1 Introduction

1.1 Models and Theories in Science

Cognitive scientists seek to understand how the mind works. That is, we want to *describe* and *predict* people's behavior, and we ultimately wish to *explain* it, in the same way that physicists predict the motion of an apple that is dislodged from its tree (and can accurately describe its downward path) and explain its trajectory (by appealing to gravity). For example, if you forget someone's name when you are distracted seconds after being introduced to her, we would like to know what cognitive process is responsible for this failure. Was it lack of attention? Forgetting over time? Can we know ahead of time whether or not you will remember that person's name?

The central thesis of this book is that to answer questions such as these, cognitive scientists must rely on quantitative mathematical models, just like physicists who research gravity. We suggest that to expand our knowledge of the human mind, consideration of the data and verbal theorizing are insufficient on their own.

This thesis is best illustrated by considering something that is (just a little) simpler and more readily understood than the mind. Have a look at the data shown in Figure 1.1, which represent the position of planets in the night sky over time.

How might one describe this peculiar pattern of motion? How would you explain it? The strange loops in the otherwise consistently curvilinear paths describe the famous "retrograde motion" of the planets—that is, their propensity to suddenly reverse direction (viewed against the fixed background of stars) for some time before resuming their initial path. What explains retrograde motion? It took more than a thousand years for a satisfactory answer to that question to become available, when Copernicus replaced the geocentric Ptolemaic system with a heliocentric model: Today, we know that retrograde motion arises from the fact that the planets travel at different speeds along their orbits; hence, as Earth "overtakes" Mars, for example, the red planet will appear to reverse direction as it falls behind the speeding Earth.

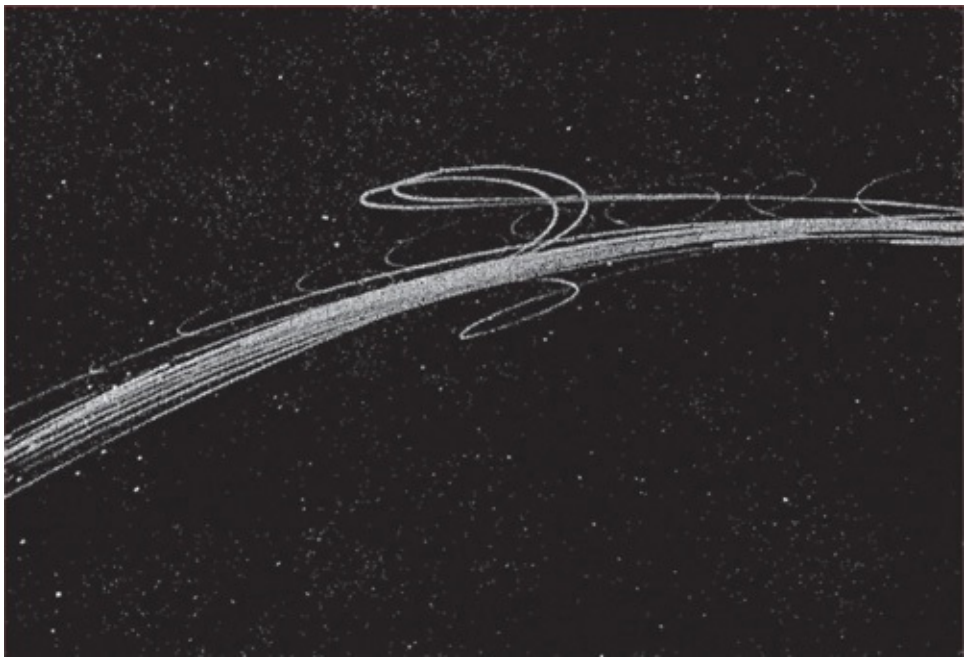

Figure 1.1 An example of data that defy easy description and explanation without a quantitative model.

This example permits several conclusions that will be relevant throughout the remainder of this book. First, the pattern of data shown in Figure 1.1 defies description and explanation unless one has a *model* of the underlying process. It is only with the aid of a model that one can describe and explain planetary motion, even at a verbal level (readers who doubt this conclusion may wish to invite friends or colleagues to make sense of the data without knowing their source).

Second, any model that explains the data is itself unobservable. That is, although the Copernican model is readily communicated and represented (so readily, in fact, that we decided to omit the standard figure showing a set of concentric circles), it cannot be directly observed. Instead, the model is an abstract explanatory device that "exists" primarily in the minds of the people who use it to describe, predict, and explain the data.

Third, there nearly always are *several* possible models that can explain a given data set. This point is worth exploring in a bit more detail. The overwhelming success of the heliocentric model often obscures the fact that, at the time of Copernicus's discovery, there existed a moderately successful alternative—namely, the geocentric model of Ptolemy shown in Figure 1.2. The model explained retrograde motion by postulating that while orbiting around the Earth, the planets also circle around a point along their orbit. On the additional, arguably somewhat inelegant, assumption that the Earth is slightly offset from the center of the planets' orbit, this model provides a reasonable account of the data, limiting the positional discrepancies between predicted and actual locations of, say, Mars to about 1° (Hoyle, 1974). Why, then, did the heliocentric model so rapidly and thoroughly replace the Ptolemaic system?[1]

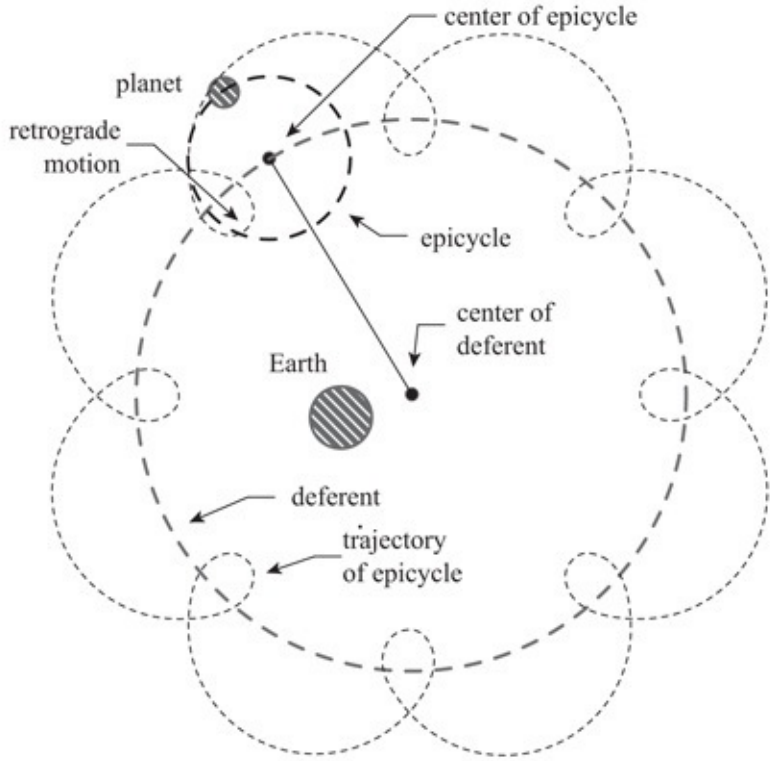

Figure 1.2 The geocentric model of the solar system developed by Ptolemy. It was the predominant model for some 1,300 years.

The answer to this question is quite fascinating and requires that we move toward a *quantitative* level of modeling.

1.2 Why Quantitative Modeling?

Conventional wisdom holds that the Copernican model replaced geocentric notions of the solar system because it provided a better account of the data.But what does "better" mean? Surely it means that the Copernican system predicted the motion of planets with less quantitative error —that is, less than the 1° error for Mars just mentioned—than its Ptolemaic counterpart? Intriguingly, this conventional wisdom is only partially correct: Yes, the Copernican model predicted the planets' motion in latitude better than the Ptolemaic theory, but this difference was slight compared to the overall success of both models in predicting motion in longitude (Hoyle, 1974). What gave Copernicus the edge, then, was not "goodness of fit" alone[2] but also the intrinsic elegance and simplicity of his model—compare the Copernican account by a set of concentric circles with the complexity of Figure 1.2, which only describes the motion of a single planet.

There is an important lesson to be drawn from this fact: The choice among competing models—and remember, there are always several to choose from—inevitably involves an *intellectual judgment* in addition to quantitative examination. Of course, the quantitative performance of a model is at least as important as are its intellectual attributes. Copernicus would not be commemorated today had the predictions of his model been *inferior* to those of

Ptolemy; it was only because the two competing models were on an essentially equal quantitative footing that other intellectual judgments, such as a preference for simplicity over complexity, came into play.

If the Ptolemaic and Copernican models were quantitatively comparable, why do we use them to illustrate our central thesis that a purely verbal level of explanation for natural phenomena is insufficient and that all sciences must seek explanations at a quantitative level? The answer is contained in the crucial modification to the heliocentric model offered by Johannes Kepler nearly a century later. Kepler replaced the circular orbits in the Copernican model by ellipses with differing eccentricities (or "egg-shapedness") for the various planets. By this straightforward mathematical modification, Kepler achieved a virtually perfect fit of the heliocentric model with near-zero quantitative error. There no longer was any appreciable quantitative discrepancy between the model's predictions and the observed paths of planets. Kepler's model has remained in force essentially unchanged for more than four centuries.

The acceptance of Kepler's model permits two related conclusions, one that is obvious and one that is equally important but perhaps less obvious. First, if two models are equally simple and elegant (or nearly so), the one that provides the better quantitative account will be preferred. Second, the predictions of the Copernican and Keplerian models cannot be differentiated by verbal interpretation alone. Both models explain retrograde motion by the fact that Earth "over-takes" some planets during its orbit, and the differentiating feature of the two models—whether orbits are presumed to be circular or elliptical—does not entail any differences in predictions that can be appreciated by purely verbal analysis. That is, although one can talk about circles and ellipses (e.g., "one is round, the other one egg shaped"), those verbalizations cannot be turned into testable predictions: Remember, Kepler reduced the error for Mars from 1° to virtually zero, and we challenge you to achieve this by verbal means alone.

Let us summarize the points we have made so far:

1. Data never speak for themselves but require a model to be understood and to be explained.
2. Verbal theorizing alone ultimately cannot substitute for quantitative analysis.
3. There are always several alternative models that vie for explanation of data, and we must select among them.
4. Model selection rests on both quantitative evaluation and intellectual and scholarly judgment.

All of these points will be explored in the remainder of this book. We next turn our attention from the night sky to the inner workings of our mind, first by showing that the preceding conclusions apply in full force to cognitive scientists and then by considering an additional issue that is of particular concern to scholars of the human mind.

1.3 Quantitative Modeling in Cognition

1.3.1 Models and Data

Let's try this again: Have a look at the data in Figure 1.3. Does it remind you of planetary motion? Probably not, but it should be at least equally challenging to discern a meaningful pattern in this case at it was in the earlier example. Perhaps the pattern will become recognizable if we tell you about the experiment conducted by Nosofsky (1991) from which these data are taken. In that experiment, people were trained to classify a small set of cartoon faces into two arbitrary categories (we might call them the Campbells and the MacDonalds, and members of the two categories might differ on a set of facial features such as length of nose and eye separation).

On a subsequent transfer test, people were presented with a larger set of faces, including those used at training plus a set of new ones. For each face, people had to make two decisions: which category the face belonged to and the confidence of that decision (called "classification" in the figure, shown on the *x*-axis), and whether or not it had been shown during training ("recognition," on the *y*-axis). Each data point in the figure, then, represents those two responses, averaged across participants, for a given face (identified by ID number, which can be safely ignored). The correlation between those two measures was found to be $r = .36$.

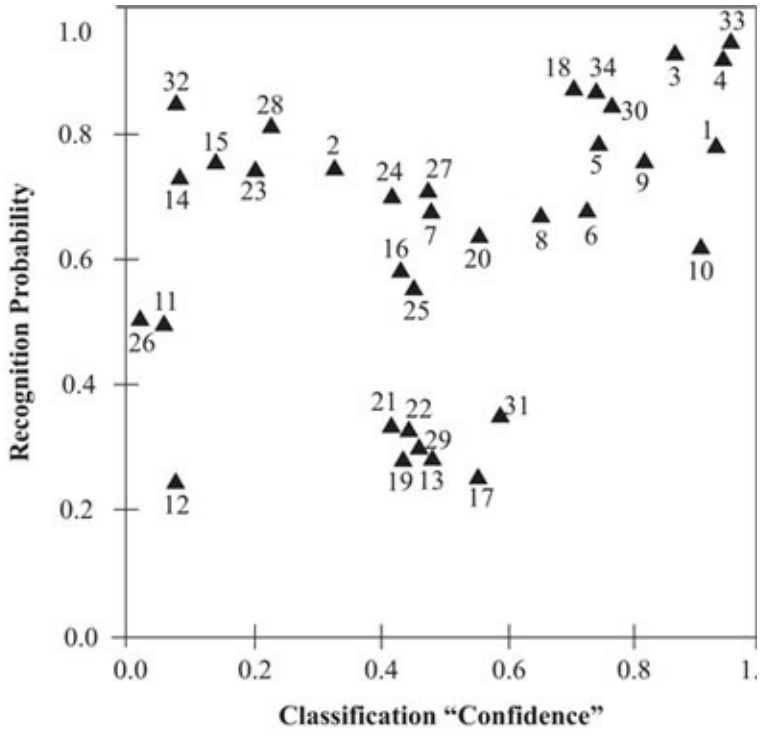

Figure 1.3 Observed recognition scores as a function of observed classification confidence for the same stimuli (each number identifies a unique stimulus). See text for details. Figure reprinted from Nosofsky, R. M. (1991). Tests of an exemplar mode for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 3–27. Published by the American Psychological Association; reprinted with permission.

Before we move on, see if you can draw some conclusions from the pattern in Figure 1.3. Do you think that the two tasks have much to do with each other? Or would you think that classification and recognition are largely unrelated and that knowledge of one response would

tell you very little about what response to expect on the other task? After all, if $r = .36$, then knowledge of one response reduces uncertainty about the other one by only 13%, leaving a full 87% unexplained, right?

Wrong. There is at least one quantitative cognitive model (called the GCM and described a little later), which can relate those two types of responses with considerable certainty. This is shown in Figure 1.4, which separates classification and recognition judgments into two separate panels, each showing the relationship between observed responses (on the *y*-axis) and the predictions of the GCM (*x*-axis). To clarify, each point in Figure 1.3 is shown twice in Figure 1.4—once in each panel and in each instance plotted as a function of the *predicted* response obtained from the model.

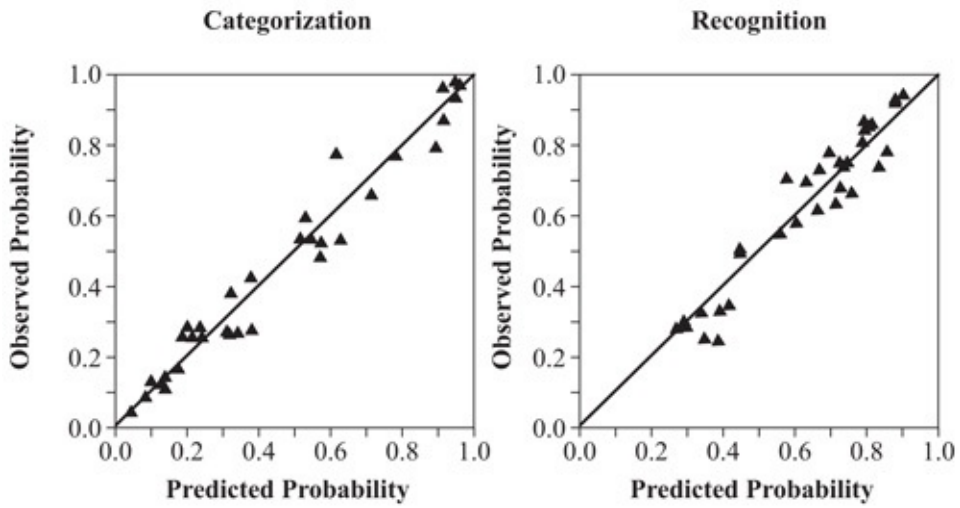

Figure 1.4 Observed and predicted classification (left panel) and recognition (right panel). Predictions are provided by the GCM; see text for details. Perfect prediction is represented by the diagonal lines. Figure reprinted from Nosofsky, R. M. (1991). Tests of an exemplar mode for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27. Published by the American Psychological Association; reprinted with permission.

The precision of predictions in each panel is remarkable: If the model's predictions were absolutely 100% perfect, then all points would fall on the diagonal. They do not, but they come close (accounting for 96% and 91% of the variance in classification and recognition, respectively). The fact that these accurate predictions were provided by the same model tells us that classification and recognition can be understood and related to each other within a common psychological theory. Thus, notwithstanding the low correlation between the two measures, there is an underlying model that explains how both tasks are related and permits accurate prediction of one response from knowledge of the other. This model will be presented in detail later in this chapter (Section 1.4.4); for now, it suffices to acknowledge that the model relies on the comparison between each test stimulus and all previously encountered exemplars in memory.

The two figures enforce a compelling conclusion: "The initial scatterplot … revealed little relation between classification and recognition performance. At that limited level of analysis, one might have concluded that there was little in common between the fundamental processes of classification and recognition. Under the guidance of the formal model, however, a unified

account of these processes is achieved" (Nosofsky, 1991, p. 9). Exactly paralleling the developments in 16th-century astronomy, data in contemporary psychology are ultimately only fully interpretable with the aid of a quantitative model. We can thus reiterate our first two conclusions from above and confirm that they apply to cognitive psychology in full force—namely, that *data never speak for themselves but require a model to be understood and to be explained* and that *verbal theorizing alone cannot substitute for quantitative analysis*. But what about the remaining earlier conclusions concerning model selection?

Nosofsky's (1991) modeling included a comparison between his favored exemplar model, whose predictions are shown in Figure 1.4, and an alternative "prototype" model. The details of the two models are not relevant here; it suffices to note that the prototype model compares a test stimulus to the *average* of all previously encountered exemplars, whereas the exemplar model performs the comparison one by one between the test stimulus and each exemplar and sums the result.[3] Nosofsky found that the prototype model provided a less satisfactory account of the data, explaining only 92% and 87% of the classification and recognition variance, respectively, or about 5% less than the exemplar model. Hence, the earlier conclusions about model selection apply in this instance as well: There were several alternative models, and the choice between them was based on clear quantitative criteria.

1.3.2 From Ideas to Models

So far, we initiated our discussions with the data and we then …poof!…revealed a quantitative model that spectacularly turned an empirical mystery or mess into theoretical currency. Let us now invert this process and begin with an idea, that is, some psychological process that you think might be worthy of exploration and perhaps even empirical test. Needless to say, we expect you to convert this idea into a quantitative model. This raises at least two obvious questions: First, how would one do this? Second, does this process have implications concerning the role of modeling other than those we have already discussed? These questions are sufficiently complex to warrant their own chapter (Chapter 2), although we briefly survey the latter here.

Consider the simple and elegant notion of rehearsal, which is at the heart of much theorizing in cognition (e.g., A. D. Baddeley, 2003). We have all engaged in rehearsal, for example, when we try to retain a phone number long enough to enter it into our SIM cards. Several theorists believe that such subvocal—or sometimes overt—rehearsal can prevent the "decay" of verbal short-term memory traces, and introspection suggests that repeated recitation of a phone number is a good means to avoid forgetting. Perhaps because of the overwhelming intuitive appeal of the notion and its introspective reality, there have been few if any attempts to embody rehearsal in a computational model. It is therefore of some interest that one recent attempt to explicitly model rehearsal (Oberauer & Lewandowsky, 2008) found it to be detrimental to memory performance under many circumstances rather than beneficial. Specifically, because rehearsal necessarily involves retrieval from memory—how else would an item be articulated if not by retrieving it from memory?—it is subject to the same vagaries that beset memory retrieval during regular recall. In consequence, repeated rehearsal is likely to first introduce and then compound retrieval errors, such as ordinal transpositions of list items, thus likely

offsetting any benefit that might be derived from restoring the strength of rehearsed information. Oberauer and Lewandowsky (2008) found that the exact consequences of rehearsal depended on circumstances—in a small number of specific conditions, rehearsal was beneficial—but this only amplifies the point we are making here: Even intuitively attractive notions may fail to provide the desired explanation for behavior once subjected to the rigorous analysis required by a computational model.[4] As noted by Fum, Del Missier, and Stocco (2007), "Verbally expressed statements are sometimes flawed by internal inconsistencies, logical contradictions, theoretical weaknesses and gaps. A running computational model, on the other hand, can be considered as a sufficiency proof of the internal coherence and completeness of the ideas it is based upon" (p. 136). In Chapter 2, we further explore this notion and the mechanics of model development by developing a computational instantiation of Baddeley's (e.g., 2003) rehearsal model.

Examples that underscore the theoretical rigor afforded by quantitative models abound: Lewandowsky (1993) reviewed one example in detail that involved construction of a model of word recognition. Shiffrin and Nobel (1997) described the long and informative behind-the-scenes history of the development of a model of episodic recognition.

Finally, theoreticians who ignore the rigor of quantitative modeling do so at their own peril. Hunt (2007) relates the tale of the 17th-century Swedish king and his desire to add another deck of guns to the *Vasa*, the stupendous new flagship of his fleet. What the king wanted, the king got, and the results are history: The *Vasa* set sail on her maiden voyage and remained proudly upright for, well, nearly half an hour before capsizing and sinking in Stockholm harbor. Lest one think that such follies are the preserve of heads of state, consider the claim in a textbook on learning: "While adultery rates for men and women may be equalizing, men still have more partners than women do, and they are more likely to have one-night stands; the roving male seeks sex, the female is looking for a better partner" (Leahey & Harris, 1989, pp. 317–318). Hintzman (1991) issued a challenge to set up a model consistent with this claim—that is, "there must be equal numbers of men and women, but men must have more heterosexual partners than women do" (p. 41). Needless to say, the challenge has not been met because the claim is mathematically impossible; the obvious lesson here is that verbal theories may not only be difficult to implement, as shown by Oberauer and Lewandowsky (2008), but may even turn out to be scientifically untenable.

1.3.3 Summary

We conclude this section by summarizing our main conclusions:

1. Data never speak for themselves but require a model to be understood and to be explained.
2. Verbal theorizing alone cannot substitute for quantitative analysis.
3. There are always several alternative models that vie for explanation of data, and we must compare those alternatives.
4. Model comparison rests on both quantitative evaluation and intellectual and scholarly

judgment.
5. Even seemingly intuitive verbal theories can turn out to be incoherent or ill-specified.
6. Only instantiation in a quantitative model ensures that all assumptions of a theory have been identified and tested.

If you are interested in expanding on these conclusions and finding out more about fascinating aspects of modeling, we recommend that you consider the studies by Estes (1975), Lewandowsky (1993), Lewandowsky and Heit (2006), Norris (2005), and Ratcliff (1998).

1.4 The Ideas Underlying Modeling and Its Distinct Applications

We have shown that quantitative modeling is an indispensable component of successful research in cognition. To make this point without getting bogged down in too many details, we have so far sidestepped a number of fundamental issues. For example, we have yet to define what a model actually is and what common ground all psychological models may share—and, conversely, how they might differ. We now take up those foundational issues.[5]

1.4.1 Elements of Models

What exactly is a model, anyway? At its most basic, a model is an abstract structure that captures structure in the data (cf. Luce, 1995). For example, a good model for the set of numbers {2, 3, 4} is their mean—namely, 3. A good model for the relationship between a society's happiness and its economic wealth is a negatively accelerated function, such that happiness rises steeply as one moves from poverty to a modest level of economic security, but further increases in happiness with increasing material wealth get smaller and smaller as one moves to the richest societies (Inglehart, Foa, Peterson, & Welzel, 2008). Those models are *descriptive* in nature, and they are sufficiently important to merit their own section (Section 1.4.2).

Needless to say, scientists want to do more than describe the data. At the very least, we want to *predict* new observations; for example, we might want to predict how much happiness is likely to increase if we manage to expand the gross national product by another zillion dollars (if you live in a rich country, the answer is "not much"). In principle, any type of model permits prediction, and although prediction is an important part of the scientific endeavor (and probably the only ability of consideration for stockbrokers and investment bankers), it is not the whole story. For example, imagine that your next-door neighbor, a car mechanic by trade, were able to predict with uncanny accuracy the outcome of every conceivable experiment on some aspect of human cognition (a scenario discussed by K. I. Forster, 1994). Would you be satisfied with this state of affairs? Would your neighbor be a good model of human cognition? Clearly the answer is no; in addition to robotic predictions, you also want an *explanation* for the phenomena under consideration (Norris, 2005). *Why* does this particular outcome obtain in that experiment rather than some other result?

It follows that most cognitive modeling goes beyond mere description and seeks to permit prediction and explanation of behavior. The latter, explanatory role is the exclusive domain of models that we refer to as providing a process characterization and process explanation, respectively.

When models are used as an explanatory device, one other attribute becomes particularly relevant: Models are intended to be simpler and more abstract versions of the system—in our case, human cognition—they are trying to explain (Fum et al., 2007). Models seek to retain the essential features of the system while discarding unnecessary details. By definition, the complexity of models will thus never match the complexity of human cognition–and nor should it, because there is no point in replacing one thing we do not understand with another (Norris, 2005).

1.4.2 Data Description

Knowingly or not, we have all used models to describe or summarize data, and at first glance, this appears quite straightforward. For example, we probably would not hesitate to describe the salaries of all 150 members of the Australian House of Representatives by their average because in this case, there is little doubt that the mean is the proper "model" of the data (notwithstanding the extra allowances bestowed upon ministers). Why would we want to "model" the data in this way? Because we are replacing the data points (N = 150 in this instance) with a single estimated "parameter."[6] In this instance, the parameter is the sample mean, and reducing 150 points into one facilitates understanding and efficient communication of the data.

However, we must not become complacent in light of the apparent ease with which we can model data by their average. As a case in point, consider U.S. President Bush's 2003 statement in promotion of his tax cut, that "under this plan, 92 million Americans receive an average tax cut of $1,083." Although this number, strictly speaking, was not incorrect, it arguably did not represent the best model of the proposed tax cut, given that 80% of taxpayers would receive less than this cut, and nearly half (i.e., some 45 million people) would receive less than $100 (Verzani, 2004). The distribution of tax cuts was so skewed (bottom 20% of income earners slated to receive $6 compared to $30,127 for the top 1%) that the median or a trimmed mean would have been the preferable model of the proposed legislation in this instance.

Controversies about the proper model with which to describe data also arise in cognitive science, although fortunately with more transparency and less disingenuousness than in the political scene. In fact, data description, by itself, can have considerable psychological impact. As a case in point, consider the debate on whether learning of a new skill is best understood as following a "power law" or is better described by an exponential improvement (Heathcote, Brown, & Mewhort, 2000). There is no doubt that the benefits from practice accrue in a nonlinear fashion: The first time you try your hands at a new skill (for example, creating an Ikebana arrangement), things take seemingly forever (and the output may not be worth writing home about). The second and third time round, you will notice vast improvements, but eventually, after some dozens of trials, chances are that all further improvements are small indeed.

What is the exact functional form of this pervasive empirical regularity? For several decades, the prevailing opinion had been that the effect of practice is best captured by a power law—that is, by the function (shown here in its simplest possible form),

$$RT = N^{-\beta}, \qquad (1.1)$$

where RT represents the time to perform the task, N represents the number of learning trials to date, and β is the learning rate. Figure 1.5 shows sample data, taken from Palmeri's (1997) Experiment 3, with the appropriate best-fitting power function superimposed as a dashed line.

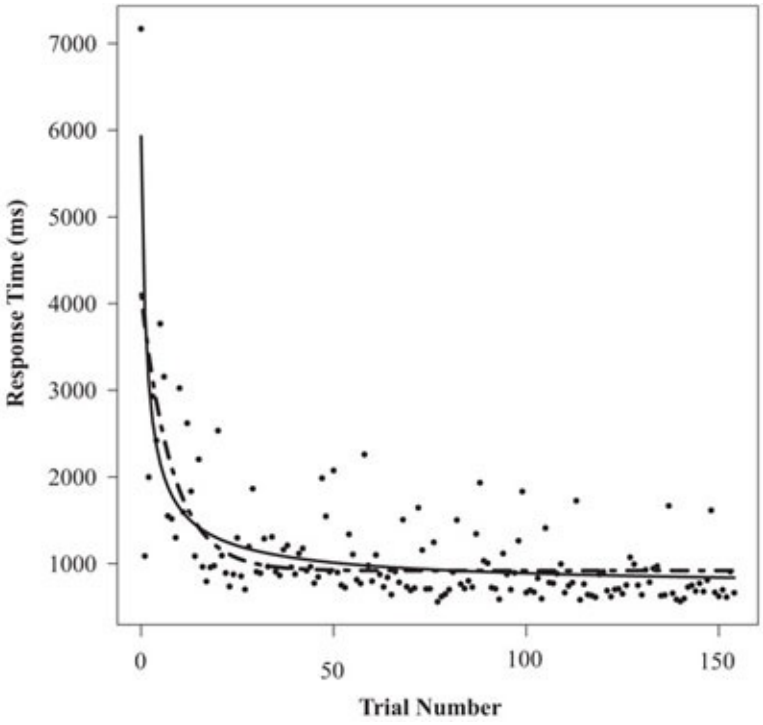

Figure 1.5 Sample power law learning function (dashed line) and alternative exponential function (solid line) fitted to the same data. Data are represented by dots and are taken from Palmeri's (1997) Experiment 3 (Subject 3, Pattern 13). To fit the data, the power and exponential functions were a bit more complex than described in Equations 1.1 and 1.2 because they also contained an asymptote (A) and a multiplier (B). Hence, the power function took the form $RT = A_P + B_P \times (N + 1)^{-\beta}$, and the exponential function was $RT = A_E + B_E \times e^{-\alpha N}$.

Heathcote et al. (2000) argued that the data are better described by an exponential function given by (again in its simplest possible form)

$$RT = e^{-\alpha N}, \qquad (1.2)$$

where N is as before and α the learning rate. The best-fitting exponential function is shown by the solid line in Figure 1.5; you will note that the two competing descriptions or models do not appear to differ much. The power function captures the data well, but so does the exponential function, and there is not much to tell between them: The residual mean squared deviation (RMSD), which represents the average deviation of the data points from the predicted function, was 482.4 for the power function compared to 526.9 for the exponential. Thus, in this instance,

the power function fits "better" (by providing some 50 ms less error in its predictions than the exponential), but given that *RT*'s range is from somewhere less than 1000 ms to 7 seconds, this difference is not particularly striking.

So, why would this issue be of any import? Granted, we wish to describe the data by the appropriate model, but surely neither of the models in Figure 1.5 mis-represents essential features of the data anywhere near as much as U.S. President Bush did by reporting only the average implication of his proposed tax cut. The answer is that the choice of the correct descriptive model, in this instance, carries important implications about the psychological nature of learning. As shown in detail by Heathcote et al. (2000), the mathematical form of the exponential function necessarily implies that the learning rate, relative to what remains to be learned, is constant throughout practice. That is, no matter how much practice you have had, learning continues by enhancing your performance by a constant fraction. By contrast, the mathematics of the power function imply that the relative learning rate is slowing down as practice increases. That is, although you continue to show improvements throughout, the rate of learning *decreases* with increasing practice. It follows that the proper characterization of skill acquisition data by a descriptive model, in and of itself, has considerable psychological implications (we do not explore those implications here; see Heathcote et al., 2000, for pointers to the background).

Just to wrap up this example, Heathcote et al. (2000) concluded after reanalyzing a large body of existing data that the exponential function provided a better description of skill acquisition than the hitherto presumed power law. For our purposes, their analysis permits the following conclusions: First, quantitative description of data, by itself, can have considerable psychological implications because it prescribes crucial features of the learning process. Second, the example underscores the importance of model selection that we alluded to earlier; in this instance, one model was chosen over another on the basis of strict quantitative criteria. We revisit this issue in Chapter 5. Third, the fact that Heathcote et al.'s model selection considered the data of individual subjects, rather than the average across participants, identifies a new issue—namely, the most appropriate way in which to apply a model to the data from more than one individual—that we consider in Chapter 3.

The selection among competing functions is not limited to the effects of practice. Debates about the correct descriptive function have also figured prominently in the study of forgetting. Does the rate of forgetting differ with the extent of learning? Is the rate of information loss constant over time? Although the complete pattern of results is fairly complex, two conclusions appear warranted (Wixted, 2004a): First, the degree of learning does not affect the rate of forgetting. Hence, irrespective of how much you cram for an exam, you will lose the information at the same rate—but of course this is not an argument against dedicated study; if you learn more, you will also retain more, irrespective of the fact that the rate of loss per unit of time remains the same. Second, the rate of forgetting *decelerates* over time. That is, whereas you might lose some 30% of the information on the first day, on the second day, the loss may be down to 20%, then 10%, and so on. Again, as in the case of practice, two conclusions are relevant here: First, quantitative comparison among competing descriptive models was required to choose the appropriate function (it is a power function, or something very close to it). Second, although the shape of the "correct" function has considerable theoretical import

because it may imply that memories are "consolidated" over time *after* study (see Wixted, 2004a, 2004b, for a detailed consideration, and see G. D. A. Brown & Lewandowsky, 2010, for a contrary view), the function itself has no psychological content.

The mere description of data can also have psychological implications when the behavior it describes is contrasted to *normative* expectations (Luce, 1995). Normative behavior refers to how people would behave if they conformed to the rules of logic or probability. For example, consider the following syllogism involving two premises (P) and a conclusion (C). P1: All polar bears are animals. P2: Some animals are white. C: Therefore, some polar bears are white. Is this argument valid? There is a 75% to 80% chance that you might endorse this conclusion (e.g., Helsabeck, 1975), even though it is logically false (to see why, replace *white* with *brown* in P2 and C). This example shows that people tend to violate normative expectations even in very simple situations. In this instance, the only descriptive model that is required to capture people's behavior—and to notice the normative violation—is a simple proportion (i.e., .75–.80 of people commit this logical error). In other, more realistic instances, people's normatively irrational behavior is best captured by a rather more complex descriptive model (e.g., Tversky & Kahneman, 1992).

We have presented several descriptive models and have shown how they can inform psychological theorizing. Before we move on, it is important to identify the common threads among those diverse examples. One attribute of descriptive models is that they are explicitly devoid of psychological *content*; for example, although the existence of an exponential practice function constrains possible learning mechanisms, the function itself has no psychological content. It is merely concerned with describing the data.

For the remainder of this chapter, we will be considering models that have increasingly more psychological content. In the next section, we consider models that characterize cognitive processes at a highly abstract level, thus going beyond data description, but that do not go so far as to explain those processes in detail. The final section considers models that go beyond characterization and explain the cognitive processes.

1.4.3 Process Characterization

What does it mean to characterize a cognitive process? There are two relevant attributes: First, models that characterize processes peek inside the "black box" that is the mind and postulate—and then measure—distinct cognitive components. Unlike descriptive models, their explanatory power thus rests on hypothetical constructs within the mind rather than within the data to be explained. Second, these models do not go beyond identification of those constructs or processes; that is, they remain neutral with respect to specific instantiations and explanations underpinning the cognitive processes they characterize. (Providing those explanations is the domain of the last class of models, to be considered in the next section.)

We illustrate this class of models using the multinomial processing tree (MPT) approach (Batchelder & Riefer, 1999; see also Riefer & Batchelder, 1988). The MPT approach makes the uncontroversial assumption that psychological data often result from multiple cognitive processes and provides a technique to disentangle and measure the relative contributions of these underlying processes. To do so, an MPT model postulates a sequence of processing

stages and connects them by a variety of paths that can give rise to the observed behavioral outcome. While this may sound complicated, it is actually quite simple once shown graphically: Figure 1.6 contains a multinomial processing tree proposed by Schweickert (1993) to characterize recall from short-term memory.

The model postulates two ways in which recall can be successful: First, if the information in memory is intact (with probability I), then the item is recalled directly. Second, if the memorial representation is not intact (probability $1 - I$), then an item might nonetheless be "redintegrated" (with probability R). The red-integration stage refers to some reconstruction process that fills in the missing bits of a partially forgotten item on the basis of, say, information in long-term memory; for example, knowledge of the word *hippopotamus* will enable you to recall a memorized item even if all you can remember is something like "h_p_ _ _ _tam_ _." Only if redintegration also fails (with probability $1 - R$), then recall will be unsuccessful.

Let us trace these possible outcomes in Figure 1.6: We enter the tree at the top, and depending on whether the trace is intact, we branch right (with probability I) or left ($1 - I$). In the former case, the item is recalled, and outcome "*C*" (for "correct" recall) is obtained. In the latter case, the second stage kicks in, and we ask whether the item—not being intact—can nonetheless be successfully redintegrated (with probability R; branch right) or not ($1 - R$; keep going left). In the former case, we score another correct response; in the latter, we commit an error (E). The overall predictions of the model—for correct responses and errors, respectively—are thus given by $C = I + (1 - I) \times R$ and $E = (1 - I) \times (1 - R)$.

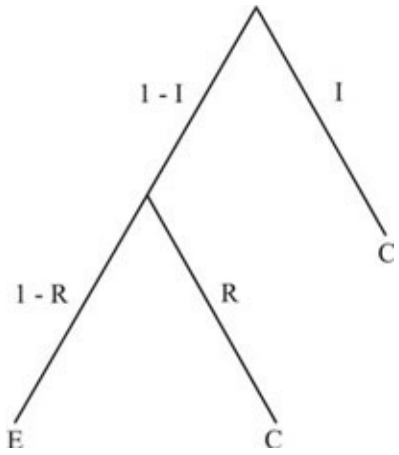

Figure 1.6 A simple multinomial processing tree model proposed by Schweickert (1993) for recall from short-term memory.

You are likely to ask at least two questions at this point: First, why are those components multiplied together, and second, how do we know what the values are of I and R?

The former question is answered by noting that each branch in the tree builds on the previous one; that is, redintegration (R) only takes place if the item was not intact ($1 - I$) in the first place. Because the two stages are assumed to be independent, their probabilities of occurrence are multiplied together (for further discussion, see first part of Chapter 4). It follows that one possible way in which a response may be correct, via the path *left-right*, is given by $(1 - I) \times R$. This outcome is then added to the other way in which one can be correct,

along the simple path *right*, which is given by I . Analogously, an error can only occur via the path *left-left*, which is thus given by $(1 - I) \times (1 - R)$.

The latter question, concerning the values of I and R, has both a simple and also a very involved answer. The simple answer is that those quantities are parameters that are estimated from the data, similar to the way in which we compute a sample mean to estimate the central tendency of the data. In contrast to the purely descriptive mean, however, the quantities I and R have psychological meaning and characterize two presumed cognitive processes—namely, memory storage (intact or not) and redintegration (successful or not). The more involved answer concerns the technical issues surrounding parameter estimation, and we will explore that answer in several of the following chapters in great detail.[7]

This is a good opportunity for recapitulation. We have presented a simple MPT model that characterizes the presumed processes operating in recall from short-term memory. Like the descriptive models in the preceding section, this model replaces the data by parameters. Unlike descriptive models, however, the parameters in the present case (I and R) have a psychological interpretation and characterize postulated cognitive processes.

To illustrate the way in which these types of models can provide a peek inside our minds, consider an application of Schweickert's (1993) model to the recall of lists containing words of different natural-language frequencies by Hulme et al. (1997). Hulme et al. compared lists composed of high-frequency words (e.g., *cat*, *dog*) and low-frequency words (*buttress*, *kumquat*) and examined performance as a function of each item's serial position in the list (i.e., whether it was presented first, second, and so on). What might the MPT model shown in Figure 1.6 predict for this experiment?

Hulme et al. (1997) reasoned that the redintegration process would operate more successfully on high-frequency words than low-frequency words because the former's representations in long-term memory are more easily accessed by partial information—and hence are more likely to contribute to reconstruction. Accordingly, R should be greater for high- than for low-frequency items. Does it follow that high-frequency items should always be recalled better than their low-frequency counterparts? No, because redintegration is only required if information in memory is no longer intact. It follows that early list items, which are less subject to degradation during recall, will be largely intact; because they thus bypass the redintegration stage, their frequency should matter little. Later list items, by contrast, are degraded more by the time they are recalled, and hence red-integration becomes more important for them—and with it, the effect of word frequency should emerge. This is precisely what Hulme et al. found: High-frequency words were recalled better than low-frequency words, but that effect was primarily confined to later list positions. The data, when interpreted within the MPT model in Figure 1.6, therefore support the notion that word frequency affects the success of reconstruction of partially degraded memory traces but not their retention in short-term memory. Given the utmost simplicity of the MPT model, this is quite an interesting insight—and not one that can be confidently inferred from inspection of the data. Instead, Hulme et al. buttressed their conclusions by quantitatively examining the correspondence between the model's predictions and the data.

That said, the limitations of the MPT model are also noteworthy—and they set the stage for

discussion of the next class of model. The MPT model may have identified and characterized a cognitive process known as redintegration, but it neither described nor explained that process. Is this even possible? Can we know more about redintegration? The answer is a clear yes, and providing that additional knowledge is the domain of process explanation models that we consider next. To wrap up this example, we briefly note that Lewandowsky (1999) and Lewandowsky and Farrell (2000) provided a detailed process account of red-integration that explains exactly how partial traces can be reconstructed. The Lewandowsky and Farrell model consists of a network of interconnected units that bounce information back and forth between them, adding bits and pieces from long-term memory to the degraded memory trace at each step, until the original item is perfectly reconstructed (instantiating *R*, in the MPT model's terminology) or another item is produced, in which case an error has occurred ($1 - R$).[8] We now consider this class of models that not only identify processes but also explain them.

1.4.4 Process Explanation

What does it mean to explain, rather than merely characterize, a cognitive process? First, explanatory models provide the most close-up view inside the "black box" that is possible with current psychological techniques. Like characterization models, their power rests on hypothetical cognitive constructs, but by providing a detailed explanation of those constructs, they are no longer neutral. That is, whereas the MPT model in the previous section identified the redintegration stage but then remained neutral with respect to how exactly that reconstruction might occur, an explanatory process model (e.g., Lewandowsky & Farrell, 2000) goes further and removes any ambiguity about how that stage might operate.

At first glance, one might wonder why not every model belongs to this class: After all, if one can specify a process, why not do that rather than just identify and characterize it? The answer is twofold. First, it is not always possible to specify a presumed process at the level of detail required for an explanatory model, and in that case, a model such as the earlier MPT model might be a valuable alternative. Second, there are cases in which a coarse characterization may be preferable to a detailed specification. For example, it is vastly more important for a weatherman to know whether it is raining or snowing, rather than being confronted with the exact details of the water molecules' Brownian motion. Likewise, in psychology, modeling at this level has allowed theorists to identify common principles across seemingly disparate areas (G. D. A. Brown, Neath, & Chater, 2007).

That said, we believe that in most instances, cognitive scientists would ultimately prefer an explanatory process model over mere characterization, and the remainder of this book is thus largely (though not exclusively) devoted to that type of model.

There are countless explanatory models of cognitive phenomena ranging from reasoning through short-term memory to categorization, and we will be touching on many of those during the remaining chapters.

We begin our discussion by presenting a close-up of the exemplar model of categorization first presented in Section 1.3.1. We choose this model, known as the generalized context model (GCM; see, e.g., Nosofsky, 1986), for three reasons: First, it is undoubtedly one of the most influential and successful existing models of categorization. Second, its basic architecture is

quite straightforward and readily implemented in something as simple as Microsoft Excel. Third, some of the GCM architecture also contributes to other important models of cognition, which we will consider in later chapters (e.g., SIMPLE in Chapter 4).

We already know that GCM is an exemplar model. As implied by that name, GCM stores every category exemplar encountered during training in memory. We mentioned an experiment earlier in which people learned to classify cartoon faces; in GCM, this procedure would be implemented by adding each stimulus to the pile of faces belonging to the same category. Remember that each response during training is followed by feedback, so people know whether a face belongs to a MacDonald or a Campbell at the end of each trial. Following training, GCM has thus built two sets of exemplars, one for each category, and all subsequent test stimuli are classified by referring to those memorized ensembles. This is where things get really interesting (and, refreshingly, a bit more complicated, but nothing you can't handle).

First, we need some terminology. Let us call a particular test stimulus i, and let us refer to the stored exemplars as the set J with members $j = 1, 2, \ldots, J$, hence $j \in J$. This notation may seem like a bit of an overkill at first glance, but in fact it is useful to clarify a few things at the outset that we will use for the remainder of the book. Note that we use lowercase letters (e.g., i, j, …) to identify specific elements of a set and that the number of elements in that set is identified by the same uppercase letters (I, J, …), whereas the set itself is identified by the "Fraktur" version of the letter (I, J, …). So, we have a single thing called i (or j or whatever), which is one of I elements of a set I.

We are now ready to consider the effects of presenting stimulus i. In a nutshell, a test stimulus "activates" all stored exemplars (remember, that's $j \in J$) to an extent that is determined by the *similarity* between i and each j. What exactly is similarity? GCM assumes that stimuli are represented in a perceptual space and that proximity within that space translates into similarity. To illustrate, consider the left panel (A) in Figure 1.7, which shows the perceptual representation of three hypothetical stimuli that differ along a single dimension —in this case, line length. The broken line labeled d represents the distance between two of those stimuli. It is easy to see that the greater this distance is, the *less* similar the two stimuli are. Conversely, the closer together two stimuli are, the greater their similarity.

Now consider Panel B. Here again we have three hypothetical stimuli, but this time they differ along two dimensions simultaneously—namely, distance and angle. Panel B again shows the distance (d) between two stimuli, which is formally given by the following equation:

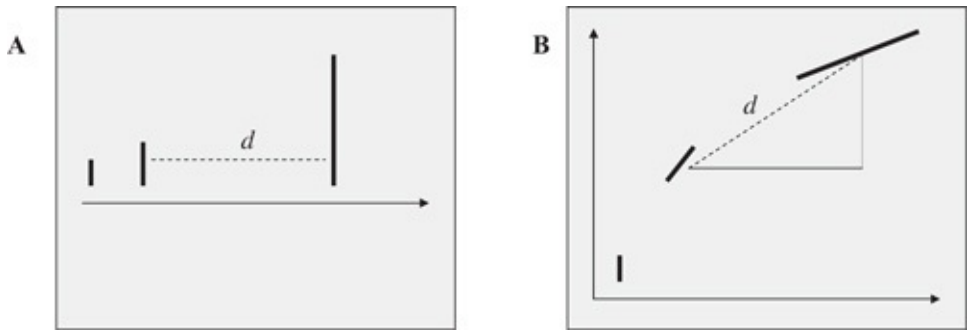

Figure 1.7 The representational assumptions underlying the generalized context model (GCM). Panel A shows stimuli that differ along one dimension only (line length), and Panel B shows stimuli that differ along two dimensions (line length and angle).

In both panels, a representative distance (d) between two stimuli is shown by the broken line.

$$d_{ij} = \left(\sum_{k=1}^{K} |x_{ik} - x_{jk}|^2 \right)^{\frac{1}{2}}, \qquad (1.3)$$

where x_{ik} is the value of dimension k for test item i (let's say that's the middle stimulus in Panel B of Figure 1.7), and x_{jk} is the value of dimension k for the stored exemplar j (say, the right-most stimulus in the panel). The number of dimensions that enter into computation of the distance is arbitrary; the cartoon faces were characterized by four dimensions, but of course we cannot easily show more than two dimensions at a time. Those dimensions were eye height, eye separation, nose length, and mouth height. [9]

An easy way to understand Equation 1.3 is by realizing that it merely restates the familiar Pythagorean theorem (i.e., $d^2 = a^2 + b^2$), where a and b are the thin solid lines in Panel B of Figure 1.7, which are represented by the more general notation of dimensional differences (i.e., $x_{ik} - x_{jk}$) in the equation.

How, then, does distance relate to similarity? It is intuitively obvious that greater distances imply lesser similarity, but GCM explicitly postulates an exponential relationship of the following form:

$$s_{ij} = exp(-c \cdot d_{ij}), \qquad (1.4)$$

where c is a parameter and d_{ij} the distance as just defined. Figure 1.8 (see page 22) visualizes this function and shows how the activation of an exemplar (i.e., s_{ij}) declines as a function of the distance (d_{ij}) between that exemplar and the test stimulus. You may recognize that this function looks much like the famous generalization gradient that is observed in most situations involving discrimination (in species ranging from pigeons to humans; Shepard, 1987): This similarity is no coincidence; rather, it motivates the functional form of the similarity function in Equation 1.4. This similarity function is central to GCM's ability to generalize learned responses (i.e., cartoon faces seen during study) to novel stimuli (never-before-seen cartoon faces presented at test only).

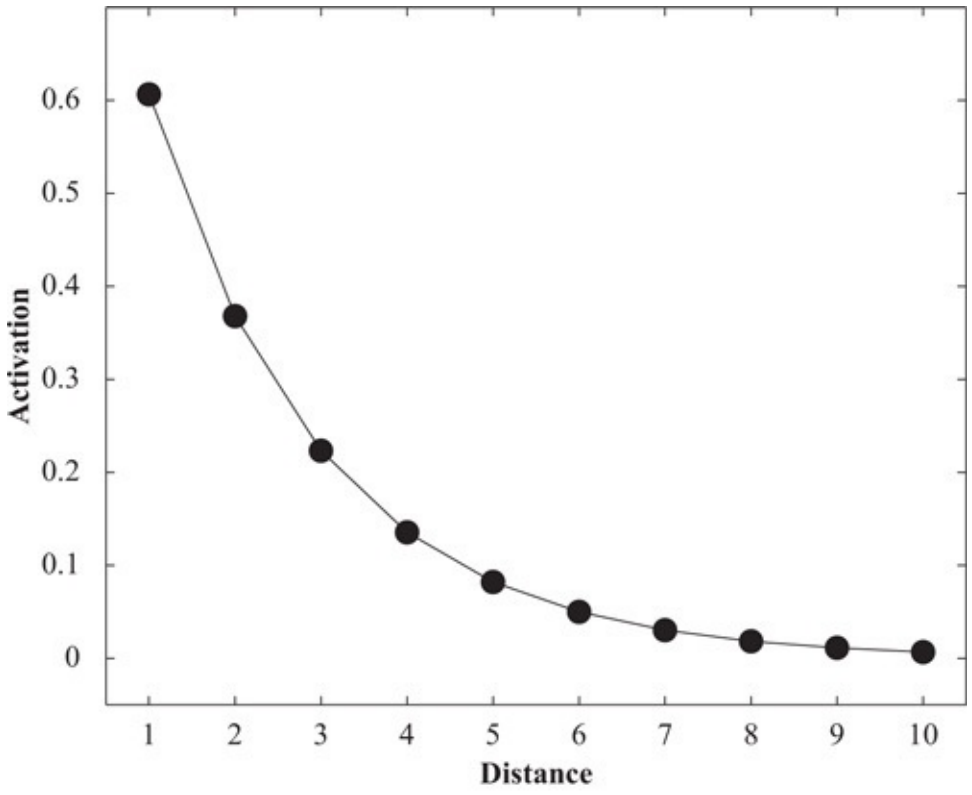

Figure 1.8 The effects of distance on activation in the GCM. Activation (i.e., s_{ij}) is shown as a function of distance (d_{ij}). The parameter c (see Equation 1.4) is set to .5.

It turns out that there is little left to do: Having presented a mechanism by which a test stimulus activates an exemplar according to its proximity in psychological space, we now compute those activations for *all* memorized exemplars. That is, we compute the distance d_{ij} between i and each $j \in J$ as given by Equation 1.3 and derive from that the activation s_{ij} as given by Equation 1.4. The next step is to convert the entire set of resulting activations into an explicit decision: Which category does the stimulus belong to? To accomplish this, the activations are summed separately across exemplars within each of the two categories. The relative magnitude of those two sums directly translates into response probabilities as follows:

$$P(R_i = A|i) = \frac{\left(\sum_{j \in A} s_{ij}\right)}{\left(\sum_{j \in A} s_{ij}\right) + \left(\sum_{j \in B} s_{ij}\right)}, \qquad (1.5)$$

where A and B refer to the two possible categories, and $P(R_i = A|i)$ means "the probability of classifying stimulus i into category A." It follows that application of Equations 1.3 through 1.5 permits us to derive classification predictions from the GCM. It is those predictions that were plotted on the abscissa (x-axis) in the left panel of the earlier Figure 1.4, and it is those predictions that were found to be in such close accord with the data.

If this is your first exposure to quantitative explanatory models, the GCM may appear daunting at first glance. We therefore wrap up this section by taking a second tour through the

GCM that connects the model more directly to the cartoon face experiment.

Figure 1.9 shows the stimuli used during training. Each of those faces corresponds to a memorized exemplar j that is represented by a set of dimensional values $\{x_{j1}, x_{j2}, \ldots\}$, where each x_{jk} is the numeric value associated with dimension k. For example, if the nose of exemplar j has length 5, then $x_{j1} = 5$ on the assumption that the first dimension (arbitrarily) represents the length of the nose.

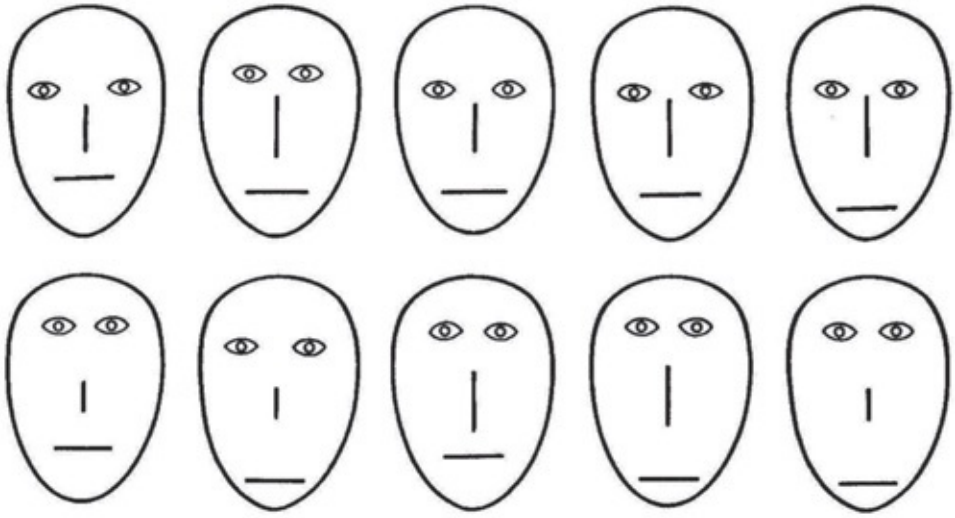

Figure 1.9 Stimuli used in a classification experiment by Nosofsky (1991). Each row shows training faces from one of the two categories. Figure reprinted from Nosofsky, R. M. (1991). Tests of an exemplar mode for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27. Published by the American Psychological Association; reprinted with permission.

To obtain predictions from the model, we then present test stimuli (those shown in Figure 1.9 but also new ones to test the model's ability to generalize). Those test stimuli are coded in the same way as training stimuli—namely, by a set of dimensional values. For each test stimulus i, we first compute the distance between it and exemplar j (Equation 1.3). We next convert that distance to an activation of the memorized exemplar j (Equation 1.4) before summing across exemplars within each category (Equation 1.5) to obtain a predicted response probability. Do this for each stimulus in turn, and bingo, you have the model's complete set of predictions shown in Figure 1.4. How exactly are these computations performed? A whole range of options exists: If the number of exemplars and dimensions is small, a simple calculator, paper, and a pencil will do. More than likely, though, you will be using a computer package (such as a suitable worksheet in Excel) or a computer program (e.g., written in a language such as MATLAB or R). Regardless of how we perform these computations, we are assuming that they represent an analog of the processes used by people. That is, we presume that people remember exemplars and base their judgments on those memories alone, without access to rules or other abstractions.

At this point, one can usefully ponder two questions. First, why would we focus on an experiment that involves rather artificial cartoon faces? Do these stimuli and the associated data and modeling have any bearing on classification of "real-life" stimuli? Yes, in several ways. Not only can the GCM handle performance with large and ill-defined perceptual

categories (McKinley & Nosofsky, 1995), but recent extensions of the model have been successfully applied to the study of natural concepts, such as fruits and vegetables (Verbeemen, Vanpaemel, Pattyn, Storms, & Verguts, 2007). The GCM thus handles a wide variety of both artificial and naturalistic categorizations. Second, one might wonder about the motivation underlying the equations that define the GCM. Why is distance related to similarity via an exponential function (Equation 1.4)? Why are responses determined in the manner shown in Equation 1.5? It turns out that for any good model—and the GCM is a good model—the choice of mathematics is not at all arbitrary but derived from some deeper theoretical principle. For example, the distance-similarity relationship in the GCM incorporates our knowledge about the "universal law of generalization" (Shepard, 1987), and the choice of response implements a theoretical approach first developed by Luce (1963).

What do you now know and what is left to do? You have managed to study your (possibly) first explanatory process model, and you should understand how the model can predict results for specific stimuli in a very specific experiment. However, a few obstacles remain to be overcome, most of which relate to the "how" of applying the model to data. Needless to say, those topics will be covered in subsequent chapters.

1.4.5 Classes of Models

We sketched out three broad classes of models. We considered descriptive models whose sole purpose it is to replace the intricacies of a full data set with a simpler representation in terms of the model's parameters. Although those models themselves have no psychological content, they may well have compelling psychological implications.

We then considered two classes of models that both seek to illuminate the workings of the mind, rather than data, but do so to a greatly varying extent. Models that characterize processes identify and measure cognitive stages, but they are neutral with respect to the exact mechanics of those stages. Explanatory models, by contrast, describe all cognitive processes in great detail and leave nothing within their scope unspecified.[10]

Other distinctions between models are possible and have been proposed (e.g., Luce, 1995; Marr, 1982; Sun, Coward, & Zenzen, 2005), and we make no claim that our classification is better than other accounts. Unlike other accounts, however, our three classes of models map into three distinct tasks that confront cognitive scientists: Do we want to describe data? Do we want to identify and characterize broad stages of processing? Do we want to explain how exactly a set of postulated cognitive processes interact to produce the behavior of interest?

1.5 What Can We Expect From Models?

We have explored some of the powerful insights that are afforded by quantitative modeling. However, all examples so far were demonstrations that one model or another could provide a good quantitative account of otherwise inexplicable data—impressive, perhaps, but is that all we can expect from models? Is a "good fit" between a model's predictions and the data the one and only goal of modeling? The answer is no; there are several other ways in which models

can inform scientific progress.

1.5.1 Classification of Phenomena

It is intuitively obvious that, at least at the current level of understanding in our science, all models will necessarily be limited in their explanatory power. Every model will be confronted sooner or later with data that it cannot accommodate. So, if every model is doomed to fail, why spend considerable time and effort on its development in the first place? One answer to this conundrum was provided by Estes (1975), who suggested that even the mere classification of phenomena into those that fall within and those that fall outside a model's scope can be very informative: "What we hope for primarily from models is that they will bring out relationships between experiments or sets of data that we would not otherwise have perceived. The fruit of an interaction between model and data should be a new categorization of phenomena in which observations are organized in terms of a rational scheme in contrast to the surface demarcations manifest in data" (p. 271).

Even if we find that it takes two different models to handle two distinct subclasses of phenomena, this need not be at all bad but may in fact crystallize an interesting question. In physics, for example, for a very long time, light was alternately considered as a wave or a stream of particles. The two models were able to capture a different subset of phenomena, with no cross-linkage between those sets of phenomena and the two theories. Although this state was perhaps not entirely satisfactory, it clearly did not retard progress in physics.

In psychology, we suggest that models have similarly permitted a classification of phenomena in categorization. We noted earlier that the GCM is a powerful model that has had a profound impact on our understanding of how people classify stimuli. However, there are also clear limits on the applicability of the GCM. For example, Rouder and Ratcliff (2004) showed that the GCM captures people's behavior only when the stimuli are few and highly discriminable. When there is a large ensemble of confusable stimuli, by contrast, people's behavior is better captured by a rule model rather than the GCM's exemplar representation (more on this in Chapter 7). Likewise, Little and Lewandowsky (2009) showed that in a complex probabilistic categorization task, some people will build an exemplar representation, whereas others will create an ensemble of partial rules; the former were described well by the GCM, but the latter were best described by a rule model. Taken together, those studies serve to delineate the applicability of two competing theoretical approaches—namely, rules versus exemplars—somewhat akin to the differentiation between wave and particle theories of light.

1.5.2 Emergence of Understanding

The models we consider in this book are, almost by definition, always implemented as a computer program. Computers, however, only do as they are programmed to do—does it not follow that our models, unlike behavioral experiments, will never generate anything truly novel or unexpected? Indeed, some time ago, this opinion appeared to reflect accepted practice (e.g., Reitman, 1965). Since then, it has become apparent that this opinion is flawed. There have been innumerable instances in which models have generated novel insights in nontrivial ways,

many of which involved artificial neural networks. (Networks contain many interconnected units that process and transmit information.) For example, Seidenberg and McClelland (1989) presented a network that could learn to pronounce both regular (*lint*) and irregular (*pint*) words from printed input: It was not at all clear prior to the modeling being conducted that a uniform architecture could handle both types of words. Indeed, a "central dogma" (Seidenberg & McClelland, 1989, p. 525) of earlier models had been that two processes were required to accommodate irregular words (via lexical lookup) and regular (non)words (via pronunciation rules).

As another example, Botvinick and Plaut (2006) recently presented a network model of short-term memory that was able to learn the highly abstract ability of "seriation"—namely, the ability to reproduce *novel random sequences* of stimuli. Thus, after learning the skill, the model was capable of reproducing short serial lists. Thus, when presented with "A K P Q B," the model would reproduce that sequence after a single presentation with roughly the same accuracy and subject to the same performance constraints as humans. This might appear like a trivial feat at first glance, but it is not: It is insufficient to learn pairwise contingencies such as "A precedes B" because in a random list, A might precede B as frequently as B precedes A. Likewise, it is insufficient to learn that "A occurs in position 1" because in fact A could occur in any position, and so on for any other specific arrangements of letters (triplets, quadruplets, etc.). Instead, the model had to learn the highly abstract ability "whatever I see I will try to reproduce in the same order" from a small subset of all possible sequences. This abstract ability, once learned, could then be transferred to novel sequences.

In summary, the point that models can yield unexpected and novel insights was perhaps best summed up by Fum et al. (2007): "New ways of understanding may assume several forms. They can derive, for instance, from the discovery of a single unifying principle that will explain a set of hitherto seemingly unrelated facts. They can lead to the emergence of complex, holistic forms of behavior from the specification of simple local rules of interaction. New ways of understanding can arise from unexpected results that defy the modelers intuition" (p. 136).

1.5.3 Exploration of Implications

Unlike people, models can quite literally be taken apart. For example, we can "lesion" models to observe the outcome on behavior of certain localized dys-functions. As a case in point, consider the model by Hinton and Shallice (1991), which was trained to map a set of orthographic representations into semantic features, so that presentation of a spelling pattern would activate the correct "word" at the semantic output level of their network. After training, Hinton and Shallice lesioned their model in various ways—for example, by removing units, by contaminating the connections between units with random noise, or by eliminating some connections altogether.

Hinton and Shallice found that virtually any such lesioning of their network, irrespective of location, led to a persistent co-occurrence of visual (*cat* read as *mat*) and semantic (*peach* read as *apricot*) errors. This generality elegantly explained why this mix of visual and semantic errors is common across a wide range of patients whose performance deficits differ

considerably in other respects.

We can draw two conclusions from this example: First, it clarifies the in-principle point that one can do things to models that one cannot do to people, and that those lesioning experiments can yield valuable knowledge. Second, the fact that the results in this instance were surprising lends further support to the point made in the previous section—namely, that models can show emergent properties that are not at all apparent by verbal analysis alone.

1.6 Potential Problems

We conclude by discussing two issues that must be considered to ensure a complete understanding of the basic principles of modeling.

1.6.1 Scope and Testability

Suppose you are a venture capitalist and a scientist approaches you for funding to develop a new theory that will revolutionize gambling. A first version of the theory exists, and it has been extremely successful because it probabilistically characterized the outcomes of 20 successive rolls of a die. In quantitative terms, the theory anticipated each individual outcome with $P = 1/6$. Would you be impressed? We trust that you are not, because any theory that predicts any possible outcome with equal facility is of little scientific interest, even if it happens to be in complete accord with the data (e.g., Roberts & Pashler, 2000). This is quite obvious with our fictitious "theory" of gambling, but it is less obvious—though nonetheless equally applicable—with psychological theories.

Let us reconsider one of the earlier examples: Nosofsky (1991) showed that an exemplar model (the GCM) can integrate people's recognition and classification responses under a common theoretical umbrella (see Figure 1.4). We considered this to be impressive, especially because the GCM performed better than a competing prototype theory, but was our satisfaction justified? What if the exemplar model could have equally explained any other possible relationship between recognition and classification and not just the one shown in Figure 1.3? Indeed, in that case, one would need to be quite concerned about the exemplar model's viability as a testable and falsifiable psychological theory.[11] Fortunately, however, these concerns can be allayed by the fact that the exemplar model is at least in principle subject to falsification, as revealed by some of the results mentioned earlier that place limits on the GCM's applicability (e.g., Little & Lewandowsky, 2009; Rouder & Ratcliff, 2004; Yang & Lewandowsky, 2004).

We are now faced with a conundrum: On the one hand, we want our theories to explain data. We want powerful theories, such as Kepler's, that explain fundamental aspects of our universe. We want powerful theories, such as Darwin's, to explain the diversity of life. On the other hand, we want the theories to be falsifiable—that is, we want to be assured that there are at least *hypothetical* outcomes that, if they are ever observed, *would* falsify a theory. For example, Darwin's theory of evolution predicts a strict sequence in which species evolved; hence, any observation to the contrary in the fossil record—for example, human bones co-

occurring with dinosaur remains in the same geological strata (e.g., Root-Bernstein, 1981)—would seriously challenge the theory. This point is sufficiently important to bear repetition: Even though we are convinced that Darwin's theory of evolution, one of the most elegant and powerful achievements of human thought, is true, we simultaneously also want it to be falsifiable—falsi*fiable*, not false.[12] Likewise, we are committed to the idea that the earth orbits around the sun, rather than the other way round, but as scientists, we accept that fact only because it is based on a theory that is falsifiable—again, falsi*fiable*, not false.

Roberts and Pashler (2000) considered the issue of falsifiability and scope with reference to psychological models and provided an elegant graphical summary that is reproduced in Figure 1.10. The figure shows four hypothetical outcome spaces that are formed by two behavioral measures. What those measures represent is totally arbitrary; they could be trials to a criterion in a memory experiment and a final recognition score or any other pair of measures of interest.

Within each panel, the dotted area represents all possible predictions that are within the scope of a psychological theory. The top row of panels represents some hypothetical theory whose predictions are constrained to a narrow range of outcomes; any outcome outside the dotted sliver would constitute contrary evidence, and only the narrow range of values within the sliver would constitute supporting evidence. Now compare that sliver to the bottom row of panels with its very generous dotted areas; the theory shown here is compatible with nearly all possible outcomes. It follows that any observed outcome that falls within a dotted area would offer greater support for the theory in the top row than the bottom row, simply because the likelihood of falsification is greater for the former than the latter, thus rendering the match between data and predictions far less likely—and hence more informative when it occurs (see Dunn, 2000, for a similar but more formalized view). Ideally, we would want our theories to occupy only a small region of the outcome space but for all observed outcomes to fall within that region—as they do for Kepler's and Darwin's theories.[13]

Another important aspect of Figure 1.10 concerns the quality of the data, which is represented by the columns of panels. The data (shown by the single black point bracketed by error bars) exhibit less variability in the left column of panels than in the right. For now, we note briefly that support for the theory is thus strongest in the top left panel; beyond that, we defer discussion of the important role of data to Chapter 6. That chapter will also provide another in-depth and more formal look at the issue of testability and falsifiability.

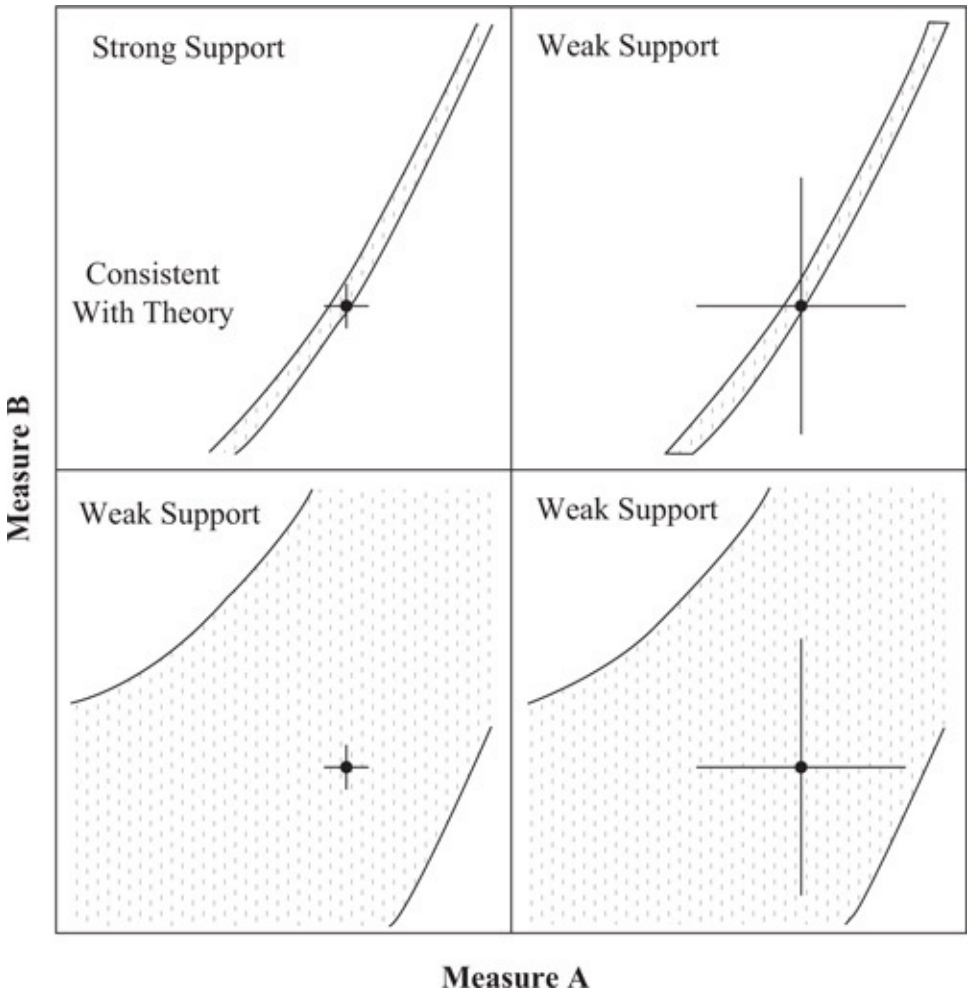

Figure 1.10 Four possible hypothetical relationships between theory and data involving two measures of behavior (A and B). Each panel describes a hypothetical outcome space permitted by the two measures. The shaded areas represent the predictions of a theory that differs in predictive scope (narrow and broad in the top and bottom panels, respectively). The error bars represent the precision of the observed data (represented by the black dot). See text for details. Figure reprinted from Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367. Published by the American Psychological Association; reprinted with permission.

Let us now turn from the abstract representation in Figure 1.10 to a specific recent instance in which two theories were compared by exploration of an outcome space. Howard, Jing, Rao, Provyn, and Datey (2009) examined the nature of associations among list items. Their study was quite complex, but their central question of interest can be stated quite simply: Are associations between list items symmetrical or asymmetrical? That is, given a to-be-memorized list such as "A B C D," is the association from A to B as strong as the association from B to A? Can you recall B when given A as a cue with equal facility as recalling A when given B? And how does the extent of symmetry vary with list position? Empirically, it turns out that adjacent associations (such as between A and B) are asymmetric and stronger in a forward direction, whereas remote associations (such as between A and D) are symmetrical. Howard et al. (2009) compared the abilities of two theories (whose identity is irrelevant in this context) to capture this pattern of symmetries; the pattern of predictions for the two rival theories is shown in Figure 1.11.

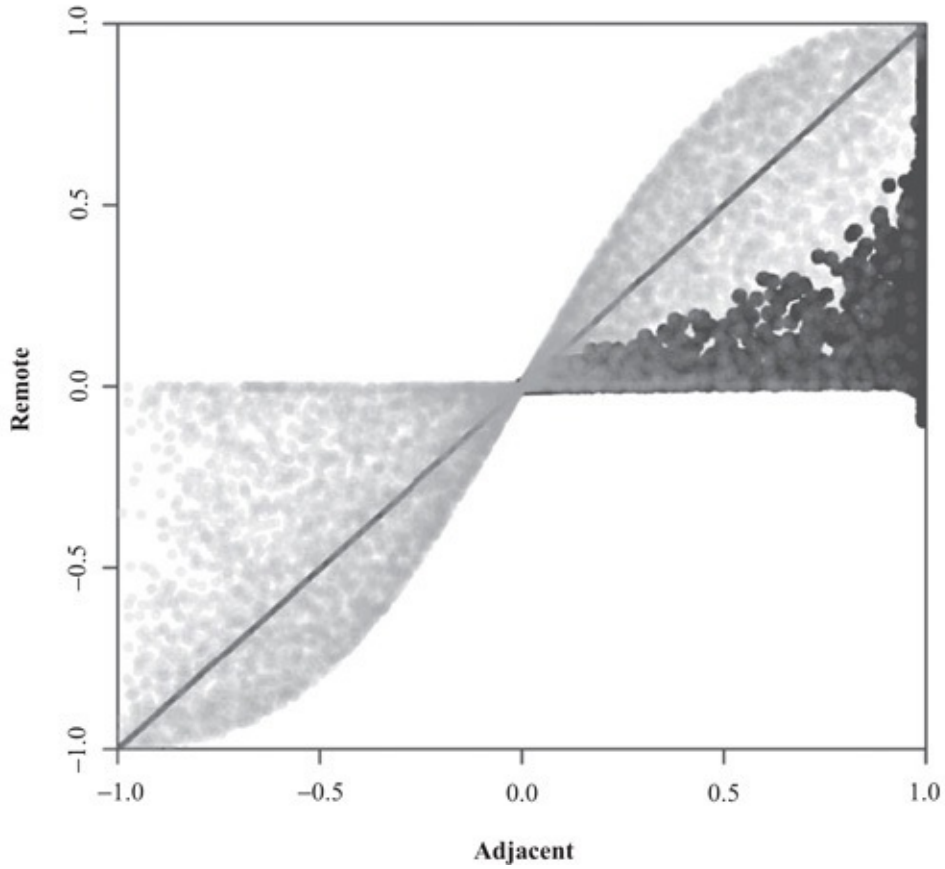

Figure 1.11 Outcome space covered by two models examined by Howard, Jing, Rao, Provyn, and Datey (2009). An index of remote asymmetry is shown as a function of an index of adjacent asymmetry for a variety of parameter values for two models (referred to here as "black" and "gray," corresponding to the color of their plotting symbols). See text for details. Figure reprinted from Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: Transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*, 391–407. Published by the American Psychological Association; reprinted with permission.

The figure shows an outcome space involving two measures—namely, indices of symmetry for adjacent and remote associations. In Howard et al.'s (2009) experiment, the observed values were .25 and .03, respectively. The dark and gray point clouds in the figure, respectively, represent the possible predictions of the two models under consideration. The figure suggests the following conclusions: First, both models can handle the data (i.e., their prediction regions contain the point .25, .03). Second, the "gray" model covers a much larger region of the outcome space than the "black" model, including regions in which remote asymmetry is greater than adjacent symmetry, something that has never been observed in the data. Third, it follows that the "black" model is supported more by these data than the "gray" model. (This conclusion is also supported by other results not shown in the figure, but for present purposes, we focus only on the trade-off between scope and falsifiability.) Note how the large area covered by the "gray" model corresponds to the hypothetical situation in the bottom panels of Figure 1.10, whereas the small area covered by the "black" model corresponds to the situation in the top panels.

1.6.2 Identification and Truth

Throughout our discussion, we have emphasized the existence of multiple alternative models to explain the same data. We considered the Ptolemaic and the Copernican system, we contrasted Nosofsky's (1986) GCM exemplar theory with a prototype model, and we repeatedly underscored the need for model selection. Our discussion entailed two tacit assumptions: first, that we can identify the "correct" model and, second, that there is such a thing as a "true" model. It turns out that both of those assumptions are most likely wrong. So why do we nonetheless advocate modeling? What are the implications of the fact that models may be neither identifiable nor true?

Let us first clarify what exactly the problem concerning model identification does and does not imply. First, it is important to realize that this problem is not unique to psychology but applies to all sciences; we noted earlier that in addition to Kepler's model, an *infinite number of equivalent models* can adequately capture planetary motion. Does this invalidate our view of the solar system? No, it does not, because as we also noted earlier, criteria other than goodness of fit help differentiate between models. So, the fact that in cognitive science, just like in astronomy, "there undoubtedly exists a very diverse set of models, but all equivalent in that they predict the behavior of humans at cognitive tasks" (J. R. Anderson, 1976, p. 4) is true in principle but not particularly troubling.

Second, the fact that there exist, in principle, many equivalent models does not imply that *all* models are equally capable. Indeed, we have shown throughout this chapter that some models handle the data better than others. It is therefore clearly possible to choose one model over another, even if (in principle) the chosen model is equivalent to many unknown others. Simply put, the fact that there are many good models out there does not prevent us from rejecting the bad ones.

Third, the mere existence of equivalent models does not imply that they have been—or indeed will be—discovered. In our experience, it is difficult enough to select a single suitable model, let alone worry about the existence of an infinite number of equivalent competitors.

Finally, even supposing that we must select from among a number of competing models of equivalent capability (i.e., equal goodness of fit), some fairly straightforward considerations have been put forward to achieve this (see, e.g., Fum et al., 2007). We revisit this issue in detail in Chapter 5.

Now let us turn to the issue concerning the "truth" of a model. Is there such a thing as one true model? And if not, what are the implications of that? The answer to the first question is strongly implied by the preceding discussion, and it was most clearly stated by MacCallum (2003): "Regardless of their form or function, or the area in which they are used, it is safe to say that these models all have one thing in common: *They are all wrong*" (p. 114). Now what?

To answer this question, we again briefly digress into astronomy by noting that Kepler's model, being based on Newtonian physics, is—you guessed it—wrong. We now know that Newtonian physics is "wrong" because it does not capture the phenomena associated with relativity. Does this mean that the earth is in fact not orbiting around the sun? No, it does not, because Kepler's model is nonetheless useful because within the realm for which it was designed—planetary motion—Newtonian physics holds to an acceptable degree. Likewise, in psychology, our wrong models can nonetheless be useful (MacCallum, 2003). We show exactly

how wrong models can still be useful at the end of the next chapter, after we introduce a few more essential tools and concepts.

Notes

1. Lest one think that the heliocentric and geocentric models exhaust all possible views of the solar system, it is worth clarifying that there is an infinite number of equivalent models that can adequately capture planetary motion because relative motion can be described with respect to *any* possible vantage point.

2. *Goodness of fit* is a term for the degree of quantitative error between a model's predictions and the data; this important term and many others are discussed in detail in Chapter 2.

3. Astute readers may wonder how the two could possibly differ. The answer lies in the fact that the similarity rule involved in the comparisons by the exemplar model is nonlinear; hence, the summed individual similarities differ from that involving the average. This nonlinearity turns out to be crucial to the model's overall power. The fact that subtle matters of arithmetic can have such drastic consequences further reinforces the notion that purely verbal theorizing is of limited value.

4. Another lesson that can be drawn from this example is a rejoinder to the popular but largely misplaced criticism that with enough ingenuity and patience, a modeler can always get a model to work.

5. Several distinctions between models have been proposed (e.g., Luce, 1995); ours differs from relevant precedents by being explicitly psychological and being driven entirely by considerations that are relevant to the cognitive researcher.

6. We will provide a detailed definition of what a parameter is in Chapter 2. For now, it suffices to think of a parameter as a number that carries important information and that determines the behavior of the model.

7. Some readers may have noticed that in this instance, there are two parameters (I and R) and two data points (proportion correct and errors; C and R), which renders the model nonidentifiable. We ignore this issue here for simplicity of exposition; for a solution, see Hulme et al. (1997).

8. This model is a connectionist model, and these are discussed further in Chapter 8.

9.For simplicity, we omit discussion of how these *psychological* distances relate to the physical measurement (e.g., line length in cm) of the stimuli; these issues are covered in, for example, Nosofsky (1986).

10. Of course, a cognitive model may leave other levels of explanation unspecified, for example, the underlying neural circuitry. However, at the level of abstraction within which the model is formulated, nothing can be left unspecified.

11. Throughout this book, we use the terms *falsifiable* and *testable* interchangeably to denote the same idea—namely, that at least in principle, there are some possible outcome(s) that are incompatible with the theory's predictions.

12. Despite its falsifiability, Darwin's theory has a perfect track record of its predictions being uniformly confirmed; Coyne (2009) provides an insightful account of the impressive list of successes.

13. It is important to clarify that, in our view, this argument should apply only with respect to a particular measurement. That is, for any given measurement, we prefer theories that could have only predicted a subset of all possible observations over theories that could have predicted pretty much any outcome. However, it does not follow that we prefer theories that are so narrow in scope that they only apply to a single experiment; on the contrary, we prefer theories that apply to a range of different situations.