How Cognitive Modeling Can Benefit from Hierarchical Bayesian Models

Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

Abstract

Hierarchical Bayesian methods provide a flexible and interpretable way of extending simple models of cognitive processes. To introduce this special issue, we discuss four of the most important potential hierarchical Bayesian contributions. The first involves the development of more complete theories, including accounting for variation coming from sources like individual differences in cognition. The second involves the capability to account for observed behavior in terms of the combination of multiple different cognitive processes. The third involves using a few key psychological variables to explain behavior on a wide range of cognitive tasks. The fourth involves the conceptual unification and integration of disparate cognitive models. For all of these potential contributions, we outline an appropriate general hierarchical Bayesian modeling structure. We also highlight current models that already use the hierarchical Bayesian approach, as well as identifying research areas that could benefit from its adoption.

Introduction

Bayesian statistics provides a compelling and influential framework for representing and processing information. Over the last few decades, it has become the major approach in the field of statistics, and has come to be accepted in many or most of the physical, biological and human sciences. This paper, and this special issue, are about what one particular niche within Bayesian statistics, in the form of hierarchical models, can contribute to cognitive modeling.

The Nature of Bayesian Statistics

It would be wrong to claim that there is complete agreement on exactly how Bayesian analyses should be conducted and interpreted. Like any powerful and fundamental idea, it can be conceived and formulated in a variety of ways. At the ba-

sic theoretical level, the 'objective Bayesian' approach expounded by Jaynes (2003) encourages a different style of thinking about Bayesian analysis than the 'subjective Bayesian' approach of de Finetti (1974). At the practical level of conducting Bayesian analyses, there is also a spectrum, ranging from work that closely follows the objective viewpoint (e.g. Gregory, 2005; Sivia, 1996), to work that is more agnostic or adopts a naturally subjective position (e.g., Congdon, 2006; Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Hill, 2007). There are many additional subtleties and perspectives in the excellent accounts provided by Bernado and Smith (2000), Lindley (1972), MacKay (2003) and others.

But Bayesian statistics is in agreement on the very basic issues. Knowledge and uncertainty about variables is represented by probability distributions, and this knowledge can be processed, updated, summarized, and otherwise manipulated using the laws of probability theory. These commitments distinguish Bayesian statistics from other competing frameworks, especially those based on frequentist views of probability, and sampling distribution approaches to handling uncertainty. What Bayesian statistics offers is a remarkably complete, coherent and intuitive method for understanding what is known, based on the assumptions being made, and the information that is available.

Three Uses for Bayesian Statistics in the Cognitive Sciences

Because Bayesian statistics provides a formal framework for making inferences, there are different ways it can be applied in cognitive modeling. One way is to use Bayesian methods as a *statistician* would, as a method for conducting standard analyses of data. Traditionally, the framework for statistical inference based on sampling distributions and null hypothesis significance testing has been used. Calls for change, noting the clear superiority of Bayesian methods, date back at least to the seminal paper of Edwards, Lindman, and Savage (1963), and have grown more frequent and assertive in the past few years (e.g., Gallistel, 2009; Kruschke, 2010a; Lee & Wagenmakers, 2005; Wagenmakers, 2007). It seems certain Bayesian statistics will play a progressively more central role in the way cognitive science analyses its data.

A second possibility is to apply Bayesian methods to cognitive modeling as a *theoretician* would, as a working assumption about how the mind makes inferences. This has been an influential theoretical position for the last decade or so in the cognitive sciences (e.g., Chater, Tenenbaum, & Yuille, 2006; Griffiths, Kemp, & Tenenbaum, 2008). Most existing work has focused on providing 'rational' accounts of psychological phenomena, pitched at the computational level within the three-level hierarchy described by Marr (1982). These models generally use Bayesian inference as an account of why people behave as they do, without trying to account for the mechanisms, processes or algorithms that produce the behavior, nor how those processes are implemented in neural hardware. More recently, however, there have also been attempts to apply computational sampling approaches from Bayesian statistics as a theoretical metaphors at the algorithmic and implementation levels. In this work,

models are developed in which people mentally sample information (e.g., Sanborn, Griffiths, & Shiffrin, 2010). These uses of Bayesian statistics as theoretical analogies have led to impressive new models, and raised and addressed a range of important theoretical questions. As with all theoretical metaphors—including previous ones like information processing and connectionist metaphors—"Bayes in the head" constitutes a powerful theoretical perspective, but leaves room for other complementary approaches.

A third way to use Bayesian statistics in cognitive science is to use them to *relate models of psychological processes to data* (e.g., Lee, 2008; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010). This is different from the data analysis approach, because the focus is not generic statistical models like the generalized linear model. Instead the goal is relate a detailed model of some aspect of cognition to behavioral or other observed data. One way to think of the distinction is that data analysis typically does inference on the measured dependent variables from an experimental design—measures of recall, learning, response times, and so on—whereas modeling applications typically do inference on latent psychological parameters—memory capacities, learning rates, decision criteria, and so on—that control the behavioral predictions of the model. It is also different from the use of Bayesian inference as a metaphor for the mind (Kruschke, 2010b). There is no requirement that the cognitive models being related to data make Bayesian assumptions. Instead, they are free to make any sort of processing claims about how cognition works. The goal is simply to use Bayesian statistical methods to evaluate the proposed model against available data.

This third approach is the focus of the current special issue. We think it is an especially interesting, important, and promising approach, precisely because it deals with fully developed models of cognition, without constraints on the theoretical assumptions used to develop the models. The idea is to begin with existing theoretically grounded and empirically successful models of cognition, and embed them within a hierarchical Bayesian framework. This embedding opens a vista of potential extensions and improvements to current modeling, because it provides a capability to model the rich structure of cognition in complicated settings.

In the remainder of this paper, we identify four major new capabilities offered by the hierarchical Bayesian extensions of cognitive models. We discuss each capability, focusing on how it can help theory and model development, and identifying places where they have already been applied, or could and should be applied soon.

Benefits of Using Hierarchical Bayes in Cognitive Modeling

Before discussing the potential contribution hierarchical Bayesian methods can make to cognitive modeling, we need to say what we mean by 'hierarchical Bayes'. We do that in the next section—by characterizing its complement, in the form of the currently dominant non-hierarchical modeling approach—and then discuss the

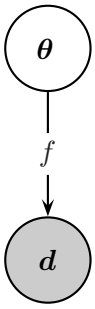

Figure 1. A general structure for non-hierarchical models of cognition.

advantages of the hierarchical approach.

Non-hierarchical Modeling

While there is consistent conceptual overlap, there does not seem to be a single formal agreed-upon definition of what makes a model "hierarchical." Some authors give a fairly formal definition in terms of the fundamental concept of exchangeability (e.g. Bernado & Smith, 2000; Schervish, 1995), while others emphasize particularly common and useful hierarchical structures like random-effects and latent mixture models, giving a sort of definition-by-example (e.g., Congdon, 2006; Koop, Poirer, & Tobias, 2007). But, there are at least some literatures that seem to regard some random-effects models as non-hierarchical (e.g., Rashbash & Browne, 2008), contradicting what other literatures advocate. More generally, it is probably possible to get caught up in an unhelpful semantic argument about whether some models are hierarchical, depending on how they are parameterized and interpreted.

To cut through these difficulties, we construe hierarchical models broadly, and with reference to their meaning as models of cognition. In particular, we treat as hierarchical any model that is more complicated than the simplest possible type of model shown[1] in Figure 1. In this model, a set of parameters $\boldsymbol{\theta}$ generate a set of data $\boldsymbol{d}$ through a likelihood function $f(\cdot)$. While this simple non-hierarchical structure seems very limiting, it could be argued to encompass the vast majority of successful and widely-used models in the current study of cognition.

As one concrete example, Figure 1 naturally accommodates Signal Detection Theory (SDT: Green & Swets, 1966; Macmillan & Creelman, 1991), which is widely used in the modeling of memory, decision-making and reasoning (e.g., Snodgrass &

[1]Throughout, we use a graphical model formalism to characterize different hierarchical modeling structures. This is a popular formalism in machine learning and, increasingly, in the cognitive sciences (e.g., Jordan, 2004; Koller, Friedman, Getoor, & Taskar, 2007; Shiffrin, Lee, Kim, & Wagenmakers, 2008). It uses a directed graph to show the relationships between unobserved (unshaded) parameters and observed (shaded) data.

Corwin, 1988; Heit & Rotello, 2005). The SDT model provides a mapping from two parameters, measuring discriminability and a response criterion or bias, to observed data in the form of hit and false-alarm counts. Formally, for model parameters giving a measure of discriminability d and a measure of bias c, Figure 1 has $\boldsymbol{\theta} = (d, c)$. For observed data counts of h hits and f false alarms out of s and n noise trials, Figure 1 has $\boldsymbol{d} = (h, f, s, n)$. Then for the likelihood function that formalizes SDT, we can write $\boldsymbol{d} \sim f(\boldsymbol{\theta})$.

As another example, a similarly simple mapping from parameters to data characterizes the Generalized Context Model (e.g., Nosofsky, 1986) of category learning, which uses psychological parameters like strength of generalization, focus of attention and response bias to model the choices made in a category learning task. As a third example, multidimensional scaling models (e.g. Borg & Lingoes, 1987) provide a mapping between latent coordinate locations representing stimuli and their observed judged pairwise similarities. As a final example, the Ratcliff diffusion model (e.g. Ratcliff & McKoon, 2008) provides a mapping from a range of parameters controlling bias, caution, evidence and a baroque menagerie of other psychological variables to the joint distribution of accuracy and response times for simple decisions. This list of cognitive models consistent with non-hierarchical mappings from parameters to data could be made quite large, and would capture many of the important contemporary models of cognition.

All of these models provide worthwhile starting points for hierarchical Bayesian development. By introducing additional structure to the simple parameter-to-data relationship shown in Figure 1, it is possible to extend existing successful models of memory, learning, decision-making and other basic cognitive phenomena. We now discuss a range of generic possible extensions, trying to highlight how they might contribute to an improved account of cognition.

Developing Deeper Theories

The most obvious hierarchical structure—and the one that intuitively warrants the label 'hierarchical'—is shown in Figure 2. In models with this form, the basic model parameters $\boldsymbol{\theta}$ are themselves generated by some other process $g(\cdot)$, parameterized by $\boldsymbol{\psi}$, which are sometimes called hyper-parameters. The impact of this extension is that it is no longer satisfactory or complete to describe how data are generated in terms of the basic parameters. In the hierarchical version, it is also theoretically important to say how these basic parameters are generated. That is, instead of just needing a theory of task performance, given by the mapping $\boldsymbol{d} \sim f(\boldsymbol{\theta})$, a theory is also needed about the parameters that control task performance, given by the mapping $\boldsymbol{\theta} \sim g(\boldsymbol{\psi})$. In this way, the hierarchical extension of basic non-hierarchical cognitive models has the potential to drive theorizing about the parameters—representing key psychological variables—to deeper and more fundamental levels of abstraction.

Perhaps the best example of the need for this sort of structure is the need to accommodate individual differences. These are ubiquitous throughout cognition,

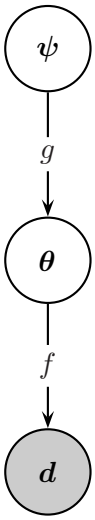

Figure 2. A general structure for the hierarchical dependence of basic data-generating process f parameterized by $\boldsymbol{\theta}$ upon a more abstract process g parameterized by $\boldsymbol{\psi}$.

but poorly handled by the non-hierarchical approach shown in Figure 1. The non-hierarchical approach has to rely on first doing separate inference for parameters and data for each person, and then trying to say something about individual differences through post-hoc analyses. In the hierarchical approach in Figure 2, the structure in individual differences is directly captured by the process g and its parameters $\boldsymbol{\psi}$.

Shiffrin et al. (2008) provide a worked tutorial example for the case of memory retention. Here the data $\boldsymbol{d}$ are counts of how often memory items are recalled at different time periods, the function f is a memory retention function like the exponential or power function (e.g., Rubin & Wenzel, 1996), and the parameters $\boldsymbol{\theta}$ are the starting points, decay rates, and other standard properties of those retention functions. In the hierarchical extension, g might be a Normal distribution, parameterized by a mean and variance in $\boldsymbol{\psi}$ that then describes the distribution over starting points and decay rates across individuals. This sort of model constitutes a deeper level of psychological theorizing, because it not only allows for individual differences, but imposes a model structure on those differences, and allows inference about parameters—like the group mean and variance—that characterize the individual differences.

Almost all of the articles in this special issue use this approach, and its generality is clear just from these applications to models of memory (Averell & Heathcote, this volume; Morey, this volume; Pooley, Lee, & Shankle, this volume; Pratte & Rouder, this volume), decision-making (Nilsson, Rieskamp, & Wagenmakers, this volume; Ravenzwaaij, Dutilh, & Wagenmakers, this volume), confidence (Merkle,

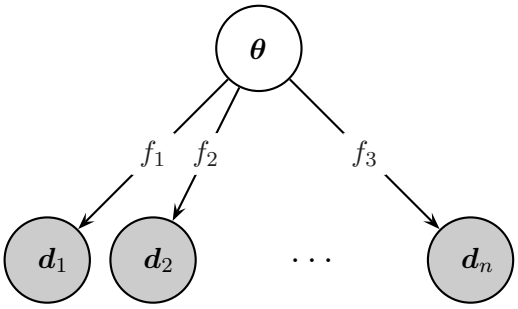

Figure 3. A hierarchical modeling approach allowing the same underlying psychological variables to generate behavior in multiple related behavioral tasks.

Smithson, & Verkuilen, this volume), and emotional states (Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, this volume). Some these papers use models that go beyond simple independent characterizations of differences in individual parameters, and begin to model the co-varying structured relationships between parameters, or their relationship to other relevant psychological variables.

While individual differences provide an intuitive example, there are many other applications of the simple hierarchical structure in Figure 2. Kemp, Perfors, and Tenenbaum (2007) use this approach in modeling of basic inductive processes in cognitive development that require learning what they term 'overhypotheses'. Essentially, their overhypotheses are the mappings g that constrain the variety seen in the basic model parameters $\boldsymbol{\theta}$. Rather than the deeper level of abstraction accommodating individual differences, it now formalizes the constrained relationship between the different problems encountered by people. As Kemp et al. (2007) argue, the ability to acquire this mapping constraint is extremely powerful developmentally, because it means what is learned in one specific situation can help improve and hasten learning in related situations. Statisticians sometimes call this basic property of hierarchical models "sharing statistical strength," and it is one of the most powerful motivations for moving beyond simple mappings of parameters to data (e.g., Gelman et al., 2004).

Linking Psychological Variables to Multiple Phenomena

A different sort of hierarchical model is shown in Figure 3. In this model, there is only one level of abstraction, but the parameters $\boldsymbol{\theta}$ are responsible for generating many sorts of data $\boldsymbol{d}_1, \ldots \boldsymbol{d}_n$ through a range of different models with likelihood functions $f_1, \ldots, f_n$. In effect, this hierarchical structure allows the same psychological variables to influence behavior on multiple tasks, through multiple cognitive processes. This sort of unification should be a basic goal for the cognitive sciences, as it is for other empirical sciences. Being able to explain a range of observed phenomena in terms of a few key variables is the hallmark of good theorizing and modeling.

It is surprisingly hard, however, to find compelling examples of this approach in modeling cognition. In the study of human memory, it has long been a goal to develop a single model of multiple tasks—recognition, free recall, serial recall, and so on—by assuming different processes operate on the same basic memory system (e.g., Gillund & Shiffrin, 1984; Norman, Detre, & Polyn, 2008). In terms of Figure 3, these unifying models would have common memory parameters represented by $\boldsymbol{\theta}$, and have f_1 formalizing the recognition process, f_2 the free recall process, and so on, with $\boldsymbol{d}_1$ being recognition data, $\boldsymbol{d}_2$ free recall data, and so on. In this special issue, Pooley et al. (this volume) make some steps towards using hierarchical methods to model recall and recognition data simultaneously.

Other areas in cognitive modeling striving for similar unification of related tasks through common psychological variables are harder to identify. Indeed, sometimes the accepted practice runs counter to the aim of unification. A good example is provided by the similarity-scaling and category learning literatures (e.g., Kruschke, 1992; Lee & Navarro, 2002; Nosofsky, 1992). In this work, similarity-scaling methods like multidimensional scaling are used to derive representations of stimuli from similarity data. Once these representations are inferred, they become part of category learning models that then attempt to account for the choices people make classifying the stimuli into categories. Conceptually, then, the causal process of the scientific inference is from similarity data, to representational parameters, and then to category learning data.

This is not at all what is depicted in Figure 3, which shows the same parameters generating multiple data. The hierarchical structure in Figure 3 would argue that the same underlying mental representation of the stimuli contributes to the generation of both the observed similarity data and the observed category learning behavior. Theoretically, this causal account seems more intuitively satisfying, and the nature of the modeling it suggests is more complete, coherent, and parsimonious. Instead of two separate inference stages—for similarity-scaling parameters and then category learning parameters—a hierarchical model would do the following: It would have the similarity data as $\boldsymbol{d}_1$, the category learning data as $\boldsymbol{d}_2$, all of the representation and category learning parameters (e.g., the coordinate location parameters from multidimensional scaling and the attention, generalization, bias and other parameters from the Generalized Context Model) in $\boldsymbol{\theta}$. Then the model f_1 would implement multidimensional scaling (i.e., not depend on category learning parameters), and f_2 would implement the GCM, using all of the parameters. Both f_1 and f_2 were given Bayesian implementations by Lee (2008), but were never combined in this hierarchical way. Recently, however, Zeigenfuse and Lee (2010a) have developed hierarchical Bayesian models of feature representations and similarity data that adopt the approach in Figure 3, and showed that is very effective.

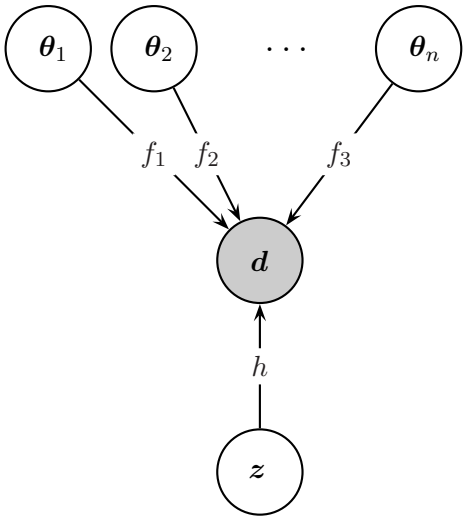

Figure 4. A hierarchical modeling approach allowing a set of different psychological processes to combine to produce observed data.

Linking Psychological Phenomena to Multiple Processes

Figure 4 shows a hierarchical model that allows for multiple cognitive processes to contribute to a single set of observed data. The different processes are represented by $f_1, \ldots, f_n$ and have different associated parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$. How these processes combine to produce the observed data $\boldsymbol{d}$ is determined by a mixing process h parameterized by $\boldsymbol{z}$. One possibility for combination is that h chooses exactly one of the processes f, according to probabilities given by $\boldsymbol{z}$. Another possibility is that h mixes together all of the processes f according to proportions given by $\boldsymbol{z}$. Of course, there are many other possibilities.

The key point, in terms of modeling cognition, is that Figure 4 does not demand a monolithic account of all of the variation seen in observed behavior in terms of a single cognitive process, or a single set of controlling psychological variables. Instead, observations are seen naturally as a mixture of potentially different processes. This sort of assumption is needed in many domains, and is perhaps best developed in the study of accuracy and response time distributions for simple decision-making. Ratcliff and Tuerlinckx (2002) pioneered a mixture approach in which a monolithic account of responding, based on the Ratcliff Diffusion model, was supplemented with the possibility of contaminant trials. These contaminant trials assumed very different distributions for accuracy and response times, to help explain the variation seen in real data.

More recently, this approach has been extended by Vandekerckhove, Tuerlinckx, and Lee (2008), using hierarchical Bayesian methods, to model a decision-making ex-

periment as a mixture of standard trials, delayed start trials, and fast guesses. The first of these types remains modeled by the Ratcliff Diffusion model, but the remaining two have detailed process accounts of their own, characterizing the accuracy and response time properties expected for various aberrant ways participants might approach some trials. As these alternative models become progressively more sophisticated, they no longer deserve the label 'contaminant', but become part of a collection of cognitive processes, potentially controlled by different psychological variables, all of which are needed to explain the observed data. In these sorts of applications, the process f_1 formalizes the Ratcliff Diffusion model, f_2 the delayed start model, and f_3 the fast guess model, with the parameters $\boldsymbol{\theta}_1$ belonging to the diffusion model, $\boldsymbol{\theta}_2$ the delayed start model, and $\boldsymbol{\theta}_3$ the fast guess model. The mixing process h identifies the accuracy and response time data in d as belonging to only one of these processes at the level of a trial, as given by the index in $\boldsymbol{z}$ for each trial.

As a modeling strategy, the use of mixtures, and the assumption of qualitatively different components, makes a lot of sense. There are some trials in any psychological experiment that just do not adhere to the interesting cognitive process that motivated the study. It should not be necessary for a cognitive model to be able account for data from these trials in order to be regarded as successful. Nor, in most cases, is it even a good idea to try and extend the model to do so. Retaining a non-hierarchical modeling approach, but complicating the basic model (through a more elaborate likelihood function f, or additional parameters in $\boldsymbol{\theta}$, or both) does not seem like the best theoretical reaction to data that do not have much to say about core aspects of intelligent human cognition. Instead, the additional complexity can come from the sort of hierarchical extension shown in Figure 4, preserving the basic model, but explaining the additional observed data through other processes. Zeigenfuse and Lee (2010b) provide a number of example of this general approach, demonstrating, among other things, how estimates of key parameters in the substantive cognitive model of interest can be affected by the assumptions made about contaminant task behavior.

Besides modeling simple decision-making, there are some other important example of multiple processes being assumed to underly data. The Topics model (e.g., Griffiths, Steyvers, & Tenenbaum, 2007) explains the generation of text documents as coming from simple word selection processes based on a mixture of different semantic topics. The mixture assumption is crucial to explain basic aspects of language like homonymy, where the same observed word can have two or more different meanings, depending upon the latent topic from which it was generated. Another recent example of mixture modeling is provided by Lee and Sarnecka (2010), in a developmental context. These authors showed that children's performance on a simple task assessing knowledge of number concepts could be described as a mixture of systematically different sorts of behavior that was dependent on underlying developmental stages.

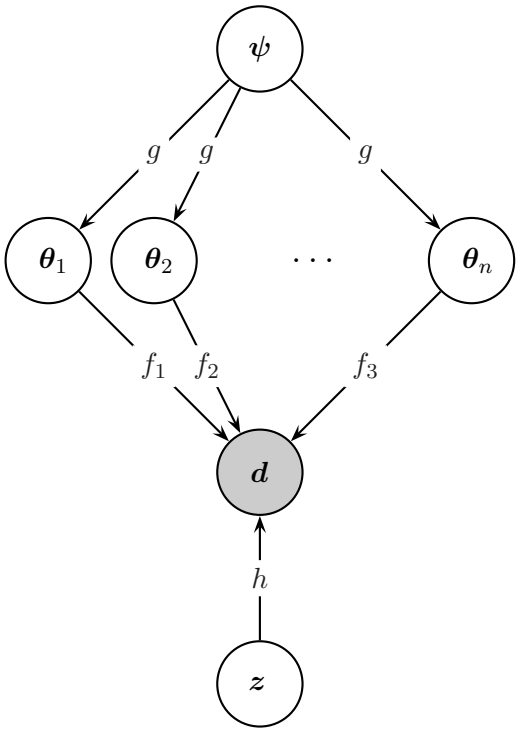

Figure 5. A hierarchical approach for treating data as being generated by a set of different models, but unifying the variation in the models themselves in terms of a common generation process.

Unifying Different Models

Figure 5 shows an ambitious hierarchical modeling structure that combines some of the approaches already discussed. Multiple different models, with data generating processes $f_1, \ldots, f_n$ contribute to the data d, according to some combination rule h governed by $\boldsymbol{z}$. This follows our discussion of Figure 3. But, in addition, the hierarchical approach in Figure 5 combines the various models, assuming they are generated by some processes g controlled by psychological variables $\boldsymbol{\psi}$. This combination is much like Figure 4, except that the unification is at the level of models rather than tasks. This means that the hierarchical approach in Figure 5 not only allows multiple models to account for observed behavior, but also provides a formal account of how those models are generated.

An excellent example of this approach in cognitive modeling is provided by Kemp and Tenenbaum (2008). These authors considered data involving inductive inferences about how binary features belong to a set of stimuli. There are many candidate structured models for explaining these sort of exemplar-by-feature inferences,

including clustering models, tree models, spatial models, and a variety of others (e.g., Shepard, 1980). Each of these models potentially involves different sorts of parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, involving psychological constructs like cluster weights, tree edge lengths, spatial coordinate locations, and so on. In addition, each model requires different processes to translate its representational formalism into observed inductive behavior, necessitating the $f_1, \ldots, f_n$ for each model.

The key contribution of the modeling presented by Kemp and Tenenbaum (2008) is to show how the different representation models can be unified by appealing to a data generating process based on graph grammars. Using some basic building blocks for assembling a graph, formalized by the process g, different choices of values for a controlling set of parameters $\boldsymbol{\psi}$ can produce cluster structures, trees, spatial grids, or a range of other rich representational possibilities. The observed inductive inferences made by people can then be given a very complete and satisfying psychological explanation, which acknowledges that different domains of knowledge are represented differently, but is able to say *how* people could learn those domain-specific representations from basic mental building blocks.

Two other examples of cognitive models following basically the same hierarchical structure as Figure 5 are provided by Lee (2006) and Lee and Vanpaemel (2008). Lee (2006) considered human performance on a type of sequential decision-making task, where different models corresponded to various decision bounds that guided observed behavior. These different decision-bound models were then unified by proposing a simple generating process for establishing a sequential set of thresholds. This overarching generative process was based on a finite state automaton, and controlled by parameters that described the probabilities of the thresholds shifting or staying fixed over the sequence. Lee and Vanpaemel (2008) developed an account of category learning behavior, relying on different models of category representation spanning the range from prototype to exemplar models. They again unified these disparate models with a simple generative mechanism based on psychological variables controlling the level of mental abstraction, and the reliance on stimulus similarity, in forming category representations. It is probably reasonable to argue that these two models—unlike the Kemp and Tenenbaum (2008) modeling—had a common data generating process linking different representational models to the observed data, and so need only a single f in Figure 5. But they do capture the key idea of needing very different models combining to account for the richness of human behavior, while also needing a theoretical unification of those models to provide a complete and coherent account of cognitive complexity.

One current debate in cognitive modeling that can naturally understood in terms of the need for explaining how models are generated comes from the decision-making literature. Here, there is a lively debate surrounding the 'fast and frugal' heuristic approach advocated by Gigerenzer and others (e.g., Gigerenzer & Todd, 1999). This is a theoretically interesting and empirically successful approach to understanding human decision-making in terms of simple heuristic processes that are tuned to en-

vironmental regularities. But, authors like Dougherty, Franco-Watkins, and Thomas (2008) and Newell (2005) have argued that relying on a repository of different simple heuristics to explain decision-making begs the question as to how those heuristics are generated in the first place. This challenge is essentially one of unifying the heuristics, by appealing to more abstract cognitive abilities that are capable of tuning mental capabilities to environments. A successful theoretical resolution would likely fit within the sort of hierarchical Bayesian modeling framework presented in Figure 5.

Perhaps most fundamentally, Vanpaemel (this volume) argues in this special issue that linking models hierarchically is one way to address the basic Bayesian need to specify theoretically meaningful priors. The key idea is that the prior predictive distribution of the hierarchical part of the model, which indexes different basic models, naturally constitutes a psychologically interpretable prior over those models. This is a powerful idea, running counter to a current prejudice for making priors as uninformative as possible, and deserves to be an active area of research in using hierarchical Bayesian methods to model cognition.

Conclusion

Non-hierarchical approaches to understanding cognitive processes dominate the current landscape. The basic approach can probably fairly be caricatured as one of identifying a psychological phenomenon (e.g., generalization, memory, decision-making), finding an interesting task relating to some aspect of the phenomenon (e.g., similarity judgments, recall, two-alternative forced-choice decisions) and building a model that can fit empirical data from the task using a few psychologically meaningful parameters. This is a very reasonable way to begin building a systematic understanding of human cognition, but has serious limitations if attempts are to made to account for its full richness and complexity.

Hierarchical Bayesian methods offer one way—although certainly not the only way (cf. Cassimatis, Bello, & Langley, 2008)—to broaden the scope of current cognitive models. This introduction has tried to identify at least four possible broadening uses of hierarchical Bayesian methods. Firstly, they allow model development to take place at multiple levels of theoretical abstraction. Secondly, they allow the same psychological variables to account for behavior over sets of related tasks. Thirdly, they permit the possibility that data from a single task are best understood as coming from a mixture of qualitatively and quantitatively different sources. And, fourthly they promise to unify disparate models, and as a consequence allow the theoretically-grounded specification of priors.

The papers in this special issue try to demonstrate concretely how hierarchical Bayesian structures can naturally extend current modeling. These extensions can be as theoretically straightforward as allowing for individual differences, or stimulus differences, or the interaction between different sorts of people and stimuli in modeling task behavior (e.g. Rouder et al., 2007; Vandekerckhove, Verheyen, & Tuerlinckx, 2010). They can allow for the more complete theoretical explanation of data from a

single task, lettering different people use different cognitive processes, or letting the same people use different processes at different times. They can force the development of better theories, by demanding that key psychological parameters and processes be identified to explain behavior on a wide range of tasks. They can force the development of new theories, answering new questions about not only what processes and parameters are involved in cognition, but also how those parameters and processes can themselves be modeled.

In short, hierarchical Bayesian approaches demand our accounts of cognition become deeper and better integrated. The aim of this special issue is to provide some concrete examples of the potential of hierarchical Bayes in practice, for models ranging from memory, to category learning, to decision-making. We hope that they are useful early exemplars of what should become an important and widespread way of building and analyzing models of cognition.

References

Averell, L., & Heathcote, A. (this volume). The form of forgetting and the fate of memories. *Journal of Mathematical Psychology*.

Bernado, J. M., & Smith, A. F. M. (2000). *Bayesian Theory*. Chichester, UK: Wiley.

Borg, I., & Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis*. New York, NY: Springer Verlag.

Cassimatis, N. L., Bello, P., & Langley, P. (2008). Ability, parsimony and breadth in models of higher-order cognition. *Cognitive Science*, *33*, 1304–1322.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Congdon, P. (2006). *Bayesian Statistical Modeling*. Chichester, UK: Wiley.

de Finetti, B. (1974). *Theory of Probability, Vol. 1 and 2*. New York: John Wiley & Sons.

Dougherty, M. R., Franco-Watkins, A., & Thomas, R. P. (2008). The psychological plausibility of fast and frugal heuristics. *Psychological Review*, *115*, 199–211.

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

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (Second ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1-67.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

Gregory, P. C. (2005). *Bayesian Logical Data Analysis for the Physical Sciences.* Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Heit, E., & Rotello, C. (2005). Are there two kinds of reasoning? In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive ScienceSociety* (pp. 923–928). Mahwah, NJ: Erlbaum.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge, UK: Cambridge University Press.

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

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

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

Koop, G., Poirer, D. J., & Tobias, J. L. (2007). *Bayesian Econometric Methods.* New York: Cambridge University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*(1), 43–58.

Lee, M. D., & Sarnecka, B. W. (2010). A model of knower-level behavior in number concept development. *Cognitive Science*, *34*, 51–67.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, *32*(8), 1403–1424.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lindley, D. V. (1972). *Bayesian Statistics, A Review*. Philadelphia (PA): SIAM.

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (this volume). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology*.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A User's Guide*. New York: Cambridge University Press.

Marr, D. C. (1982). *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

Merkle, E., Smithson, M., & Verkuilen, J. (this volume). Using beta-distributed hierarchical models to examine simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*.

Morey, R. (this volume). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*.

Newell, B. R. (2005). Re-visions of rationality. *Trends in Cognitive Sciences*, *9*(1), 11–15.

Nilsson, H., Rieskamp, J., & Wagenmakers, E. (this volume). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*.

Norman, K. A., Detre, G. J., & Polyn, S. M. (2008). Computational models of episodic memory. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 189–224). New York: Cambridge University Press.

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

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53.

Pooley, J. P., Lee, M. D., & Shankle, W. R. (this volume). Understanding Alzheimer's using memory models and hierarchical bayesian analysis. *Journal of Mathematical Psychology*.

Pratte, M., & Rouder, J. (this volume). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*.

Rashbash, J., & Browne, W. J. (2008). Non-hierarchical multilevel models. In J. de Leeuw (Ed.), *Handbook of Multilevel Analysis* (pp. 301–334). New York: Springer.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two–choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing withcontaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Ravenzwaaij, D. van, Dutilh, G., & Wagenmakers, E. (this volume). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review*, *12*, 195–223.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*, 63–106.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *214*(24), 390-398.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Sivia, D. S. (1996). *Data analysis: A Bayesian tutorial*. Oxford: Clarendon Press.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34-50.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision-making. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1429–1434). Austin, TX: Cognitive Science Society.

Vandekerckhove, J., Verheyen, S., & Tuerlinckx, F. (2010). A cross random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica*, *133*, 269–282.

Vanpaemel, W. (this volume). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*.

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

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

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

Zeigenfuse, M. D., & Lee, M. D. (2010b). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362.