

**Mathematical Modeling**

Daniel R. Cavagnaro, Jay I. Myung, and Mark A. Pitt

The Ohio State University

August 26, 2010

To appear in Todd D. Little (ed.), *The Oxford Handbook of Quantitative Methods*.

Oxford University Press: New York, NY.

Running head: Mathematical Modeling

Word count: 10,600

Corresponding author and address:

Dr. Jay Myung  
Department of Psychology  
Ohio State University  
1835 Neil Avenue  
Columbus, OH 43210-1351  
Email: [myung.1@osu.edu](mailto:myung.1@osu.edu)  
Phone: 614-292-1862

### Abstract

Explanations of human behavior are most often presented in a verbal form, as theories. Psychologists can also harness the power and precision of mathematics by explaining behavior quantitatively. This chapter introduces the reader to how this is done and the advantages of doing so. It begins by contrasting mathematical modeling with hypothesis testing to highlight how the two methods of knowledge acquisition differ. The many styles of modeling are then surveyed, along with their advantages and disadvantages. This is followed by an in-depth example of how to create a mathematical model and fit it to experimental data. Issues in evaluating models are discussed, including a survey of quantitative methods of model selection. Particular attention is paid to the concept of generalizability and the trade-off of model fit with model complexity. The chapter closes by describing some of the challenges for the discipline in the years ahead.

**Keywords:** Cognitive modeling; model testing; model evaluation; model comparison.

## Introduction

Psychologists study behavior. Data, acquired through experimentation, are used to build theories that explain behavior, which in turn provide meaning and understanding. Because behavior is complex, a complete theory of any behavior (e.g., depression, reasoning, motivation) is likely to be complex as well, having many variables and conditions that influence it.

Mathematical models are tools that assist in theory development and testing. Models are theories, or parts of theories, formalized mathematically. They complement theorizing in many ways, as discussed in the following pages, but their ultimate goal is to promote understanding of the theory, and thus behavior, by taking advantage of the precision offered by mathematics. Although they have been part of psychology since its inception, their popularity began to rise in the 1950s and has increased substantially since the 1980s, in part due to the introduction of personal computers. This interest is not an accident or fad. Every style of model that has been introduced has had a significant impact in its discipline, and sometimes far beyond that. After reading this chapter, the reader should begin to understand why.

This chapter is written as a first introduction to mathematical modeling in psychology for those with little or no prior experience with the topic. Our aim is to provide a good conceptual understanding of the topic and make the reader aware of some of the fundamental issues in mathematical modeling, but not necessarily to provide an in-depth step-by-step tutorial on how to actually build and evaluate a mathematical model from scratch. In doing so, we assume no more of the reader than a year-long course in graduate-level statistics. For related publications on the topic, the reader is directed to

Busemeyer and Diederich (2010), Fum, Del Missier and Stocco (2007), and Myung and Pitt (2002). In particular, the present chapter may be viewed as an updated version of the last of these. The focus of the first half of the chapter is on the advantages of mathematical modeling. By turning what may be vague notions or ideas into precise quantities, significant clarity can be gained revealing new insights that push science forward. In the next section, we highlight some of the benefits of mathematical modeling relative to the method of investigation that currently dominates psychological research: verbal modeling. After that, we provide a brief overview of the styles of mathematical modeling. The second half of the chapter focuses on algebraic models, discussing in detail how to build them and how to evaluate them. We conclude with a list of recommended readings for many of the topics covered.

### **From Verbal Modeling to Mathematical Modeling**

#### **Verbal Modeling**

To understand the importance and contribution of mathematical modeling, it is useful to contrast it with the way scientific investigation commonly proceeds in psychology. The typical investigation proceeds as follows. First, a hypothesis is generated from a theory in the form of differences across conditions. These could be as general as higher ratings in the experimental condition compared to a control condition, or a V-shaped pattern of responses across three levels of an independent variable such as task difficulty (e.g., low, medium, high). The hypothesis is usually coarse grained and expressed verbally (e.g., “memory will be worse in condition A compared with condition B,” or “one’s self-image is more affected by negative than positive reinforcement”), hence it is referred to as a verbal model. To test the hypothesis, it is contrasted with the

hypothesis that there is absolutely no difference among conditions. After data collection, inferential statistics are used to pass judgment on only this latter, "null" hypothesis. A statistically significant difference leads one to reject it (which is not the same as confirming the hypothesis of interest), while on the other hand, a difference that is not statistically significant leads one to fail to reject the null, effectively returning one to the same state of knowledge as before the experiment was conducted.

This verbal modeling ritual is played out over and over again in the psychology literature. It is usually the case that a great deal of mileage can be gained from it when testing a new theory because correctly predicting qualitative differences (e.g.,  $A > B$ ) can be decisive in keeping a theory alive. However, a point of diminishing returns will eventually be reached once a majority of the main claims have been tested. The theory must expand in some way if it is to be advanced. After all, models should provide insight and explain behavior at a level of abstraction that goes beyond a redescription of the data. Moreover, although the data collected are analyzed numerically using statistics, numerical differences are rarely predicted nor of primary interest in verbal models, which predict qualitative differences among conditions. To take the theory a step further and ask the degree to which performance should differ between two conditions goes beyond the level of detail provided in verbal models.

Mathematical modeling offers a means for going beyond verbal modeling by using mathematics in a very direct manner, to *instantiate* theory, rather than a supplementary manner, to test simple, additive effects predicted by the theory. In quantifying a theory, the details provided in its mathematical specification push the theory in new directions and make possible new means of theory evaluation. In a mathematical model, hypotheses

about the relations between the underlying mental processes and behavioral responses are expressed in the form of mathematical equations, computer algorithms, or other simulation procedures. Accordingly, mathematical models can go beyond qualitative predictions such as “performance in condition A will be greater than performance in condition B,” to make quantifiable predictions such as “performance in condition A will be two times greater than in condition B,” which can be tested experimentally.

Furthermore, using mathematics to instantiate theory opens the door to models with nonlinear relationships and dynamic processes, which are capable of more accurately reflecting the complexity of the psychological processes that they are intended to model.

### **Shifting the Scientific Reasoning Process**

Mathematical modeling also aids scientific investigation by freeing it from the confines of null hypothesis significance testing (NHST) of qualitative predictions in verbal models. The wisdom of NHST has been criticized repeatedly over the years (Rozeboom, 1960; Bakan, 1966; Lykken, 1968; Nickerson, 2000; Wagenmakers, 2007). In NHST, decisions pertain only to the null hypothesis. Decisions about the accuracy of the experimental hypothesis the researcher is interested in are not made. Statistically significant results merely keep the theory alive, making it a contender among others. In the end, the theory should be the only one standing if it is correct, but with NHST, commitment to one’s theory is never made and evidence is only indirectly viewed as accumulating in favor of the theory of interest. This mode of reasoning makes NHST very conservative.

Although the landscape of statistical modeling in psychology is changing to make increasing use of NHST of quantitative predictions in conjunction with mathematical

models, as in *Structural Equations Modeling* (SEM) and *Multilevel Modeling* (MLM), the dominant application of NHST is still to test qualitative predictions derived from verbal models. Continuous use of NHST in this way can hinder scientific progress by creating a permanent dependence on statistical techniques such as linear regression or ANOVA, rather than at some point switching over to using mathematics to model the psychological processes of interest. Furthermore, statistical tests are used in NHST in a way that gives the illusion of being impartial or objective about the null hypothesis, when in fact all such tests make more explicit assumptions about the underlying mental process, the most obvious one being that behavior is linearly related to the independent variables. If one is not careful, theories can end up resembling the statistical procedures themselves. Gigerenzer (1991) refers to this approach to theory building as tools-to-theories. Researchers take an available statistical method and postulate it as a psychological explanation of data. However, unless one thinks that the mind operates as a regression model or other statistical procedure, these tools should not be intended to reflect the inner workings of psychological mechanisms (Marewski & Olsson, 2009).

When engaged in mathematical modeling, there is an explicit change in the scientific reasoning process away from that of NHST-based verbal modeling. The focus in mathematical modeling is on assessing the viability of a particular model, not rejecting or failing to reject the status quo. Correctly predicted outcomes are taken as evidence in favor of the model. Although it is recognized that alternative models could potentially make the same predictions (this issue is discussed more thoroughly below), a model that passes this “sufficiency test” is pursued and taken seriously until evidence against it is generated or a viable contender is proposed.

## Types of Mathematical Models

This section offers a brief overview of the various types of mathematical models that are used in different subfields of psychology.

### Core Modeling Approaches

The styles of modeling listed under this heading were popularized before the advent of modern computing in the 1980s. Far from being obsolete, the models described here comprise the backbone of modern theories in psychophysics, measurement, and decision making, among others, and important progress is still being made with these methods.

### *Psychophysical Models*

The earliest mathematical models in psychology came from psychophysicists, in their efforts to describe the relationship between the physical magnitudes of stimuli and their perceived intensities (e.g., does a 20 pound weight feel twice as heavy as a 10 pound weight?). One of the pioneers in this field was Ernst Heinrich Weber (1795-1878). Weber was interested in the fact that very small changes in the intensity of a stimulus, such as the brightness of a light or the loudness of a sound, were imperceptible to human participants. The threshold at which the difference can be perceived is called the *just-noticeable difference*. Weber noticed that the just-noticeable difference depends on the stimulus' magnitude (e.g., 5%) rather than being an absolute value (e.g., 5 grams). This relationship is formalized mathematically in terms of the differential equation known as Weber's Law:  $\Delta_{JND}x = k_w x$ , where,  $\Delta_{JND}x$  is the just-noticeable difference (JND) in the physical intensity of the stimulus,  $x$  is the current intensity of the stimulus, and  $k_w$  is an empirically determined constant known as the Weber fraction. That is, the

just-noticeable difference is equal to a constant times the physical intensity of the stimulus. For example, a Weber fraction of 0.01 means that participants can detect a 1% change in the stimulus intensity. The value of the Weber fraction varies depending on the nature of the stimulus (e.g., light, sound, heat).

Gustav Fechner (1801-1887) rediscovered the same relationship in the 1850s and formulated what is now known as Fechner's law:  $\psi(x) = k \cdot \ln(x)$  where  $\psi(x)$  denotes the perceived intensity (i.e., the perceived intensity of the stimulus is equal to a constant times the log of the physical intensity of the stimulus). Since Fechner's law can be derived from Webers Law as an integral expression of the latter, they are essentially one and the same and are often referred to collectively as the Weber-Fechner Law. For more details on these and other psychophysical laws, see Stevens (1975).

The early psychophysical laws were extended by Louis Thurstone (1887–1955), who considered the more general question of how the mind assigns numerical values to items, even abstract items such as attitudes and values, so that they can be meaningfully compared. He published his paper on the "law" of paired comparisons in 1927. Although Thurstone referred to it as a law, it is more aptly described as a model since it constitutes a scientific hypothesis regarding the outcomes of pairwise comparisons among a collection of objects. If data agree with the model, it is possible to produce a scale from the data. Thurstone's model is the foundation of modern psychometrics, which is the general study of psychological measurement. For more details, see Thurstone (1974).

### ***Axiomatic Models***

The axiomatic method of mathematical modeling involves replacing the phenomenon to be modeled with a collection of simple propositions, or "axioms," which

are designed in such a way that the observed pattern of behavior can be deduced logically from them. Each axiom by itself represents a fundamental assumption about the process under investigation, and often takes the form of an ordinal restriction or existence statement, such as "The choice threshold is always greater than zero" or "there exists a value  $x$  greater than zero such that a participant will not be able to distinguish between  $A$  units and  $A+x$  units." Taken together, a set of axioms can constrain the variables sufficiently for a model to be uniquely identified.

Axiomatic models are especially prevalent in the field of judgment and decision making. For example, the Expected Utility model of decision making under uncertainty (Morgenstern & Von Neumann, 1947) states that any decision maker's preferences can be characterized according to an internal utility function that they use to evaluate uncertain prospects. This utility function has the form of an expected utility in the sense that a gamble  $G$  offering  $x$  dollars with probability  $p$  and  $y$  dollars with probability  $(1 - p)$ , would have expected utility  $U(G) = pv(x) + (1 - p)v(y)$ , where  $v(x)$  represents the subjective value of money to the participant. That is, the utility of the gamble is equal to a weighted sum of the possible payoffs, where the weight attached to each payoff is its probability of occurring. The model predicts that a decision maker will always choose the gamble with higher expected utility.

On the face of it, the existence of such a utility function that fully defines a decision maker's preferences over all possible gambles is a difficult assumption to justify. However, its existence can be derived by assuming the following three, reasonable axioms (see, e.g., Fishburn, 1982):

1. *Ordering*: Preferences are weak orders (i.e., rankings with ties);

2. *Continuity*: For any choice B such that choice A is preferred to choice B, which is in turn preferred to choice C, there exists a unique probability  $q$  such that one is indifferent between choice B and a gamble composed of  $q$  chance of A and a  $(1 - q)$  chance of C, in which A is chosen with probability  $q$  and C is chosen with probability  $(1 - q)$ ;

3. *Independence*: If choices A and B are equally preferable, then a gamble composed of a  $q$  chance of A and a  $(1 - q)$  chance of C is equally preferable to a gamble composed of a  $q$  chance of B and a  $(1 - q)$  chance of C for any choice C and all  $q$  ( $0 < q < 1$ ).

The axiomatic method is very much the “slow-and-steady” approach to mathematical modeling. Progress is often slow in this area because of the mathematical complexities involved in constructing coherent and justifiable axioms for psychological phenomena of interest. However, since all of the assumptions are spelled out explicitly in behaviorally verifiable axioms, axiomatic models are highly transparent in how they generate predictions. Moreover, due to the logical rigor of their construction, axiomatic models are long-lasting. That is to say, unlike other types of mathematical models that we will discuss later, axiomatic models are not prone to being deposed by competing models that perform “just a little bit better.” For these reasons, many scientists consider the knowledge gained from axiomatic modeling to be of the highest quality. For more details on axiomatic modeling, the reader is referred to Luce (2000).

### *Algebraic Models*

Algebraic models are probably what come to mind first for most people when they think of mathematical models. An algebraic model is essentially a generalization of the standard linear regression model in the sense that it describes exactly how the input stimuli and model parameters are combined to produce the output (behavioral response), in terms of a closed-form algebraic expression. Algebraic models are usually easy to understand due to this tight link between the descriptive (verbal) theory and its computational instantiation. Further, their assumptions can usually be well justified, often axiomatically or through functional equations (e.g., Aczel, 1966).

The simplest example of an algebraic models is the general linear model, which is restricted to linear combinations of input stimuli, such as  $y = ax + b$ , in which the tunable, free parameters (a, b) measure the relative extent to which the output response y is sensitive to the input stimulus dimension x. In general, however, algebraic models may include nonlinear terms and parameters that can describe various psychological factors.

For example, it is well known among memory researchers that a person's ability to retain in memory what was just learned (e.g., a list of words) drops quickly at first and then levels off. The exponential model of memory retention (e.g., Wixted & Ebbesen, 1991) states this relationship between time and amount remembered with the equation  $p = ae^{-bx}$ , where p is the probability of a participant being able to correctly recall the learned item (e.g., a word), x is the length of time since learning it, and a and b are model parameters. This means that the probability of correct recall is found by first multiplying the length of time since learning by  $-b$ , exponentiating the result the constant e to that power, and then multiplying the resulting value by a. When  $x = 0$ , the value of this equation is a, which means that the parameter a ( $0 < a < 1$ ) represents the baseline

retention probability before any time passed. The parameter  $b$  ( $b > 0$ ) represents the rate at which retention performance drops with time, which is a psychological process. We could of course entertain other model equation that can capture this decreasing trend of retention memory, such as power ( $p = a(x+1)^{-b}$ ), hyperbolic ( $p = 1/(a+bx)$ ), or logarithmic models, to name a few (see, e.g., Rubin & Wenzel, 1996).

Other examples of algebraic models include the Diffusion Model of Memory Retrieval (Ratcliff, 1978), Generalized Context Model of category learning (Nosofsky, 1986), Multinomial Processing Tree models of source monitoring (Batchelder & Riefer, 1999), and the Scale-Independent Memory, Perception and Learning model (SIMPLE) of memory retrieval (Brown, Neath & Chater, 2007).

### **Computational Modeling Approaches**

Modern day mathematical models are characterized by an increased reliance on the computational power provided by the rise of modern computing in the 1980s.

#### ***Algorithmic Models***

An algorithmic model is defined in terms of a simulation procedure that describes how specific internal processes interact with one another to yield an output behavior. The processes involved are often so complicated that the model's predictions can not be obtained by simply evaluating an equation at the appropriate values of the parameters and independent variables, as in algebraic models. Instead, deriving predictions from the model requires simulating dynamic processes on a computer with the help of random number generators. The process begins with an activation stimulus, and then runs through a sequence of probabilistic interactions that are meant to represent corresponding mental

activity, finally yielding an output value that usually corresponds to a decision or an action taken by a participant.

When building an algorithmic model, the primary concern is that the system accurately reproduces human data. In contrast to the axiomatic modeling approach, in which each assumption is well-grounded theoretically, algorithmic models often make many assumptions about the mental processes involved in a behavior, which cannot be verified empirically because they are not directly observable. This gives scientists considerable leeway to tweak the internal structure of a model and quickly observe its behavior.

One advantage of this approach is that it allows scientists to work with ideas that cannot yet be expressed in precise mathematical form (Estes, 1975). This extends the domain of what can be modeled to include very complex cognitive and neural processes. Moreover, this type of model can provide a great deal of insight into the mental processes that are involved in behavior. For example, an algorithmic model such as the Decision Field Theory model of decision making (Busemeyer & Townsend, 1993) predicts not only the final action taken by a participant, but also the amount of time elapsed before taking that action. Another excellent example of this type of model is the retrieving-effectively-from-memory (REM) model of recognition memory (Shiffrin & Steyvers, 1997).

The main drawback of algorithmic modeling is a lack of transparency between the parts of the model and their mental counterparts. The same flexibility that allows them to be built and tested quickly also allows them to create a host of assumptions that often serve no other purpose than simply to fit the data. To minimize this problem, algorithmic

models should be designed with as few assumptions as possible, and care should be taken to ensure that all of the assumptions are well-justified and psychologically plausible.

### *Connectionist Models*

Connectionist models make up a class of cognitive models in which mental phenomena are described by multi-layer networks of interconnected units, or “nodes.” Model predictions are generated by encoding a stimulus in the activation of a set of “input nodes”, which then pass the activation across a series of “hidden nodes”, which transform the original stimulus into new codes or features, until the activation finally reaches an “output node” representing a response. This structure is often meant to simulate the way the brain works, with the nodes representing neurons and the connections between nodes representing synapses, but other interpretations are also possible. For example, in a connectionist model of language acquisition, the nodes could represent words with connections indicating semantic similarity. Examples of connectionist models include the TRACE model of speech perception (McClelland & Elman, 1986), the ALCOVE model of category learning (Kruschke, 1992), the Connectionist Model of Word Reading (Plaut, McClelland, Seidenberg & Patterson, 1996), and the Temporal Context Model of episodic memory (Howard & Kahana, 2002).

Connectionist models can be characterized as a particular subclass of algorithmic models. The key difference is that connectionist models make even fewer explicit assumptions about the underlying processes, and instead focus on learning the regularities in the data through training. Essentially, the network learns to produce the correct data pattern by adapting itself from experience with the input, strengthening and weakening connections in a manner similar to the way learning occurs in the human brain. This

flexibility allows connectionist models to predict highly complex data patterns. In fact, certain connectionist models have been proved by mathematicians to have literally unlimited flexibility. That is, a connectionist model with a sufficiently large number of hidden units can approximate any continuous nonlinear input-output relationship to any desired degree of accuracy (Hornik, Stinchcombe & White, 1989, 1990). Unfortunately, this means that connectionist models are prone to fit not only the underlying regularities in the data, but also spurious, random noise that has no psychological meaning.

Consequently, care must be taken to make sure that the model learns only the underlying regularities and does not degenerate into a mere re-description of the idiosyncrasies in the data, which would provide little insight into mental functioning.

### ***Bayesian Modeling***

"Bayesian model" has become somewhat of a buzz phrase in recent years, and it is now used very broadly in reference to any model that takes advantage of the Bayesian statistical approach to processing information (Chater, Tenenbaum & Yuille, 2006; Kruschke, 2010; Lee, in press). However, since the Bayesian approach can be utilized in diverse ways to the aid of mathematical modeling, there are actually a few different classes of models, all of which are referred to as Bayesian models.

Briefly, a Bayesian model is defined in terms of two components: (a) the *prior distribution*, which is a probability distribution -representing the investigator's initial uncertainty about the parameters before the data are collected and; (b) the *likelihood function*, which specifies the likelihood of the observed data as a function of the parameters. From these, the *posterior distribution*, which is a probability distribution that expresses an updated uncertainty about the parameters in light of the data, is obtained by

applying Bayes rule. A specific inference procedure is then constructed or performed on the basis of the posterior distribution depending upon the inference problem at hand. For further details of Bayesian inference, the reader is directed to other sources (e.g., Gill, 2008; Gelman, Carlin, Stern & Rubin, 2004).

Two types of Bayesian models that we will briefly discuss here are Bayesian statistical models -- those that use Bayesian statistics as a tool for data analysis, and Bayesian theoretical models -- those that use Bayesian statistics as a theoretical analogy for the inner workings of the mind. Bayesian statistical models often use Bayesian statistics as a method of conducting standard analyses of data, as an alternative to frequentist statistical methods such as null hypothesis significance testing (NHST) (for a review, see, e.g., Kruschke, 2010). Bayesian hypothesis testing using the Bayes factor, for example, extends NHST to allow accumulation of evidence in favor of a null hypothesis (Wetzels, Raaijmakers, Jakab & Wagenmakers, 2009; Wagenmakers, Lodewyckx, Kuriyal & Grasman, 2010). It also provides the necessary machinery for doing inference on unobservable, or "latent," psychological parameters, as opposed to just measured dependent variables such as recall rate and response time. This style of Bayesian modeling, called Hierarchical Bayes, accounts for additional sources of variation, such as individual differences, in a rigorous way using latent parameters (Rouder & Lu, 2005; Rouder, Sun, Speckman, Lu & Zhou, 2003; Lee, 2008; Lee, in press). Due to their popularity, specialized software packages have been developed for building and testing them (Lunn, Thomas, Best & Spiegelhalter, 2000).

Bayesian theoretical models, on the other hand, utilize Bayesian statistics as a working assumption for how the mind makes inferences. In this style of modeling,

Bayesian inference is used to provide a rational account of why people behave the way they do, often without accounting for the cognitive mechanisms that produce the behavior. Bayesian statistics as a theoretical analogy has been an influential position for the last decade or so in cognitive science, and it has led to the development of impressive new models addressing a wide range of important theoretical questions in psychological science (e.g., Chater et al., 2006; Tenenbaum, Griffiths & Kemp, 2006; Griffiths, Steyvers & Tenenbaum, 2007; Steyvers, Lee & Wagenmakers, 2009; Xu & Griffiths, 2010; Lee & Sarnecka, 2010).

### **How to Build and Evaluate Mathematical Models**

Just as verbal models are built from interpretation of past data and intuitions about the psychological process of interest, mathematical models require one to make more of these same decisions, but at a much finer level of precision. This can make a first-time modeler uncomfortable because of the many decisions that must be made, which force the practitioner to make important choices and think about critical issues at a high level of specificity. However, the process can be tremendously insightful and cause the practitioner to rethink past assumptions, viewpoints, and interpretations of data. In this section, we walk through the process of mathematical modeling, from model specification through fitting data, model comparison, and finally model revision. Before that, though it is important to explain the logic of modeling.

#### **Logic of Model Testing**

The generally accepted criterion for a model to be “correct” is that it is both necessary and sufficient for its predictions about the data to be true. Estes (2002) succinctly illustrates how this criterion can be scrutinized more carefully by considering

it in the framework of formal logic, some key points of which we review here. Following the standard logical notation (Suppes, 1957), let  $P$  denote the model of interest, collectively referring to the assumptions the model makes, and let  $Q$  denote the predictions being made about possible observations in a given experimental setting. The sufficiency of the model can be assessed by examining the logical statement  $P \rightarrow Q$ , which reads “ $P$  implies  $Q$ ,” and the necessity can be assessed by examining the logical statement  $\sim P \rightarrow \sim Q$ , which reads “not  $P$  implies not  $Q$ .”

The sufficiency condition,  $P \rightarrow Q$ , is equivalent to the informal statement that under the assumptions of the model, the predictions of the data follow. What this means for model testing is that if the predictions are shown to be accurate (i.e., confirmed by observed data), then the model is said to be sufficient to predict the data. On the other hand, if the predictions are shown to be inaccurate and thus unconfirmed, then the model must be false (incorrect). In short, the model can be tested, and possibly falsified, by observing experiment data (Estes, 2002, p. 5).

It is important to emphasize that confirming sufficiency alone does not validate the model. This is because one might be able to construct another model, with a different set of assumptions from those of the original model, that may also make exactly the same predictions, that is,  $P' \rightarrow Q$ , where  $P'$  denotes the competing model. Consequently, confirming  $Q$  does not constitute the unequivocal confirmation of the model  $P$ . To establish the model as valid, the necessity of the model in accounting for the data must also be established.

The necessity condition,  $\sim P \rightarrow \sim Q$ , is equivalent to the informal statement that every possible deviation from the original model (e.g., by replacing the assumptions of

the model with different ones) fails to generate the predictions of the data. If this condition is satisfied, then the model is said to be necessary to predict the data.

The reality of model testing is that establishing the necessity of a model is generally not an achievable goal in practice. This is because testing it requires individual examinations of the assumptions of the model, which are not typically amenable to empirical verification. This means that in model testing one is almost always restricted to confirming or disconfirming the sufficiency of a model.

### **Model Specification**

Modeling can be a humbling experience because it makes one realize how incomplete the corresponding theory is. Given how little is actually known about the psychological process under study (how many outstanding questions have yet to be answered) could it be any other way? This state of affairs highlights the fact that models should be considered to be only approximations of the “true” theory. To expect a model to be correct on the first try is not only unrealistic, but impossible.

In contrast to predictions of verbal models, which are qualitative in nature and expressed verbally, the predictions made by mathematical models characterize quantitative relationships that clearly specify the effect on one variable that would result from a change in the other, and are expressed in, of course, mathematical language, i.e., equations. Translating a verbal prediction into a mathematical language is one of the first challenges of creating mathematical models.

To illustrate the process, we will examine a model of lexical decision making. The lexical decision task is a procedure used in many psychology and psycholinguistics experiments (Perea, Rosa & Gomez, 2002). The basic procedure involves measuring how

quickly participants can classify stimuli as words or non-words. It turns out that speed of classification depends on the frequency with which the stimulus word is used in the English language. A simple, verbal prediction for this task could be stated as “the higher the word frequency, the faster the response time.” This verbal prediction describes a qualitative relationship between word frequency and response time, whereby response time decreases monotonically as a function of word frequency. This qualitative relationship could be captured by many different mathematical functions, but different functions make different quantitative predictions that can be tested empirically.

There are many different models related to this task (see, e.g., Adelman, 2008). One example of a model for the lexical decision task is a power function, which uses the equation

$$RT = a(WF + 1)^{-b} + c$$

where RT is the response time measured in an appropriate unit, WF is the word frequency and a, b, and c are parameters (a, b, c > 0). That is, the response time is found by adding one to the word frequency, raising that value to the  $-b$  power, multiplying the result by the parameter a, and then adding the parameter c. Like all algebraic models, this one can be broken down into "observables", whose values are *a priori* known or obtained from an experiment, and "non-observables", which must be inferred from the observables. Here, the observables are RT and WF, while the non-observables are the three parameters a, b and c. A typical prediction of this model is illustrated in Figure 1.

Writing the model equation is an important first step in specifying the model, but it is not the end of the process. The next step is to account for random variability in the data. A naïve view of modeling is that the data would directly and perfectly reveal the

underlying process, but this view is unrealistic because people are neither perfect nor identical, which means that experiment data will inevitably contain random variability between participants and even within the data for individual participants. It is therefore important that a mathematical model specify not only the hypothesized regularity behind the data but also the error structure of the data. For example, the above power function for the lexical decision task could be made into a probabilistic model by adding an error term,  $e$ , yielding

$$RT = a(WF + 1)^{-b} + c + e$$

The error term  $e$  is a random variable whose value is drawn from a probability distribution, often a normal distribution, centered at zero and with variance  $\sigma^2$ . With the error term  $e$ , the model now predicts a data pattern in which the response times are not identical on every trial even with the same word frequency, but rather normally distributed with mean  $a(WF+1)^{-b} + c$ , and with the variance  $\sigma^2$ , as shown in Figure 1. Other error specifications are of course possible.

Technically speaking in more formal terms, a model is defined as a parameterized family of probability distributions  $M = \{f(y | w), w \in W\}$  where  $y = (y_1, \dots, y_n)$  is the data vector of  $n$  observations,  $w$  is the parameter vector defining model parameters (e.g.,  $w = (a, b, c)$  for the above power model), and  $f(y|w)$  is the probability density function (pdf) specifying the probability of observing  $y$  given  $w$ , and finally,  $W$  is the parameter space. From this viewpoint, the model consists of a collection of probability distributions indexed by its parameters so that each parameter value is associated with a probability distribution of responses.

### **Model Fitting**

Once a model has been fully specified with a model equation and an error structure, the next step is to assess its descriptive adequacy. The descriptive adequacy of a model is measured by how closely its predictions can be aligned with the observed pattern of data from an experiment. Given that the model can describe a range of data patterns by varying the values of its parameters, the first step in assessing the descriptive adequacy of a model is to find the set of parameter values for which the model fits the data “best” in some defined sense. This step is called parameter estimation.

There are two general methods of parameter estimation in statistics, *least-squares estimation* (LSE) and *maximum likelihood estimation* (MLE). Both of these methods are similar in spirit but differ from one another in implementation (see Myung, 2003, for a tutorial).

Specifically, the goal of LSE is to identify the parameter values that most accurately describe the data, whereas in MLE the goal is to find the parameter values that are most likely to have generated the data. LSE is tied with familiar statistical concepts in psychology such as the sum of squares error, the percent variance accounted for, and the root mean squared deviation. Formally, the LSE estimate, denoted by  $w_{LSE}$ , that minimizes the sum of squares error between observed and predicted data is obtained using the formula

$$w_{LSE} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n (y_{obs,i} - y_{prd,i}(w))^2$$

where the symbol ‘argmin’ stands for the argument of the minimum, referring to the argument value (i.e.,  $w$ ) that minimizes the given expression. The expression is a sum over  $n$  observations, indexed by  $i$ , of the squared difference between the value predicted

by the model and the actual observed value. LSE is primarily a descriptive measure, often associated with linear models with normal error.

On the other hand, MLE is the standard method of parameter estimation in statistics and forms a basis for many inferential statistical methods such as the chi-square test and several model comparison methods (described in the next section). The central idea in MLE estimation is the notion of the likelihood of the observed data given a parameter value. For each parameter value of a model, there is a corresponding likelihood that the model generated the data. Together, these likelihoods constitute the likelihood function of the model. The MLE estimate, denoted by  $w_{MLE}$ , is obtained by maximizing the likelihood function,

$$w_{MLE} = \underset{w}{\operatorname{argmax}} f(y_{obs} | w),$$

which entails finding the value of  $w$  that maximizes the likelihood of  $y_{obs}$  given  $w$ .

Figure 2 displays a hypothetical likelihood function for the power model of lexical decision, highlighting the model likelihoods of three parameter values.

It is not generally possible to find an analytic form solution (i.e., single equation) for the LSE or MLE estimate. As such, the solution must be sought numerically using search algorithms implemented on computer, such as the Newton-Raphson algorithm and the gradient descent algorithm (e.g., Press, Teukolsky, Vetterling & Flannery, 1992).

### **Model Comparison**

Specifying a mathematical model and justifying all of its assumptions is a difficult task. Completing it, and then going further to show that it provides an adequate fit to a set of experimental data, is a feat worthy of praise (and maybe a journal publication).

However, these steps are only the beginning of the journey. The next question to ask of

this model is, why ~~should anyone~~ should anyone should use it instead of someone else's model that also has justifiable assumptions and also fits the data well? This is the problem of model comparison, and it arises from what we discussed earlier about the logic of model testing, namely, that it is almost never possible to establish the necessity of a model (only the sufficiency), because someone can almost always come up with a competing model based on different assumptions that produces exactly the same predictions, and hence an equally good fit to the data. Given the difficulty in establishing the sufficiency of a model, how should we choose between differing explanations (i.e., models) given a finite sample of noisy observations?

The ultimate goal of model comparison is to identify, among a set of candidate models, the one that actually generated the data you are fitting. This is, however, not possible in general due to at least two difficulties in practice: (1) there are never enough observations in a data set to pin down the truth exactly and uniquely; and (2) the truth may be quite complex and beyond the descriptive power of any of the models under consideration. Given these limitations, a more realistic goal is to choose the model that provides the closest approximation to the truth in some defined sense.

In defining the “best” or “closest approximation”, there are many different model evaluation criteria to choose from (e.g., Jacobs & Grainger, 1994). Six of them are summarized in Table 1. Among these six criteria, three are qualitative and the other three are quantitative. In the rest of this section, we focus on the three quantitative criteria: goodness of fit, complexity or simplicity, and generalizability.

### ***Goodness of Fit, Complexity, and Generalizability***

The goodness of fit criterion (GOF) is defined as a model's best fit to the observed data, obtained by searching the model's parameter space for the best-fitting parameter values that maximize or minimize a specific objective function. The common measures of GOF include the *root mean squared error* (MSE), the *percent variance accounted for*, and the *maximized likelihood* (ML).

One cannot use GOF alone for comparing models because of what is called the over-fitting problem (Myung, 2000). Over-fitting arises when a model captures not only the underlying regularities in a dataset, which is good, but also random noise, which is not good. It is inevitable that behavioral data include random noise from a number of sources, including sampling error, human error, and individual differences, among others. A model's ability to fit that noise is meaningless because, being random, the noise pattern will be different from one data set to another. Fitting the noise reveals nothing of psychological relevance, and can actually hinder the identification of more meaningful patterns in the data.

Since GOF measures the model's fit to both regularity and noise, properties of the model that have nothing to do with its ability to fit the underlying regularity can improve GOF. One such property is complexity. Intuitively, complexity is defined as a model's inherent flexibility in fitting a wide range of data patterns (Myung & Pitt, 1997). It can be understood by contrasting the data-fitting capabilities of simple and complex models. A simple model will have few parameters and make clear and easily falsifiable predictions. A simple model predicts that a specific pattern will be found in the data, and if this pattern is found then the model will fit well, otherwise it will fit poorly. On the other hand, a complex model will have many more parameters, making it more flexible and

able to predict with high accuracy many different data patterns by finely tuning those parameters. A highly complex model is not easily falsifiable because its parameters can be tuned to fit almost any pattern of data including random noise. As such, a complex model can often provide superior fits by capitalizing on random noise, which is specific to the particular data sample, but not necessarily by capturing the regularity underlying the data.

What is desired in model comparison is a yardstick by which a model is measured by its ability to capture the underlying regularity only, not idiosyncratic noise. This is the *generalizability* criterion (Pitt, Myung & Zhang, 2002). Generalizability refers to a model's ability to fit the current data sample (i.e., actual observations) and all 'future' data samples (i.e., replications of the experiment) from the same underlying process that generated the current data. Generalizability is often called predictive accuracy or generality (Hitchcock & Sober, 2004). An important goal of modeling is to identify hypotheses that generate accurate predictions; hence the goal of model comparison is to choose the model that best generalizes, not the one that provides the best fit to a single data set.

The relationship between complexity and generalizability is illustrated in Figure 3, which shows the fits of three different models to a dataset from a lexical decision experiment. The linear model (top left graph) under-fits the data because it does not have sufficient complexity to capture the underlying regularities. When under-fitting occurs, increasing the complexity of the model not only improves GOF, it will also improve generalizability because the additional complexity captures unaccounted-for, underlying regularities in the data. However, too much complexity, as in the Spline model (top right

graph), will cause the model to pick up on not just the underlying regularities, but also idiosyncratic noise that does not generalize to future datasets (bottom graph). This will result in overfitting and reduce generalizability. Thus the dilemma in trying to maximize generalizability is a delicate balance between complexity and goodness of fit.

To summarize, what is needed in model comparison is a method that estimates a model's generalizability by taking into account the effects of its complexity. Various measures of generalizability have been proposed in statistics, which we discuss next. For more thorough treatments of the topic, the reader is directed to two Journal of Mathematical Psychology special issues (Myung, Forester & Browne, 2000; Wagenmakers & Waldorp, 2006) and a recent review article (Shiffrin, Lee, Kim & Wagenmakers, 2008).

### ***Methods of Model Comparison***

Akaike Information Criterion and Bayesian Information Criterion: The *Akaike Information Criterion* (AIC; Akaike, 1973) and the *Bayesian Information Criterion* (BIC; Schwartz, 1978) address the most salient dimension of model complexity, the number of free parameters, and are defined as

$$AIC = -2 \ln f(y_{obs} | w_{MLE}) + 2k$$

$$BIC = -2 \ln f(y_{obs} | w_{MLE}) + k \ln n.$$

The first term in each of these expressions assesses the model's goodness of fit (as -2 times the natural logarithm of the value of the likelihood function at the MLE estimate), while the second term penalizes the model for complexity. Specifically, the second term includes a count of the number of parameters,  $k$ . AIC and BIC penalize a model more as the number of parameters increases. Under each criterion, the smaller the criterion value is, the better the model is judged to generalize. Consequently, to be selected as more

generalizable, a more complex model must overcome this penalty with a much better goodness of fit to the data than the simpler model with fewer parameters.

Bayesian Model Selection and Minimum Description Length: Another feature that affects model complexity is functional form, which refers to the way in which the model's parameters are combined in the model equation. More sophisticated selection methods, such as *Bayesian Model Selection* (BMS; Kass & Raftery, 1995; Wasserman, 2000) and *Minimum Description Length* (MDL; Rissanen, 1996; Pitt, Myung & Zhang, 2002; Hansen & Yu, 2001) are sensitive to a model's functional form as well as the number of parameters, and are defined as

$$BMS = -\ln \int f(y_{obs} | w) \pi(w) dw$$

$$MDL = -\ln f(y_{obs} | w_{MLE}) + \frac{k}{2} \ln \left( \frac{n}{2\pi} \right) + \ln \int \sqrt{|I(w)|} dw$$

Where  $\pi(w)$  is the parameter prior and  $I(w)$  is the Fisher information matrix. The effects of functional form on model complexity are reflected in the third term of the MDL equation, while in BMS it is hidden inside the integral.

Cross-Validation and Accumulative Prediction Error: Two other measures, *Cross-Validation* (CV; Browne, 2000) and *Accumulative Prediction Error* (APE; Dawid, 1984; Wagenmakers, Grunwald & Steyvers, 2006) assess generalizability by actually evaluating the model's performance against "future" data. The basic idea of CV is to partition the data sample into two complementary subsets. One subset, called the training or calibration set, is used to fit the model via LSE or MLE. The other subset, called the validation set, is treated as a "future" dataset and is used to test the estimates from the training set. If the parameters estimated from the training set also provide a good fit to the validation set, then the conclusion is that the model generalizes well.

APE is similar to CV in spirit but differs from it in implementation. In APE, the size of the training set is increased successively one observation at a time while maintaining the size of the validation set fixed to one. The litmus test for generalizability is performed by assessing how well the model predicts the next “unseen” data point  $y_{\text{obs},j+1}$  using the best-fit parameter value obtained based on the first  $j$  observations  $\{y_{\text{obs},1}, y_{\text{obs},2}, \dots, y_{\text{obs},j}\}$  for  $j = k+1, \dots, n-1$ . APE then estimates the model’s generalizability by the sum of the prediction errors for the validation data.

Both CV and APE are thought to be sensitive to number of parameters as well as functional form.

### **Model Revision**

When a model is found to be inappropriate, in terms of a lack of fit or lack of generalizability, steps must be taken to revise it, perhaps substantially, or even replace it with a new model (Shiffrin & Nobel, 1997, p. 7). This could be emotionally difficult for the investigator, especially if the person has invested substantial resources into developing the model (e.g., years of work). In these situations, it is best to put aside personal attachment and make the goals of science paramount.

In the words of Meehl (1990), "Even the best theories are likely to be approximations of reality." However, mathematical models can still be very useful, even in this limited capacity. Many people have heard the famous quote "All models are false, but some are useful" credited to George E. P. Box, (1975). The nature of that usefulness is summed up by Samuel Karlin (1983), who says, "The purpose of models is not to fit the data but to sharpen the questions". In a sense, a model is only as valuable as the insights it provides and the research hypotheses that it generates. This means that

mathematical models are not ends in themselves, but rather steps on the road to scientific understanding. We will always need new models to expand on the knowledge and insights gained from previous models.

One viable approach in model revision is to selectively add and remove relevant features to and from the model. In taking this course of action, one should be mindful of the important but often neglected issues of *model faithfulness* (Myung et al., 1999) and *irrelevant specification* (Lewandowsky, 1993). Model faithfulness refers to the question of whether a model's success in mimicking human behavior due to the theoretical principles embodied in the model or merely due to its computational instantiation. In other words, even if a model provides an excellent description of human data in the simplest manner possible, it is often difficult to determine whether the theoretical principles that the model originally intended to implement are critical for its performance, or if less central choices in model instantiation are instead responsible for good performance.

Irrelevant specification, which is similar to the concept of model faithfulness, refers to the case in which a model's performance is strongly affected by irrelevant modeling details that are theoretically neutral and fully interchangeable with any viable alternatives. Examples of irrelevant details include input coding methods, the specification of error structure, and idiosyncratic features of the simulation schedule (Fum et al, 2007).

### **Conclusion**

The science of mathematical modeling involves converting the ideas, assumptions, and principles embodied in psychological theory into mathematical

abstraction. Mathematics is used to craft precise representations of human behavior. The specificity inherent in models opens up new avenues of research. Their usefulness is evident in the rapid rate at which models are appearing in psychology, as well as in related fields such as human factors, behavioral economics, and cognitive neuroscience. Mathematical modeling has become an essential tool for understanding human behavior, and any researcher with an inclination toward theory building would be well served to begin practicing it.

### **Future Directions**

Mathematical modeling has contributed substantially to advancing the study of mind and brain. Modeling has opened up new ways of thinking about problems, provided a framework for studying complex interactions among causal and correlational variables, provided insight needed to tie together seemingly inconsistent findings, and increased the precision of prediction in experimentation.

Despite these advances, for the field to move forward and beyond the current state of affairs, there remain many challenges to overcome and problems to be solved. Below we list four challenges for the next decade of mathematical modeling:

1. At present, mathematical modeling is confined to a relatively small group of mostly self-selected researchers. To impact the mainstream of psychological science, an effort should be made to ensure that frontline psychologists learn to practice the art of modeling. Examples of such efforts include writing tutorial articles in journals and publishing graduate-level textbooks.

2. Modeling begins in a specific domain, whether it be a phenomenon, task, or process. Modelers eventually face the challenge of expanding the scope of their models to explain performance on other tasks, account for additional phenomena, or to bridge multiple levels of description (e.g., brain activity and behavior responses). Model expansion is difficult because the perils of complexity multiply. The development of methods for doing so will be an important step in the discipline.
3. Model faithfulness, discussed above, concerns determining what properties of a model are critical for explaining human performance and what properties serve lesser roles. Failure to make this distinction runs the risk of erroneously attributing a model's behavior to its underlying theoretical principles. In the worst case, computational complexity is mistaken for theoretical accuracy. A method should be developed to formalize and assess a model's faithfulness such that the relative contribution of each modeling assumption to the model's data-fitting ability is quantified in some justifiable sense.
4. Models can be difficult to discriminate experimentally because of their complexity and the extent to which they mimic each other. A method for identifying an "optimal" experimental design that would produce the most informative, differentiating outcome between the models of interest needs to be developed. Related to this, quantitative methods of model comparison have their limits. Empirical data alone may not be sufficient to discriminate highly similar models. Modeling would benefit from the introduction of new and more

powerful measures of model adequacy. In particular, it would be desirable to quantify the qualitative dimensions described in Table 1.

**Author Note**

This research is supported in part by National Institute of Health Grant R01-MH57472.

### References

- Aczel, J. (1966). *Lectures on Functional Equations and their Applications*. Academic Press.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Caski, F., editors, *Proceedings of the Second International Symposium on Information Theory*, pages 267-281, Budapest. Akademiai Kiado.
- Bakan, D. (1966). Statistical significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Batchelder, W. H. and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review*, 6, 57-86.
- Brown, G. D. A., Neath, I. and Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114, 539-576.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108-132.
- Busemeyer, J. R. and Diederich, A. (2010). *Cognitive Modeling*. Sage Publications.
- Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100, 432-459.
- Chater, N., Tenenbaum, J., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287-291.
- Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147, 278-292.

- Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, 9: 3-25.
- Fishburn, P. (1982). *The Foundations of Expected Utility*. Springer.
- Fum, D., Del Missier, F., and Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, 8, 135-142.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis (2<sup>nd</sup> edition)*. Chapman & Hall/CRC.
- Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, 98, 254-267.
- Gill, J. (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach (2<sup>nd</sup> edition)*. Chapman & Hall/CRC.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244
- Hansen, M. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96, 746-774.
- Hitchcock, C. and Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55, 1-34.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-368.
- Hornik, K., Stinchcombe, M., and White, H. (1990). Universal approximations of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3, 551-560.

- Howard, M. and Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269-299.
- Jacobs, A. and Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1311-1334.
- Karlin, S. (1983). The 11<sup>th</sup> R. A. Fisher Memorial Lecture given at the Royal Society 20 Meeting in April, 1983.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Kruschke, J. K. (1992). Alcové: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14, 293-300.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1-15
- Lee, M. D. (in press). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*.
- Lee, M. D., and Sarnecka, B. W. (2010). A model of knower-level behavior in number-concept development. *Cognitive Science*, 34, 51-67.
- Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, 4, 236-243.
- Luce, R. D. (2000). *Utility of Gains and Losses: Measurement-theoretical and Experimental Approaches*. Lawrence Erlbaum.

- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modeling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10, 325 - 337.
- Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159.
- Marewski, J. N. and Olsson, H. (2009). Beyond the null ritual. *Zeitschrift fur Psychologie*, 217, 49-60.
- McClelland, J. L. and Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18, 1-86.
- Meehl, P. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108-141.
- Morgenstern, O. and Von Neumann, J. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44, 190-204.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90-100.
- Myung, I., J. Brunsmann IV, A., and Pitt, M. A. (1999). True to thyself: Assessing whether computational models of cognition remain faithful to their theoretical principles. In M. Hahn and S. C. Stoness (eds.), *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, pp. 462-467. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Myung, I. J., Forster, M., and Browne, M. W. (2000). Special issue on model selection.

*Journal of Mathematical Psychology*, 44, 1-2.

Myung, I. J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A

Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79-95.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In J. Wixted (ed.), *Stevens'*

*Handbook of Experimental Psychology (Third Edition), Volume IV*

(*Methodology*), pp. 429-459. New York, NY: John Wiley & Sons.

Nickerson, R. (2000). Null hypothesis statistical testing: A review of an old and

continuing controversy. *Psychological Methods*, 5, 241-301.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization

relationship. *Journal of Experimental Psychology: General*, 115, 39-57.

Perea, M., Rosa, E., and Gomez, C. (2002). Is the go/no-go lexical decision task an

alternative to the yes/no lexical decision task? *Memory & Cognition*, 30, 34-45.

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among

computational models of cognition. *Psychological Review*, 109, 472-491.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. and Patterson, K. (1996).

Understanding normal and impaired word reading: Computational principles in

quasi-regular domains. *Psychological Review*, 103, 56-115.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. R. (1992). *Numerical*

*Recipes in C: The Art of Scientific Computing (2<sup>nd</sup> edition)*. Cambridge University

Press.

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

- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transaction on Information Theory*, 42, 40-47
- Rouder, J. and 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., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68, 589-606.
- Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Rubin, D. and Wenzel, A. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103, 734-760.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: Remembering effectively from memory. *Psychonomic Bulletin & Review*, 4, 145-166.
- Shiffrin, R. M., Lee, M. D., Kim, W. and Wagenmakers, E-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.
- Stevens, S. (1975). *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley & Sons, New York.

- Steyvers, M., Lee, M. D., and Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168-179.
- Suppes, P. (1957). *Introduction to Logic*. Dover Publications, Mineola, N.Y.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309-318
- Thurstone, L. (1974). *The Measurement of Values*. The University of Chicago Press, Chicago.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.
- Wagenmakers, E.-J., Grunwald, P., and Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149-166.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H. and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158-189.
- Wagenmakers, E.-J. and Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, *50*, 99-100.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92-107.
- Wetzels, R., Raaijmakers, J., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WINBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review*, *16*, 752-760.

Wixted, J., and Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, 2, 409-415.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, 60, 107-126.

### Figure Captions

**Figure 1:** Behavior predicted by the power model of lexical decisions with  $a = 0.78$ ,  $b = 0.50$ ,  $c = 0$ , and  $\sigma^2 = 0.10$ .

**Figure 2:** A hypothetical likelihood function. The curve indicates the likelihood of the model (y-axis) for each possible parameter value (x-axis). In this case, the parameter ranges from zero to 12. This likelihood function has local maxima at 2.8, 9.2, and 12. The global maximum, (MLE) is at 6.5. When using an automated search algorithm to find the MLE, it is important to avoid getting stuck in a local maximum.

**Figure 3:** Top row: three models of the lexical decision task with their fits to a fictitious dataset -- from left to right: linear model, power model, Spline model. Bottom: generalizability of the fits in the top row to the data from a different participant.

**Table 1:** Criteria for comparing models.

<b>Criterion</b>	<b>Description</b>	<b>Measurement</b>
Falsifiability	Do potential observations exist that would be incompatible with the model?	Qualitative
Plausibility	Does the theoretical account of the model make sense of established findings?	Qualitative
Interpretability	Are the components of the model understandable and linked to known processes?	Qualitative
Goodness of fit	Does the model fit the observed data sufficiently well?	Quantitative
Complexity	Is the model's description of the data achieved in the simplest possible manner?	Quantitative
Generalizability	Does the model provide a good prediction of future observations?	Quantitative

Figure 1.

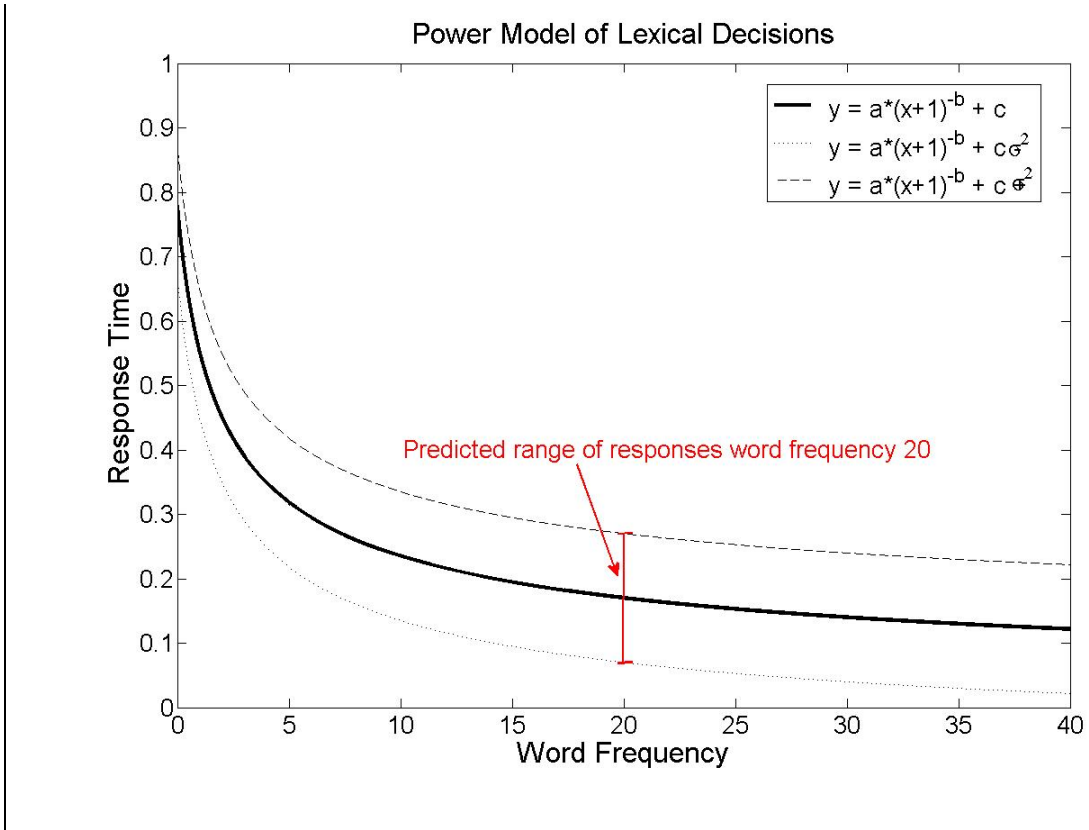


Figure 2.

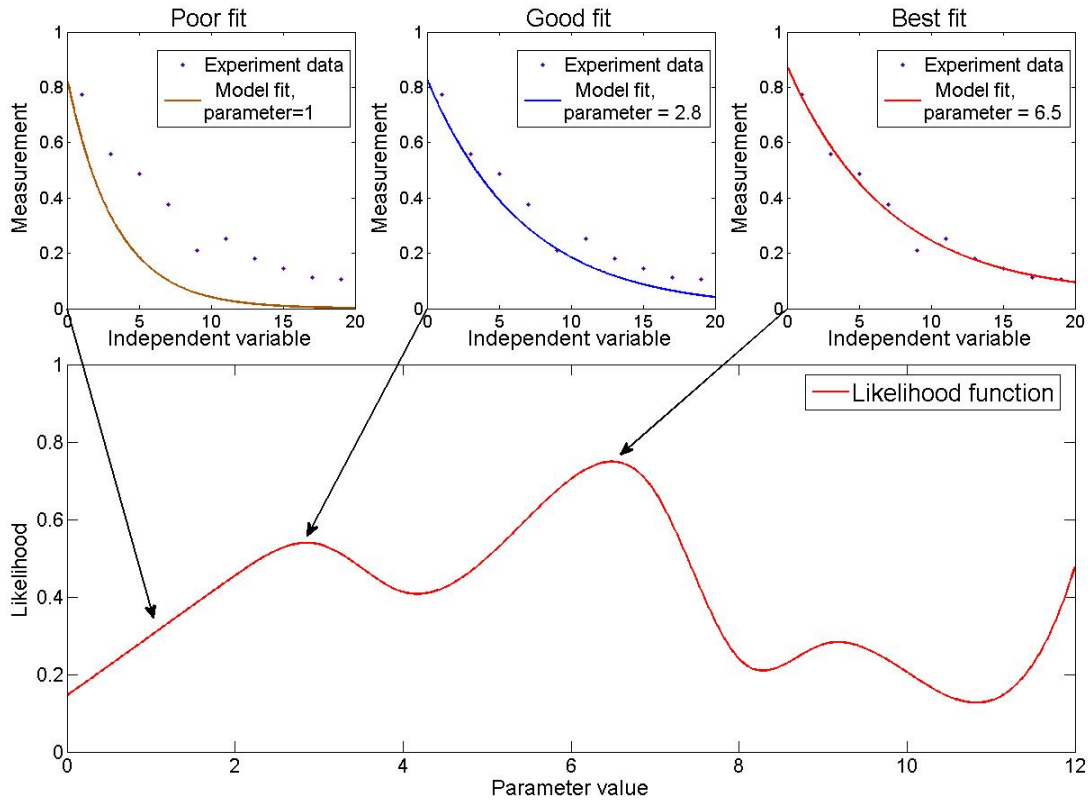


Figure 3.

