CHAPTER 1

Computational Modeling in Cognition and Cognitive Neuroscience

STEPHAN LEWANDOWSKY AND KLAUS OBERAUER

Scientific reasoning rests on two levels of inferences (see Figure 1.1). On the first level, we draw inferential links between data and empirical generalizations. Empirical generalizations, sometimes boldly called *laws*, are statements that pertain to a regular relationship between observable variables. On this level, we make inductive inferences from data in individual studies to empirical generalizations, and deductive inferences using established or hypothesized empirical generalizations to make predictions for further studies. For instance, it is well established that performance on almost any cognitive task improves with practice, and there is widespread agreement that this improvement is best described by a power function (Logan, 1988) or by an exponential function (Heathcote, Brown, & Mewhort, 2000). This regularity is sufficiently well established to enable strong predictions for future studies on practice effects and skill acquisition.

On the second level of inference, we link empirical generalizations to theories. Theories differ from empirical generalizations in that they make assumptions about unobservable variables and mechanisms, and their connections to observable variables. On this second level, we use inductive reasoning to infer theoretical constructs from empirical generalizations. For example, the empirical relationship between practice and performance has been used to infer the possibility that people are remembering every instance of stimuli they encounter (Logan, 1988). To illustrate, this theory proposes that repeated exposure to words in a lexical-decision task results in multiple memory traces of those words being laid down, all of which are accessed in parallel during further trials. With practice, the increasing number of traces permits increasingly fast responding because it becomes increasingly more likely that one of the traces will be accessed particularly quickly. (We expand on this example later.)

Scientists use deductive reasoning to derive predictions of empirical regularities from theoretical assumptions. For instance, the notion that practice effects result from the encoding of additional memory traces of specific stimuli gives rise to the prediction that those performance benefits should not transfer to new items that have never been seen before. This prediction has been confirmed (Logan & Klapp, 1991).

The two levels of inference differ in the degree of formalization that has evolved

Preparation of this paper was facilitated by a Wolfson Research Merit Award from the Royal Society to the first author.

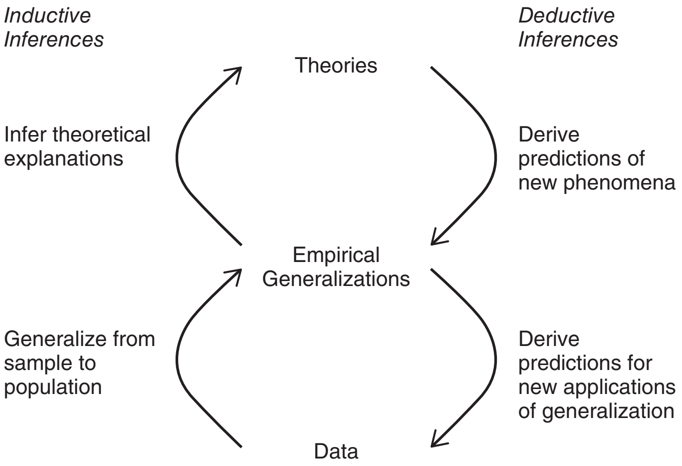
Inductive Inferences
Deductive Inferences
Theories
Infer theoretical explanations
Derive predictions of new phenomena
Empirical Generalizations
Generalize from sample to population
Derive predictions for new applications of generalization
Data

Figure 1.1 Two levels of inferences in science.

over time. Many decades ago data analysis in psychology became highly formalized: As a result, it is now nearly inconceivable for contemporary empirical research to be presented without some supporting statistical analysis. Thus, on the first level of inference—involving data and empirical regularities—psychology has adapted rigorous tools for reducing bias and ambiguity in the inferential process. This process continues apace to this date, with new developments in statistics and methodology coming online at a rapid rate (e.g., Cramer et al., 2015; Wagenmakers, Verhagen, & Ly, 2015).

On the second level of inference—between theories and empirical generalizations—the picture is less homogeneous: Although there are several areas of enquiry in which rigorous quantitative and computational models are ubiquitous and indispensable to theorizing (e.g., in decision making, psychophysics, and categorization), in other areas more informal and purely verbal reasoning has retained a prominent role. When theorizing is conducted informally, researchers derive predictions from a theory by a mixture of deduction, mental simulation, and plausibility judgments. The risks of such informal reasoning about theories and their relation to data has long been known and repeatedly illustrated (Farrell & Lewandowsky, 2010; Lewandowsky, 1993; Lewandowsky & Farrell, 2011).

This chapter surveys the solution to those risks associated with informal theorizing—namely, the use of mathematical or computational models of memory and cognition. We begin by showing how the use of models can protect researchers against their own cognitive limitations, by serving as a kind of “cognitive prosthesis.” We next differentiate between different classes of models, before we discuss descriptive models, measurement models, and explanatory models in some detail. We then survey several cognitive architectures, large-scale endeavors to build models of human cognition.

MATHEMATICAL MODELS AS COGNITIVE PROSTHESIS

Models of Choice Reaction Time Tasks

Imagine an experiment in which participants are shown a cluster of 300 lines at various orientations and their task is to decide whether the lines slant predominantly to the left or to the right. This is a difficult task if the orientations of individual lines within the cluster are drawn from a distribution with high variance (e.g., Smith & Vickers, 1988).

The data from such "choice-reaction-time" experiments are strikingly rich: There are two classes of responses (correct and incorrect), and each class is characterized by a distribution of response times across the numerous trials of each type. To describe performance in a choice-reaction-time experiment would therefore involve both response accuracy and latency, and the relationship between the two, as a function of the experimental manipulations (e.g., variations in the mean orientation of the lines or in how participants are instructed to trade off speed and accuracy). There are a number of sophisticated models that can describe performance in such tasks with considerable accuracy (S. D. Brown & Heathcote, 2008; Ratcliff, 1978; Wagenmakers, van der Maas, & Grasman, 2007), all of which are based on the premise that when a stimulus is presented, not all information is available to the decision maker instantaneously. Instead, the models all assume that the cognitive system gradually builds up the evidence required to make a decision, although they differ with respect to the precise mechanism by which this accumulation can be modeled.

For the present illustrative example, we assume that people sample evidence in discrete time steps and keep summing the evidence until a decision is reached. At each step, a sample nudges the summed evidence toward one decision or another until a response threshold is reached. When deciding whether the 300 lines are predominantly slanted to the right or the left, each sampling step might involve the processing of a small number of lines and counting of the left-slanted vs. right-slanted lines. The sample would then be added to the sum of all previous samples, nudging the overall evidence toward the "left" or "right" decision. Figure 1.2 illustrates this "random walk" model with a number of illustrative sampling paths. Each path commences at time zero (i.e., the instant the stimulus appears) with zero evidence. Evidence is then sampled until the sum of the evidence is sufficient for a response, which occurs when the evidence exceeds one or the other response threshold, represented by the dashed horizontal lines (where the top line arbitrarily represents a "left" response and the bottom a "right" response).

The top panel shows what happens when the 300 lines in the stimulus are scattered evenly to the left and right. In that case, information is equally favorable to the two response alternatives, and hence the sampling paths are erratic and end up crossing each threshold (roughly) equally often. We would also expect the two response types to have identical response times on average: Sampling starts with zero evidence, and if the stimulus is noninformative, then each sample is equally likely to nudge the path up or down. It follows that if the boundaries for the two responses are equidistant from the origin, response times—that is, the point along the abscissa at which a sampling path crosses the dashed line—should be equal. With the small number of trials shown in the figure this cannot be ascertained visually, but if a large number of trials were simulated then this fact would become quite obvious.

What would happen if the evidence instead favored one decision over the other, as expected when an informative stimulus is present? Suppose most of the 300 lines were slanting to the left; in that case most of the evidence samples would be positive and as a result, this so-called drift would increase the probability of the evidence crossing the upper boundary. The bottom panel of Figure 1.2 illustrates this situation. All but one sampling

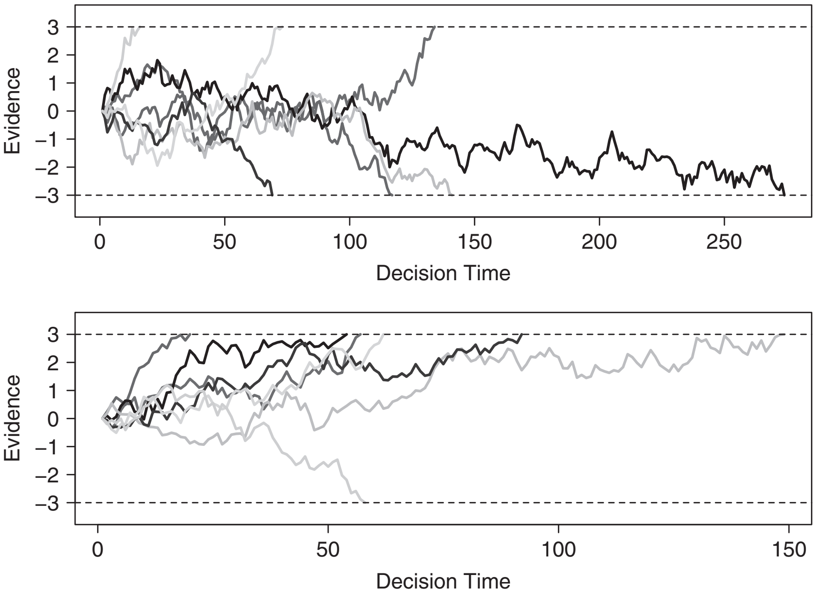

Figure 1.2 Graphical illustration of a simple random walk model. The top panel plots seven illustrative sampling paths when the stimulus is noninformative. The bottom panel plots another seven sampling paths with a drift rate toward the top boundary (representing a "left" response in the line-orientation task). Note the difference in the horizontal scale between panels. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.

paths cross the "left" boundary at the top, and only a single "right" response occurs. It is also apparent that the speed of responding is quicker overall for the bottom panel than the top. This not surprising, because having an informative stimulus permits more rapid extraction of information than a random cluster of 300 lines at varying orientations.

This brings us to the question of greatest interest: When an informative stimulus is present, what happens to the decision times for the less likely responses—that is, "right" responses that cross the bottom boundary—as the drift rate increases? Suppose there are many more trials than shown in the bottom panel of Figure 1.2, such that there is ample opportunity for errors ("right" responses) to occur. How would their response latencies compare to the ones for the correct ("left") responses in the same panel? Think about this for a moment, and see if you can intuit the model's prediction.

We suspect that you predicted that the decision time would be slower for the less likely responses. The intuition that an upward drift must imply that it will take longer for a random walk to (rarely) reach the bottom boundary is very powerful. You might have thought of the erroneous responses as a person struggling against a river current, or you might have pictured the sampling paths as rays emanating from the starting point that are rotated counterclockwise when drift is introduced, thereby producing slower responses when the lower boundary is accidentally crossed.

Those intuitions are incorrect. In this model, the mean response times—and indeed the entire distribution of response times—for both response types are identical, irrespective of drift rate. This property of the random walk model has been known for decades (Stone, 1960), but that does not keep it from being counterintuitive. Surely that swimmer would have a hard time reaching the bottom against the current that is pushing her toward the top? The swimmer analogy, however, misses out on the important detail that the only systematic pressure in the model *is the drift*. This is quite unlike the hypothetical swimmer, who by definition is applying her own counterdrift against the current. The implication of this is that paths that hit the bottom boundary do so only by the happenstance of collecting a series of outlying samples in a row that nudge the path against the drift. If there were additional time, then this would merely give the path more opportunity to be bumped toward the top boundary by the drift. It follows that the only errors the model can produce are those that occur as quickly as a correct response.

We argue that the behavior of this basic random-walk model is not at all obvious from its description. In our experience, most people resort to analogies such as the swimmer or the rays emanating from the origin in order to predict how the model will behave, therefore almost invariably getting it wrong. This example is a good illustration of the risks associated with relying on mental simulation to presage the behavior of models: Even very simple models can fool our unaided thinking.

Models of Rehearsal in Short-Term Memory

The potential for intuition to lead us astray is even greater when the processes involved are accessible to introspection. We illustrate this with the notion of maintenance rehearsal in short-term or working memory. From an early age onward, most people spontaneously rehearse (i.e., recite information subvocally to themselves) when they have to retain information for brief periods of time. When given the number 9671111, most people will repeat something like "967–11–11" to themselves until they report (or dial) the number. There is no question that rehearsal exists. What is less clear is its theoretical and explanatory status. Does rehearsal causally contribute to recall performance? Does it even "work"—that is, does rehearsal necessarily improve memory?

At first glance, those questions may appear unnecessary or indeed adventurous in light of the seemingly well-supported link between rehearsal and memory performance (e.g., D. Laming, 2008; Rundus, 1971; Tan & Ward, 2000). In a nutshell, many studies have shown that recall can be predicted by how often an item has been recited, and by the position of the last rehearsal. On closer inspection, however, those reports all involved free recall—that is, situations in which participants were given a list of words to remember and were then able to recall them in any order. This protocol differs from the serial recall that is commonly required in short-term memory situations: When trying to remember a phone number (such as 9671111), there is a distinct difference between dialing 9671111 (which earns you a pizza in Toronto) and dialing 1179611 (which gets you nowhere). Under those circumstances, when the order of items is important above and beyond their identity, does rehearsal support better memory performance?

Many influential theories that are formulated at a verbal level state that rehearsal is crucial to memory even in the short term. For example, in Baddeley's working memory model (e.g., Baddeley, 1986; Baddeley & Hitch, 1974), memories in a phonological short-term store are assumed to decay over time unless they are continually restored through rehearsal. Although there is no logical necessity for rehearsal to be accompanied by decay, models of short-term or working memory that include a rehearsal component are also presuming that unrehearsed memories decay inexorably over time (Baddeley, 1986; Barrouillet, Bernardin, & Camos, 2004; Burgess & Hitch, 1999; Daily, Lovett, & Reder, 2001; Kieras, Meyer, Mueller, & Seymour, 1999; Page & Norris, 1998; Oberauer & Lewandowsky, 2011). A sometimes tacit but often explicit claim in those models is that rehearsal is beneficial—that is, at the very least, rehearsal is seen to offer protection against further forgetting, and at its best, rehearsal is thought to restore memory to its original strength.

The implications of this claim are worth exploring: For rehearsal to restore partially decayed memory representations to their original strength when serial order is important implies that the existing trace must be retrieved, boosted in strength, and *re-encoded into the same position* in the list. If errors arise during retrieval or encoding, such that the boosted trace is assigned to a different position, then rehearsal can no longer be beneficial to performance. Recall of 9671111 can only be facilitated by rehearsal if the “9” is strengthened and re-encoded in position 1, the “6” remains in position 2 after rehearsal, the “7” in position 3, and so on.

It turns out that this successful rehearsal is difficult to instantiate in a computational model. We recently examined the role of rehearsal within a decay model in which items were associated to positions, and those associations decayed over time (Lewandowsky & Oberauer, 2015). We found that conventional articulatory rehearsal, which proceeds at a pace of around 250 ms/item, rarely served its intended purpose: Although the model reproduced the pattern of overt rehearsals that has been observed behaviorally (Tan & Ward, 2008), it was unable to simulate the associated recall patterns. Specifically, the model performed *worse* with additional time for rehearsal during encoding, whereas the data showed that performance increases with additional rehearsal opportunity.

Analysis of the model’s behavior revealed that this departure from the data arose for reasons that are not readily overcome. Specifically, rehearsal turns out to introduce a large number of “virtual” repetition errors (around 50% of all rehearsal events) into the encoded sequence. (As no items are overtly recalled during rehearsal, the errors are virtual rather than actual.) This contrasts sharply with observed recall sequences, which exhibit repetition errors only very infrequently (i.e., around 3% of responses; Henson, Norris, Page, & Baddeley, 1996). The excessive number of repetition errors is a direct consequence of the fact that rehearsal, by design, boosts the memory strength of a rehearsed item substantially.

The consequences of this strengthening of memory traces are outlined in Figure 1.3, which also outlines the model’s architecture. Items are represented by unique nodes (shown at the top of each panel) that are associated to preexisting position markers when an item is encoded. Multiple units represent the position markers, and the position markers partially overlap with each other. At retrieval (or during rehearsal), the position markers are used as retrieval cues. Recall errors arise from the overlap between markers, and also because the associations between the position markers and items decay over time.

Panel A shows the state of memory after two hypothetical items have been encoded and before rehearsal commences. Rehearsal commences by cueing with the first set of context markers. This cue retrieves the correct item (panel B), permitting the strengthening of the associations between it and the corresponding context markers (panel C). When the model next attempts to retrieve the second item for rehearsal, the overlap between adjacent position markers implies that the first item is again partially cued (panel D). Because the association of the first item to its position markers has just been strengthened, it may be activated more than the second item

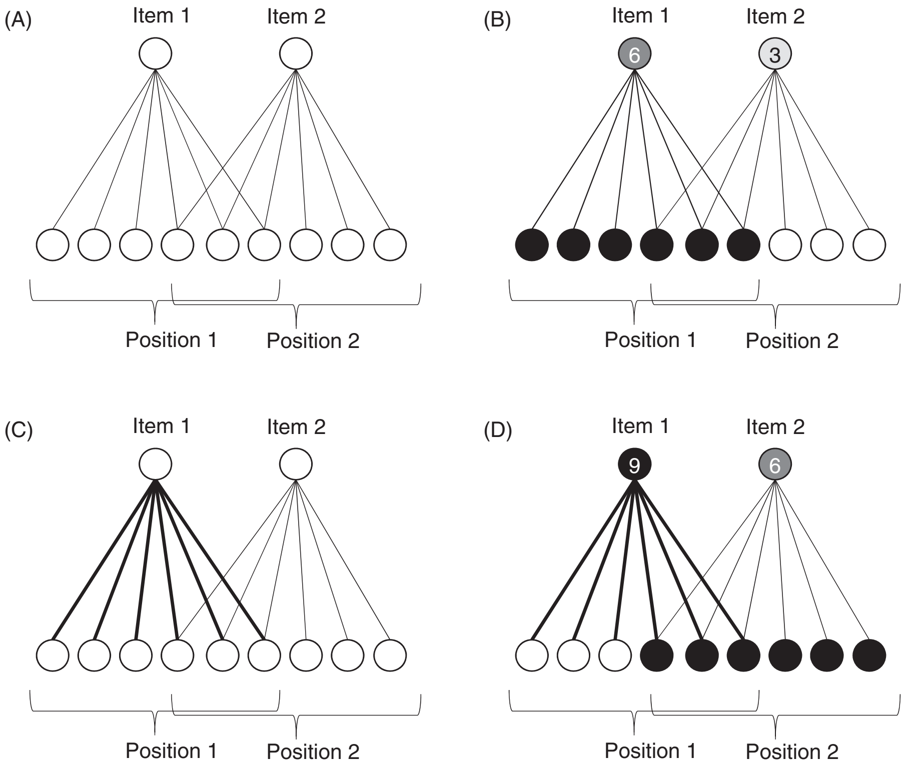

Figure 1.3 Effects of articulatory rehearsal on strengthening of two list items in a decay model that includes rehearsal. Shading of circles and superimposed numbers refers to the extent of activation of each item or context element (on an arbitrary scale), and thickness of lines indicates strength of association weights between an item and its context markers. Items are shown at the top and use localist representations; context is shown in the bottom and involves distributed representations. The layers are connected by Hebbian associations that are captured in the weights. Weights decay over time. Panel A shows the state of memory before rehearsal commences. Both items are associated to their overlapping context markers. Panel B: First item is cued for rehearsal by activating the first context marker. Item 1 is most active and is hence retrieved for rehearsal. Panel C: Item 1 is re-encoded and the context-to-item associations are strengthened (by a factor of 3 in this example). Panel D: The second item is cued for rehearsal but Item 1 is more active because of its recent rehearsal.
SOURCE: From Lewandowsky and Oberauer (2015). Reprinted with permission.

when the second item is cued, as is indeed the case in panel D.

In general, when item n has just been rehearsed, there is a high risk of retrieving item n again in position $n + 1$. The resultant encoding of a second copy of item n in position $n + 1$ introduces a virtual repetition error that subsequent rehearsal sweeps will likely reinforce. This problem is an inevitable consequence of the fact that rehearsal boosts items one at a time, thereby introducing an imbalance in encoding strength that often overpowers the cueing mechanism.[1]

[1] One might wonder why rehearsal does not involve the failsafe, nearly instant, and simultaneous amplification of all contents of memory. This alternative conception of rehearsal is ruled out by the fact that overt or covert articulation is necessarily sequential in nature and is known to proceed at a relatively slow pace. It is logically impossible for a slow sequential process to restore all list items.

This analysis implies that a reflexive verbal appeal to rehearsal in order to explain a memory phenomenon is not an explanation—it can only be the beginning of a process of examination that may or may not converge on rehearsal as an underlying explanatory process. That process of examination, in turn, cannot be conducted outside a computational model: Decades of verbal theorizing about rehearsal has continued to advance fairly undifferentiated claims about its effectiveness that eventually turned out to be overstated.

The Need for Cognitive Prostheses

The preceding two examples converge on two conclusions: First, no matter how carefully we may think about a conceptual issue, our cognitive apparatus may fail to understand the workings of even simple models, and it may readily misinterpret the implications of constructs that are specified at a verbal level. This can occur to any researcher, no matter how diligent and well intentioned.

There has been much emphasis recently on improvements to the way in which science is conducted, spurred on by apparent difficulties to replicate some findings in psychology and other disciplines (e.g., Munafò et al., 2014; see also Chapter 19 in this volume). Measures such as open data and preregistration of experiments have become increasingly popular in recognition of the fact that scientists, like all humans, may be prone to fool themselves into beliefs that are not fully supported by the evidence (Nuzzo, 2015). Researchers are not only prone to errors and biases in interpreting data—we argue that they are equally prone to make mistakes in interpreting theories. Computational models are one particularly useful tool to prevent theoreticians from making inconsistent assumptions about psychological mechanisms, and from deriving unwarranted predictions from theoretical assumptions. As we show next, models can serve this purpose in a variety of ways.

CLASSES OF MODELS

All models are comprised of an invariant structure and variable components, known as parameters, which adapt the structure to a particular situation. For example, the random-walk model considered earlier has a fixed structural component involving the sampling mechanism: The model is committed to repeatedly sampling evidence from a noisy source, and to accumulate that evidence over time until a decision threshold is reached. This invariant structural component is adapted to the data or experiment under consideration by adjusting parameters such as the location of the response thresholds. For example, if experimental instructions emphasize speed over accuracy, the response thresholds in the model are moved closer to the origin to produce faster (but likely less accurate) responses, without however altering the basic sampling structure. Similarly, if the stimuli contain a stronger signal (e.g., all lines are slanted in the same direction), this would be reflected in a higher drift rate but it would not alter the sampling structure.

One way to classify models is by considering the role of data in determining a model's structure and parameters. For example, in the physical sciences, a model's structure, as well as its parameters are specified a priori and without reference to data. Thus, the structure of models used for weather or climate forecasting is determined by the physics of heat transfer (among other variables) and their parameters are well-known physical constants, such as the Boltzmann constant, whose value is not in question. Both structure and parameters are known independently of the data and do not depend on the data (i.e., the historical climate or today's weather). There are few, if any, such well-specified models in psychology.

At the other end of the extreme, regression models that describe, say, response times as a function of trials in a training study, are entirely constructed in light of the data. Their structure—that is, the number and nature of terms in the model—as well as the parameters—that is, the coefficients of those terms—are estimated from the data. If the data are better characterized by a curvilinear relationship, then a quadratic or logarithmic component would be added to the model without hesitation to improve its fit with the data. We call those types of models *descriptive* models, and although they are most often associated with data analysis, they do have their theoretical uses as we show in the next section.

Most cognitive models, however, lie somewhere in between those extremes. Their structure is determined a priori, before the data for an experiment are known, based on theoretical or conceptual considerations. For example, the random-walk model's development was influenced by theoretical statistics, in particular the optimal way to conduct a sequential hypothesis test (Wald, 1945). The model's structure, therefore, remains invariant, irrespective of which data set it is applied to (which is not to ignore that other variants of sampling models have been developed; e.g., Smith & Ratcliff, 2015, but their development was not a simple result of data fitting). We call those models *theoretical* models later because their structure incorporates theoretical commitments that can be challenged by data.

Descriptive Models

We already noted that descriptive models do not have an a priori structure that is defined before the data are known. They may, therefore, appear to be mere statistical tools that, at best, provide a summary of an empirical regularity. This conclusion would be premature: Even though descriptive models are, by definition, devoid of a priori structure, this does not mean they cannot yield structural insights. Indeed, one of the aims of applying descriptive models to data may be the differentiation between different possible psychological structures.

To illustrate, consider the debate on whether learning a new skill is best understood as following a "Power Law" or is better described by an exponential improvement (Heathcote et al., 2000). There is no doubt that the benefits from practice accrue in a nonlinear fashion: Over time and trials, performance becomes more accurate and faster. What has been less clear is the functional form of this empirical regularity. For decades, the prevailing opinion had been that the effect of practice is best captured by a "Power law"; that is, a function that relates response speed (RT) to the number of training trials (N); thus, $RT = N^{-\beta}$. The parameter β is the learning rate, and when both sides of the equation are transformed logarithmically, the power function becomes a nice linear relationship: $\log(RT) = -\beta \times \log N$.

An alternative view, proffered by Heathcote et al. (2000), suggests that practice effects are better described by an exponential function: $RT = e^{-\alpha \times N}$, where the parameter α again represents a learning rate. Why would it matter which function best describes practice data? It turns out that the choice of descriptive model carries implications about the psychological nature of learning. The mathematical form of the exponential function implies that the proportional improvement, relative to what remains to be learned, is constant throughout practice—no matter how much you have already practiced, learning continues apace. By contrast, the mathematics of the power function imply that the relative learning rate is slowing down as practice increases. Although performance continues to improve, the rate of that improvement *decreases* with further practice. It follows that the proper characterization of skill acquisition data by a descriptive model, in and of itself, has psychological implications: If the exponential function is a better descriptor of learning, then any explanation of practice effects has to accommodate this by postulating a practice-invariant underlying process. Conversely, if the power function is a better descriptor, then the underlying process cannot be practice-invariant.

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, in particular the question whether the rate of forgetting differs with retention interval. The issue is nuanced, but it appears warranted to conclude that the rate of forgetting *decelerates* over time (Wixted, 2004a). That is, suppose 30% of the information is lost on the first day, then on the second day the loss may be down to 20% (of whatever remains after day 1), then 10%, and so on. Again, as in the case of practice, the function itself has no psychological content but its implications are psychological: The deceleration in forgetting rate may imply that memories are "consolidated" over time *after* study (e.g., Wixted, 2004a, 2004b).

Theoretical Models

Within the class of theoretical models, we find it helpful to differentiate further between what we call "measurement models," which capture a complex pattern of data and replace those data by estimates of a small number of parameters, and what we call "explanatory models," which seek to provide a principled explanation of experimental manipulations. As we show next, the difference between those two types of theoretical models revolves around the role of the parameters.

MEASUREMENT MODELS

The problem appears simple: Suppose there are two participants in the earlier experiment involving the detection of the predominant slant of a cluster of 300 lines. Suppose that across a wide range of stimuli, participant A performs at 89% accuracy, with a mean response latency (for correct responses) of 1,200 ms. Participant B, by contrast, performs at 82% with a mean latency of 800 ms. Who is the better performer? Equivalently, suppose the preceding example involved not two participants but two experimental conditions, A and B, with the mean across participants as shown earlier. Which condition gives rise to better performance?

This problem does not have a straightforward solution because speed and accuracy are incommensurate measures. We cannot determine how many milliseconds a percentage point of accuracy is worth. There is no independently known transformation that converts accuracy into speed. We can express response times variously in seconds, minutes, milliseconds, or even nanoseconds, but we cannot express response times in terms of accuracy or vice versa. We therefore cannot readily compare two individuals or experimental conditions that differ in accuracy and speed but in opposite directions.[2]

Enter the measurement model. The solution to the problem is to re-express both accuracy and speed of responding within the parameter space of a model that can describe all aspects of performance in the experiment.

[2] If a person or condition is slower *and* less accurate than another person or condition, then we can at least make an ordinal inference about which is worse without having to worry about scale incommensurability.

Translating Data Into Parameters

We illustrate the basic idea of reexpressing complex data as parameters within the random-walk model discussed at the outset. We noted already that the model can provide information about the accuracy as well as the speed of responding, and we noted that the drift rate was a crucial parameter that determined which response boundary the model would, on average, approach, and at what speed (Figure 1.2). We will use this type of model architecture to reexpress the observed speed and accuracy of responding by a participant (or in an experimental condition) within the model's parameter space. To foreshadow, we understand the drift rate to be an indicator of performance, as it "characterizes the quality of evidence accumulation and can be influenced by stimulus characteristics as well as by individual differences in processing efficiency" (Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007, p. 416). Hence, if person A has a greater drift rate than person B, then we can say that A performs the task better than B.

Measurement Models Are Falsifiable

We begin our exploration of measurement models by revisiting one of the properties of the random-walk model presented at the outset. It will be recalled that the model in Figure 1.2 predicts identical latencies for errors and correct responses. This prediction is at odds with the empirical fact that errors can be either fast or slow, but are rarely equal in speed to correct responses (Ratcliff, Van Zandt, & McKoon, 1999). (As a first approximation, fast errors occur when the subject is under time pressure and discriminability is high, whereas errors are slow when the task is more difficult and time pressure is relaxed; Luce, 1986.)

The random-walk model, in other words, fails to capture an important aspect of the data. In the present context, this "failure" is welcome because it highlights the difference between a descriptive model and a theoretical measurement model: A descriptive model can never fail to capture (nonrandom) data, because its structure can be revised on the basis of the same data until it matches the observations. A theoretical measurement model, by contrast, is committed to certain structural properties, and like the simple random-walk model it can in principle be falsified by a failure to fit the data.

The "failure" of the simple random-walk model to handle error response times has been known for over half a century (Stone, 1960), and the model has evolved considerably since then. Modern theories of choice response times have inherited the sequential-sampling architecture from the random-walk model, but they have augmented in other important ways that enable them to provide a convincing account of accuracy and response times.[3]

Measurement Models of Decision Latencies

The key to the ability of sequential-sampling architectures to handle error latencies turns out to be trial-to-trial variability in some parameter values. This trial-to-trial variability differs from the noise (i.e., variability) that is inherent in the accumulation process, and which in Figure 1.2 showed up as the jitter in each accumulation trajectory toward one or the other boundary.

Trial-to-trial variability is based on the plausible assumption that the physical and psychological circumstances in an experiment never remain invariant: Stimuli are encoded more or less well on a given trial, people may pay more or less attention, or they

[3] Those modifications and extensions have not imperiled the model's falsifiability (Heathcote, Wagenmakers, & Brown, 2014; Ratcliff, 2002).

may even jump the gun and start the decision process before the stimulus is presented.

There are two parameters whose variability across trials has been considered and has been found to have powerful impact on the model's prediction: Variability in the starting point of the random walk, and variability in the drift rate (e.g., Ratcliff & Rouder, 1998; Rouder, 1996).

Returning briefly to Figure 1.2, note that all random walks originate at 0 on the ordinate, instantiating the assumptions that there is no evidence available to the subject before the stimulus appears and that sampling commences from a completely neutral state. But what if people are pressed for time and sample "evidence" before the stimulus appears? In that case the starting point of the random walk—defined as the point at which actual evidence in the form of the stimulus becomes available—would randomly differ from 0, based on the previous accumulation of (nonexistent) "evidence" that is being sampled prematurely.

Introducing such variability in the starting point drastically alters the model's predictions. Errors are now produced much more quickly than correct responses (D. R. J. Laming, 1968). This outcome accords with the observation that under time pressure, people's errors are often very quick. It is easy to see why errors are now faster than correct responses. Suppose that there is a high drift rate that drives most responses toward one boundary (e.g., the upper boundary as shown in the bottom panel of Figure 1.2). Under those conditions it requires an unlucky coincidence for any random walk to cross the lower boundary. The opportunity for this unlucky coincidence is enhanced if the starting point, by chance, is below the midpoint (i.e., < 0). Thus, when errors arise, they are likely associated with a starting point close to the incorrect boundary and hence they are necessarily quick. Of course, there is a symmetrical set of starting points above the midpoint, but those fast responses constitute a much smaller proportion of correct responses compared to the errors.

We next consider introducing variability in the drift rate from trial to trial, to accommodate factors such as variations in encoding strength between trials. Thus, on some simulated trials the drift will be randomly greater than on others. When this variability is introduced, error responses are now slower than those of correct responses (Ratcliff, 1978). To understand the reasons for slow errors, we need to realize that drift rate affects both latency and the relative proportions of the two response types. Suppose we have one drift rate, call that d1, which yields a proportion correct of 0.8 and, for the sake of the argument, average latencies of 600 ms. Now consider another drift rate d2, which yields proportion correct 0.95 with a mean latency of 400 ms. If we now suppose that d1 and d2 are (the only) two samples from a drift rate with trial-to-trial variability, then we can derive the latency across all trials (presuming there is an equal number with each drift rate) by computing the probability-weighted average. For errors, this will yield $(.05 \times 400 + .20 \times 600)/.25 = 560$ ms. For correct responses, by contrast, this will yield $(0.95 \times 400 + 0.80 \times 600)/1.75 = 491$. (To form a weighted average we divide not by the number of observations but by the sum of their weights.) It is easy to generalize from here to the case where the drift rate is randomly sampled on each trial. Errors will be slower than correct responses because drift rates that lead to faster responses will preferentially yield correct responses rather than errors and vice versa.

When both sources of trial-to-trial variability are combined, modern random-walk models can accommodate the observed relationship between correct and error latencies (Ratcliff & Rouder, 1998). Specifically, a continuous version of the random-walk model, known as the diffusion model (e.g., Ratcliff, 1978), can quantitatively accommodate the fast errors that subjects show in a choice task under speeded instructions, as well as the slow errors they exhibit when accuracy is emphasized instead. Further technical details about this class of models are provided in Chapter 9 in this volume.

To summarize, we have shown that theoretical measurement models, unlike descriptive models, are in principle falsifiable by the data (see also Heathcote, Wagenmakers, & Brown, 2014). We therefore abandoned the simple random-walk model—which fails to handle error latencies—in favor of contemporary variants that include trial-to-trial variability in starting point and drift rate. Those models handle the empirically observed relationship between correct and error latencies, thereby allowing us to map complex data into the relatively simple landscape of model parameters.

Using Measurement Models to Illustrate Performance Differences

We illustrate the utility of measurement models by focusing on the diffusion model (e.g., Ratcliff, 1978). The model has a long track record of explaining variation in performance, either between different groups of people or between individuals.

For example, the literature on cognitive aging has been replete with claims that older adults are generally slower than young adults on most tasks (Salthouse, 1996). This slowing has been interpreted as an age-related decline of all (or nearly all) cognitive processes, and because many everyday tasks entail time limitations, the decline in speed may also translate into reduced accuracy. Contrary to this hypothesis, when young and old participants are compared within a diffusion-model framework, the observed response time differences across a number of decision tasks (e.g., lexical decision) are found to be due primarily to the older adults being more cautious than the younger adults: What differs with age is the boundary separation but, in many cases, not the drift rate (Ratcliff, Thapar, & McKoon, 2010). That is, in Figure 1.2 the horizontal dashed lines would be further apart for older participants than younger people, but the average slopes of the accumulation paths in the bottom panel would be identical across age groups. (There are some exceptions, but for simplicity we ignore those here.)

By contrast, when performance is compared across people with different IQs, then irrespective of their age, drastic differences in drift rate are observed. Boundary separation is unaffected by IQ (Ratcliff et al., 2010). Thus, whereas aging makes us more cautious, our ability to quickly accumulate information for a decision is determined not by age but by our intelligence.

The impact of those two results can be highlighted by noting that at the level of mean response times, the effects of aging are one of general slowing (Ratcliff, Spieler, & Mckoon, 2000; Salthouse, 1996), as are the effects of (lower) IQ (Salthouse, 1996; Sheppard & Vernon, 2008). Looking at mean response time alone might therefore suggest that aging and (lower) IQ have similar effects. It is only by application of a measurement model that the striking differences become apparent within the landscape of model parameters.

Using Measurement Models to Understand Neural Imaging

Measurement models have proven to be particularly useful in the neurosciences. The basic objective of the cognitive neurosciences is to understand cognitive processes; however, this understanding is often hampered because the relationship between behavioral data and their neural correlates is typically opaque. For example, a correlation between response times and activation in a certain brain region has unclear implications without further theory. Conversely, the failure to observe a correlation between response times and brain activation in regions of interest may arise because mean differences in response times obscure some substantive differences in cognitive processes that become apparent only through application of a model.

The importance of measurement models in the neurosciences can again be illustrated through the diffusion model. At the lowest level of analysis, it has repeatedly been shown that in areas known to be implicated in decision making (lateral intraparietal cortex and other parts of the prefrontal cortex in monkeys and rats; for a detailed discussion see Forstmann, Ratcliff, & Wagenmakers, 2016), activity in single neurons increases over time to a constant maximum that is unaffected by decision-relevant variables such as difficulty of the choice. This observation is compatible with the idea that evidence is accumulated until a relatively invariant decision threshold is reached. Remarkably, the buildup of activation can be modeled by the evidence-accumulation process in the diffusion model, using parameters that were estimated from the behavioral data (Ratcliff, Cherian, & Segraves, 2003). Thus, the accumulation trajectories shown in the bottom panel of Figure 1.2 are not just abstract representations of a decision process but appear to have a direct analog in neural activity.

Although the results from single-cell recordings in animals are promising, it is unclear whether humans approach choice tasks in the same way as animals (Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015). Moreover, single-cell recordings provide only a microscopic snapshot of neural activity, and the linkage between single cells and complex behavior is often difficult to ascertain. Those problems can be circumvented by using functional imaging with humans.

The use of functional magnetic resonance imagery (fMRI) to augment purely behavioral data has become almost routine in cognitive science. Henson (2005) provides an eloquent case for the use of fMRI data, arguing convincingly that it can contribute to our understanding of cognition under some reasonable assumptions. Most relevant in the present context is the fact that brain activity in certain key regions has been systematically related to parameters within decision models. For example, if people's time to respond is curtailed experimentally, they become less cautious and responses are faster but less accurate (e.g., Forstmann et al., 2008). If that variability in behavior can be captured by changes in a model parameter, and if those parameter estimates in turn are correlated with activity in specific brain regions, then inferences about neural substrates of decision making become possible that could not have been detected by analyzing the raw data alone.

Mulder, van Maanen, and Forstmann (2014) reviewed the available relevant studies and found that task manipulations that affect speed and accuracy of responding involve regions of the frontobasal ganglia network. Specifically, a number of studies have shown that the anterior cingulate cortex (ACC), the pre-supplementary motor area (pre-SMA), and striatal regions are associated with people's setting of the decision boundaries. It has been argued that those regions, in particular the ACC, serve as a "control unit to adjust the response threshold via the striatum" (Mulder et al., 2014, p. 878).

Summary

In summary, measurement models can serve as an intermediate conceptual layer that bridges behavioral data with theoretical constructs or their neural substrates via the model's parameters. These parameters can serve as dependent variables in experiments and as correlates of other behavioral or neural variables.

The defining attribute of measurement models is that they are applied separately to each experimental condition or each individual, estimating separate parameter values for each condition and each person. Thereby, the models translate the variability across conditions or across individuals from the initial, purely descriptive scales of measurement that are often incommensurate (e.g., milliseconds, proportion correct) into a theoretically interpretable scale (e.g., drift rate as a measure of information processing efficiency).

At the same time, measurement models do not aim to explain that variability. For example, drift rates differ between different set-sizes in short-term recognition tasks (Ratcliff, 1978), and between people with different IQs (Ratcliff et al., 2010), but a measurement model cannot explain how these differences come about—it can only characterize them. In contrast, the aim of explanatory models is to explain performance differences between experimental conditions by reproducing these differences with a common set of parameters across conditions.

EXPLANATORY MODELS

What does it mean to explain anything? In modern science, an "explanation" is commonly interpreted as identifying causes for an event or phenomenon of interest (Sun, Coward, & Zenzen, 2005). In psychology this generally implies that we seek to identify the psychological processes that cause an observed outcome. The fact that those processes are unobservable is not necessarily of concern; contemporary physics, too, relies on unobservable constructs such as quarks, leptons, or mesons. More specifically, when we seek explanations within computational models, we want those explanations to "fall out" of the model's structure, rather than being the result of variations in parameter values. The reason for this is simple: If we estimate parameters for each condition in an experiment, then our "explanation" for differences between those conditions is informed by the very data that we seek to explain. To avoid this circularity, explanatory models generally do not allow the estimated parameters to vary between conditions that are to be explained.

Explaining Scale Invariance in Memory

We illustrate explanatory models with SIMPLE (scale-invariant memory, perception and learning); a memory model that has been successfully applied to a wide range of phenomena in short-term and long-term memory (G. D. A. Brown, Neath, & Chater, 2007). SIMPLE explains accuracy of memory retrieval based on a target item's discriminability from other potential recall candidates. SIMPLE's primary claim is that list items are represented in memory along the temporal dimension; when we recall something, we look back along that temporal dimension and try to pick out the target memory from other memories that occurred at around the same time. This means that the separation of events in time determines the accuracy of their recall. Items that are crowded together in time (a specific daily commute to work among many other such commutes) are more difficult to recall than isolated events (your annual holiday).

Another assumption of SIMPLE is that the temporal dimension is logarithmically compressed: As items recede into the past, they become more squashed together, just as equidistant telephone poles appear to move closer together as they recede into the distance when viewed from the rear of a moving car (Crowder, 1976). Taken together, these two assumptions of SIMPLE give rise to a property that is known as "scale invariance"; that is, the model predicts that what should determine memory performance is the *ratio* of the times at which two items are presented, not their absolute separation in time. Specifically, two items that were presented 2 and 1 second ago, respectively, are as discriminable as two items that were presented 20 and 10 seconds ago. This scale invariance arises because any ratio of temporal distances is equivalent to a difference in distance in logarithmic space. Specifically, in logarithmic temporal space the separations within the pair presented 2 and 1 seconds ago (log(2) − log(1)) and within the items from 20 and 10 seconds ago (log(20) − log(10)) are identical.

It follows that the presumed distinctiveness process embodied in SIMPLE entails the strong prediction that performance should be invariant across different time scales, provided the ratio of retention intervals is equal. SIMPLE is therefore committed to making a prediction across different conditions in an experiment: Any experimental condition in which two items are presented 1 and 2 seconds, respectively, before a memory test must give rise to the same performance as a condition in which the two items are presented 10 and 20 seconds before the test. Note how this prediction differs from the ability of measurement models discussed earlier, which cannot express a strong commitment to equality between conditions. At best, measurement models such as the diffusion model or other sequential-sampling models can make ordinal predictions, such as the expectation that instructions emphasizing speed should accelerate responding at the expense of accuracy (but even that expectation requires a theoretical interpretation of the model; namely, that instructions translate into boundary placement).

To illustrate the role of explanatory models, we present a test of this prediction of SIMPLE that was reported by Ecker, Brown, and Lewandowsky (2015). Their experiment involved the presentation of two 10-word lists that were separated in time, the first of which had to be recalled after a varying retention interval (the second list was also tested, but only on a random half of the trials, and performance on that list is of no interest here.) The crucial manipulation involved the temporal regime of presentation and test, which is shown in Figure 1.4. The regime shown in the figure instantiates the ratios mentioned earlier: In the LL condition, the first list (L1) was presented 480 s before the test (we ignore the few seconds to present L2), and 240 s before L2. In the SS condition, L1 appeared 120 s before the test and 60 s before L2. According to SIMPLE, the temporal discriminability of L1 is therefore identical in both conditions because log(480) − log(240) = log(120) − log(60).

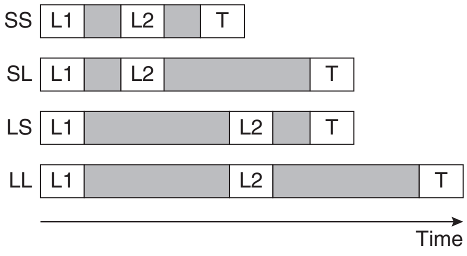

Figure 1.4 A schematic summary of the four experimental conditions used by Ecker et al. (2015). L1 and L2 denote the two study lists. T denotes the recall test, which always targeted L1. The temporal intervals were either 60 s (short gray bars) or 240 s (long gray bars). The four conditions are labeled SS (short L1–L2 interval, short L2–T interval), SL (short–long), LS (long–short), and LL (long–long).

Source: From Ecker, Brown, and Lewandowsky (2015). Reprinted with permission.

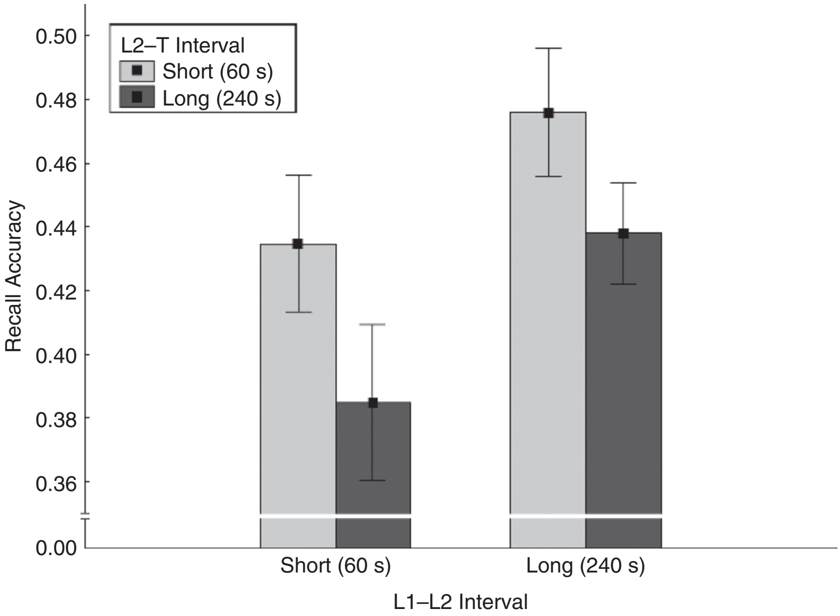

Figure 1.5 Recall accuracy for L1 in the study by Ecker et al. (2015). Error bars represent standard errors, and L1 and L2 refer to the first and second list presented for study, respectively. See Figure 1.4 for explanation of the temporal regime.
SOURCE: From Ecker, Brown, and Lewandowsky (2015). Reprinted with permission.

The results of Ecker et al. (2015) are shown in Figure 1.5. Here we are particularly concerned with the comparison between the SS condition (light gray bar on the left) and the LL condition (dark gray bar on the right). It is apparent that performance in those two conditions is nearly identical, exactly as predicted by SIMPLE. This result is quite striking, given that in the LL condition, the retention interval for L1 was 4 times greater than in the SS condition (480 s vs. 120 s). Any memory model that relies on absolute durations to predict performance can be expected to have difficulty with this result.

We conclude that SIMPLE explains the results of the study by Ecker et al. (2015) because it predicts that performance should be equal across the SS and LL conditions, and this prediction arises as a logical implication of the model's basic assumptions. The flipside of this explanation is that alternative empirical outcomes could falsify the model—if performance had not been equal between the SS and LL conditions, then SIMPLE would have great difficulty explaining that outcome.

Explanatory Necessity Versus Sufficiency

The fact that a model fits the data implies that it is *sufficient* to explain those data. However, it does not follow that the model is also *necessary*. That is, the fact that SIMPLE successfully predicted the SS and LL conditions to yield equal performance does not rule out the possibility that other models might also explain that equality. Indeed, the existence of such alternative models can be taken for granted (Anderson, 1990).

This is an in-principle problem that cannot be side-stepped by improving the quality of the data or of the model, and at first glance it might call into question the logic and utility of modeling. However, upon closer inspection we suggest that the problem is not quite that serious: First, the fact that many potentially realizable alternative models exist does not imply that any of those models are easy to come by. Quite on the contrary! Constructing cognitive models is an effortful and painstaking process whose success is not always ensured. Second, the existence of an unknown number of potential alternative models that reproduce empirical data patterns does not prevent us from comparing a limited set of *known* models and selecting the best one from that set.

This model-selection process can again be illustrated using the study by Ecker et al. (2015).

Model Selection and Model Complexity

The broader purpose of the study by Ecker et al. (2015) was to pit the distinctiveness approach embodied in SIMPLE against the notion of consolidation of memories. Consolidation is a presumed process that occurs after encoding of memories and serves to strengthen them over time—in particular during sleep or periods of low mental activity. Memories are said to become increasingly resistant to forgetting as they are being consolidated (Wixted, 2004b, 2004a).

The consolidation view is supported by the fact that recall of a list is poorer when a second, interfering list follows closely in time rather than when the second list is delayed. Müller and Pilzecker first reported this result more than a century ago (1900). In terms of the design in Figure 1.4, the consolidation view expects L1 recall to be better in condition SL than in condition LS, even though the overall retention interval is identical across both conditions. Indeed, Ecker et al. (2015) obtained this result; compare the dark gray bar on the left with the light gray bar on the right in Figure 1.5.

However, could the consolidation view accommodate the fact that the LL and SS conditions yielded identical performance? Given that L1 has less time to consolidate in the SS condition than in the LL condition, it is unclear how the consolidation view would accommodate these results. To explore whether consolidation might contribute to explaining their results, Ecker et al. (2015) created more than 30 models that combined the distinctiveness notion in SIMPLE with several presumed consolidation processes. Because consolidation as a computational process has not been well-specified in the literature (Ecker & Lewandowsky, 2012), multiple different variants of consolidation had to be compared. All variants shared, however, one characteristic: They increased the distinctiveness of L1 in memory after encoding, to reflect the assumption that memories become more retrievable over time as they are being consolidated.

Table 1.1 shows the results for the six top models in their study. The top entry

Table 1.1 Best-Fitting Models in Experiment 1 of Ecker et al. (2015)

Model	*N* (pars)	Deviance	*AICc* wt	*BIC* wt
1d SIMPLE (no consolidation)	4	4569	0.33	0.38
2d SIMPLE (equally weighted dimensions, no consolidation)	4	4560	0.36	0.42
2d SIMPLE (free dimension weight, no consolidation)	5	4552	0.12	0.09
1d SIMPLE (linear consolidation)	5	4552	0.13	0.09
2d SIMPLE (free dimension weight, linear consolidation)	6	4548	0.04	0.02
2d SIMPLE (free dimension weight, nonlinear consolidation)	7	4523	0.02	0.01

NOTE: 1d, one-dimensional; 2d, two-dimensional; *N* (pars), number of free model parameters; Deviance, summed deviance across all participants; *AICc* and *BIC* wt, information criterion weights

(1d SIMPLE, no consolidation) refers to the unmodified version of SIMPLE described earlier: All items are represented along a temporal dimension that is logarithmically transformed, and retrieval is a sole function of discriminability along that dimension. The entries labeled 2d SIMPLE add a second representational dimension that permits the two lists to be further differentiated by a change in context. That is, in addition to time, the memory representation is organized by the context that accompanies each list. As before, items are retrieved based on how easily they can be differentiated from their neighbors, except that in this instance the differentiation occurs in two-dimensional space rather than along a single temporal dimension. That is, the two lists are not just separated along the temporal axis, but also offset along an orthogonal abstract context dimension that takes the same value within each list but differs between lists. Because the lists are offset along that second dimension, they are separated further from each other more than the temporal dimension alone would suggest, similar to the way in which the distance between your home and a neighbor's home a fixed distance down the road is greater if the neighbor's driveway is excessively long. The two dimensions are either equally weighted when discriminability is computed, or their respective contributions can be freely estimated. Finally, the models that contain consolidation additionally enhance the discriminability of L1 over time by sharpening its representation in space: All items in SIMPLE have a "fuzzy" position along the temporal dimension (and others if they are present), and the extent of that fuzz was gradually decreased over time when consolidation was present.

To interpret the results in Table 1.1, it must be noted that the models differed with respect to the number of parameters that had to be estimated from the data. The standard SIMPLE had four parameters and the most complex consolidation version had seven. Although the parameters did not differ between conditions—that is, irrespective of the duration of the L1–L2 interval or the retention interval, all parameter values were the same—in general any model will accommodate the data with greater precision if it has access to more parameters (for details, see Lewandowsky & Farrell, 2011). This basic fact is reflected in the Deviance column, which presents the discrepancy between the data and the model's prediction (the scale of the deviance measure is somewhat arbitrary and need not concern us here). It can be seen that as the number of parameters increases, the deviance is reduced—that is, the more flexible models fit better than the simpler ones. The most flexible model with two dimensions and nonlinear consolidation yields a deviance of 4,523, compared to the unmodified SIMPLE whose deviance is 4,569.

At first glance, one might therefore prefer the most complex model because it fits the data better than any of the others, and one might therefore interpret the modeling as providing evidence for the existence of consolidation in memory. This conclusion would be premature because it does not consider the trade-off between a model's goodness-of-fit (the deviance in Table 1.1) and model complexity (the number of parameters). This trade-off is often called the bias-variance trade-off (e.g., Forster, 2000) and refers to the necessary fact that if the model is underspecified (i.e., not complex enough), we will miss accounting for important effects in the data and our model will be biased. Conversely, if our model has too many parameters, we will overfit the data and will be explaining noise as well as real effects. Thus, a good fit by itself does not support a model's viability if it arises from fitting statistical noise in addition to capturing the real effects in the data.

Table 1.1 contains two additional statistics—AIC weights and BIC weights—that deal with this trade-off and permit a more informed model selection. Both AIC and BIC pit goodness-of-fit (represented by the deviance) against model complexity (estimated by the number of parameters). Introducing more parameters will improve the fit by reducing the deviance, but it will also increase the size of the penalty term for complexity. The BIC and AIC, therefore, instantiate the principle of parsimony: to find the *best* and *simplest* model. The chapter by Myung and Pitt in this volume presents the AIC and BIC in greater detail and also addresses the issue of model selection in depth.

Here, it suffices to point out that when a set of models are compared, the values of AIC and BIC can be turned into weights that represent the probabilities of each model being the best model among the set of candidates, given the data at hand. It is these weights that are shown in the final two columns of Table 1.1. The AIC and BIC weights permit a straightforward interpretation: the two versions of SIMPLE that do not incorporate consolidation and do not estimate a weighting parameter between representational dimensions are the "best" models for the results of Ecker et al. (2015). None of the models involving consolidation have a notable chance of being the best when the trade-off between goodness-of-fit and complexity is considered. We therefore conclude that the data support the role of distinctiveness rather than consolidation in memory.

Quantitative Fit and Qualitative Predictions

A good quantitative fit, as indexed by AIC, BIC, and other fit indicators, is not the only yardstick by which to assess and compare models. A model that reproduces a large number of findings across many different experimental paradigms in a coarse, qualitative fashion arguably contributes more to our theoretical understanding of the human mind than a model that makes very precise, accurate predictions in a narrow domain of data, such as the findings from a single experimental paradigm. For instance, in the field of memory, much intellectual energy has been invested into determining whether so-called receiver-operating characteristic (ROC) curves from recognition tests are better fit by signal-detection models, high-threshold models, or dual-process models (Bröder & Schütz, 2009; Wixted, 2007; Yonelinas & Parks, 2007). If ever a model emerges to win this battle, it will have conquered only a small corner of the empirical landscape of memory research, because the ROC curve is just one dependent measure from one paradigm for studying human episodic memory. More comprehensive models of memory such as REM (Shiffrin & Nobel, 1997) or the temporal-clustering and sequencing model of recall (Farrell, 2012) usually do not even take part in the competitions for quantitative model fit, but they offer explanations for a broad range of findings by reproducing their qualitative pattern. There is arguably a trade-off between achieving a good quantitative account of one or a few data patterns with a minimal set of assumptions and parameters on the one hand, and accounting comprehensively for a broad range of benchmark findings in a research area in a less precise manner, and with a larger number of assumptions.

Whereas sophisticated methods have been developed to select between models that compete for quantitative fit of a given data set (see Chapter 3 in this volume), there exists no established method for adjudicating between competing comprehensive models that aim to provide an integrated explanation for a large set of findings in a domain, if only qualitatively. The problem is that these models are built to explain sets of findings that only partially overlap. For instance, some memory models, such as SIMPLE and the temporal-clustering and sequencing model, account for detailed patterns of data from recall tests but have been applied only sparsely, if at all, to phenomena from recognition tests, whereas other models such as REM cover much ground in recognition but have only begun to be applied to recall. In addition, each model has idiosyncratic strengths from successfully predicting new findings that the competing models cannot readily account for, and the authors of models have a natural inclination to emphasize the findings that their model predicts as particularly diagnostic. A fair competition between models that vie for a comprehensive, integrated explanation of findings in a broad domain of investigation requires a consensus on which findings in that domain count as benchmark findings that all models should aim to explain. Sets of benchmark findings have been proposed in some domains, such as eye movements in reading (Rayner, 2009), word reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), and immediate serial recall (Lewandowsky & Farrell, 2008) but so far there are no established criteria for determining which findings qualify as benchmarks in a field.

A second important role for qualitative model predictions is to discriminate between models, or classes of models, that are difficult to discriminate quantitatively. The competition between signal-detection, dual-process, and high-threshold models of recognition offers an example: These models all give reasonably close quantitative fits to ROC curves from numerous variants of item and associative recognition experiments, and the differences between their predictions for these data are subtle. To compound the problem, the original models in each class (e.g., the standard equal-variance signal-detection model) can be made much more flexible by relaxing auxiliary assumptions, such as the assumption that signal and noise distribution are equal, or that signals are normally distributed. These model classes can nevertheless be distinguished by qualitative predictions that follow from their core properties independent of auxiliary assumptions. For instance, high-threshold models assume that sometimes memory simply fails, leaving the person in a state of no memory information at all. This assumption entails the principle of conditional independence: If the person is in a memory-failure state, their behavior is independent of any memory variable such as the strength of the memory representation they tried to retrieve (Kellen & Klauer, 2015; Province & Rouder, 2012; see also Chapter 5 in this volume).[4]

Summary

We have shown that explanatory models make testable predictions and are therefore subject to falsification. We have also shown that competing theoretical notions can be instantiated in different models, which can then be compared with respect to their ability to explain the data from an experiment. The "best" model is not always the one that fits the data best, but it is the model that achieves the best possible fit with the least degree of flexibility possible. At the same time, training our microscope exclusively onto

[4] It does not follow that a continuous signal-detection model cannot also be in a state of failed memory. However, except for some special and unlikely circumstances, the signal-detection model will always assume the presence of some residual memory, at least when averaged across trials in an experimental condition. This residual memory ensures that the model will violate conditional independence, thereby permitting empirical test.

subtle differences in quantitative model fit can induce a short-sighted perspective on a narrow set of phenomena on which the model competition is played out. Searching for new qualitative predictions that distinguish between models can help to solve an impasse in model selection. In the same way that the availability of a robot does not preclude use of a hammer when it is appropriate, there are situations in which simple experiments involving very plain statistics can play a useful role. Moreover, at our present state of ignorance, a broad, integrative explanation of phenomena in a field of research can be more decisive than advanced model-fitting techniques. The perhaps broadest and most encompassing kind of computational models are known as cognitive architectures, to which we turn next.

COGNITIVE ARCHITECTURES

So far we have been concerned with models that aim to explain data patterns in a particular domain of research in cognitive science, such as episodic memory or perceptual decision making. In this section, we turn our attention to models of the cognitive architecture. The cognitive architecture is conceptualized as the relatively stable system of structures and mechanisms that underlies cognition in general. An architecture model does not aim to explain a particular pattern of behavior but rather to explain how cognition works in general. As Anderson (2007) puts it—quoting one of the founding fathers of cognitive architecture models, Allen Newell—an architecture model aims to explain "how the human mind can occur in the physical universe." Hence, architectures do not make assumptions about which representations and processes generate behavior, but rather describe the cognitive system in which such representations and processes operate, and the constraints it places on these processes. Explanatory models of specific processes can be built within an architecture. To that end, architectures are implemented as programming environments for building and running simulations of cognitive processes. The primary aim of architectures is to integrate models of specific phenomena into a consistent theory of the cognitive system as a whole. As Newell (1973) commented, experimental psychology is at risk of amassing an ever-growing pile of unrelated phenomena, and the same can be said for process models: Even if we had a successful process model for each and every experimental finding to date, we would still be left with a fractionated picture of the human mind. Architectures aim to explain how all the mechanisms and processes assumed in process models act together.

Two families of architectures have been developed, production-system architectures and neural-network architectures. Production systems have emerged from the understanding of cognition as symbolic computation that has dominated cognitive science between 1950 and 1980. At their core lies the distinction between declarative representations—symbolic structures representing facts—and procedural representations—rules for manipulating symbolic structures, which are called productions. Neural-network architectures aim to model the cognitive system by modeling the brain. They consist of networks of interacting units that are more or less abstract, simplified models of neuronal networks. Each family of architectures has many members—here we will present but one example for each family.

Production Systems: ACT-R

The ACT* and ACT-R architecture has been developed by John Anderson and colleagues over several decades (e.g., Anderson, 1983;

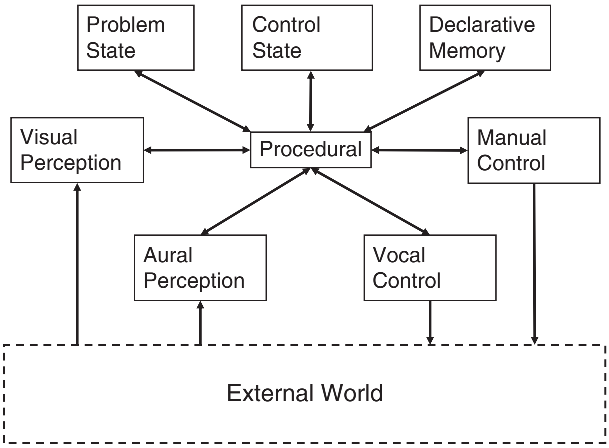

Figure 1.6 Overview of the ACT-R architecture.

Anderson & Lebiere, 1998; Anderson, 2007). In its current version, it consists of eight modules (see Figure 1.6).

Four modules—two perceptual and two motor modules—serve to interact with the environment. Three further modules—declarative, goal state, and problem state—handle declarative representations for different purposes: The declarative module serves as the system's episodic and semantic long-term memory, holding a vast number of declarative knowledge structures; the goal state represents the current goal, and the problem state represents the current state of the object of ongoing thought. Declarative representations are chunks that represent facts in the form of propositions. A chunk is a structure of slots that can be filled with elementary symbolic representations of concepts or with other chunks. For instance, arithmetic facts such as "the sum of 3 and 4 equals 7" are represented as chunks in declarative memory. The goal-state module holds chunks representing the system's goals, such as "solve equation." The problem-state module might hold a representation of the current state of the equation to be solved, such as "$3x \times 4 = 48$." Each module has a buffer through which it communicates with its environment. The buffer can hold only one chunk at a time; only that chunk is directly accessible to other modules. The problem-state module is nothing but a buffer, so its capacity is constrained to a single chunk.

The procedural module lies at the heart of the system. It holds a large number of productions, which represent rules connecting a condition to an action (where the condition and the action can consist of multiple components). For instance, a production involved in mental arithmetic could be: "IF the goal is to solve the equation, and the current state of the equation is $\mathrm{A}x \times \mathrm{B} = \mathrm{C}$, then divide both sides by B." The procedural module makes cognition happen: It compares the current contents of all other modules' buffers to the condition parts of all its productions. This comparison process is a form of pattern matching that proceeds in parallel and instantaneously for all productions. Those productions whose conditions match all buffer contents become candidates for execution ("firing"). Only one production can fire at any time. Productions compete for execution based on their utility value, which reflects the reward history of each production. Productions with partial matches also become candidates but enter the competition with a mismatch penalty. Firing of a production takes 50 ms and results in execution of the action component, which could make changes to the chunk in the problem state or the goal state, send a retrieval request to the declarative module, or an action command to one of the motor buffers.

Within the ACT-R, architecture models for individual tasks can be implemented by specifying a set of productions and a set of chunks in declarative memory. For instance, an ACT-R model for solving a class of algebraic equations consists of a set of arithmetic facts in declarative memory, together with a handful of productions for reading equations off the screen (i.e., controlling the visual module to scan the string of symbols and placing it into the problem state), retrieving the necessary facts from declarative memory, transforming equations in the problem state, and producing an answer through one of the motor modules. This model can be used to simulate equation-solving behavior of people. The simulation produces a sequence of states in the buffers, together with predictions for their distribution of durations. The durations of processing steps in a model are governed by a set of assumptions about the dynamics of the basic mechanisms of the architecture: Only one production can fire at any time, and its execution takes 50 ms. Retrieval from declarative memory takes time, and its duration—as well as its probability of success—is a function of the level of activation conferred to the target chunk. That level of activation depends on the chunk's baseline activation, which decays at a fixed rate, and activation through associations to retrieval cues available in the buffers at the time of the retrieval request.

Other assumptions in the architecture model pertain to principles of learning. ACT-R acquires new declarative knowledge by keeping a copy of every chunk in a buffer in the declarative module. ACT-R acquires new production rules by compilation of existing productions: Two productions executed in immediate succession can be unified into one production with a more specific set of conditions. The new rule is initially a very weak competitor but gains strength when it is created again and again as its components are repeatedly executed together with success. ACT-R has principles for attributing successful completion of a task to the productions contributing to it, by which the utility value of each production is updated after task completion. In this way ACT-R learns the relative utilities of its productions.

Many of the principles in ACT-R are informed by rational analysis, that is, considerations of optimal computation under certain general constraints of the cognitive system and conditions in the environment (Anderson, 1990). For instance, the baseline activation of chunks in declarative memory decays over time but is increased every time the chunk is retrieved. The resulting baseline activation mirrors closely the probability that a piece of knowledge will be needed, given its history of use (Anderson & Schooler, 1991).

Process models implemented in ACT-R make predictions for the behavioral responses in a task, for their latencies, and—as a recent addition—for the level of neural activity in brain areas that serve as indicators for the work of each module. Simulation runs of an ACT-R model yield a profile of work intensity of each module over time. Borst and Anderson (2013) generated such profiles from process models of five tasks and correlated them with the BOLD signals recorded while participants carried out these tasks. In this way they identified for each ACT-R module one cortical area that correlated highest with that module's profile of use over time. Based on these links between modules and brain areas other process models in ACT-R can be used to predict the time course of BOLD signals in each area. These predictions place further constraints on the ACT-R process models. Nijboer, Borst, van Rijn, and Taatgen (2016) demonstrated the benefits of these additional constraints: They developed a process model for a multitasking experiment that fit the behavioral data well, but found that it mispredicts the BOLD data. This observation motivated a revision of the model, upon which it also accommodated the neural data.

Neural-Network Architectures: Spaun

The seminal work of Rumelhart and McClelland (1986) has sparked renewed interest in connectionist models of cognition. In these models, behavior arises from the interaction of simple units, which can be interpreted as model neurons or neural populations. Each unit receives input (in the form of scalar signals) from many other units, carries out a nonlinear transformation on its summed input, and sends it to other units. Intelligent behavior arises from tuning the connection weights—interpretable as synaptic weights—through learning rules. The rehearsal simulations that we reported earlier were carried out within a connectionist architecture.

More recently, interest has shifted toward neural network models using more realistic neuron models. The currently most advanced effort toward building a neural-network architecture is Spaun (Eliasmith et al., 2012; Eliasmith, 2013). Spaun implements a broad range of cognitive functions in simulated spiking neurons. As such, Spaun is a model of the cognitive system and a model of the brain. It is able to carry out eight different tasks, responding to visual stimuli—among them symbols coding the desired task—and controlling a mechanical arm to produce motor responses.

Spaun builds on the Neural Engineering Framework (NEF), and is implemented in the Nengo simulation framework (Bekolay et al., 2014). The NEF (Eliasmith & Anderson, 2003) is a generic method for implementing representations and their transformations in populations of neurons. Representations are conceptualized as vectors in a vector space of arbitrary dimensionality. As a simple example, think of the orientation of an arrow on a computer screen, represented as a two-dimensional vector, $[\sin(\theta), \cos(\theta)]$. This vector can be encoded in a population of neurons, each with a different, randomly selected tuning curve. A neuron's tuning curve is a nonlinear function relating the encoded vector state into the neuron's firing rate. Tuning curves differ between neurons, such that each neuron has its own "preferred" orientation to which it responds maximally, and they differ in the sensitivity and specificity of their responses. Hence, the population code of our arrow orientation will be a pattern of firing rates across the neurons in the population. The orientation can be decoded by a linear combination of the firing rates of all neurons in the population. The optimal weights for this linear combination can be found by minimizing the difference between the encoded orientation $[\sin(\theta), \cos(\theta)]$ and the decoded orientations $[\sin(\theta), \cos(\theta)]$ over all possible values of θ. This comes down to finding the best weights for a linear regression model, and it can be solved analytically.

Transformations of representations can be implemented by the same principle. Suppose, for instance, that we want the model to mentally rotate a given arrow by 90 degrees to the right. We can implement this operation by decoding, instead of the original orientation θ, the transformed orientation $\phi = \theta + 90°$. That is, we need to find the optimal weights for decoding ϕ from the population that is coding θ, for all possible values of θ. Then we connect the population coding θ to a second population coding ϕ, with connection weights set to reflect the optimal decoding weights for the desired transformation. It helps to think of the operation as two steps: The first is to decode ϕ from the population representing θ, and the second is to encode ϕ into the second population of neurons (in the same way as θ was encoded into the first population). In reality, the two steps are wrapped into one, mediated by a single matrix of connection weights between the two populations. In principle, this method allows implementing any function on vectors in a network of spiking neurons, although some functions (in particular addition and subtraction) are much easier to implement than others.

Hence, the NEF can be used as a compiler for translating models initially developed in vector space into neuronal space. Models in vector space are very common on cognitive psychology; connectionist networks, for instance, model cognitive processes as transformations of patterns of activity over sets of units, and, mathematically, these activation patterns are vectors with one dimension for each unit. Many models of memory and categorization on a more abstract level of description, such as the Generalized Context Model (Nosofsky, 1984), SIMPLE (G. D. A. Brown et al., 2007), and Latent Semantic Analysis (Landauer & Dumais, 1997), also use representations that can be described as vectors. The states of sequential-sampling models of decision making are time-varying vectors with one dimension for each accumulator. In principle, the NEF enables researchers to implement any such model in spiking neurons. In practice, the neural implementation does impose constraints in two ways: First, neural computation is only an approximation of the mathematical functions implemented in the network, and not all functions can be approximated equally well. Second, biologically realistic neuron models have features such as time constants and limited dynamic range that have consequences for the model's speed and accuracy. For instance, a model of immediate serial recall (Eliasmith, 2013, Chapter 6.3) produced serial-position curves much at odds with the data when simulated in vector space, but reproduced the empirical serial-position curves well when simulated in spiking neurons.

The second pillar of Spaun besides the NEF is the idea of a semantic pointer architecture (SPA). A semantic pointer is a high-dimensional vector representation that fulfills the role of a symbol in production-system architectures. To that end it must meet two requirements: It must have meaning, and it must be flexibly usable in a way that endows the cognitive system with the powers of symbolic computation. Semantic pointers have meaning because they are compressed representations that point to other representations. For instance, the representations of the numbers 1 to 9 in Spaun are semantic pointers generated through several steps of compression of the visual input (i.e., images of hand-written digits). The compression can be reversed to regenerate a prototypical visual image of a written digit. A second route of decompression is the generation of a pattern of motor commands for writing the digit with the mechanical arm.

The power of symbolic computations rests on the recursive combination of symbols into structures, such as propositions. We can combine representations of "cat," "dog," and "bite" into structures representing the fact that "the dog bit the cat" or "the cat bit the dog," and we can recursively use such structures as elements in other structures, such as "Peter saw that the dog bit the cat." This requires a mechanism for ad-hoc binding of semantic pointers. In Spaun, vector representations are bound by an operation called circular convolution (Plate, 2003).

For instance, the proposition “the dog bit the cat” requires three bindings of concepts to roles through circular convolution (denoted by $\otimes$), the results of which are superimposed (i.e., added, denoted by $+$):

$$\begin{aligned} P = &\ \text{AGENT} \otimes \text{CAT} + \text{THEME} \otimes \text{DOG} \\ &+ \text{ACTION} \otimes \text{BITE} \end{aligned}$$

The elements of that structure can be extracted by de-convolution—for instance, the question “Who bit the dog?” can be answered by convolving the inverse of AGENT with P, which produces a noisy approximation of CAT.

Circular convolution returns a new vector of the same length as the two bound vectors, thereby facilitating recursive binding without increasing the demands on neural resources. In this way, complex concepts can be formed from simpler ones—for instance, the concept “cat” can be created by binding compressed representations of perceptual features of that creature with more abstract features such as “is a mammal,” and the resulting semantic pointer can in turn be bound into propositions involving cats.

Symbolic computation involves applying rules to symbol structures—such as applying productions to declarative knowledge chunks in ACT-R and other production systems. Spaun includes an action-selection mechanism that implements the functionality of productions in a spiking-neuron model of the basal-ganglia-thalamus-cortex loop (e.g., Stewart, Bekolay, & Eliasmith, 2012). This mechanism monitors semantic pointers in several buffers and selects the action with the highest utility in the context of these representations. Actions include routing representations from one buffer to another, thereby controlling which computations are carried out on them. The action-selection mechanism gives Spaun the flexibility of carrying out different tasks on the same stimuli depending on instructions for instance, given a series of images of hand-written digits, it can copy each digit immediately, do a digit-span task (i.e., write down the digits in order at the end of the list), or do matrix reasoning (interpreting each set of three digits as a row of a matrix, and finding the rules governing rows and columns to determine the ninth digit).

Relating Architectures to Data

Models need to be tested against data. To that end, we need to determine what they predict. Earlier we emphasized as one of the strengths of computational models that they facilitate the generation of unambiguous predictions. Doing so is relatively straightforward for models for a well-defined set of tasks and experimental paradigms, but less so for architecture models. The assumptions defining an architecture model do not, by themselves, entail testable predictions. Architecture models generate predictions for behavioral or brain data only in conjunction with process models that are implemented in them. Therefore, assumptions about the architecture must be tested indirectly through tests of the process models built in the architecture: When an empirical finding appears to challenge one of the assumptions about the architecture, proponents of the architecture model can defend the assumption by building a process model that accommodates the finding.

For instance, ACT-R is committed to the sequential firing of productions, which imposes a strict bottleneck for all processes that involve production firing. Whereas there is much evidence for a bottleneck for central processes (Pashler, 1994), there is also a growing number of demonstrations that people can—after a substantial amount of practice—carry out two simple tasks in parallel without dual-task costs (Hazeltine, Teague, & Ivry, 2002; Oberauer & Kliegl, 2004; Schumacher et al., 2001).

Anderson, Taatgen, and Byrne (2005) demonstrated that a process model of the task combination studied by Hazeltine et al. (2002), together with the learning principles of ACT-R, can achieve dual-task performance with vanishingly small time costs after extensive practice by compiling multiple productions into a single production per task, and scheduling demands on the procedural module—as well as the buffers of other modules that also create bottlenecks—so that temporal overlap is minimized. It remains to be seen whether the results of Oberauer and Kliegl (2004), who found that highly practiced young adults could carry out two operations in working memory simultaneously without costs, can also be accommodated by ACT-R.

The preceding example shows that, strictly speaking, it is impossible to put an architecture to an empirical test: Testable predictions always arise from the conjunction of assumptions about the architecture and about the specific processes for doing a task, and the empirical success or failure of these predictions cannot be attributed unambiguously to one or the other set of assumptions. When such a prediction fails, it is in most cases more rational to revise the process model than the architecture, because revising the architecture has more far-reaching implications: Any change to assumptions about the architecture could sabotage the empirical success of other process models built within the architecture.

Yet, in a less strict sense, architecture models are testable, if only indirectly: Confidence in an architecture model increases with the number of empirically successful process models that were developed with it, and decreases as the number of empirical challenges mounts, and as modelers find it difficult to develop process models within the architectures constraints that fit the data. Assumptions about the architecture are related to data indirectly, mediated by process models, but the weakness of each such link can be compensated by a large number of such links, because the architecture must work in conjunction with many process models. To use an analogy, the data of each experiment pull at the architecture model on a rubber leash: A single problematic finding will not make a large impression on the architecture, but many findings pulling in the same direction will make a change increasingly inevitable.

In some sense, the relation of architecture models to specific process models is analogous to the relation between higher-level and lower-level parameters in hierarchical regression models: Group-level parameters are informed by lower-level parameters (e.g., those characterizing individual persons), and in turn place constraints on them. In the same way, assumptions about the cognitive architecture are informed by the successes and failures of process models built within an architecture, and the architecture in turn places constraints on process models. Process models built outside an architecture are constrained only by the data (together with considerations of parsimony and interpretability). Process models built within an architecture are also constrained by the assumptions of the architecture model, such as the duration of processing cycles, the time and success chance for retrieving a representation, and the restrictions on what information is available for which kind of operation at which time.

THE USE OF MODELS IN COGNITIVE NEUROSCIENCE

Throughout this chapter we have reviewed several applications of computational models in cognitive neuroscience. In this section we revisit the three ways in which models can be related to data from neuroscience and point to future challenges and opportunities.

First, we can search for neural correlates of model parameters. For instance, as we noted earlier, research has identified the brain networks that correlate with the caution parameter in sequential-sampling models of perceptual decision making (Mulder et al., 2014). Model parameters can be correlated with neural signals over participants or over experimental manipulations. Second, we can search for neural correlates of cognitive states or processes predicted by a model. This use of models is exemplified by the recent work with ACT-R (Anderson, 2007; Borst & Anderson, 2013). ACT-R models predict which hypothetical modules are active at which time during a complex task, and these predictions can be correlated with neural signals over time. ACT-R models can also be used to predict at which time modules communicate with each other. Van Vugt (2014) made a first step toward testing the hypothesis that increased communication is reflected in stronger coherence between pairs of EEG electrodes in the theta frequency band. On a much more fine-grained temporal and neuronal scale, Purcell et al. (2010) related the predicted dynamics of several sequential-sampling models of perceptual decision making to the spike rate of neurons in the monkey frontal eye field (FEF). They distinguished neurons whose firing pattern reflected stimulus information and neurons whose firing pattern reflected the response (i.e., a saccade toward the stimulus). The firing rates of stimulus-related neurons were used as inputs for the models to drive the evidence accumulation, and the time course of accumulation in the models was used to predict the time course of firing rates of the response-related neurons. Purcell et al. (2010) found that none of the standard sequential-sampling models fit the neural data, and therefore proposed a new variant in which the accumulation process was delayed until sensory processing provided a sufficiently strong input to overcome a threshold.

Third, we can look for neural correlates of the representations that a model predicts to be used during a task. In recent years several techniques have been developed for decoding information about stimuli or intended actions from multivariate patterns of neural activity. These techniques use signals from multiple single-cell recordings (Georgopoulos, Schwartz, & Kettner, 1986; Stokes et al., 2013) or much more aggregated multivariate signals from fMRI, EEG, or MEG (Chan, Halgren, Marinkovic, & Cash, 2011; Haynes & Rees, 2006; Haynes, 2015). Decoding of information from these signals usually involves training a pattern classifier (e.g., an artificial neural network or a machine-learning algorithm) to classify patterns of neural activity into classes of contents that the person currently processes or holds in working memory (e.g., animate vs. inanimate nouns, Chan et al., 2011; or different orientations of motion; Emrich, Riggall, LaRocque, & Postle, 2013). To the extent that the algorithm classifies new patterns not used for training with above-chance accuracy, the neural activity patterns must carry information about which content class is being processed. There are multiple ways in which multivariate pattern analyses can be used to test model predictions about mental representations. One approach is to test model assumptions about the similarity structure of representations against the similarity matrix of neural patterns measured while people engage these representations (Kriegeskorte, 2011). Another approach is to use process models to predict which category of representation a person is using at which interval during task performance, and testing how well a pattern classification algorithm can detect the predicted category (Polyn, Kragel, Morton, McCluey, & Cohen, 2012).

All three approaches need to be mindful of the risk of circularity in linking cognitive models to neuroscience, as noted by Wixted and Mickes (2013): When a computational model is used to identify and interpret a neural correlate of some construct of the model, then that endeavor cannot at the same time provide an empirical validation of the model. An alternative model would result in the detection of other correlates of its constructs, and other interpretations of the neural data, which would necessarily be more consistent with that alternative model. That is, “the validity of the interpretation lives and dies with the validity of the cognitive theory on which it depends” (Wixted & Mickes, 2013, p. 106).

One way out of the risk of such circularity is to competitively test alternative models against the same neuroscience data, in the same way as we competitively fit models to behavioral data. One challenge on this route will be to decide, in a way that is fair to the competing models, which of the myriad of possible neuroscience variables to use: Each model is likely to identify different neural correlates of its constructs, thereby marking different neural variables as relevant.

Those challenges are beginning to be addressed. One promising development was reported by Turner et al. (2013), who proposed a Bayesian framework for joint modeling of behavioral and neural data. In their approach, a model of one’s choice is first fit to the behavioral data and another model to the neural data. For example, some behavioral data on categorization might be accommodated by the generalized context model (Nosofsky, 1984), and the hemodynamic response function in an fMRI might be estimated by Bayesian means (Friston, 2002). The parameters of both models are then combined into a single joint model whose hyperparameters are estimated by joint fitting of the neural and behavioral data. Turner et al. (2013) illustrate the utility of the approach by showing that the behavioral data of individual participants can be predicted from knowledge of the hyperparameters estimated by fitting a joint model to the remaining behavioral and neural data. Turner, van Maanen, and Forstmann (2015) extended the approach to trial-to-trial variability in tasks that are captured by the diffusion model. We expect this approach to become increasingly popular in the future.

CONCLUSION

Computational models provide an important tool for researchers in cognition and the cognitive neurosciences. We close by highlighting two aspects of computational models that we find particularly useful and exciting: First, their role as “cognitive prosthesis.” The field is currently undergoing a period of critical reflection and self-examination in light of widespread concerns about the replicability of basic phenomena (e.g., Shanks et al., 2015). Part of this critical reflection should also focus on the state of our theorizing. We suggest that purely verbal theorizing in cognition is increasingly inadequate in light of the growing richness of our data: whereas several decades ago decision-making tasks might have yielded only simple accuracy measures, we now have access not only to accuracy but also to the latencies of all response classes and their distributions. This richness defies verbal analysis but presents an ideal landscape for computational modeling. Indeed, we suggest that models also help avoid replication failures because the likelihood that an experiment will yield a quantitatively predicted intricate pattern of results involving multiple dependent variables by chance alone is surely lower than that a study might, by randomness alone, yield simple pairwise differences between conditions that happen to mesh with a verbally specified theoretical notion.

Second, we consider the increasingly tight connection between modeling and the cognitive neurosciences to be a particularly promising arena. Measurement models, explanatory models, and cognitive architectures are now either directly neurally inspired, or they provide a conceptual bridge between behavioral data and their neural underpinnings. There is little doubt that this trend will continue in the future.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford, United Kingdom: Oxford University Press.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Anderson, J. R., Taatgen, N. A., & Byrne, M. D. (2005). Learning to achieve perfect timesharing: Architectural implications of Hazeltine, Teague, and Ivry (2002). *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 749–761.

Baddeley, A. D. (1986). *Working memory*. New York, NY: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*, 83–100.

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., . . . Eliasmith, C. (2014). Nengo: A Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*. doi:10.3389/fninf.2013.00048

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences, USA*, *110*, 1628–1633. doi:10.1073/pnas.1221572110

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, *54*, 3028–3039. doi:10.1016/j.neuroimage.2010.10.073

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascade model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

Cramer, A., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R., . . . Wagenmakers, E.-J. (2015). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*, 640–647. doi:10.3758/s13423-015-0913-5

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Daily, L., Lovett, M., & Reder, L. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*, 315–353.

Ecker, U. K. H., Brown, G. D. A., & Lewandowsky, S. (2015). Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive Science*, *39*, 1570–1593. doi:10.1111/cogs.12214

Ecker, U. K. H., & Lewandowsky, S. (2012). Computational constraints in cognitive theories of forgetting. *Frontiers in Psychology*, *3*, 400. doi:10.3389/fpsyg.2012.00400

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*, 1202–1205. doi:10.1126/science.1225266

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience*, *33*, 6516–6523.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*, 223–271.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*, 329–335.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205–231.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramond, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences, USA*, *105*, 17538–17542.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology 67*, 641–666.

Friston, K. (2002). Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, *16*, 513–530. doi:10.1006/nimg.2001.1044

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*, 1416–1419.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *Journal of Neuroscience*, *35*, 2476–2484. doi:10.1523/JNEUROSCI.2410-14.2015

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*, 257–270. doi:10.1016/j.neuron.2015.05.025

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534. doi:10.1038/nrn1931

Hazeltine, E., Teague, D., & Ivry, R. B. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 527–545.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Heathcote, A., Wagenmakers, E.-J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review*, *121*, 676–678. doi:10.1037/a0037771

Henson, R. N. A. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*, *58A*, 193–233.

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory:

Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology*, *49A*, 80–115.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, *122*, 542–557.

Kieras, D. E., Meyer, D. E., Mueller, S., & Seymour, Y. (1999). Insights into working memory from the perspective of the EPIC architecture for modeling skilled perceptual-motor and cognitive human performance. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and control* (pp. 183–223). New York, NY: Cambridge University Press.

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, *56*, 411–421. doi:10.1016/j.neuroimage.2011.01.061

Laming, D. (2008). An improved algorithm for predicting free recalls. *Cognitive Psychology*, *57*, 179–219.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. London, United Kingdom: Academic Press.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, *4*, 236–243.

Lewandowsky, S., & Farrell, S. (2008). Phonological similarity in serial recall: Constraints on theories of memory. *Journal of Memory and Language*, *58*, 429–448.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the non-existent problem of decay. *Psychological Review*, *122*, 674–699.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 179–195.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Mulder, M., van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences—A model-based review. *Neuroscience*, *277*, 872–884. doi:10.1016/j.neuroscience.2014.07.031

Müller, G. E., & Pilzecker, A. (1900). Experimentelle Beiträge zur Lehre vom Gedächtnis [Experimental contributions to the science of memory]. *Zeitschrift für Psychologie*, *1*, 1–300.

Munafò, M., Noble, S., Browne, W. J., Brunner, D., Button, K., Ferreira, J., . . . Blumenstein, R. (2014). Scientific rigor and the art of motorcycle maintenance. *Nature Biotechnology*, *32*, 871–873.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York, NY: Academic Press.

Nijboer, M., Borst, J. P., van Rijn, H., & Taatgen, N. A. (2016). Contrasting single and multi-component working-memory systems in dual tasking. *Cognitive Psychology*, *86*, 1–26.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 104–114.

Nuzzo, R. (2015). Fooling ourselves. *Nature*, *526*, 182–185. doi:10.1038/526182a

Oberauer, K., & Kliegl, R. (2004). Simultaneous cognitive operations in working memory after dual-task practice. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 689–707.

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the time-based resource-sharing theory. *Psychonomic Bulletin & Review*, *18*, 10–45.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761–781.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244.

Plate, T. A. (2003). Convolution-based memory models. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (pp. 824–828). London, United Kingdom: Nature Publishing Group.

Polyn, S. M., Kragel, J. E., Morton, N. W., McCluey, J. D., & Cohen, Z. D. (2012). The neural dynamics of task context in free recall. *Neuropsychologia*, *50*, 447–457.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences, USA*, *109*, 14357–14362. doi:10.1073/pnas.1103880109

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143.

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

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*, 278–291.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of Neurophysiology*, *90*, 1392–407.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., Spieler, D., & Mckoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin & Review*, *7*, 1–25. doi:10.3758/BF03210723

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 12–157. doi:10.1016/j.cogpsych.2009.09.001

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*, 1457–1506. doi:10.1080/17470210902816461

Rouder, J. N. (1996). Premature sampling in random walks. *Journal of Mathematical Psychology*, *40*, 287–296. doi:10.1006/jmps.1996.0030

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63–77.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429. doi:10.1037/0096-3445.136.3.414

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science*, *12*, 101–108. doi:10.1111/1467-9280.00318

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., . . . Puhlmann, L. M. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General*, *144*, 142–158.

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*, 535–551.

Shiffrin, R. M., & Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments & Computers*, *29*, 6–14.

Smith, P. L., & Ratcliff, R. (2015). Diffusion and random walk processes. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 6, pp. 395–401). Oxford, United Kingdom: Elsevier.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*, 135–168.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in Neuroscience*, *6*. doi:10.3389/fnins.2012.00002

Stokes, M., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*, 364–375. doi:10.1016/j.neuron.2013.01.039

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251–260.

Sun, R., Coward, A., & Zenzen, M. J. (2005). On levels of cognitive modeling. *Philosophical Psychology*, *18*, 613–637.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1589–1625.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review*, *15*, 535–542.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. doi:10.1016/j.neuroimage.2013.01.048

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*, 312–336. doi:10.1037/a0038894

van Vugt, M. K. (2014). Cognitive architectures as a tool for investigating the role of oscillatory power and coherence in cognition. *NeuroImage*, *85*(Part 2), 685–693. doi:10.1016/j.neuroimage.2013.09.076

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*. doi:10.3758/s13428-015-0593-0

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*, 117–186.

Wixted, J. T. (2004a). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review*, *111*, 864–879.

Wixted, J. T. (2004b). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.

Wixted, J. T., & Mickes, L. (2013). On the relationship between fMRI and theories of cognition: The arrow points in both directions. *Perspectives on Psychological Science*, *8*, 104–107. doi:10.1177/1745691612469022

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*, 800–832.