Computational models as aids to better reasoning in psychology

Simon Farrell[1]

University of Bristol and University of Western Australia

and

Stephan Lewandowsky

University of Western Australia

Word count: 2500

Address correspondence to:

Simon Farrell

Department of Experimental Psychology

University of Bristol

12a Priory Road

Clifton Bristol BS8 1TU

UK

ABSTRACT

Scientists can reason about natural systems—including the mind and brain—in many ways, each being associated with its own set of limitations. It follows that the process of reasoning about theories and communicating those theories will be error-prone; we must therefore be concerned about the reproducibility of theories whose very nature is shaped by constraints on human reasoning. This problem can be alleviated by computational modelling, which maximises correspondence between the behaviour of a posited system and the behaviour inferred through reasoning; and which increases the fidelity of communication of our theories to others.

Science depends on reproducibility. However much we may debate about theories, scientists tacitly assume that we are all reasoning on the same terms and from a shared understanding about data. Data that are not reproducible, or that can only be reproduced under certain conditions, should rightly be given little weight in reasoning about natural systems. Without this shared understanding, progress in science would be impossible, as different scientists could reach different conclusions simply because the same data are analyzed and interpreted inconsistently.

Given this emphasis on the reproducibility of experimental methods and data analysis, it is striking that little—if any—consideration is given to the fidelity and reproducibility of another core aspect of science; namely, scientific reasoning. Like it or not, science communication resembles a game of "telephone" whereby theories are formulated, recorded on paper, read by the next scientist who needs to understand them, and so on. Each step in this chain involves *reasoning*, and is thus subject to known cognitive limitations. Numerous experiments have established those limitations which indubitably extend to scientific reasoning.

LIMITATIONS ON THINKING

A worryingly long list of limitations on human thinking was presented by Hintzman (1991), who argued that aspects of cognition such as the *confirmation bias* (see, e.g., Evans, 1989) and the limited capacity of working memory (e.g., Engle & Kane, 2004) have profoundly negative implications for scientific reasoning. Equally worryingly, it is easy to extend Hintzman's list. For example, when given the sequence "2-4-6", people tend to develop overly specific and baroque theories of the generating rule (see, e.g., Evans, 1989); few discover the experimenter's rule of "any increasing series". This finding arguably has parallels in science: for example,

putatively specific deficits in grammar may instead reflect more general cognitive deficits (e.g., Christiansen & Ellefson, 2002).

Similarly, scientists often draw *analogies* between a source domain (in which the inter-relations between elements are known) and a target domain (where the relationships are unknown). Unfortunately, analogies can be misapplied. Gentner and Gentner (1983) identified two common analogies to understand electricity: water flowing through pipes, and crowds of people running through passageways. Gentner and Gentner found that errors made on electrical circuit problems depended on which analogy an individual adopted. This has implications for scientific reasoning, where relying on analogies could potentially produce misunderstanding of a psychological system. As a case in point, consider the popular spreading activation theory which postulates that concepts are represented by an interconnected network of nodes. Nodes are activated upon stimulus presentation, and activation spreads through the connections to neighboring nodes. To understand and communicate the notion of spreading activation, several analogies might be used; for example, electricity passing through wires (e.g., Radvansky, 2006, p. 220) or water passing through pipes. Those analogies help determine our understanding of the model's behavior. The water analogy necessarily implies a relatively slow spread of activation, contrary to the data which imply activation of distal concepts to be almost instant (Ratcliff & McKoon, 1981). Conversely, the electricity analogy can handle the instantaneity but places the explanatory burden on the links between nodes—akin to circuits being closed—rather than on their activation. Although analogies are often drawn within a scientific domain (e.g., "this data pattern resembles one I've seen recently"; e.g., Dunbar & Fugelsang, 2005), many psychological mechanisms have no known equivalence within psychology that support analogies, and theorists will therefore be required to draw analogies to systems outside psychology where they—and other psychologists—may not be experts.

A second problem, related to these vagaries of reasoning, is that the *shared* understanding of a system by multiple individuals depends on the extent to which all involved reason identically. A group of scientists can have a shared (though perhaps flawed) understanding of a system if all scientists adopt the same analogy. If analogies are heterogeneous, mutual misunderstandings are inevitable: Two scientists may reach mutually incompatible hypotheses about spreading activation if they adopt different analogies (water vs. electricity). Furthermore, scientists may use any number of modes of thinking or representations to conceptualize a theory and to derive predictions (Markman & Gentner, 2001): Scientific reasoning may be accomplished by constructing mental models (Gentner & Gentner, 1983); using visual representations such as sketches and graphs (Oestermeier & Hesse, 2000); and mental simulation (Trickett & Trafton, 2007). Not only are these different modes imperfect, but there is also no guarantee that any two scientists will use the same mode of reasoning and thus share an understanding of a system.

In light of these concerns and others (see left column of Table 1), do we really have a deep understanding of the theories we reason about? Do we really know what our theories predict? Do our colleagues understand our theories the same way we do?

COMPUTATIONAL MODELS AS AN AID TO REASONING

Those reasoning problems can be alleviated by implementing one's theoretical principles as a computational model (or equations in a mathematical model). A principal advantage of computational modeling is that we are forced to specify *all* parts of our theory. In the case of spreading activation, we must answer such questions as: Can activation flow backwards to immediately preceding nodes? Is the amount of activation unlimited? Is there any leakage of activation from nodes? These further specifications, which verbal theories omit altogether, render

Table 1. Limitations on reasoning that affect our understanding of our theories and their communication to other scientists (left column); and ways in which computational modeling can address these failures (right column). For examples of other constraints on reasoning and the solutions offered by modeling, see Hintzman (1991) and Lewandowsky (1993).

Problem	**Solution from computational modelling**
REASONING	
Confirmatory bias: tendency to seek out evidence that confirms (rather than disconfirms) a hypothesis	Emphasis on comparing multiple models; model selection allows us to find evidence for and against models
Analogical reasoning may rely on different analogues; contents of source domain leak in to inference about target domain	Formal system means theory will behave the same way regardless of back-pocket conceptual model
Reasoning about complicated, distributed or massively parallel structures is restricted by working memory limitations, and bias to interpret such networks unidirectionally (e.g., White, 2008)	Models not limited by WM limitations; many models are highly distributed or complicated in nature (e.g., connectionist models)
Incompleteness of reasoning: there can be more in the data than might be inferred from standard analyses of performance (e.g., proportion correct)	Emergent phenomena: fitting a model can reveal hidden structure or processes that are not directly inferable from standard analyses of performance
SHARED UNDERSTANDING	
Precise communication	Computer code (commented to make links with textual description) can be shared between researchers in a similar fashion to data and analysis files
"That's not what I meant" problem: Shared understanding of theories, and falsifiable theories, difficult to achieve when specification of theory is fuzzy	With clear, computationally formulated definition, theories make unambiguous predictions (at least within target domain) and are more falsifibiable.

our theory more readily communicable and more falsifiable. A summary of such advantages of computational modeling is provided in the right-hand column of Table 1.

Computational models thus check whether our intuitions about the behavior of a theorized system match what actually arises from its realization. To illustrate, consider the "random walk"

model of binary decisions, which assumes that people sample evidence from their environment in discrete steps, and then *sum* this sampled evidence to make a decision. At each step, a sample can nudge the summed evidence towards one decision or another until a threshold is reached. Depending on which of two thresholds the evidence trail crosses, one decision (e.g., “The light is green”), or the other decision (“The light is red”) is reached. Figure 1 shows some illustrative sampling paths. The model predicts both the probability and the time taken to make a decision. The left panel shows a case in which the information is equally favorable to the two alternatives. As one would expect, the probabilities of the two decisions are equal and they are predicted to have identical response time characteristics (because this setup is symmetric around the starting point). Now imagine how the model will behave when the evidence favors one decision over the other, as expected when an informative stimulus is present. This introduces some “drift” towards the favoured threshold by “biasing” the sampled information, as depicted in the right panel. What do you expect will happen to the choice probabilities? What about the time taken to make each choice?As you probably predicted, under these circumstances the drift will increase the probability of the evidence crossing the upper boundary (as is apparent in the figure). You probably also predicted that the decision time would be slower for the less likely response. The latter prediction is actually incorrect; in fact, the mean response times for the two alternatives are identical, and as shown in Figure 2, the histogram of the latencies remains absolutely unchanged (Stone, 1960)! This seems a little strange—after all, doesn’t the upward drift mean that it will take longer for a random walk to reach the bottom boundary, like a person struggling against a river current? Or maybe you (like us) pictured “rays” emanating from the starting point representing some reasonable range of average trends, and imagined this rotating around in the case where some drift is introduced to produce slow “bottom” responses (see Figure 3). The swimmer analogy misses out the important detail that the only systematic pressure is the drift

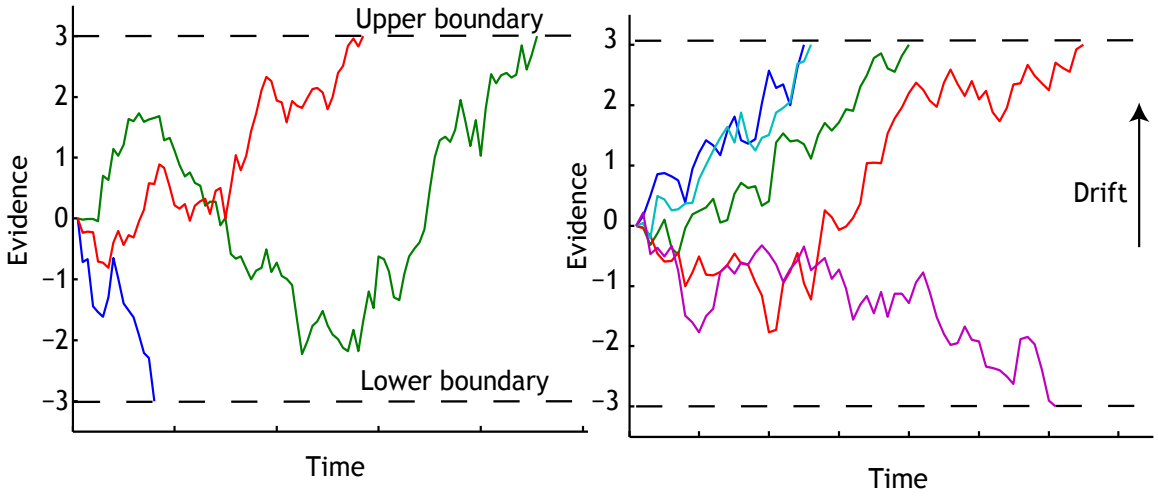

Figure 1. The random walk model of decision making. At each time step, evidence is sampled from an information source, and the evidence is accumulated over time. When the information exceeds one of the two decision boundaries, the associated decision is made. *Left panel*: The summed evidence for three different trials is shown. The three paths all rely on the same source of information, and differ only in the random sampling of the information over time. The red and green paths have reached the upper boundary and would indicate one decision (e.g., "The light is green"), while the blue path has reached the bottom boundary corresponding to the other decision (e.g., "The light is red"). *Right panel*: Five sample paths are shown for the case where there is a tendency to drift towards the upper boundary because the information favors that decision on average.

(unlike the swimmer, who by definition is applying their own "counter-drift" against the current). This means that paths that hit the bottom boundary do so only by the chance happenstance of collecting a series of samples that work against the drift. This also explains why the "ray" analogy in Figure 3 also fails—having a slower rate of approach to the bottom boundary slows those responses down in the analogy, but in reality, any additional time gives those paths more time to be bumped towards the top boundary. We can also reject this analogy by a logical consideration, as follows. In the ray model, if we have reached a certain level of evidence (suppose -2), our remaining time to hit the bottom boundary will be quicker the less time it took us to get to this point, as can be seen by considering the horizontal arrow drawn through the two bottom paths at -2, and noting that the lines diverge. However, the random walk model is by definition agnostic to the time that has already passed; if we have reached an evidence level of -2

we have a constant probability of taking a downwards (vs. an upwards) step. The behavior of the basic random walk model is not at all obvious from its description, and shows up the limitations on our reasoning about such processes.[2]

By revealing the "real" behavior of a system, modeling can generate insights that conventional reasoning processes may fail to uncover. For example, in the random walk model, we can differentiate the drift rate (the quality of information) from the separation between the boundaries (the amount of evidence needed to make a decision). Schmiedek, Oberauer, Wilhelm, Süß, and Wittmann (2007) did exactly that by fitting a variant of the random walk model to data

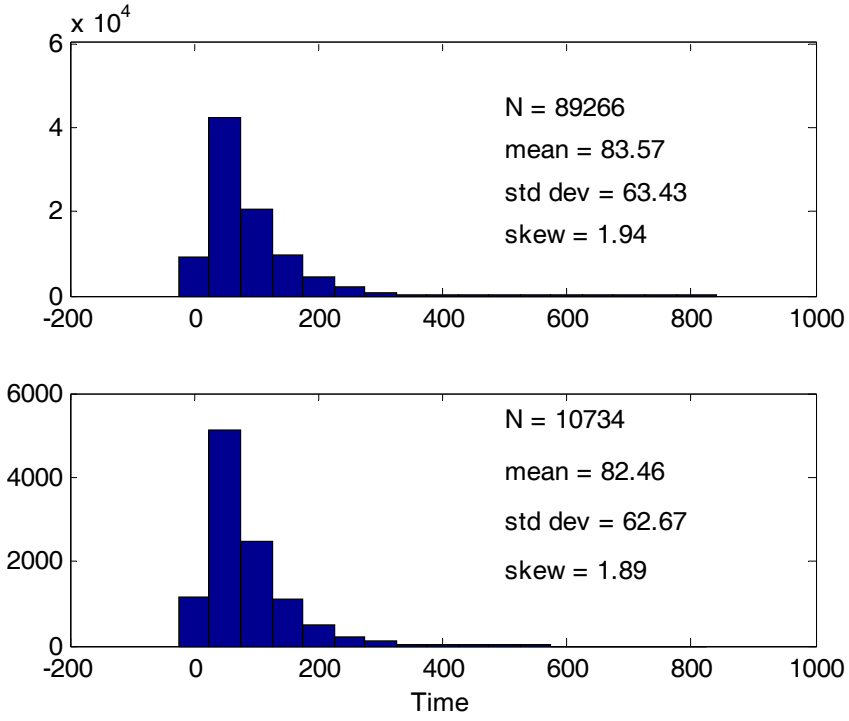

Figure 2. Histograms of decision times for decisions corresponding to the upper boundary (top panel) and lower boundary (bottom panel) from 100 000 simulation runs of the random walk model. Although upper boundary decision is more likely to be made (around 89% of the time, indicated by N), the two latency histograms have an identical appearance, with the same mean, standard deviation, and skew (the slight differences are due to the randomness in the simulations).

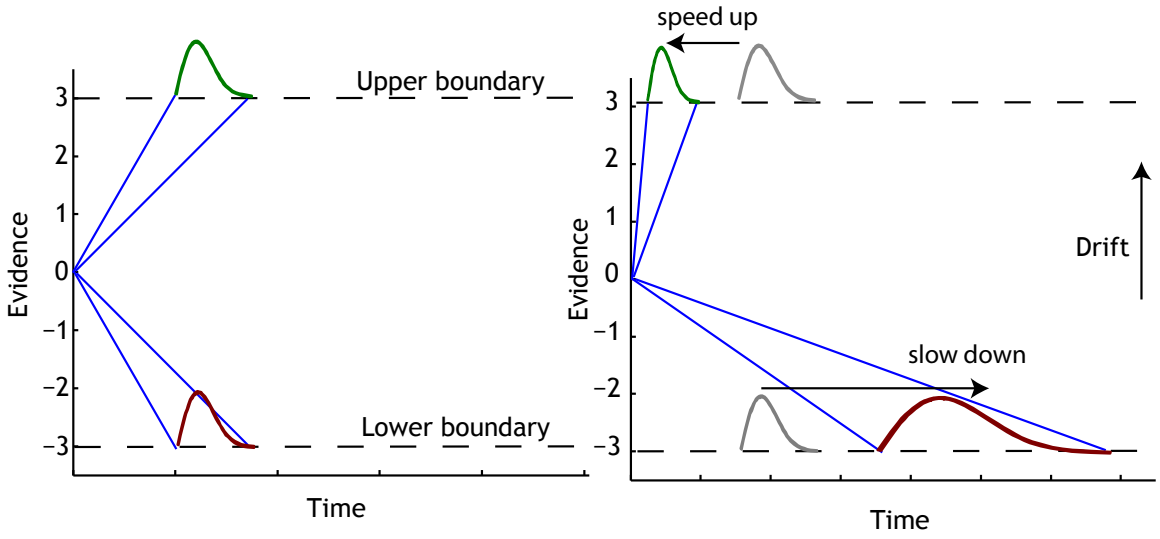

Figure 3. A schematic depiction of an intuitively reasonable but *incorrect* mental simulation of the effects of drift in the random walk model. The left panel depicts some representative rays emanating from the starting point in the situation where drift = 0. When drift towards the positive (top) boundary is introduced, the rays might incorrectly be assumed to rotate in that direction, producing a speed-up of top responses, and a slow-down of bottom responses.

from several choice reaction time tasks, and found that measures of general intelligence were more strongly related to drift rate than to boundary separation, suggesting that information extraction is a fundamental aspect of intelligence. Moreover, we frequently observe behavior emerging from a model in which no such behavior is specified up front. Elman (1990) found that when he trained a particular type of model, called a recurrent network, on the sequential statistics in a large corpus of natural language, the model formed representations of nouns, verbs, and adjectives, even though these grammatical types were not specified in the input or in the model itself.

Finally, computational modeling helps ensure reproducibility in scientific thinking. By implementing a model as a computer program or a set of equations, another researcher can take our model and exactly reproduce our predictions. For example, you may still doubt whether the random walk model really produces the same latency distributions for the different decisions in

the presence of drift. With a little experience in a computer language such as R or MATLAB you would be able to simulate the model as described and confirm that prediction for yourself. As an additional step, publishing the model code on the web facilitates testing and exploration by others. Use of the same formalized description ensures that we ultimately derive the same predictions from a theory, and assists in forming a shared conceptual understanding.

To illustrate the last point, consider the concept of inhibition that is often invoked to explain differences between individuals in their ability to perform a task (see MacLeod, Dodd, Sheard, Wilson, and Bibi, 2003, for a review). For example, the phenomenon of negative priming—whereby a response to a stimulus is slowed when that stimulus recently appeared as distracting information—is taken as evidence for an inhibitory component of selective attention (Tipper, 1985). However, MacLeod and colleagues suggested that “inhibition” is a vague term that often amounts to little more than the renaming of an observed difference between two mean latencies, giving fertile ground for confusion or lack of shared understanding between researchers. This problem is overcome within a computational model, such as the choice model of Brown and Heathcote (2005), in which representations of response alternatives compete for selection. The model provides several mechanisms by which inhibition can be implemented, including: a) reduced external input to one or more alternatives; b) increased “leakage” from alternatives; c) a reduction in the resting baseline of an alternative; and d) a tendency for more active units to reduce the activation of other units, called “lateral inhibition”. By specifying the operation of the model (see, e.g., Equation 1 of Brown & Heathcote, 2005), the source of inhibition is made explicit, leaving other researchers (and ourselves!) with a better grasp of the meaning of “inhibition”.

CURRENT DIRECTIONS

Contemporary theorizing increasingly involves quantitative comparison of competing models to weigh the evidence for and against various theories in light of a particular data set. Whenever a model provides a better quantitative explanation of the data, it receives further support. These comparisons can additionally be corrected for the complexity of a model. One widely accepted dictate in science is "Occam's Razor:" we should prefer the simplest theory that adequately explains the data. Determining the proper level of complication of a theory is sometimes difficult, and techniques for the quantification of complexity continue to be developed (Pitt & Myung, 2002). Another advance is the incorporation of psychological models into common statistical frameworks. For example, individual-differences research often employs structural-equation modeling, in which relationships between observed variables are captured via "latent" variables that represent psychologically meaningful constructs (e.g., working memory capacity). Recent research has used psychologically meaningful model parameters (e.g., the strength of the drift of a random walk) as variables that can be used in structural equation models (Schmiedek et al., 2007). Furthermore, computational modeling can inform research from cognitive neuroscience, where model parameters can be linked to brain activity (as recorded by EEG or fMRI), and can be used to make inferences about changes in cognitive mechanisms that would not otherwise be possible (e.g., Ho, Brown, & Serences, 2009).

In closing, we should note that modeling is not a panacea for all scientific ills. In some situations the extent to which an algorithm or equation is mandated by (and itself mandates) a psychological interpretation may be questionable. For example, much work has been dedicated to comparing different mathematical functions in their ability to account for the extent of forgetting from memory over time (e.g., Rubin & Wenzel, 1996); however, there is an arguable limit on how much these functions can tell us in the absence of a model of the underlying mechanisms.

Bearing in mind this caveat, computational modeling should have a place in any psychologist's toolkit—alongside experimental design and statistics—as a way of developing, understanding, and communicating of theories.

ACKNOWLEDGEMENTS AND END NOTE

1 Address correspondence to Simon Farrell, Department of Experimental Psychology, University of Bristol, 12a Priory Road, Clifton Bristol BS8 1TU UK; email: Simon.Farrell@bristol.ac.uk. Preparation of this paper was facilitated by a Discovery Grant from the Australian Research Council and an Australian Professorial Fellowship to the second author. We thank the anonymous reviewers for their helpful comments.

2 In fact, other types of evidence-summing models do predict different latencies for the two response classes, but this is by effectively making the random walk model more like the ray analogy.

REFERENCES

Botvinick, M. M. & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113*, 201-233.

Brown, S. & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112*, 117-128.

Christiansen, M. H. & Ellefson, M. R. (2002). Linguistic adaptation without linguistic constraints: The role of sequential learning in language evolution. In A. Wray (Ed.), *Transitions to language* (pp. 335–358). Oxford: Oxford University Press.

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. Morrison (Eds.), *Cambridge Handbook of Thinking & Reasoning* (pp. 705-726). Cambridge: Cambridge University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.). *The psychology of learning and motivation* (Volume 44, pp. 145-199). NY: Elsevier.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, UK: Lawrence Erlbaum Associates.

Gentner, D. & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99–129). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39-56). Hillsdale, NJ: Lawrence Erlbaum.

Ho, T. C., Brown, S. & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience, 29*, 8675-8687.

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

MacLeod, C. M., Dodd, M. D., Sheard, E. D., Wilson, D. E. & Bibi, U. (2003). In opposition to inhibition. In B. H. Ross (ed.), *The psychology of learning and motivation* (Vol. 43), pp. 163-214. San Diego, CA: Elsevier.

Markman, A. B. & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, *52*, 223-247.

Oestermeier, U., & Hesse, F. W. (2000). Verbal and visual causal arguments. *Cognition, 75*, 65-104.

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science, 6*, 421-425.

Radvansky, G. (2006). *Human memory*. Boston, MA: Pearson.

Ratcliff, R., & McKoon, G. (1981). Does activation really spread? *Psychological Review, 88*, 454-462.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734-760.

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.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.

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

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology, 37*, 571 – 590.

Trickett, S. B. & Trafton, J. G. (2007). “What if...”: The use of conceptual simulations in scientific reasoning. *Cognitive Science*, 31, 843–875.

White, P. A. (2008). Beliefs about interactions between factors in the natural environment: A causal network study. *Applied Cognitive Psychology, 22*, 559-572.

RECOMMENDED READINGS

Carruthers, P., Stich, S., & Siegal, M. (2002). *The cognitive basis of science*. Cambridge: Cambridge University Press. An edited volume that provides a good overview of cognitive approach to scientific reasoning, and the social and epistemic context in which scientific reasoning takes place.

Hintzman, D. L. (1991). See reference list. A classic chapter arguing for the widespread formal use of computational models in psychology.

Lewandowsky, S., & Farrell, S. (in press). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage. A new textbook that explains the logic behind computational modeling, and works through the steps of developing and testing models in cognitive psychology.