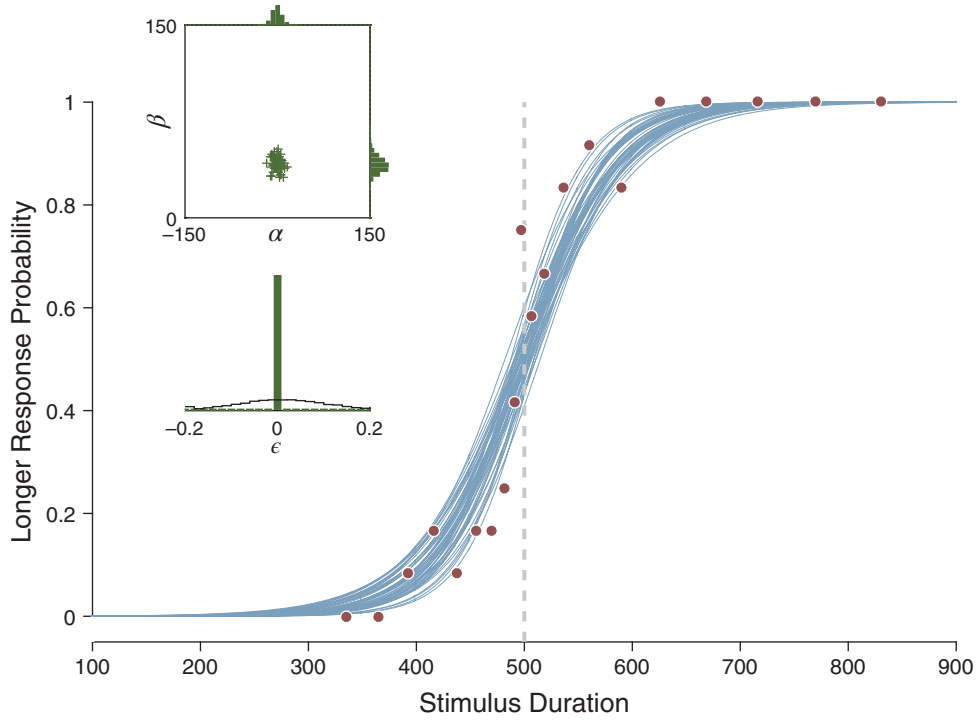


Figure 2.19 The posterior distribution for the logistic psychophysical model allowing for sequential effects, based on the visual task data from subject A. The lines show sample psychophysical functions from the posterior, and the circular markers summarize the behavioral response data. The upper inset panel shows the joint and marginal posterior distributions for the shift α and scale β parameters. The lower inset panel shows the prior and posterior marginal distributions for the sequential effect ϵ parameter.

The lower inset panel in Figure 2.19 shows the prior and posterior distribution over the ϵ parameter. This is inferred to be near zero, consistent with the subject not having the sorts of sequential effects assumed by the model. A Savage-Dickey analysis of the prior and posterior results of a Bayes factor of about 10 against the presence of sequential effects, since when $\epsilon = 0$ the sequential model reduces to the original model.

As was the case for the sensitivity analysis of priors, the consideration of the sequential model is just one of many that could be considered. There are obviously other possible ways a sequential dependency might affect behavior. Different assumptions could be made about the nature of the ϵ effect parameter, or more than just the previous trial could be assumed to affect responding. More generally, there are other possible overly simple assumptions in the likelihood of the original model, such as the assumption that the scale and shift are fixed to the same values on every trial. It is straightforward to specify and apply alternative models that test the sensitivity of inferences to these assumptions.

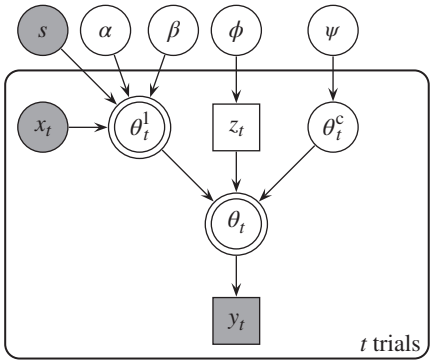
The sensitivity analyses presented here make it clear that there is a close relationship between sensitivity analysis and model comparison. Each time a modified prior or likelihood is examined, inference is being done with a different model. The Bayes factor comparing the model with sequential dependencies to the original model could validly be interpreted as the result of a model selection exercise. Conceptually, the difference is that a sensitivity analysis considers a set of models that are all consistent with a single theory of the cognitive phenomenon being considered. The different models are designed to cover a space of plausible models consistent with the theory, necessitated by the theory not being complete enough to formalize a single model. Model comparison, in contrast, intends to evaluate different models that correspond to competing theoretical assumptions, using the available data to evaluate the merits of each.

Latent-Mixture Modeling

The assumption that a single model generates all of the behavior in a task, even for a single subject, is a very strong one, and often seems implausible. One of the most obvious general exceptions in cognitive modeling comes in the form of contaminant trials. These are trials in which a subject uses some cognitive process other than the one of interest to generate their behavior. While understanding people's behavior on these trials is usually not an important goal in itself, contaminant trials can corrupt inferences about the behavior that is the focus. With Bayesian methods, as for classical methods, it is possible for a single contaminant trial to change inferences about a parameter drastically. This impact is not a shortcoming of the inference method itself, but rather a consequence of the overly strong assumption that the model of interest generated the observed behavior on every trial.

One way to overcome this problem is by using latent-mixture models (Zeigenfuse & Lee, 2010). The key idea is that the basic model is extended to include a separate model of the contaminant process, and it is now assumed that behavior on each trial is generated by one or the other of these models. The presence of these two possibilities makes the model a mixture model. The fact that it is not known which trial belongs to which possibility makes the model a latent-mixture model.

Figure 2.20 shows a latent-mixture extension of our model to include a very simple contaminant model. The "longer" response probability θ_t for the trial t trial can now be generated in one of two ways. It may come from a psychophysical function, formalized exactly as in the original model, which provides the probability θ_t^l. Alternatively, whether the response is "longer" or "shorter" can be decided by a contaminant process, which chooses "longer" according to a base rate. This base rate is assumed to be unknown and is represented by the parameter $\psi \sim \text{Uniform}(0, 1)$, so that $\theta_t^c \sim \text{Bernoulli}(\psi)$ is the potential contaminant response. Which of these alternatives is used depends on a



$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$
$$\theta_t^1 = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$
$$\psi \sim \text{Uniform}(0,1)$$
$$\theta_t^c \sim \text{Bernoulli}(\psi)$$
$$\phi \sim \text{Uniform}(0,1)$$
$$z_t \sim \text{Bernoulli}(\phi)$$
$$\theta_t = \begin{cases} \theta_t^1 & \text{if } z_t = 0 \\ \theta_t^c & \text{if } z_t = 1 \end{cases}$$
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Figure 2.20 Graphical model representation of the logistic psychophysical model with a trial-level contaminant process, implemented as a latent-mixture model.

binary indicator variable z_t, which follows a contamination base rate represented by the parameter $\phi \sim \text{Uniform}(0, 1)$. Thus, the model infers for each trial whether the response is generated by the psychophysical model or by a general contaminant process. The model simultaneously infers the base rate or proportion of contaminant trials and the bias in those contaminant trials toward "longer" or "shorter" responses.

We applied the contaminant model to the data from subject F in the visual task. The results are summarized in Figure 2.21. The joint and marginal posterior distributions of the α shift and β scale parameters, and the posterior distribution over the psychophysical function they imply, are shown as before. The lower-left inset panel shows the joint and marginal posterior distributions over the ϕ and ψ base-rate parameters associated with the contaminant process. The marginal distribution of ϕ shows that the inferred probability of any individual trial being a contaminant trial is very low. The marginal distribution of ψ shows much greater uncertainty about whether there is a bias toward contaminant trials being "longer" or "shorter" responses. This makes sense, since the lack of contaminant trials means the data provide little information about their nature.

The lower-right inset panel shows the expected marginal posterior probability for each z_t indicator variable, corresponding to the probability that each trial is a contaminant trial. On two trials in close proximity a little more than halfway through the experiment—trials 138 and 146—the posterior probability of contamination is greater than 0.5, and is highlighted. These trials both presented a target stimulus of 830 ms, much longer than the 500 ms standard, but the subject's response was that they were both perceived as shorter. It seems intuitively reasonable that these trials are inferred to be contaminant trials.

Importantly, the joint posterior for the shift and scale parameters is sensitive to the identification of the contaminant trials. To the extent that a trial is inferred to be a contaminant trial, the behavioral data for that trial do not influence the inference of α and β. The impact of this property is shown in the upper-left inset panel showing the joint posterior distribution of the shift and scale parameters. The lighter joint samples and

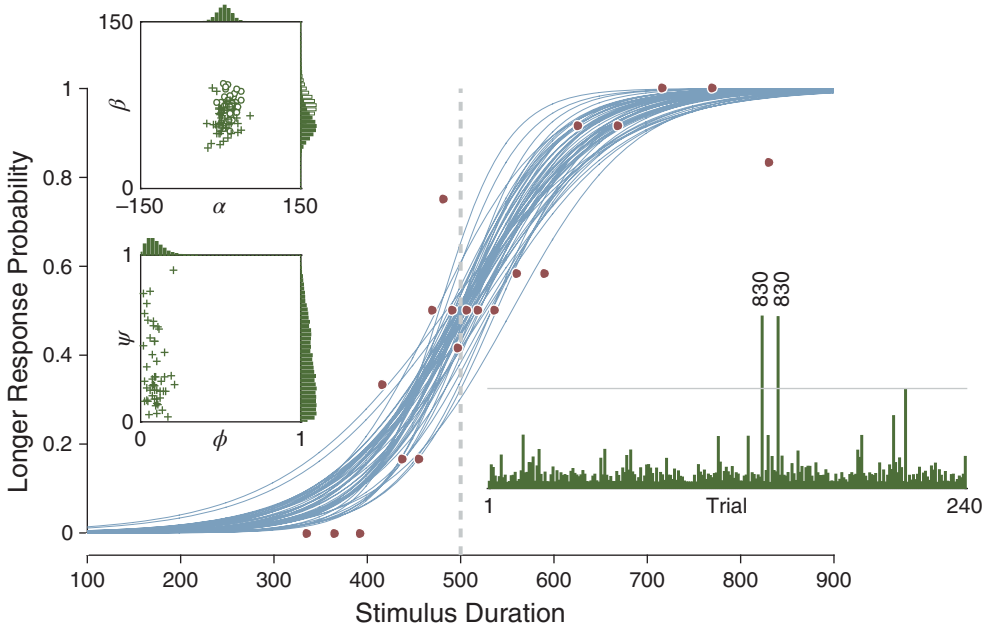
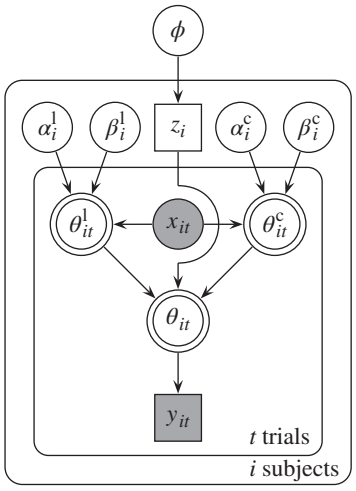


Figure 2.21 The posterior distribution for the logistic psychophysical model allowing for sequential effects, based on the visual task data for subject F. The lines show sample psychophysical functions from the posterior, and the circular markers summarize the behavioral response data. The upper-left inset panel shows the joint and marginal posterior distributions for the shift α and scale β parameters. The cross markers and filled histograms show these distributions for the model with the contaminant process included, while the circular markers and unfilled histograms show the distribution inferred when the contaminant process is not included in the model. The lower-left inset panel shows the joint and marginal distributions for the ϕ and ψ base-rate parameters of the contaminant process. The lower-right inset panel shows the posterior probability of contaminant for each of the 240 trials as a histogram, with a 50% probability cutoff shown by the solid line. Two trials inferred as most likely to be contaminant, both for target stimuli with 830 ms duration, are highlighted.

marginal posterior for the β scale parameter correspond to the inferences made for the same behavioral data using the basic model in Figure 2.12. It is clear that the inferences about the scale are different when the contaminant process is included in the latent-mixture model. In particular, a steeper psychophysical function is inferred when contaminant trials—especially the two at 830 ms emphasized in Figure 2.21—are not assumed to be generated by the same underlying psychophysical function.

A more general application of latent-mixture modeling is to consider two or more cognitive models that are of theoretical interest. The same basic idea holds of allowing different models to contribute to explaining the behavioral data, but none of the models needs to be considered as a contaminant model. Figure 2.22 shows a graphical model that implements this approach for the logistic and Cauchy models. The model considers multiple subjects and assumes that each subject's behavioral data is generated by either the logistic or the Cauchy model. That is, each subject is assumed to use one model or the other, and the same subject is assumed to use the same model for all of the trials. The z_i



$$\alpha_i^1 \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta_i^1 \sim \text{TruncatedGaussian}_+(0,\ 1/100^2)$$
$$\theta_{it}^1 = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha_i^1}{\beta_i^1}\right)\right)$$
$$\alpha_i^c \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta_i^c \sim \text{TruncatedGaussian}_+(0,\ 1/100^2)$$
$$\theta_{it}^c = \arctan\left(\frac{x_t - s - \alpha_i^c}{\beta_i^c}\right)/\pi + 0.5$$
$$\phi \sim \text{Uniform}(0,\ 1)$$
$$z_i \sim \text{Bernoulli}(\phi)$$
$$\theta_{it} = \begin{cases} \theta_{it}^1 & \text{if } z_i = 0 \\ \theta_{it}^c & \text{if } z_i = 1 \end{cases}$$
$$y_{it} \sim \text{Bernoulli}(\theta_{it})$$

Figure 2.22 Graphical model representation of the latent-mixture model that assumes each individual subject uses a model based on either a logistic or Cauchy psychophysical function. The latent indicator z_i determines which model is used by the ith subject, and the base-rate ϕ determines the proportion of subjects using the logistic model.

indicator parameter indexes which model the ith subject uses, controlling whether the response θ_{it} on the tth trial follows that predicted by the logistic or Cauchy model. The model also infers a latent base-rate ϕ of the proportion of subjects using the logistic model.

The results of applying this latent-mixture model to all six subjects for the visual task are summarized in Figure 2.23. The left panel shows the marginal posterior distribution for the z_i indicator parameter for each subject, quantifying the probability of the logistic rather than the Cauchy model. It is clear the logistic model is more likely for all of the subjects, ranging from about 65% for subject B to 85% for subject C. The right panel of Figure 2.23 shows the posterior for the ϕ base-rate parameter. There remains a large degree of uncertainty about the base rate of logistic model subjects, although the data provide evidence this base rate may be higher rather than lower.

Parameter Estimation as Model Selection

The inferences made for the z_i parameters can naturally be interpreted as a type of model selection, choosing whether the logistic or Cauchy model is a better account for each subject. In fact, there is a close relationship between the posterior expectation of the indicator parameter and the Bayes factor that evaluates the logistic and Cauchy models. This relationship hinges on an alternative conception of the Bayes factor from that presented in Equation (5), as the factor that uses the data to update prior odds to posterior odds for two models:

$$\overbrace{\frac{p(M_a \mid y)}{p(M_b \mid y)}}^{\text{Posterior odds}} = \overbrace{\frac{p(y \mid M_a)}{p(y \mid M_b)}}^{\text{Bayes factor}} \times \overbrace{\frac{p(M_a)}{p(M_b)}}^{\text{Prior odds}}. \quad (6)$$

In a simple latent-mixture model without the base-rate parameter ϕ, the prior placed directly on a z indicator parameter corresponds to setting prior odds. The posterior expectation of z_i then estimates the posterior

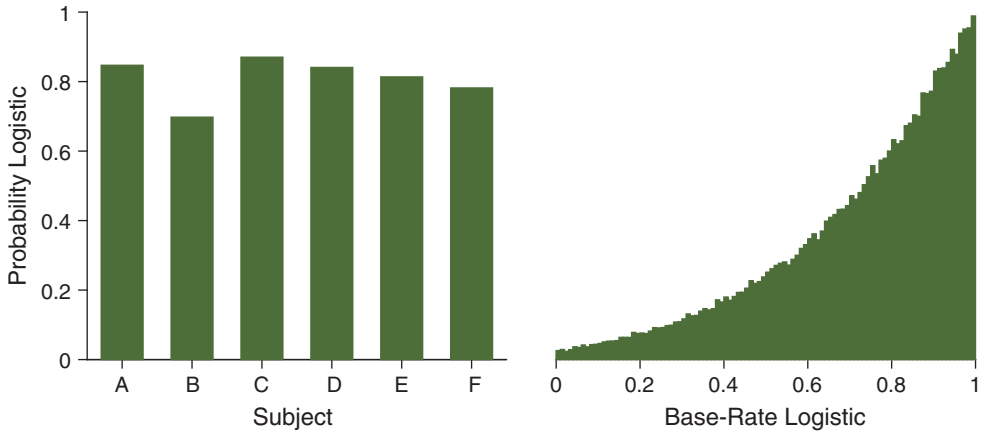


Figure 2.23 Results for the latent-mixture model allowing individual differences between a logistic and a Cauchy model at the level of individual subjects. The left panel shows the inferred posterior probability that each of the six subjects A to F uses the model based on the logistic rather than Cauchy psychophysical function. The right panel shows the inferred base-rate ϕ with which the logistic model is used across a population of subjects.

odds. Thus, the ratio between the inferred posterior and the known prior provides an estimate of the Bayes factor. This is not quite the case for the graphical model in Figure 2.22, because the dependence of each z_i indicator parameter on a common base rate complicates things, but there are applications in the cognitive modeling literature where latent mixture modeling effectively does provide Bayes factors between models. A good example is provided by Lee (2016), who uses a simple latent-mixture model to evaluate five decision-making models at the individual-subject level. More generally, the product-space method for estimating Bayes factors is based exactly on the idea of inferring a single latent indicator parameter that controls which of two or more alternative models generates observed data (Lodewyckx et al., 2011).

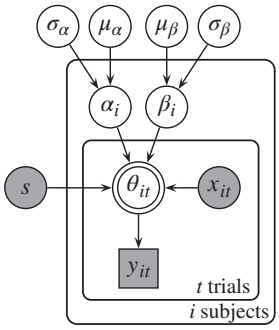
Hierarchical Modeling

One of the most powerful modeling possibilities that Bayesian methods allow for cognitive psychology involves hierarchical models. The term *hierarchical* is widely and imprecisely used (Lee, 2011), but intuitively refers to the situation in which some key psychological variables in a model are themselves modeled as the outcomes of other cognitive processes and parameters. An important application of hierarchical models is to individual differences (Pratte & Rouder, 2011; Rouder et al., 2009), allowing for more realistic assumptions than all subjects being identical or all subjects being completely independent of each other (Shiffrin et al., 2008).

Figure 2.24 shows a graphical model that uses a simple approach to modeling individual differences in the shift and scale parameters of the logistic psychophysical model. Subjects are assumed to have their own shift and scale parameters, but the parameters are drawn from an overarching group distribution. For the *i*th subject, this means that

$$\begin{aligned} \alpha_i &\sim \text{Gaussian}(\mu_\alpha, 1/\sigma_\alpha^2) \\ \beta_i &\sim \text{TruncatedGaussian}_+(\mu_\beta, 1/\sigma_\beta^2) \end{aligned} \quad (7)$$

where μ_α and μ_β are group-level mean parameters, and σ_α and σ_β are group-level standard



$$\mu_\alpha \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\sigma_\alpha \sim \text{Uniform}(0,\ 50)$$
$$\mu_\beta \sim \text{TruncatedGaussian}_+(0,\ 1/100^2)$$
$$\sigma_\beta \sim \text{Uniform}(0, 100)$$
$$\alpha_i \sim \text{Gaussian}(\mu_\alpha, 1/\sigma_\alpha^2)$$
$$\beta_i \sim \text{TruncatedGaussian}_+(\mu_\beta, 1/\sigma_\beta^2)$$
$$\theta_{it} = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha_i}{\beta_i}\right)\right)$$
$$y_{it} \sim \text{Bernoulli}(\theta_{it})$$

Figure 2.24 Graphical model representation of a hierarchical model, using a logistic psychophysical function, in which the shift α_i and scale β_i parameters for the ith subject are independently drawn, respectively, from overarching Gaussian and truncated Gaussian group-level distributions.

deviation parameters that correspond to the heterogeneity across subjects for the shift and scale, respectively.

Figure 2.25 shows the posterior psychophysical functions for all six subjects inferred by applying the model to their visual task data. The ability of the model to capture individual differences is clear. For example, the inferred psychophysical functions for subjects A and D are very different, reflecting the very different judgments these subjects made about target stimuli. These inferences are very similar to what is achieved by assuming the subjects are independent by, for example, applying the graphical model in Figure 2.12 independently to each subject's data.

Figure 2.26 highlights the additional inferences made possible by the hierarchical approach. The top-left panel shows the joint posterior distribution for the group-level mean shift and scale parameters, μ_α and μ_β. This joint distribution corresponds to the inferences made about the average shift and scale for subjects in the visual task. Although not shown, the joint posterior distribution also includes the standard deviation parameters, so it is also possible to make inferences about the variability of individual differences in both shift and scale.

The top-right panel of Figure 2.26 shows individual-level inferences about the shift and scale parameters. For each of the six subjects, the 95% credible intervals, based on 2.5% and 97.5% percentile bounds, are shown. The individual similarities and differences between subjects are clear with, for example, subjects A and B being similar to each other, but different from subjects D and F. The predicted credible intervals for a new subject, labeled "N," are also shown. This subject can be conceived as the next subject to take the experiment, with a prediction based on the model and the information provided by the data from the six observed subjects. The ability to make this prediction stems directly from the hierarchical nature of the model. The joint posterior distribution of the shift and scale parameters for the new subject is found by averaging over all the possible group distributions, weighted by their posterior densities, as inferred from the observed subjects. The prediction for the new subject spans the range of shift and scale parameterizations inferred for the observed subjects. Intuitively, the prediction for the new subject captures the commonalities in the observed subjects, but maintains uncertainty consistent with the differences between them.

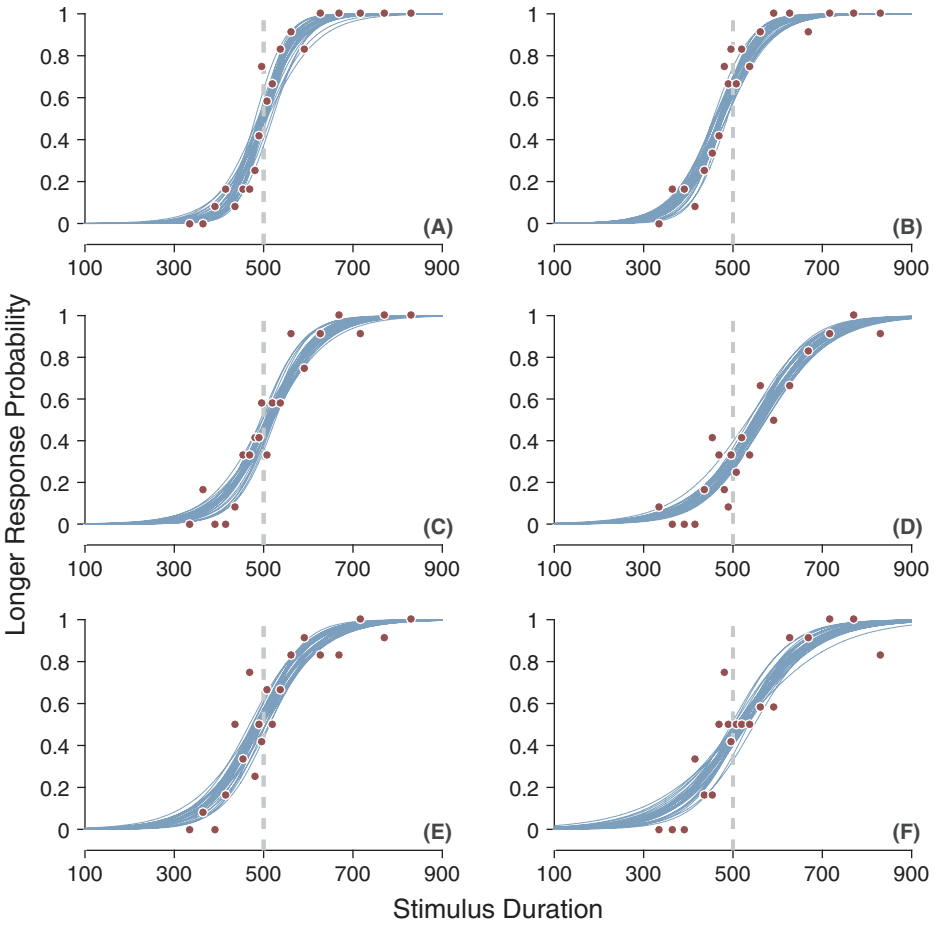


Figure 2.25 The posterior distribution for the visual task data for each of the six subjects A to F, inferred by the hierarchical logistic psychophysical model. In each panel, the lines show sample psychophysical functions from the posterior, and the circular markers summarize the behavioral response data.

The lower-left panel of Figure 2.26 shows the predicted psychophysical function for the new subject. This is simply a reexpression of the predicted joint distribution over the shift and scale parameters. It can usefully be compared to the prior distribution of the psychophysical function for the original model, shown in Figure 2.5. The difference between the two distributions corresponds to what has been learned from the six observed subjects. That is, the additional certainty and specificity of prediction in the lower-left panel of Figure 2.26 arises from the additional information provided by the behavioral data of the observed subjects.

Posterior and prior distributions for other functions of the model parameters can be found in the same way. As an example, the lower-right panel of Figure 2.26 shows the posterior distributions of a standard just noticeable difference (JND) measure. This is the difference between a target stimulus duration and standard at which a subject just notices a difference in the duration between the two stimuli. One common approach uses a response probability of 84% as a critical

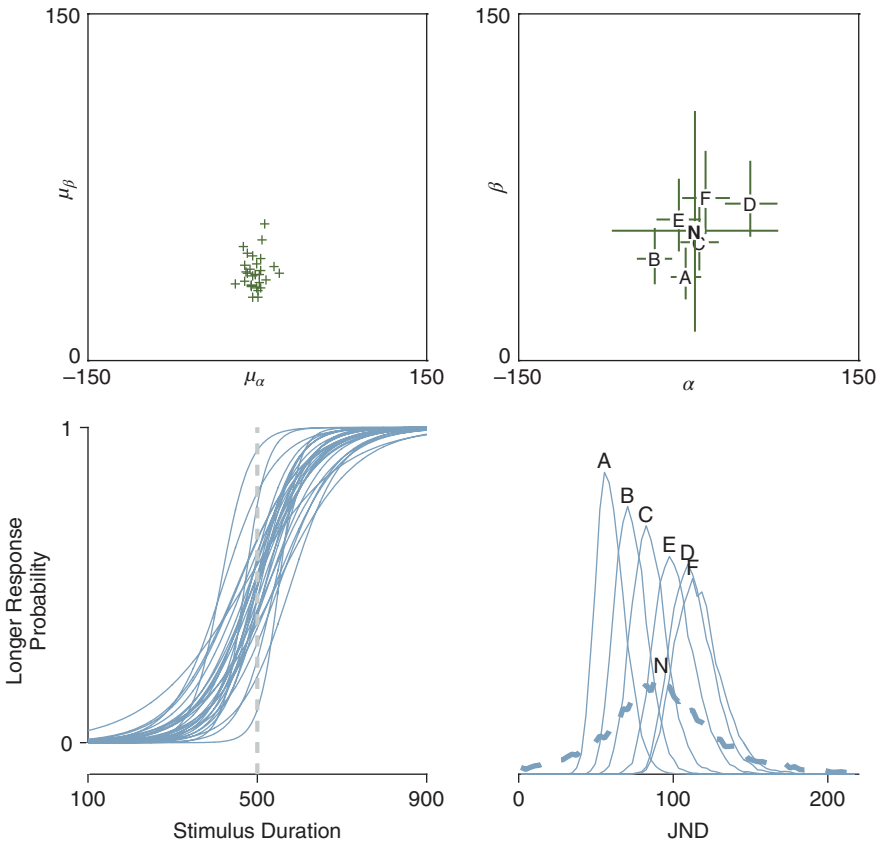


Figure 2.26 Group-level and new-subject inferences made by applying the hierarchical logistic model to the visual data data from all six subjects. The top-left panel shows the joint posterior distribution over the group mean parameters for shift μ_α and scale μ_β. The top-right panel shows the expected value and 95% credible interval for subject-level shift α and scale β, for the six observed subjects A to F, and a new subject N. The bottom-left panel shows the predicted distribution of the psychophysical function for the new subject. The bottom-right panel shows the posterior distribution of a just noticeable difference (JND) measures for the observed subjects, and a predicted JND distribution for a new subject.

level, and so defines the JND to be the difference in duration that makes response performance rise from 50% to 84% (Ernst, 2005). The lower-right panel of Figure 2.26 shows the posterior distributions for JND defined this way for the six subjects, as well as the new predicted subject. Once again, the inferences are interpretable for the observed subjects, and the predictions for the new subject are sensitive to the range of possibilities evidenced in the observed subjects.

Cognitive and Statistical Modeling of Individual Differences

The hierarchical approach of assuming that subject-level parameters come from simple group-level distributions, such as beta or Gaussian distributions, has been widely and profitably used in cognitive modeling (e.g., Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Rouder & Lu, 2005). Statistically, this hierarchical approach is consistent with random-effects models that

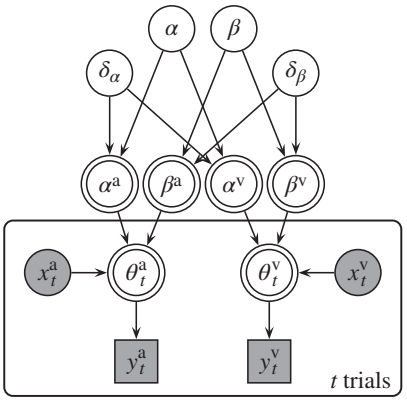
are well established and widely used in both classical and Bayesian settings. Besides theoretically allowing for a form of individual differences, it has a number of attractive statistical properties—including those involving the pooling of subject-level data, sometimes called "sharing statistical strength" or "shrinkage"—that make it especially useful for experimental designs in which many subjects each contribute relatively few data. It is worth pointing out, however, that letting subject parameters come from a Gaussian distribution falls well short of all that can or should be done to model individual differences in cognition. One simple extension is to assume that subject-level parameters come from hierarchical latent-mixture distributions, with the latent-mixture component capturing large qualitative individual differences, and the hierarchical component continuing to capture more minor variation within these subgroups (Bartlema, Lee, Wetzels, & Vanpaemel, 2014). More fundamentally, however, all of these approaches to modeling individual differences are largely statistical and lack guiding psychological theory. The goal should be to model the relationships between groups and individuals in the same way the relationships between individuals and their behavior are currently modeled, through the development of theory that creates psychologically meaningful variables and processes. One recent important step in this direction has been the development of cognitive latent variable modeling, which couples cognitive models with standard psychometric factor theories within a hierarchical Bayesian framework (Vandekerckhove, 2014).

Finding Invariances

The within-subjects design of the current experiment, in which each subject does both an auditory and a visual task, naturally raises the question as to whether there are between-modality differences in the perception of duration and response behavior. Finding invariances, such as the same psychophysical model accounting for both modalities, is one of the most important goals in empirical science. The modeling of invariances often identifies regularities, guiding principles, or laws that lay the groundwork for theoretical development. The compilation of 20 "great scientific experiments" presented by Harré (2002) reflects the importance of invariants. Of the 18 relevant experiments—two are included because of the apparatus, rather than the scientific discovery—more than half are clearly focused on invariances, including experiments under the headings "finding the form of a law," "null results," and "the demonstration of underlying unity within apparent variety." From this perspective, the focus of some areas of psychology with finding differences, especially when the emphasis is on demonstrating surprising effects that are unlikely to be true, seems counterproductive.

One way to formalize the research question of whether the auditory and visual tasks are guided by the same psychological process is to test whether the same psychophysical function at the same parameterization can account for a subject's behavior in both tasks. This "same" model needs to be compared to one or more alternative "different" models. In this example, we choose a single simple different model that assumes separate parameterizations are needed for each task, but the same basic psychophysical function applies.

Figure 2.27 shows a graphical model that allows for the Bayes factor between the "same" and "different" models to be estimated. The condition-specific shift and scale parameters are defined in terms of overall shift and scale parameters α and β, using difference parameters δ_α and δ_β. For the shift parameters, it is assumed that $\alpha^{\mathrm{a}} = \alpha + \frac{1}{2}\delta_\alpha$ and $\alpha^{\mathrm{v}} = \alpha - \frac{1}{2}\delta_\alpha$, and an analogous assumption is made for the scale parameters. Thus, the differences in the condition-specific parameters are δ_α and δ_β, which are given priors corresponding to assumptions about the sorts of changes across tasks that seem



$$
\begin{aligned}
\alpha &\sim \text{Gaussian}(0, 1/50^2) \\
\beta &\sim \text{TruncatedGaussian}_{+}(0, 1/100^2) \\
\delta_\alpha &\sim \text{Gaussian}(0, 1/20^2) \\
\delta_\beta &\sim \text{Gaussian}(0, 1/40^2) \\
\alpha^{\text{a}} &= \alpha + \frac{1}{2}\delta_\alpha \\
\alpha^{\text{v}} &= \alpha - \frac{1}{2}\delta_\alpha \\
\beta^{\text{a}} &= \beta + \frac{1}{2}\delta_\beta \\
\beta^{\text{v}} &= \beta - \frac{1}{2}\delta_\beta \\
\theta_t^{\text{a}} &= 1/\left(1 + \exp\left(-\frac{x_t^{\text{a}} - s - \alpha^{\text{a}}}{\beta^{\text{a}}}\right)\right) \\
\theta_t^{\text{v}} &= 1/\left(1 + \exp\left(-\frac{x_t^{\text{v}} - s - \alpha^{\text{v}}}{\beta^{\text{v}}}\right)\right) \\
y_t^{\text{a}} &\sim \text{Bernoulli}(\theta_t^{\text{a}}) \\
y_t^{\text{v}} &\sim \text{Bernoulli}(\theta_t^{\text{v}})
\end{aligned}
$$

Figure 2.27 A graphical model for comparing a model that allows for different shift and scale parameterizations of a logistic psychophysical function to account separately for a subject's auditory and visual task data against a model that assumes the same parameterization accounts for both sets of behavior. The difference parameters δ_α and δ_β quantify the differences in the shift and scale, respectively, between the auditory and visual tasks.

plausible. In this example, they are given zero-centered Gaussian priors with standard deviations of 20 and 40, respectively, based on a visual analysis of the impact of these magnitudes of effects on the psychophysical function, along the lines of Figure 2.3. The remainder of the graphical model assumes the condition-specific parameters generate the behavioral data using the logistic psychophysical model.

The key inference is the joint posterior of the δ_α and δ_β parameters. When $(\delta_\alpha, \delta_\beta) = 0$, the "different" model reduces to the "same" model, since the shift and scale parameters for both conditions will be the same. Thus, the Savage-Dickey procedure can be applied to the two-dimensional joint posterior to estimate the required Bayes factor. Figure 2.28 shows the joint prior and posterior for δ_α and δ_β for subjects A and B. The prior distribution is the same for both subjects, following its definition in the graphical model in Figure 2.27. It is clear that the posterior for subject A has significant density near $(\delta_\alpha, \delta_\beta) = 0$. The estimate of the Bayes factor is about 9 in favor of the "same" model, corresponding to the posterior density being about 9 times greater than the prior at the origin.[2] Thus, the data provide evidence in favor of subject A using the same model in both the auditory and visual tasks. This is consistent with the visual impression from the behavioral data in Figure 2.2, in which the response proportions for matching target durations appear very similar for the auditory and visual data.

The posterior for subject B, in contrast, has most of its density for values with $\delta_\alpha > 0$ and $\delta_\beta < 0$ consistent with the visual

[2]Technical details: Results are based on eight chains of 100,000 samples each collected after 1,000 burn-in samples. Various estimates of the Bayes factor were obtained by counting all the samples in the joint prior within ϵ of the origin 0, varying ϵ from 10 down to 1 in steps of 1. The Bayes factors reported use $\epsilon = 1$, but very similar estimates were obtained for other values.

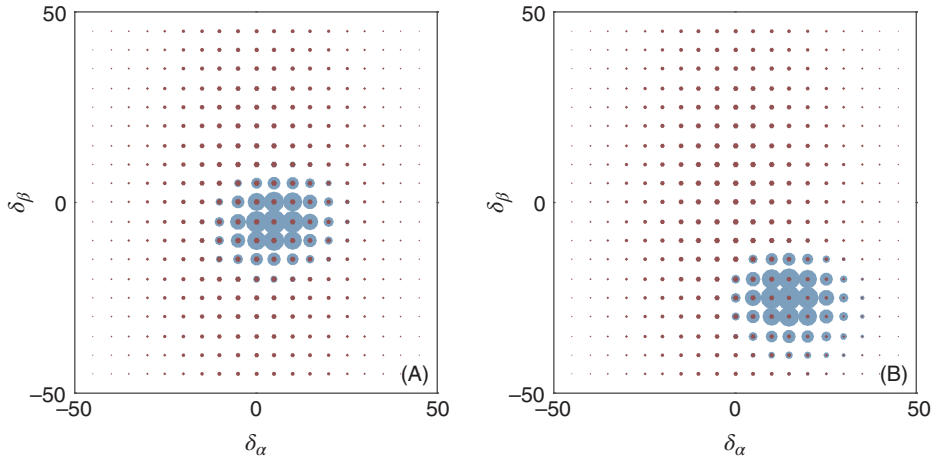


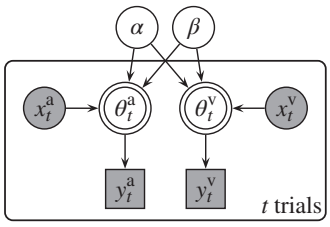
Figure 2.28 The joint prior and posterior distribution for the difference parameters, for subjects A (*left panel*) and B (*right panel*). The prior is shown by lighter circles and the posterior is shown by darker circles. The area of a circle represents the density of the prior or posterior distribution for the combination of difference parameters.

condition having a smaller shift but greater scale than the auditory condition. This is consistent with the visual impression from the behavioral data in Figure 2.2, where the curve corresponding to the visual data appears to be shifted to the left and shallower relative to the curve for the auditory data. In terms of comparing the "same" and "different" models, it is clear from Figure 2.28 that the joint posterior has almost no density at the origin. The estimate of the Bayes factor is about 600 in favor of the "different" model.

This example raises an important technical issue. It might seem tempting to estimate the Bayes factor by examining the difference in parameters for the two conditions directly, rather than constructing the parameters for each condition as in Figure 2.27. The idea would be that, by inferring α^{a} and α^{a} separately and then considering the deterministic difference $\delta_\alpha^{\mathrm{a}} = \alpha^{\mathrm{a}} - \alpha^{\mathrm{v}}$, the Bayes factor could be found by examining the ratio of the posterior and prior density of the derived at the critical value of zero. This approach suffers from the so-called Borel-Kolmogorov paradox (Jaynes, 2003, Chapter 15; Wetzels et al., 2010). Intuitively, the problem is that equality depends on the limiting process that produces the equality—such as subtraction yielding zero, or division yielding one—but there is no reason to prefer any limiting process over another. This arbitrariness is circumvented by making the difference process explicit in the model and placing a prior on the difference parameter or parameters, as done in Figure 2.27.

Common-Cause Modeling

Evidence for invariance naturally leads to common-cause modeling, in which the same psychological variables and processes are assumed to contribute to multiple sorts of observed behaviors in multiple tasks or contexts. Figure 2.29 shows a common-cause model of the auditory and visual task behavior for a single subject, assuming the same psychophysical model generates behavior in both modalities.



$$\alpha \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0,\ 1/100^2)$$
$$\theta_t^a = 1/\left(1+\exp\left(-\frac{x_t^a - s - \alpha}{\beta}\right)\right)$$
$$\theta_t^v = 1/\left(1+\exp\left(-\frac{x_t^v - s - \alpha}{\beta}\right)\right)$$
$$y_t^a \sim \text{Bernoulli}(\theta_t^a)$$
$$y_t^v \sim \text{Bernoulli}(\theta_t^v)$$

Figure 2.29 Graphical representation of a common-cause model for visual and auditory behavioral data, based on the same underlying logistic psychophysical model.

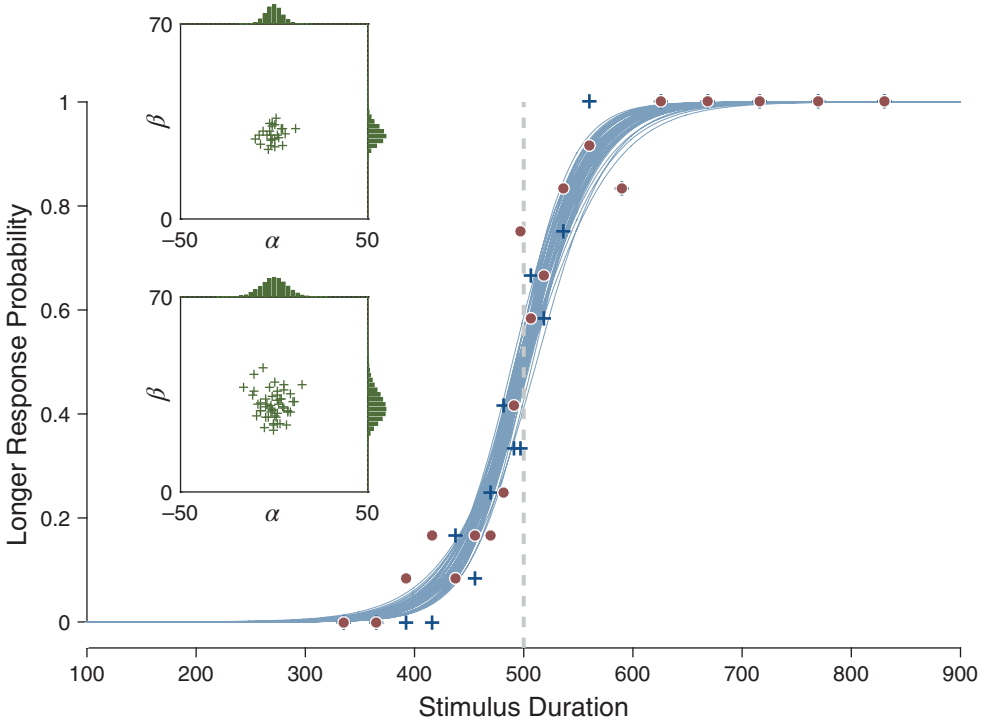


Figure 2.30 The posterior distribution for the common-cause psychophysical model, based on the auditory and visual task data from subject A. The lines show sample psychophysical functions from the posterior, and the circles and crosses summarize the behavioral response data from both tasks. The upper inset panel shows the joint and marginal posterior distributions for the shift α and scale β parameters for the common-cause model. The lower inset panel shows the joint and marginal posterior distributions for the shift α and scale β parameters when inferred independently for each task and then combined.

Figure 2.30 summarizes the inferences from applying the common-cause model to the auditory and visual task behavioral data for subject A. The posterior psychophysical function is shown, with data from both tasks overlain. The upper inset panel shows the joint and marginal posterior distribution for the shift and scale parameters. The lower

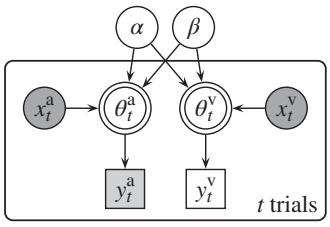
inset panel provides a comparison, showing the inferences obtained by applying the original model in Figure 2.12 to both tasks separately and then combining the inferences about the shift and scale. It is clear that both the joint and the independent models make similar inferences about α and β, which makes sense since they are based on the same basic psychophysical model and the same data. The common-cause model, however, is more certain in its inferences, as shown by the tighter or narrower joint and marginal distributions. This also makes sense, since the assumption that the same shift and scale generate both data sets means inferences about α and β for the common-cause model are based on stronger theoretical assumptions and more empirical information.

The Generality and Paucity of Common-Cause Models

Identifying common causes is a powerful and general approach to building psychological models. The idea that the same psychological variable—a learning rate, memory capacity, risk propensity, or whatever—influences observed behavior in multiple cognitive phenomena is an appealing one. The approach is much more general than the model in Figure 2.29. It is possible to use different psychophysical functions—a logistic for the auditory task and a Cauchy for the visual task, for example—if that modeling assumption was sensible, and still infer jointly common underlying shift and scale parameters. It would also be possible to consider more fundamentally different tasks, as long as one or more of the same psychological variables influenced behavior. Surprisingly, there seem to be few fully developed psychological models that jointly account for multiple sources of behavioral data (see Guan, Lee, & Vandekerckhove, 2015; Lee & Sarnecka, 2011; Selker, Lee, & Iyer, 2017, for some exceptions). Rather, the field often effortfully misses opportunities for productive common-cause models. A good example of this is provided by the use of multidimensional scaling methods to infer stimulus representations from similarity or other choice data, and the subsequent use of these representations to model cognitive processes like identification and category learning (e.g., Kruschke, 1992; Nosofsky, 1986; Romney, Brewer, & Batchelder, 1993). Typically, this is done in a two-step process, where, for example, the representation is derived from the similarity judgment data and then used as part of the categorization model. It seems more natural, however, to conceive of the same latent mental representation contributing to the generation of both the similarity data and the categorization data. A common-cause model formulated this way would have advantages in inferring the representation from all of the available relevant data, and still allow the models of similarity judgment and categorization processes to be developed, evaluated, and applied. The state of affairs for common-cause modeling is considerably better in model-based neuroscience, where the common-cause modeling of neural and behavioral data is a productive and maturing approach (Turner, Dennis, & Van Zandt, 2013; Turner et al., 2016).

Prediction and Generalization

An important strength of Bayesian methods in general, and of their application to hierarchical and common-cause modeling in particular, is the ability to support prediction and generalization. Although both terms are used loosely, it is conceptually helpful to distinguish predictions as being for data not observed, but arising from the current task, and generalizations as being for unseen data from one or more different tasks (Ahn, Busemeyer, Wagenmakers, & Stout, 2008). The results, based on the hierarchical model in Figure 2.24 for the new subject, as shown in Figure 2.26, are a good example of prediction. Based on the behavior of observed subjects and the assumptions of a model of their individual differences, it is possible to predict the behavior of a yet-to-be-observed subject.



$$\alpha \sim \text{Gaussian}(0,\ 1/50^2)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0,\ 1/100^2)$$
$$\theta_t^a = 1/\left(1 + \exp\left(-\frac{x_t^a - s - \alpha}{\beta}\right)\right)$$
$$\theta_t^v = 1/\left(1 + \exp\left(-\frac{x_t^v - s - \alpha}{\beta}\right)\right)$$
$$y_t^a \sim \text{Bernoulli}(\theta_t^a)$$
$$y_t^v \sim \text{Bernoulli}(\theta_t^v)$$

Figure 2.31 Graphical representation of a common-cause model for visual and auditory behavioral data based on the same underlying logistic psychophysical model, with partially observed auditory data and unobserved visual data.

Figure 2.31 shows a graphical model that demonstrates both prediction and generalization in the context of a joint model. It is a natural extension of the model in Figure 2.29, introducing the notion of partial observability. In particular, the node for auditory task behavioral data y_t^a is lightly shaded, indicating that it is partially observed. This means that the behavioral data for some trials are observed, but for other trials are unobserved. The graphical model in Figure 2.31 also differs from the common-cause model in Figure 2.29 by assuming that all of the behavioral data for the visual task are unobserved, as indicated by the unshaded y_t^v node.

The partially observed and unobserved data can be thought of as missing data. Bayesian methods naturally deal with missing data because they are inherently generative. Technically, this means they model the joint distribution of the data and parameters (Lasserre, Bishop, & Minka, 2006). Intuitively, it means the model is capable of generating behavioral data and so, in a sense, is capable of producing the behavior required of a subject in the experiment. Thus, inferences about missing data are treated in the same principled ways as inference about latent parameters, by representing the possible values using probability distributions based on the model and available data. For a model like Figure 2.31, these inferences about missing data effectively are probabilistic predictions about unseen data in the auditory task and generalizations to unseen data on the different visual task.

Figure 2.32 summarizes the main results of applying the graphical model in Figure 2.31 to subject A, with only the first 60 trials in the auditory condition observed. The left panel shows the joint and marginal posterior distributions for the shift and scale parameters inferred from the 60 observed trials. The top-right panel shows the posterior predictive accuracy of the model for the auditory data, giving the posterior probability of the decision made by the subject on each trial. For the first 60 trials, this is a standard posterior predictive analysis of descriptive adequacy, since behavior on these trials is data presented to the model. For trials 61 to 240, however, the data are not observed, so the posterior predictive distribution of the model is a genuine prediction. The bottom-right panel shows the posterior predictive accuracy of the model for the visual task. The model is given no data for this task, and it is a different task from the one for which data are provided, so this is a test of the model generalization. The results make it clear that the model fares well in both prediction and generalization.

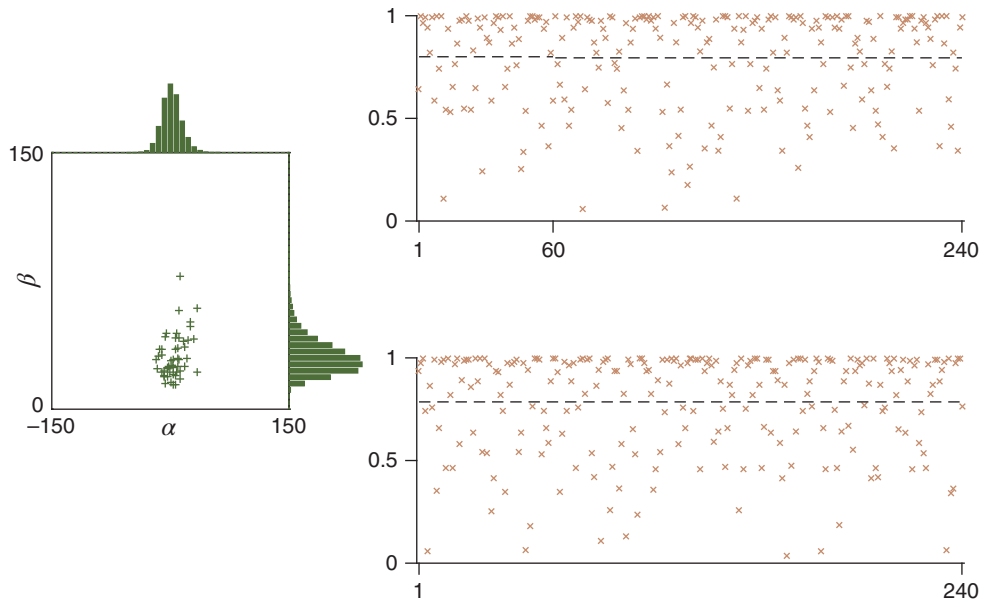


Figure 2.32 Prediction and generalization results from applying the common-cause model to the visual and auditory task data for subject A. The left panel shows the joint and marginal posterior distributions for shift α and scale β, based on the observed behavior for the first 60 trials of the auditory task. The top-right panel shows the posterior predictive accuracy of trials 1–60 and then the predictive accuracy for trials 61–240 in the auditory task. The bottom-right panel shows the generalization accuracy for all of the trials in the visual task.

Average accuracy, shown by the broken line, is 0.80 for the first 60 observed trials, and 0.79 for both the prediction and generalization trials.

Inference Is Not Inversion

All of our worked examples have dealt directly with the empirical data, and are justified on the basis of the interpretability of inferred parameters, the ability to predict and describe the data, and the ability to generalize to new and different data. None of the examples involve model or parameter recovery studies, which are common in the literature. In these studies, data are simulated by known models and parameter values, and evaluation hinges on the ability of a model or method to recover the ground truth. These simulation studies are useful in some ways, especially for sanity checking the accuracy of model implementation or for exploring the informativeness of experimental designs. But the use of simulation studies to evaluate the accuracy of parameter recovery in a fine-grained way (e.g., Ratcliff & Childers, 2015) or to evaluate methods of estimation or model selection themselves (e.g., Pitt, Myung, & Zhang, 2002) fundamentally confuses the concepts of inference and inversion. Inference finds what follows from the available information, whereas inversion aims to recover the truth. To get the intuition for the difference, consider a model-recovery study in which an extremely simple model and an extremely complicated model are the alternative ground truths, and a single datum is generated from the complicated model. Assuming, as is likely, this datum has some reasonable probability under the simple model, the correct inference is that it was generated by the simple model, even though that is not the ground truth. As more

data are generated from the complicated model, it becomes likely that there will be enough information to infer this model, but for a single datum the simple account may well be the appropriate inference. It is what is justified by the available information. Thus, evaluation in terms of recovering the ground truth can be misleading. Of course, model and parameter recover studies usually involve greater numbers of data, but that just makes the basic logical problem more difficult to identify and does not fundamentally negate it. The bottom line is that the correct inference is the one that is justified by the data, not a ground truth for which insufficient evidence exists. From the Bayesian perspective, the joint posterior distribution contains all of the relevant and valid inferential information. If these inferences differ from a ground truth in a simulation study, that is a message about the setup and outputs of the simulation, not the method of inference.

CONCLUSION

The goal of the case study was to demonstrate the usefulness of Bayesian methods as a way to relate cognitive models to behavioral data. The range of research questions addressed—from individual differences to underlying invariants, from latent parameters to predictions about data, from mixture models to common-cause models, and the consideration of sensitivity to contamination or sequential effects—highlights the flexibility of Bayesian methods. Our experience is that the research questions asked in cognitive modeling applications usually have natural and straightforward translations as model-based inference made possible by Bayesian methods. The use of the same principles and methods throughout the case study—namely the representation of uncertainty via joint, marginal, and conditional distributions, and the updating of these distributions to incorporate new information using probability theory—highlights the principled, complete, and coherent foundations for statistical inference offered by Bayesian methods. As we argued from the outset, this combination of creative freedom and methodological soundness makes Bayesian methods extremely useful for developing, testing, and applying models of cognition.

Accordingly, it is no surprise that Bayesian methods are quickly becoming common in all areas of cognitive modeling, well beyond the psychophysical modeling that was the focus of the case study. There are substative applications of Bayesian methods in models spanning perception (e.g., Rouder, Yue, Speckman, Pratte, & Province, 2010); representation (e.g., Okada & Lee, 2016); memory (e.g., Horn, Pachur, & Mata, 2015; Osth & Dennis, 2015); learning (e.g., Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010); development (e.g., Bäumler et al., 2014; Lee & Sarnecka, 2011); response times (e.g., Rouder, Lu, Morey, Sun, & Speckman, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011); judgment and choice (e.g., Nilsson, Rieskamp, & Wagenmakers, 2011; Vincent, 2016); decision making (Lee & Newell, 2011; Scheibehenne, Rieskamp, & Wagenmakers, 2013); and problem solving (e.g., Guan et al., 2015; Lee, Zhang, Munro, & Steyvers, 2011); and they include applications from clinical detection and diagnosis (e.g., Pooley et al., 2011; Steingroever, Wetzels, & Wagenmakers, 2013) to collective cognition and the wisdom of the crowd (e.g., Batchelder & Anders, 2012; Lee, Steyvers, & Miller, 2014).[3]

We expect that Bayesian methods will continue to become increasingly important, widespread, and useful in cognitive modeling. They allow models to be considered

[3]Many of these representative papers are taken from a more complete list provided at http://bayesmodels.com/bugs-models.

that are ambitious, and they allow them to be evaluated carefully. Our hope is that Bayesian methods will serve a narrow but critical role in furthering our understanding of human cognition, providing a bridge between theory and models on the one side and the behavior they attempt to describe, explain, and predict on the other.

REFERENCES

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, *32*, 1376–1402.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Batchelder, W. H., & Alexander, G. E. (2013). Discrete-state models: Comment on Pazzaglia, Dube, and Rotello (2013). *Psychological Bulletin*, *139*, 1204–1212.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332.

Bäumler, D., Voigt, B., Miller, R., Stalder, T., Kirschbaum, C., & Kliegel, M. (2014). The relation of the cortisol awakening response and prospective memory functioning in young children. *Biological Psychology*, *99*, 41–46.

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins University Press.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, 7–29.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*, 240–261.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, *42*, 204–223.

Donkin, C., Taylor, R., & Le Pelley, M. (2017). Evaluating models of visual working memory using a ranking task. Manuscript submitted for publication.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, J. Grosjean, & M. Shiffrar (Eds.), *Human body perception from the inside out* (pp. 105–131). New York, NY: Oxford University Press.

Feynman, R. (1994). *The character of physical law*. New York, NY: Modern Library/Random House.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6.

Gigerenzer, G. (2016). Taking heuristics seriously. In *The behavioral economics guide* (pp. v–xi). Behavioral Science Solutions.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goodman, N. D., & Stuhlmuller, A. (2014). *The design and implementation of probabilistic programming languages*. http://dippl.org (accessed December 17, 2015).

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

Guan, H., Lee, M. D., & Vandekerckhove, J. (2015). A hierarchical cognitive threshold model of human decision making on different length optimal stopping problems. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 824–829). Austin, TX: Cognitive Science Society.

Harré, R. (2002). *Great scientific experiments: Twenty experiments that changed our view of the world*. New York, NY: Dover.

Hemmer, P., Tauber, S., & Steyvers, M. (2014). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review*, *22*, 614–628.

Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, *21*, 1431–1443.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Horn, S. S., Pachur, T., & Mata, R. (2015). How does aging affect recognition-based inference? A hierarchical Bayesian modeling approach. *Acta Psychologies,*, *154*, 77–85.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, United Kingdom: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability*. Oxford, United Kingdom: Oxford University Press.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–231.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Kievit, R. A. (2011). Bayesians caught smuggling priors into Rotterdam harbor. *Perspectives on Psychological Science*, *6*, 313–313.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nut-shell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658–676.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573.

Kuss, M., Jakel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478–492.

Lasserre, J., Bishop, C. M., & Minka, T. (2006). Principled hybrids of generative and discriminative models. In *Proceedings 2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York*.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*, 29–41.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*, 259–273.

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making*, *6*, 832–842.

Lee, M. D., & Sarnecka, B. W. (2011). Number knower-levels in young children: Insights from a Bayesian model. *Cognition*, *120*, 391–402.

Lee, M. D., Steyvers, M., & Miller, B. J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE*, *9*, 1–9.

Lee, M. D., & Vanpaemel, W. (in press). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review.*

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, United Kingdom: Cambridge University Press.

Lee, M. D., Zhang, S., Munro, M. N., & Steyvers, M. (2011). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research*, *12*, 164–174.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition*, *39*, 914–923.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia, PA: SIAM.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Review of General Psychology*, *8*, 265.

Lodewyckx, T., Kim, W., Tuerlinckx, F., Kuppens, P., Lee, M. D., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, United Kingdom: Cambridge University Press.

Marr, D. C. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *404*, 1023–1032.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 1–21.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgment: A nonparametric Bayesian approach. *Neural Computation*, *20*, 2597–2628.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Nilsson, H., Rieskamp, J., & Wagenmakers, E. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*, 84–93.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35–44.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*, 260–311.

Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. *Handbook of defeasible reasoning and uncertainty management systems: Quantified representation of uncertainty and imprecision*, *1*, 367–389.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*.

Pooley, J. P., Lee, M. D., & Shankle, W. R. (2011). Understanding Alzheimer's using memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, *55*, 47–56.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

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

Reisberg, B. (1988). Functional assessment staging (FAST). *Psychopharmacology Bulletin*, *24*, 653–659.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review*, *109*, 605.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*, 28–34.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Rouder, J. N., Haaf, J., & Vandekerckhove, J. (2017). *Bayesian inference for psychology, Part IV: Parameter estimation and Bayes factors*. Manuscript submitted for publication.

Rouder, J. N., Yue, Y., Speckman, P. L., Pratte, M. S., & Province, J. M. (2010). Gradual growth versus shape invariance in perceptual decision making. *Psychological Review*, *117*, 1267.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*, 39.

Schönbrodt, F. (2015, April 17). Grades of evidence—A cheat sheet [Blog post]. Retrieved from http://www.nicebread.de/grades-of-evidence-a-cheat-sheet/

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. *Decision*, *4*, 87–101.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*, 1–25.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Sprenger, J. (2015, December). *The objectivity of subjective Bayesian inference*. Retrieved from http://philsci-archive.pitt.edu/11797/

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). Validating the PVL-Delta model for the Iowa gambling task. *Frontiers in Psychology*, *4*, 898.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*, 410–441.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review*, *120*, 667–678.

Turner, B. M., Rodriguez, C. A., Norcia, T., McClure, S. M., & Steyvers, M. (2016). Why more *is* better: A method for simultaneously modeling EEG, fMRI, and behavior. *NeuroImage*, *128*, 96–115.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

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

van Driel, J., Knapen, T., van Es, D. M., & Cohen, M. X. (2014). Interregional alpha-band synchrony supports temporal cross-modal integration. *NeuroImage*, *101*, 404–415.

Vanpaemel, W. (2016). Prototypes, examplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, *72*, 183–190.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Vincent, B. (2016). Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. *Behavior Research Methods*, *48*, 1608–1620.

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

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

Wagenmakers, E., Lee, M. D., Rouder, J. R., & Morey, R. (2017). Another statistical paradox. Manuscript submitted for publication.

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

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis*, *54*, 2094–2102.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the expectancy valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, *54*, 14–27.

Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, *133*, 283–295.