Psychon Bull Rev (2015) 22:614–628
DOI 10.3758/s13423-014-0725-z

THEORETICAL REVIEW

Moving beyond qualitative evaluations of Bayesian models of cognition

Pernille Hemmer · Sean Tauber · Mark Steyvers

Published online: 19 September 2014

Abstract Bayesian models of cognition provide a powerful way to understand the behavior and goals of individuals from a computational point of view. Much of the focus in the Bayesian cognitive modeling approach has been on qualitative model evaluations, where predictions from the models are compared to data that is often averaged over individuals. In many cognitive tasks, however, there are pervasive individual differences. We introduce an approach to directly infer individual differences related to subjective mental representations within the framework of Bayesian models of cognition. In this approach, Bayesian data analysis methods are used to estimate cognitive parameters and motivate the inference process within a Bayesian cognitive model. We illustrate this integrative Bayesian approach on a model of memory. We apply the model to behavioral data from a memory experiment involving the recall of heights of people. A cross-validation analysis shows that the Bayesian memory model with inferred subjective priors predicts withheld data better than a Bayesian model where the priors are based on environmental statistics. In addition, the model with inferred priors at the individual subject level led to the best overall generalization performance, suggesting that individual differences are important to consider in Bayesian models of cognition.

P. Hemmer (✉)
Department of Psychology, Rutgers University, 512 Frelinghuysen Road, Piscataway, NJ 08854, USA
e-mail: pernille.hemmer@rutgers.edu

S. Tauber
Department of Psychology, University of Adelaide, Adelaide, SA, Australia
e-mail: sean.tauber@adelaide.edu.au

M. Steyvers
Department of Cognitive Science, University of California, Irvine, CA, USA
e-mail: mark.steyvers@uci.edu

Keywords Bayesian models of cognition · Bayesian data analysis · Episodic memory · Individual differences

Introduction

Bayesian models of cognition (BMCs) have experienced a recent upsurge in popularity in the cognitive sciences. These models have made significant theoretical contributions to cognitive science in their accounts of *why* people behave as they do. The strength of BMCs (sometimes also referred to as rational models) is that they can be used to characterize the computational problems people face when trying to make sense of the world given the sparse and noisy input from our senses. Assuming that the mind solves inference problems in a Bayesian way, BMCs give a principled account of how we update our beliefs about the world given observed data, and how our prior knowledge about the world influences our judgment. These models have been applied to a broad range of areas in human cognition (Anderson, 1990) and specific areas such as reasoning (Oaksford & Chater, 1994), generalization (Tenenbaum & Griffiths, 2001), number concepts in children (Lee & Sarnecka, 2010), categorization (Huttenlocher, Hedges & Vevea, 2000), episodic memory (Shiffrin & Steyvers, 1997; Steyvers & Griffiths, 2008), and semantic memory (Hemmer & Steyvers, 2009; Steyvers, Griffiths, & Dennis, 2006). For an overview of BMCs of cognition see Oaksford & Chater (1998), but also see Mozer, Pashler, & Homaei (2008), Jones & Love (2011), Marcus & Davis (2013), and Bowers & Davis (2012a,b) for critiques of BMCs.

Traditionally, the focus in the Bayesian cognitive modeling approach has been on qualitative model evaluations, where predictions from the models are compared to data that is averaged over participants. At this qualitative level, BMCs can provide useful insights into the cognitive goals and computational mechanisms of average individuals. In many cognitive tasks, however, there are pervasive individual differences, e.g., in working memory (e.g., Unsworth, 2007), judgment and decision making (e.g., Vickers et al., 2006), and reinforcement learning (Steyvers, Lee, & Wagenmakers, 2009). While averaging across participants provides a powerful tool for analysis, it might also lead to mischaracterizations of the behavior of individuals, and of the underlying cognitive goals and processes. Estes (1956) cautioned against the uncritical use of averaged curves to determine effects of experimental treatments in the study of learning. Reliance on average data can obscure important individual differences stemming from different psychological parameters within a model or even from different psychological models (Lee & Webb, 2005; Navarro et al., 2006). In this paper, our goal is to develop new approaches in a BMC framework that allow us to use Bayesian procedures to estimate individual difference parameters. The individual differences can relate to process parameters that regulate internal cognitive processes as well as differences in the nature and use of prior knowledge relevant to the cognitive task. Instead of assuming that prior knowledge used in the BMC is based on environmental statistics, our proposed approach allows researchers to estimate subjective prior knowledge that is adopted by individual subjects. Therefore, instead of making strong assumptions about the source of prior knowledge, we simply estimate the prior (subjective) knowledge for individual observers that accounts, in a quantitative fashion, for the observed cognitive behavior.

One consequence of our approach is that incorporating individual differences and subjective prior knowledge in BMCs will increase the flexibility of these models to account for data, which can raise concern about the relative flexibility of these models compared to some non-BMCs (e.g., Bowers & Davis, 2012a). However, it is important to note that the quantitative framework we propose allows for standard methods of model selection that can guard against selecting overly flexible models. Our argument here is not that all BMCs will necessarily benefit from the individual differences framework. Our main contributions are (1) to describe the modeling framework that allows researchers to estimate the subjective priors and potentially include individual differences in the subjective priors, and (2) to provide a demonstration on a reconstructive memory task that a BMC with subjective priors and individual differences is able to generalize to unseen data better than a corresponding model without subjective priors or individual differences.

While the approach to modeling individual differences is not new in terms of the structure of individual differences that we assume, the idea of inferring individual differences within a BMC, based on human data, is novel. Several recent lines of research have been introduced in this area. One approach explored by Griffiths and colleagues (e.g., Martin, Griffiths, & Sanborn, 2012; Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Shiffrin, 2010) is based on a new technique to estimate subjective distributions using the rational framework called ‘MCMC with people’. This technique is an iterative procedure used to infer the subjective probability distributions over categories. The standard approach for investigating mental representations is to experimentally assess participant judgments on fixed stimuli. In MCMC with people, however, the stimuli are not predetermined but rather are adaptively selected. This procedure is often applied between subjects but has also been applied within subjects (Xu & Griffiths, 2010). One advantage of this approach is that it is a simple procedure to get samples from subjective distributions and does not require parametric assumptions. The disadvantage is that these procedures might require non-trivial modifications to existing experimental designs because of their iterative nature. Furthermore, MCMC with people cannot be applied after the fact and must explicitly be a part of the experimental design. Thus, it might be challenging to (re)interpret existing data with this method.

A second approach, termed ‘Doubly Bayesian’, seeks to unify Bayesian models of the mind and Bayesian data analysis (Huszár, Noppeney, & Lengyel, 2010). This approach assumes participants to be quasi-ideal observers using Bayesian inference in the mind. The ideal observer models are then placed in a Bayesian data analysis framework to infer the participants’ subjective distributions over stimuli. The advantage of this approach is that it allows for the inference of the mental representations of humans while eliminating experimental constraints and confounds.

A third approach, termed ‘meta-Bayesian’ or ‘observe the observer’ is based on Bayesian decision theory applied to perceptual inference problems. This approach seeks to simultaneously model perceptual inference at the level of the subject and statistical inference about the perceptual inference processes at the level of the experimenter (Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan et al., 2010). The strength of this meta-Bayesian approach is that it extends the Inverse Bayesian decision Theoretic (IBCT) problem to experimental psychology, neuroscience, and other experimental paradigms not suitable for the IBCT problem.

In this research we develop an integrative framework similar to the Doubly Bayesian and meta-Bayesian approaches and focus our analysis on the problem of estimating subjective priors. As the name suggests, the Bayesian approach is used in two distinct ways – one as an estimation procedure to analyze individual differences and the other to motivate the inference process within a cognitive model. In this way, we apply ideas of Bayesian Data Analysis (Kruschke, 2010; Lee, 2008) to learn about the underlying psychological variables of the BMC, and ideas from rational analysis to motivate a model of the mind of the observer. The key point of our approach, and that of Huszár et al. (2010) and Daunizeau et al. (2010), is that combining these two approaches involves an application of Bayesian inference at the level of the researcher and observer. The observer is trying to make the best use of information to make a decision or update an internal representation, and the researcher is trying to infer what is going on in the inference process in the observer's mind. Both the observer's model and the researcher's model of the observer's model are being evaluated concurrently.

The plan for this paper is as follows. We will first sketch out how to integrate the observer and experimenter models by introducing each step – from the experimenter designing the stimuli, and the BMC of the observer in the experimental task, to the Bayesian account of the experimenter analyzing the observed behavior. We will then illustrate this model on behavioral data from an experiment for the recall of the height of people and contrast three different approaches to determine the subjective priors. Finally, we will assess the generalization performance of the three models in a cross-validation procedure.

An integrated observer-researcher analysis

We will illustrate our approach with a relatively simple Bayesian model of reconstructive memory. Our approach can, however, in principle, be applied to other areas of memory, as well as any area of cognition. In the area of memory, starting with the pioneering approach by Anderson (1990), a number of BMCs have been developed in episodic and semantic memory (Shiffrin & Steyvers, 1997; Steyvers & Griffiths, 2008; Steyvers, Griffiths, & Dennis, 2006; Xu & Griffiths, 2010). There are likely better and more sophisticated Bayesian models of memory than the one instantiated here. We chose this model because it is the simplest possible BMC and provides a good qualitative fit to the data. One could easily imagine more complicated models as we have in the past. In a previous work, two of the current authors have developed Bayesian memory models for the reconstruction of events from memory using multiple sources of information – episodic and semantic (Hemmer & Steyvers, 2009). In this approach, episodic memory is viewed as a problem of extracting and storing information from noisy signals presented to our senses, which need to be combined with prior knowledge about the environment. Specifically, in the experiments by Hemmer and Steyvers, observers were presented with a series of stimuli (e.g., vegetables and fruits) during a study phase and were instructed to retrieve from memory an attribute (e.g., the size) of specific stimuli at a later time. The results showed that the memory estimation errors that observers make can be explained by the use of prior knowledge – smaller objects were later recalled to be larger and larger objects were later recalled to be smaller, as predicted by a memory system that uses prior knowledge at the category level to help the recall of instance-specific attributes.

In this paper, we will apply this Bayesian modeling approach to data from a memory task where people estimate the height of men and women. We chose to illustrate our approach with recall for height, because people are known to have strong prior expectations for height (Nelson, Biernat, & Manis, 1990), and are quite accurate at estimating height based on accessible gender information (Kato & Higashiyama, 1998). Furthermore, the distribution of heights in the population is well documented, and known to be normally distributed. As in the Hemmer and Steyvers (2009) study, having categorical knowledge leads to two clear predictions about the effect of prior knowledge on episodic memory. First, that there will be an overall regression to the mean, where recalled height for people at heights below the mean population height will be overestimated while recalled height for people at heights above the mean population height will be underestimated (see Fig. 1, panel A). The second and critical prediction is that when two people (a short male and a tall female) have the exact same height, recall will be differentially biased towards the height distributions that are gender specific. The tall female will be underestimated towards the mean of female height and the short male (objectively the same height as the female) will be overestimated towards the mean of male height. In other words, prior knowledge will differentially affect the memory of two people originally presented at the same height. Figure 1, panel B illustrates this effect.

In the height estimation experiment, the goal of the observer is to reconstruct the original study events (the heights of people shown sequentially, either male or female) as best as possible given their noisy episodic memory content and their prior knowledge of the stimulus attribute (i.e., general knowledge of heights of males and females). The goal of the researcher who analyzes the data from the memory experiment, on the other hand, is to estimate potential individual differences. In our scenario this could relate to differences in priors or memory noise that are part of the BMC.

To analyze the goals of the researcher and observer simultaneously requires an integrative application of Bayesian inference. The conceptual challenge in this approach is to carefully distinguish between observed and latent variables[1] because what is known and what is latent depends on whose perspective we are considering. In the memory task, the observer and researcher have access to different information. For example, the observer knows their own memory content, but the researcher does not. The researcher knows the original stimulus that was provided, whereas the observer is trying to infer the original study stimulus. Table 1 shows how the observer and the researcher differ on what is known and what is latent.

We will adopt graphic models as a way to visualize the structure of the model. There are different graphic models that need to be considered – the model from the observer's perspective, who is trying to reconstruct from memory the original study event, and the model from the researcher's

[1] From now on, we will refer to *observed* variables as *known* variables to avoid confusion with the term *observer*.

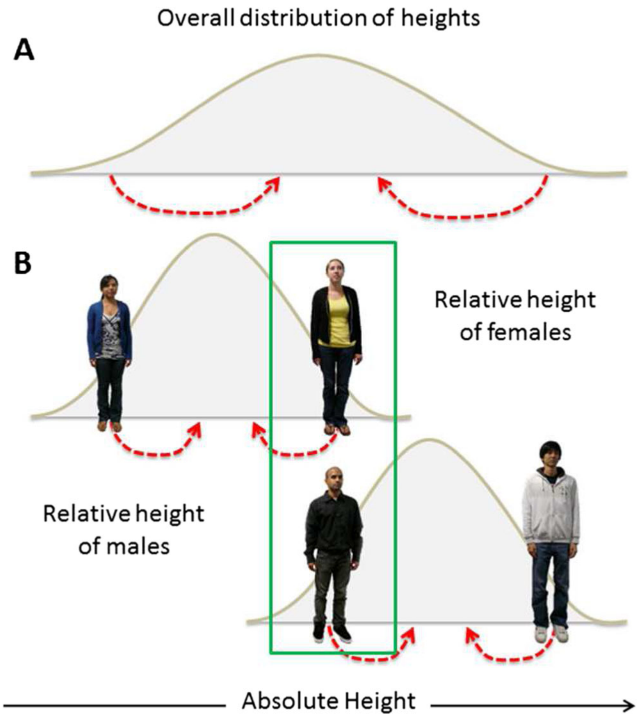

Fig. 1 Predicted biases in recall as a function of (**A**) the overall distribution of height in the population, and (**B**) gender information for the height of females and males

perspective, who is trying to infer what is going on in the observer's mind. Figure 2 shows the graphic models from these different perspectives. In the graphic models shaded nodes indicate known variables and unshaded nodes indicate latent variables. The arrows in the model indicate the conditional dependencies. For simplicity and visual clarity we eliminate any indices in the graphic model that represent the repetitions in our sampling steps (across stimuli, categories, individuals, etc.). The next sections detail these assumptions at each stage of analysis.

Generating the stimulus from the researcher's perspective

We first consider the question of how the stimulus is generated in the experimental task. The left panel of Fig. 2 shows the graphic model for the researcher who generates stimulus θ to be studied by the observer in the task. The variable θ measures an attribute of the stimulus, such as the height of a person that needs to be remembered. In the height memory experiment we will report later, the stimulus is drawn from the true environmental distribution with a mean of μ^* and a precision of τ^*, which is known to the researcher. For simplicity, we assume here that the environmental distribution of the stimulus is Gaussian, $\mu \sim \mathrm{N}\ (\mu^*,\ \tau^*)$. While the true environmental distribution is known to the researcher, note that the observer might not have perfect knowledge of the environmental statistics of the attribute.

Table 1 Known Variables for Observer and Researcher in Memory Task

	Original stimulus	Memory content	Subjective prior
Observer	No	Yes	Yes
Researcher	Yes	No	No

A Bayesian account from the observer's perspective

We next consider how the observer in the memory task might reconstruct the studied stimulus. The goal for the observer is to reconstruct the original attribute θ using the content of episodic memory and their prior knowledge of the study attribute. We first need to make some assumptions on how the episodic memory contents are created during study. We assume that a single memory sample y is drawn from a Gaussian distribution centered on the original stimulus value[2]:

$$y \sim \mathrm{N}(\theta, \psi) \tag{1}$$

The precision parameter ψ is a parameter that governs memory noise – it controls the degree to which the stored episodic representation resembles the attribute of the original studied stimulus. Suppose the observer also has some subjective knowledge about the general distribution of attributes in the form of a Gaussian distribution:

$$\theta \sim N\ (\mu, \tau) \tag{2}$$

This distribution corresponds to the prior in the observer's memory model. There are a number of approaches to determine the parameters μ and τ of this prior distribution. One approach, consistent with a typical rational analysis of a cognitive task, would be to assume that the observer has learned the environmental statistics and that the actual environmental statistics can be used as a proxy for the prior in the observer's model. In our experimental setting, we would assume that μ=μ^* and τ=τ^*. Another approach is to assume that the observer uses a subjective prior where the parameters μ and τ might not accurately reflect the true environmental statistics. In this case, the parameters μ and τ need to be estimated from the behavior of the observer. In this paper, we will explore both approaches to determine the observer's prior.

The goal for the observer is to reconstruct the original attribute from memory. Bayes' rule gives us a principled way of combining prior knowledge and evidence from memory to calculate the posterior probability, $p(\theta|y, \psi, \mu, \tau) \propto p(y|\theta, \psi)p(\theta|\mu, \tau)$. The posterior probability $p(\theta|y, \psi, \mu, \tau)$

[2] We can also extend the approach and assume that multiple samples are stored depending on the amount of study time. Since study time is not a relevant factor in the current experimental approach, we have restricted the model to a single sample.

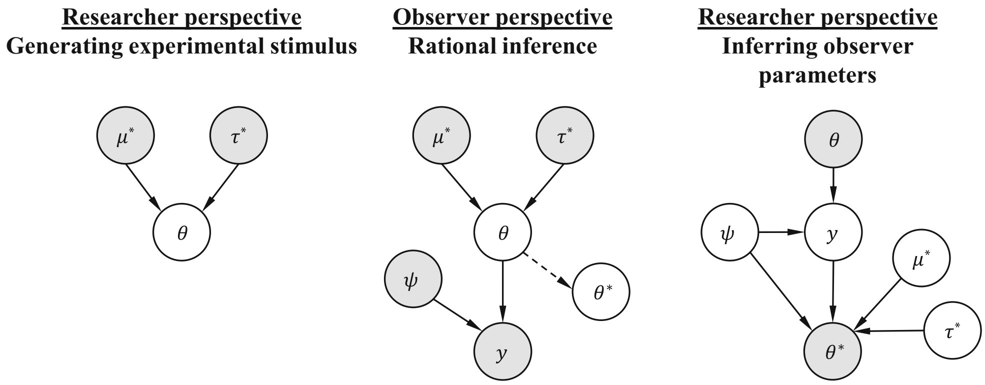

Fig. 2 Graphic models that relate the generation of experimental stimuli, and models from the observer's and the researcher's perspectives

describes how likely attribute values θ are given the noisy memory contents y and prior knowledge of the attributes. Standard Bayesian techniques (Gelman et al., 2003) can now be used to calculate the posterior distribution:

$$\theta \,\Big|\, y, \psi, \mu, \tau \sim \mathrm{N}\left(\frac{\psi y + \mu\tau}{\psi + \tau}, \psi + \tau\right) \tag{3}$$

Note that the mean in Eq. (3) is a weighted linear combination of the prior mean μ and memory content y. The prior mean μ is weighted more heavily in recall when the prior has a higher precision τ and when the memory noise increases – which is equivalent to a decrease in the memory precision, ψ. This corresponds to the intuitive notion that if the prior is strong, it will have a strong influence on recall. Similarly, if memory contents are very noisy, the prior will also exert a strong influence on recall. This BMC predicts systematic biases toward the category center, or prior category mean, at reconstruction. Note also that the solution in Eq. (3) assumes that the memory precision, ψ is known to the observer. The solution where the memory precision is latent for the observer was presented in Hemmer and Steyvers (2009).

For purposes of notation it is important to distinguish between two conceptual usages of the variable θ. This variable can refer to the actual stimulus value used during study but also the value inferred by the observer at test. We will use θ to refer to the actual stimulus and introduce a new variable θ^* to refer to the response produced by the observer. The response can be based on the posterior distribution in Eq. (3) in a number of ways, such as the mean, the mode, or a sample. We will assume that θ^* is based on a sample of the posterior distribution and therefore follows the same distribution as Eq. (3). In the graphic model in the middle panel of Fig. 2, we use the somewhat unconventional notation of a dashed line between nodes θ and θ^* to indicate that θ^* is based on a sample from the posterior of θ.

A Bayesian account from the researcher's perspective

To explore individual differences in recall performance we need to consider how to combine Bayesian data analysis with our Bayesian model of reconstructive memory. Now that we have specified the BMC for the observer, we can assume that both the observer and the researcher use the same BMC, but with different unknowns, as detailed in Table 1. To the observer at test, the goal is to infer the original stimulus θ, which is unknown at the time of test and must be inferred from the episodic memory contents y. To the researcher, however, the observer's response is the observed data in the experiment, and the unknowns are the memory contents y, memory noise ψ and subjective priors μ and τ in the mind of the observer.

Each perspective is associated with different but interrelated graphic models. In the observer model, recall θ^* is a sample from the posterior distribution of θ given the observed information. From the perspective of the researcher, both the original studied attribute θ, as well the observer's response θ^*, are part of the known (observed) data. We can restructure the graphic model such that the dashed line (using the unconventional notation to indicate a sample from a posterior distribution) is removed and instead describe the process that generates the observer's response θ^*. This is shown in the right panel of Fig. 2. Note that although the graphic models in the middle and right panels of Fig. 2 look very different, they are in fact closely related and depend on which person is doing the Bayesian analysis: the observer or the researcher.

The goal of the researcher is now to infer the observer parameters in terms of the subjective priors μ and τ as well as memory noise ψ in the mind of the observer. In other words, the researcher wants to know the posterior distribution of $p(y, \psi, \mu, \tau|\theta, \theta^*)$, which using Bayes' rule is given by $p(y, \psi, \mu, \tau|\theta, \theta^*) \propto p(\theta^*|y, \psi, \mu, \tau)\, p(y|\theta, \psi)\, p(\psi)\, p(\mu)\, p(\tau)$. The terms $p(y|\theta, \psi)$ and $p(\theta^*|y, \psi, \mu, \tau)$ can be evaluated using Eqs. 1 and 3, respectively. To complete the Bayesian analysis from the researcher's perspective, we need to specify the prior distributions for the subjective observer variables, $p(\psi)$, $p(\mu)$, *and* $p(\tau)$. In the next section, we will describe a reconstructive memory experiment and a number of modeling approaches to estimate parameters in the Bayesian memory model.

Overall, the novelty in this approach is that the posterior distribution of the BMC (Eq. 3) is used as the likelihood function for the human responses given the parameter values. This enables us to use Bayesian methods to infer parameter values (e.g., subjective priors, process level parameters) that can best explain the human data. This is different from the usual approach in Bayesian cognitive modeling, where the posterior distribution is used only to simulate responses from the model but where the BMC is not adjusted to the human data. For the particular BMC we consider in this paper, the posterior distribution is in an analytic form which makes Bayesian estimation particularly straightforward. However, Bayesian estimation can also be applied to BMCs where the posterior does not come in analytic form. Overall, the advantage of formulating the BMC at the researcher's perspective is that model evaluation methods such as Bayesian model selection and generalization tests can be applied to quantitatively assess the BMC.

A reconstructive memory experiment

We will now describe the experiment involving memory for heights of people, and show a number of approaches to estimate subjective priors in the Bayesian memory model. In the experiment participants viewed images of males and females that were representative of the general population. Notably, the selected images for each gender were distributed with the same range and frequency as the heights in the general population.

Participants

Twenty-two undergraduate students at the University of California, Irvine, CA, USA participated in the experiment. The participants were not involved in the stimulus development phase. They were compensated with course credit.

Stimuli

We developed naturalistic stimuli in the form of photographs of real people. We photographed 212 randomly selected male (68) and female (144) students at the University of California, Irvine. They participated in exchange for course credit. All images were taken against a white wall and next to a blue door. The subjects were required to stand up straight, have their hands to the side, maintain a neutral expression, and look at a designated fixation point on the opposite wall. Women were required to have their hair up. All participants were instructed to stand in a fixed position relative to the camera. The distance from the camera to the participants' heels was exactly 230 cm. The camera was maintained at 120 cm of elevation (which was roughly the center of the photographic frame) from the floor using a tripod. Each individual was measured and their height was recorded in whole inches. See Fig. 1 for sample images.

To ensure that our sample of individuals was representative of the general population, we compared our sample distribution to height statistics obtained from the Center for Disease Control (McDowell et al., 2008). Our sample ranged in height from 147.3 to 182.9 cm for females and from 162.6 to 193 cm for males. The sample was normally distributed around a mean of 162.9 cm for females and 175 cm for males. This is comparable to the range and distribution in the US population over age 20 years with a mean of 162.1 cm for females and 176.3 cm for males.

From the 212 photographs we selected 48 images, 24 female and 24 male, to be used as the experimental stimuli. Importantly, the selected images for each gender were distributed with the same range and frequency as the heights in the general population based on the CDC data. The purpose of the experimental stimuli was that it retained all aspects of the figure, allowing participants to use prior knowledge of height when interacting with the stimuli. All original stimuli measured 456 × 1229 pixels, a resolution sufficiently high such that when we filled the whole screen with the image, no artifacts due to pixilation were visible.

Procedure

Participants completed a continuous recall paradigm where the study and test trials were randomly intertwined. The study and test stimulus was presented sequentially on the right side of the computer screen. On all trials an image of a door was displayed as a comparison image on the left side of the screen. The door in the image was the door to the experimental room, and before entering the participants were asked to familiarize themselves with their height relative to the door.

Each study image was presented for 2 seconds at the true height of the figure relative to the door. The true height of the door remained fixed at 80 % of the total screen size. For each test trial the height of the person in the stimulus was initially presented at a random size between 50 % and 100 % of the total screen size, which means that each test figure was initially presented between 50/80 = 62 % and 100/80 = 125 % of their true size relative to the door. The task for the participant was to rescale the height to address the question: "What was the height of this person, compared to the door on the left, when you saw them at study? If you are not sure, make a best guess." To reconstruct the studied size, the participants used the computer mouse to move a slider on the right edge of the screen. Once they had scaled the figure to the size they recalled from study, they clicked on a button labeled "OK" and proceeded to the next trial until the experiment ended.

Experimental results

To evaluate performance we measured recall error as the difference between the recalled size and the studied size. Positive numbers mean that the stimuli are overestimated at recall and negative numbers mean that the stimuli are underestimated at recall. The empirical results, shown in Fig. 3, indicated a systematic regression to the mean effect. There was a bias toward the mean height such that short people were overestimated and tall people were underestimated. Furthermore, there were two separate regression effects, one for male and one for female stimuli. In other words, participants regressed to two different mean heights: one for each gender. These differences are compatible with the idea that people use their prior knowledge of height in the population and use this information when recalling studied heights of men and women.

To assess the influence of this gender-level prior knowledge, a linear regression model was fitted separately for each subject. The regression model contained three parameters: two intercept parameters, corresponding to male and female, and a single slope parameter. Table 2 shows the mean estimated slopes and intercepts across categories.[3]

These differences are consistent with an influence of prior knowledge at a more fine-grained level of knowledge. The intercept differences confirm the prediction that when a male and a female are studied at the same height, reconstruction is differentially biased depending on their relative height.

Bayesian model analyses

We next describe three model-based analyses of the empirical results. First, we describe a qualitative Bayesian model where the priors are not estimated and instead are based directly on the environmental statistics. Next, we describe a model that estimates a single subjective prior for all observers. Finally, we present a hierarchical model in which each observer is associated with their own subjective prior.

Priors determined by environmental statistics

Our first step is to apply the standard analysis in Eqs. 1, 2, and 3 to the experiment without directly estimating any parameters. Instead, we assume that the observers adopt a prior that corresponds exactly to environmental statistics. The goal of this analysis is to compare the predictions of the BMC and the empirical data at a qualitative level.

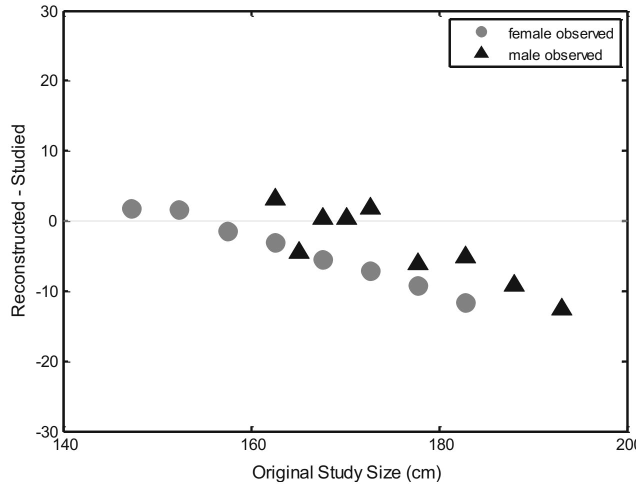

Fig. 3 Empirical results. Recall bias is shown as a function of original study size and stimulus category

For the prior in Eq. 3 we use a mean μ female of 162.14 cm, and μ male of 176.3 cm, and a precision τ female of 0.009 and τ male of 0.007 (corresponding to standard deviations of 10.5 cm and 12.0 cm respectively). These are based on environmental priors obtained from the CDC. They also correspond to the distribution of the stimuli in the experiment (μ^* and σ^* in Fig. 2, 'Researchers Perspective') and are centered on the mean height for each gender. We used the model to simulate exactly the same trials that we used in the experiment – including the same sizes for study stimuli. For the memory precision ψ, we used a value 0.0149, which is based on the variance in the experimental stimuli, although the exact value does not influence the qualitative results. The parameters and priors used in the model are shown in the top panels of Fig. 4. The two right panels at the top compare the simulated responses to those of human observers. For both simulated and observed responses, the results show effects of the category prior. The heights of people that are relatively short for their gender are overestimated while the heights of people that are relatively tall for their gender are underestimated.

Overall, when using environmental priors, the model produces results that are qualitatively consistent with the responses given by human observers. However, it is also clear that there are quantitative differences between the observed and predicted results. The model predictions for the female

Table 2 Mean slopes and intercepts by category

	Females		Males	
	Mean	SD	Mean	SD
Slope	−0.39	0.28	−0.39	0.27
Intercepts	59.88	4.08	65.89	3.74

[3] The results show that the intercept for female was smaller than that for male. This difference in intercepts by relative study size supports the prediction of gender-level prior effects. A one-way ANOVA with two levels (female, male) found a significant effect of category [F (1, 42) = 25.83, $p < .001$].

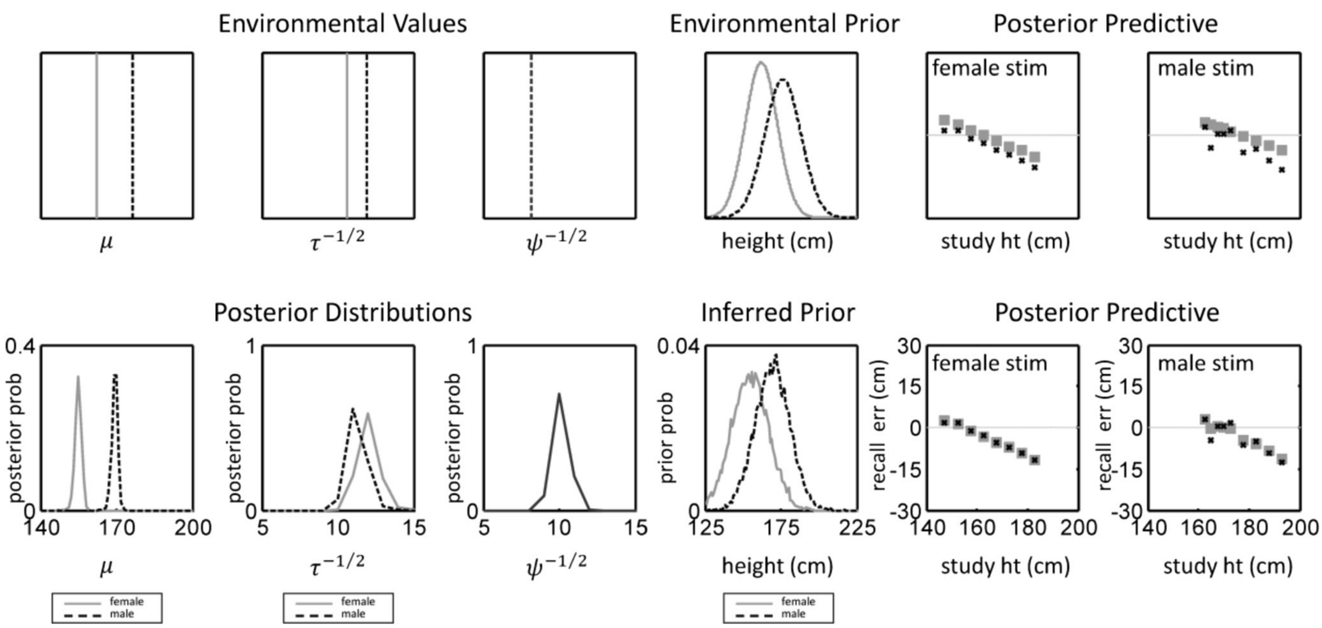

Fig. 4 Comparison of the parameters and model predictions for the model where the priors are determined by environmental statistics (top row) or estimated without assuming individual differences (bottom row). The left three columns show the point-estimates (top panel) and posterior distributions (bottom panel) for the means (μ), standard deviations ($\tau^{-1/2}$), and memory noise parameter ($\psi^{-1/2}$). The fourth column shows the environmental prior and inferred prior corresponding to these parameters. The two right columns show the mean posterior predictives for the recall errors in the memory experiment, separately for the female and male stimuli. *Ht* height, *stim* stimuli, *prob* probability

stimuli consistently lie above the observed values, and the predicted slope for the male stimuli is too shallow compared to the human observers. Both results suggest that there might be other factors at work than observers simply using the true environmental information, e.g., either that the observers use a different prior to the environmental prior, or that some subjects use the environmental prior while others use a different prior, and that the observed pattern averages out the individual differences in our pool of subjects.

Estimating subjective priors – no individual differences

In the next modeling approach, we use the integrated observer-researcher analysis to estimate the subjective priors. This model is less restrictive than the previous model because the subjective prior does not have to correspond to the environmental statistics. We start with a simple model that assumes that all observers use the same subjective prior for μ and τ (one each for the female and male category). Therefore, the model assumes no individual differences. In addition, to keep the model simple we assume a single memory noise parameter across conditions (although this assumption can be relaxed).

To introduce notation, let N and M represent the number of observers and experimental trials per observer respectively. Let K represent the number of categories in the experiment (K = 2 in the current experiment). Let $1 \leq i \leq N$ index the observers, $1 \leq j \leq M$ index the trials in the experiment, and $1 \leq k \leq K$ index the stimulus category. In the experiment, each observer receives a different ordering of trials in the experiment. To keep track of the stimulus and category assignment for each trial, we introduce the variable $c_{i,j}$, where $1 \leq c_{i,j} \leq K$, to indicate whether a given stimulus is male or female for the i-th observer on the j-th trial.

The no-individual difference model assumes that there is only a single subjective prior for the male and female stimuli across individuals. The prior distribution parameters μ and τ are vectors that each contain two values for the male and female stimuli respectively. Therefore, μ_1 and μ_2 represent the means of the subjective prior distribution for male and female heights respectively.

The no-individual differences model extends Eq. 3 as follows:

$$\theta^*_{i,j} \Big| y_{i,j}, c_{i,j}, \psi, \mu, \tau \sim \mathrm{N}\left(\frac{\psi y_{i,j} + \mu_{c_{i,j}} \tau_{c_{i,j}}}{\psi + \tau_{c_{i,j}}}, \psi + \tau_{c_{i,j}}\right) \tag{4}$$

where we have assumed that the observer response $\theta^*_{i,j}$ has the same distribution as the posterior distribution $\theta_{i,j}$ in the observer model. We extend Eq. 1 with our new notation to get the distribution of the memory trace for a specific observer-trial:

$$y_{i,j} \Big| \theta_{i,j}, \psi \sim \mathrm{N}\left(\theta_{i,j}, \psi\right). \tag{5}$$

To complete the model, we place vague prior distributions on all of the parameters that are unobserved from the perspective of the researcher:

$$\psi \sim \text{Gamma}(0.001,\ 0.001),\ \mu_k \sim \text{Unif}(90, 300),\ \ \tau_k \sim \text{Gamma}(1, 1) \tag{6}$$

The resulting graphic model is shown in Fig. 5. In the simulations of the model, we used JAGS (Plummer, 2003) to estimate the joint posterior distribution of the model parameters $\{\psi, \mu, \tau, y\}$. For each model, we obtained 2,500 samples from the joint posterior after a burn-in period of 2,500 samples, and we also collapsed across four chains.

Figure 4, bottom panel shows the model results. The model provides a better fit of the empirical data relative to the qualitative rational analysis (we will provide a quantitative assessment of model fit in the next section). This suggests that it is important to estimate the subjective prior. The panel labeled 'inferred prior' shows the estimated prior distribution that observers have for the heights of females and males. We generated these estimates by drawing samples from a normal distribution, where the parameters of this distribution were resampled on each draw. The parameters for each sample of the estimated priors were drawn from the inferred posterior distributions of u and τ. Figure 6 shows the same results for the environmental and inferred prior distributions but overlays them to better illustrate the differences. The results suggest that the prior used in the observer's memory model is based on an underestimation of heights in the population. Note that it is possible that all our subjects have accurate knowledge of the environmental statistics yet still use the wrong distribution in their memory model. On the basis of the current experimental data, we cannot conclude anything about the locus of the mismatch between the environmental statistics and the estimated priors. Our results simply suggest that if the subjects are following Bayesian inference procedures to derive their memory responses, they appear to do this on the basis of distributions that are similar but not equivalent to the environmental distributions.

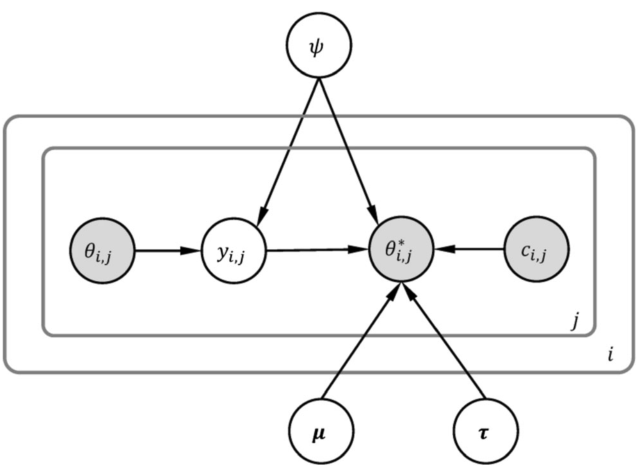

Fig. 5 Complete graphic model from the researcher's perspective with no individual differences. The plates indicate repeated sampling steps for observers (indexed by i) and trials (indexed by j)

Estimating subjective priors with individual differences

In the previous section we assumed that all observers have the same prior distributions and memory precision. Here, we take this a step further and allow for individual differences such that each observer has an individual prior distribution with mean μ_i and precision τ_i – we assume that each observer has a single precision parameter that is used for both the female and the male prior distributions. The individual differences model extends Eq. 4 as follows:

$$\theta^*_{i,j}\Big| y_{i,j}, c_{i,j}, \psi_i, \mu_i, \tau_i \sim N\left(\frac{\psi_i y_{i,j} + \mu_{i,c_{i,j}}\tau_i}{\psi_i + \tau_i}, \psi_i + \tau_i\right) \tag{7}$$

Similarly, we extend Eq. 5 with individual differences to get the distribution of the memory trace for a specific observer-trial:

$$y_{i,j}\Big|\theta_{i,j}, \psi_i \sim N\left(\theta_{i,j}, \psi_i\right) \tag{8}$$

As in the previous model, we place vague prior distributions on all of the parameters that are unobserved from the perspective of the researcher:

$$\psi_i \sim \text{Gamma}(0.001, 0.001), \mu_{i,k} \sim \text{Unif}(90, 300), \tau_i \sim \text{Gamma}(1, 1) \tag{9}$$

We applied standard Bayesian inference techniques to infer the unknown parameters ψ_i, μ_i and τ_i for each observer i. Similar to the no-individual differences model, we obtained 2,500 samples from the joint posterior after a burn-in period of 2,500 samples, and we also collapsed across four chains.

Although we were able to infer posterior distributions for the prior parameter values and memory precision for every observer individually, we show the complete set of results for only a representative subset of the observers in Fig. 7. The right two columns in Fig. 7 show the posterior predictive distributions for the responses for each of the four selected observers. These are the distributions of future (unknown) responses that the model predicts the observer will make for new presentations of the stimuli. Each observer's responses fall within high-probability regions of the posterior predictive distribution, which indicates that the model provides a good description of the data. Figure 8 shows the *mean* posterior predictive responses for all observers using both the individual differences model and the environmental model. Whereas the posterior predictive plots in Fig. 7 only show predictions for stimuli that were used in the experiment, Fig. 8 shows mean predictions over a wider range of stimuli. The panel labeled

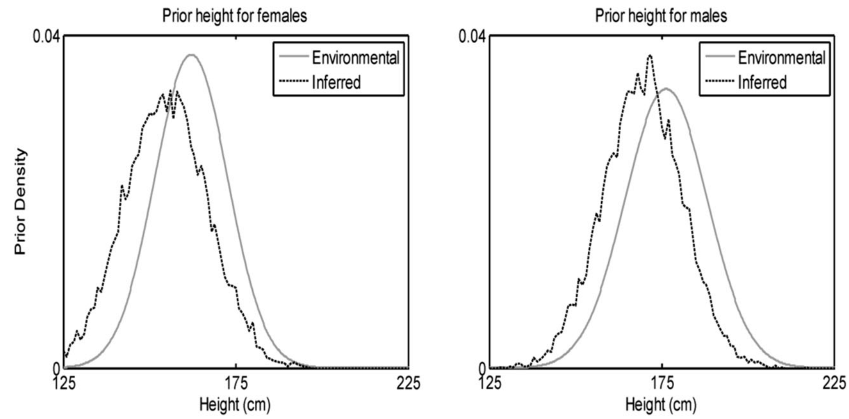

Fig. 6 Comparison of the estimated prior distributions for the height of females and males, and the true environmental distributions based on data from the CDC

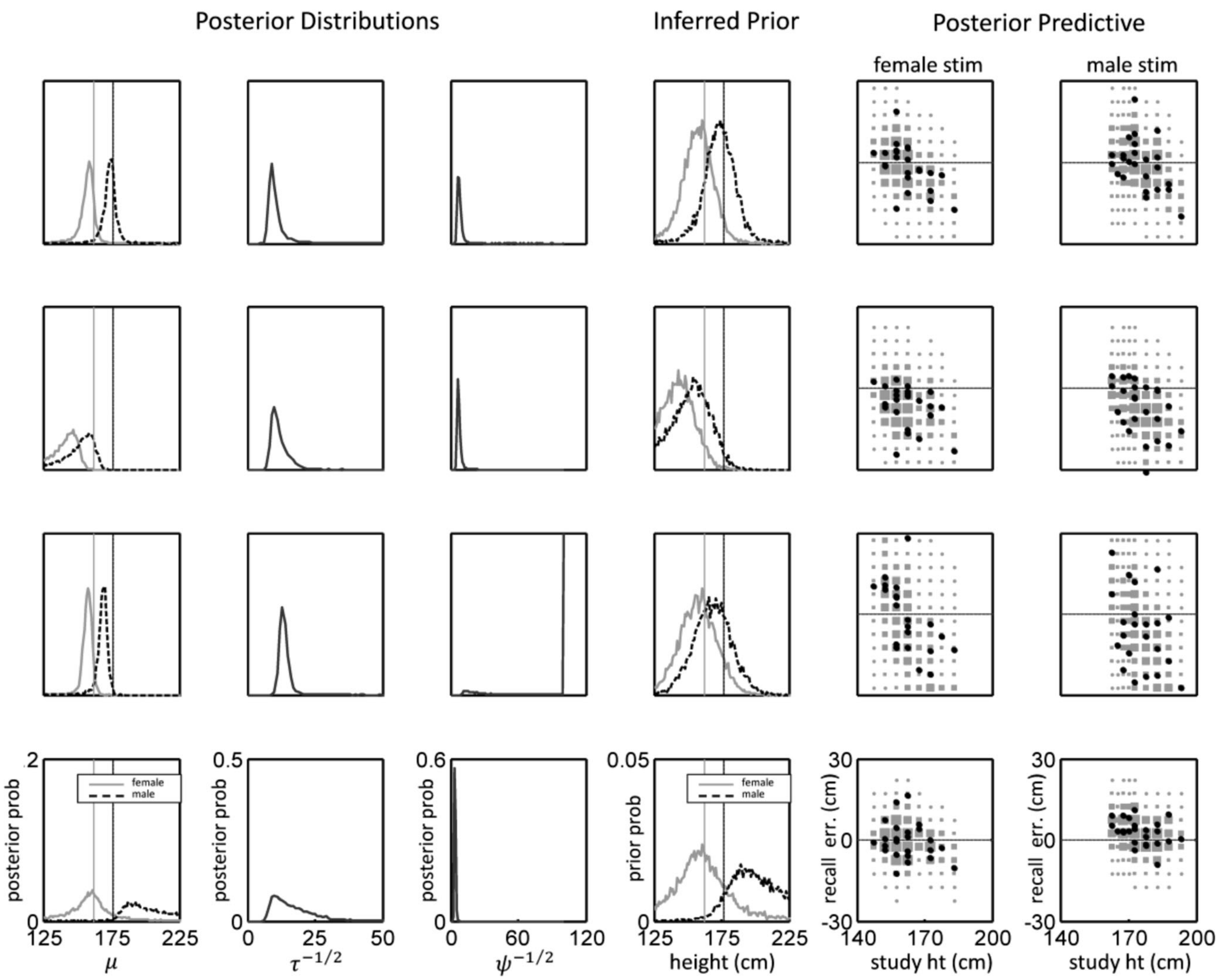

Fig. 7 Results for the individual differences model for observers 14, 8, 10, and 17 (from top to bottom). The first three columns show the posterior distribution of model parameters for each observer. The fourth column is a reconstruction of each observer's prior distribution for the height of females and males. The vertical lines in the first and fourth columns show the actual mean height of females and males in the environment (based on data from the CDC). The last two columns show the posterior predictive distribution of future, unobserved, responses (probability is proportional to the area of the grey squares) against individual response data (black dots) for female (column 5) and male (column 6) stimuli. *Stim* stimuli, *ht* height

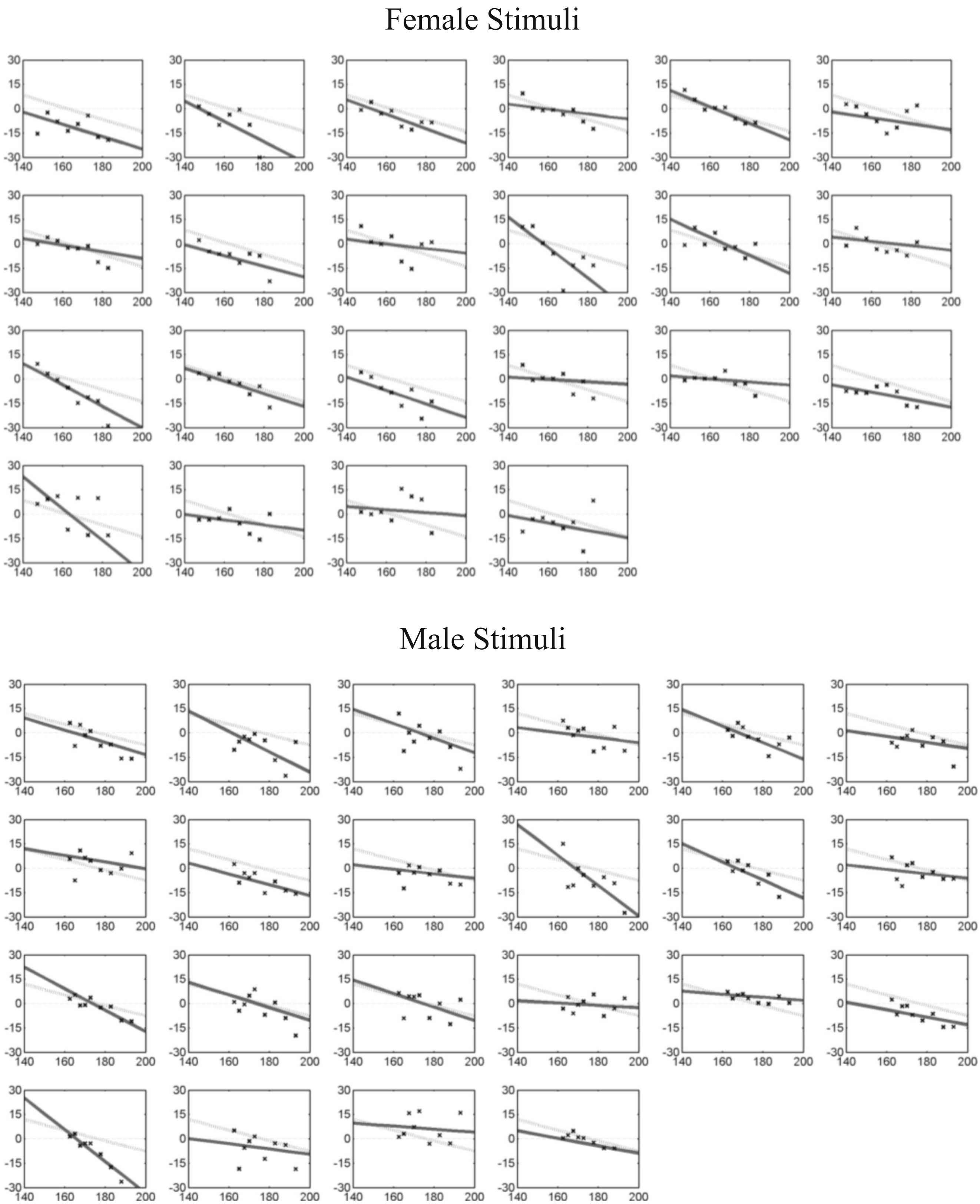

Fig. 8 Mean posterior predictive for the individual differences model (dark line) and the environmental prior model (light line) against averaged response data (black dots) for each participant

'inferred prior' (Fig. 7) shows the estimated subjective prior distributions for the height of males and females that were used by each observer. We estimated these subjective priors for observer i by taking samples from a normal distribution where, for each sample, the mean and precision were sampled from the posterior distributions for μ_i and σ_i.

Our results suggest that there are individual differences between observers' prior distributions for height, as well as for memory noise. It appears that some observers use a model similar to the environmental model. For example, the estimated subjective priors for observer 14 (Fig. 7, top row) are very similar to the environmental distributions. The mean posterior predictive responses for the individual differences model for this observer are very similar to those of the environmental model (Fig. 8, second column of the third row).

Observer 8 is representative of several observers who appear to have subjective priors that underestimate the heights of men and women. Figure 7 (second row) shows the estimated subjective priors for this observer. Figure 8 (second column of the second row) shows that the mean posterior predictive responses for observer 8 follow the same pattern as the environmental model predictions, but are consistently lower.

Observer 10 (Fig. 7, third row) had a very high estimated memory noise. This corresponds with posterior predictive responses that deviate more extremely from the studied heights. Although they were more extreme, the observer's responses showed the same general pattern of overestimating below-average heights and underestimating above-average heights. This pattern is visible in the model predictions (Fig. 8, fourth column of the second row).

Lastly, the results for some observers were not consistent with a systematic regression to the mean effect. For example, the responses of observer 17 (Fig. 7, bottom row) are not suggestive of a bias toward the mean height such that short people were overestimated and tall people were underestimated. Instead, the dispersion of responses is relatively symmetrical around the true height of the stimuli that were being recalled. The average responses for observer 17 (Fig. 8, fifth column of third row) are very close to the true height of the test stimuli. The averaged model predictions in Fig. 8 form a line with a slope near zero, which captures this observer's tendency to provide responses that are unbiased by the height of stimuli. The only systematic bias appears to be an overestimation of the heights of males, regardless of height.

Model comparison

The strength of the approach we have presented here using the posterior distribution of the BMC as a likelihood function for the purpose of Bayesian data analysis is that it allows us to compare models quantitatively using approaches such as Bayesian model selection techniques and generalization tests. We pursue a simple generalization test based on cross-validation where the data is partitioned into a training set used to estimate the model parameters and a validation set (with data unseen by the model) to test the generalization performance of the model. There are a number of ways to set up partitioning of the data into training and validation sets. One could leave out a subset of participants and test which models can best generalize to new data from new participants. Alternatively, one can leave out a subset of trials from each participant and test which models can best generalize to new data from existing participants. Because one of our goals is to compare BMCs with and without individual differences, we have focused on the latter approach. This allows us not only to test whether learned priors are better than empirical priors but also whether learned priors at the individual subject level generalize better than learned priors without individual differences.

We used a tenfold cross-validation procedure where the human data were split into ten randomly generated training and validation sets. Each of the validation sets contained approximately 10 % of the data. Each data point was in a validation set exactly one time across the ten folds. In each fold, we used Bayesian inference to infer posterior distributions for the unknown variables in a model based only on the training data. For each individual, we computed the likelihood of each of their responses, given the posteriors that were inferred when the response was out-of-sample, i.e., the response was in the validation set and did not contribute to the inference. This resulted in a distribution of likelihoods for each response consisting of likelihood values for each sample of the posterior. We combined these distributions across all of an individual's responses, to obtain a distribution of the out-of-sample likelihood of their responses.

We performed the above cross-validation procedure in three different ways, which we will refer to as *self prior*, *other priors*, and *group prior*. The results are shown as three distributions for each individual in Fig. 9. For each individual, the first bar (self prior) shows the distribution of the likelihoods of that individual's out-of-sample responses, given their individual posteriors in the individual differences model. The second bar (other priors) shows the distribution of the likelihoods of that individual's out-of-sample responses, given all of the *other* individuals' posteriors, but excluding their own, in the individual differences model. The third bar (group prior) shows the distribution of the likelihoods of that individual's out-of-sample responses, given the group posteriors in the no individual differences model. The dotted lines show the joint likelihood of individuals' responses under the assumption that they used the environmental prior distributions.

The self prior condition had the highest score for six individuals, the group prior for 13 individuals, and the environmental prior for three individuals. The self prior condition

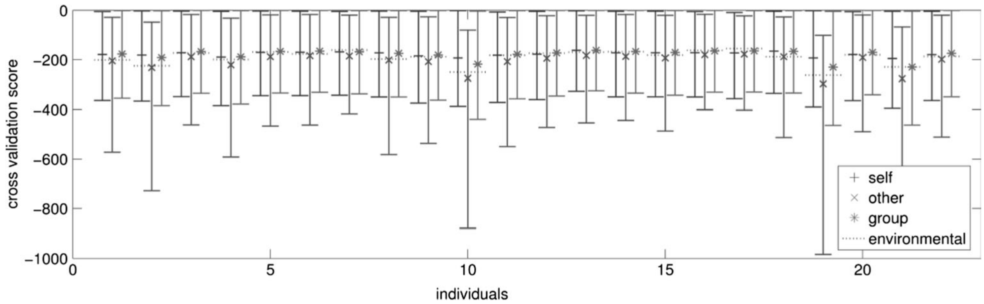

Fig. 9 Cross-validation results

had the best overall score − 177 when taking the average of all individuals' mean scores. The other conditions had scores of −179 (group prior), −190 (environmental prior), and −207 (other priors). The subjective prior models, both with individual and without individual differences, outperformed the environmental prior model on average score per individual, with at least one of the subjective models outperforming the environmental model for 19 of 22 individuals.

The additional flexibility of the individual differences model allows it to have overall better generalization to unobserved data from the same individual than the no-individual differences model. The poor performance of the *other priors* condition, including the finding that this condition had the lowest score for every individual, provides evidence that inferring an individual's prior provides information about that individual that leads to better predictions about their future behavior.

Although the *self prior* condition had a higher overall score than the *group prior* condition, it only had the best score for six of 22 individuals. This suggests that the inferred individual priors did not always provide information that went beyond the inferred group prior. However, the improved generalization for some of the individuals, in addition to the overall performance of the self condition, validates the idea that the individual differences model does lead to useful insights about some individuals' prior knowledge.

Discussion

The traditional approach to Bayesian cognitive modeling requires the experimenter to specify the observers' prior knowledge (usually by collecting normative data) in order to predict observer responses using the model. The majority of studies demonstrating the plausibility of BMCs use a qualitative approach to model fitting generally the available data is not used in an updating process to learn about latent model parameters or prior distributions. In this article we have illustrated the integrated observer-researcher analysis with a Bayesian model of memory. By applying Bayesian data analysis in a way that encapsulates the BMC we can infer posterior distributions for any model parameters and priors given the available data.

We first implemented a qualitative BMC of behavioral data on memory for the height of people. When assuming that the observers in the task use the environmental prior, the model produces results that are qualitatively consistent with the responses given by human observers. However, the results also suggest that some subjects might not use the true environmental prior. It is important to estimate the subjective prior, as this provides a better fit to the behavioral data. Extending the analysis with individual differences suggests that there are individual differences between observers' prior distributions for height, as well as for memory noise. For example, some observers appear to use a prior consistent with the environmental statistics whereas other observers appear to have subjective priors that systematically underestimate the environmental statistics.

A major advantage of our integrative Bayesian approach is that is also allows us to quantitatively evaluate and contrast models. We illustrated this using a cross-validation procedure in order to evaluate whether or not there are true individual differences. By inferring the BMC to a subset of the data, cross-validation allows us to assess how well the model generalizes to withheld data. We found that inferred priors outperformed environmental priors, which validates the idea that we have learned something useful about our subjects that leads to better predictions of unobserved responses. The out-of-sample likelihood of responses given the inferred BMC with individual priors was highest for six of 22 individuals. The individual difference BMC had the highest (out-of-sample) likelihood overall suggesting that individual differences might be important to include, at least in the memory domain we considered. While we have focused on cross-validation to compare BMCs, any model selection techniques could be applied as well (Pitt, Myung, & Zhang, 2002), including Bayes factors (e.g. Lodewyckx, et al., 2011), Savage–Dickey density ratio tests (Wagenmakers, et al., 2010), and DIC (Spiegelhalter, et al., 2002).

Our work illustrates the usefulness of going beyond qualitative BMC evaluations, and how reliance on average data can obscure important individual differences stemming from different psychological parameters within a model or even from different psychological models. Our integrated observer-researcher analysis provides a way to explore individual differences that can relate to process parameters that regulate internal cognitive processes as well as differences in the nature and use of prior knowledge relevant to the cognitive task.

The idea of applying Bayesian data analysis to cognitive models is, in itself, not unusual. There are several examples in the literature where cognitive models are analyzed with Bayesian data analysis. For example, Lee (2008) showed how three different cognitive models (MDS representation for stimulus similarity, the generalized context model of category learning, and a SDT account of deductive and inductive reasoning) can be evaluated using Bayesian inference. The goal of Bayesian data analysis, then, is to relate given models of psychological processes to observed behavior. This approach provides, among other things, a powerful tool for assessing individual differences within the cognitive model, and as such provides the natural extension to go beyond qualitative evaluations of the BMC. The novelty of our analysis is that we combine two different styles of Bayesian inference, from the viewpoint of the observer who is drawing rational inferences about data coming into the senses, and from the experimenter who receives data from observer and applies Bayesian inference to draw conclusions about latent parameters in the observer's mind.

While it could be argued that our approach of fitting priors to data makes the models too flexible, the cross-validation procedure clearly demonstrates that the inferred priors generalize better to unobserved data. Furthermore, we assume that the inferred priors must correspond to people's actual subjective knowledge, but what we ascribe to priors could actually reflect biases in perception or other processing stages. When this sort of criticism is presented against a standard BMC questioning its assumptions it is difficult to address one way or the other. This further elucidates one of the advantages of our approach: unlike with standard BMCs, the existence of the likelihood in our models allows them to be directly compared to competing models Bayesian or non-Bayesian. In this sense, the assumptions of our model are falsifiable. It is possible for another researcher can propose an alternative model with different assumptions that involve perception or other processing stages, and then compare their model to ours using standard model comparison methods. We welcome such a discussion, and believe this would make for would be great future research.

There are a number of extensions we can pursue in the integrative Bayesian approach. We assumed that people's responses were based on a sample from the posterior distribution. One direction for future research is to investigate alternative response processes from the BMC. The model could include parameters to allow for different response strategies such as probability matching (taking a single sample) and maximizing (taking the mode of the posterior distribution). One challenge is that the level of determinism in the response process will trade off with the uncertainty in subjective priors – an increase in the noise in the response process can be counteracted by an increase in the precision of the priors making it difficult to identify the subjective priors independent from response level parameters. A direction for research is to develop parametrizations of BMCs where a joint inference of response and prior parameters lead to meaningful interpretations.

Another course for future research is to better understand how we can interweave the two Bayesian inference schemes especially when inference at the observer level does not lead to an analytical solution. In this case, it might become challenging to do Bayesian inference at two levels simultaneously. For example, Huszár et al. (2010) point out that calculating the posterior over subjective distributions in their model is intractable. Daunizeau et al. (2010) also express concern that the optimal policy in experimental decision measures lacks a closed form solution. However, we plan to investigate approximate inference techniques such as Markov chain Monte Carlo simultaneously at the two levels such that posterior samples at the observer level become data at the experimenter level.

Lastly, another important direction is to expand the scope of models and also include non-Bayesian models in the model selection approach. At the researcher level, we could allow for the possibility that the observer is using a non-BMC (such as those illustrated in the Lee 2008, examples). For example, one could conceive of a simple mixture model allowing observers to use the BMC with an informative prior or with an uninformative prior. This mixture model could be extended to include other models as well, e.g., if we do not believe that a given observer is well fit by the Bayesian model but is instead better fit by a simple heuristic. It is also possible to mix different levels of analysis (Marr, 1982) within the model. For example, one can mix models proposed at the computational level with models proposed at the algorithmic level. Our overall goal is to develop a framework in which multiple types of models can be compared and investigated, thereby freeing the researcher from making a particular commitment to one type of model.

References

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

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138,* 389–414.

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget 2012. *Psychological Bulletin, 138,* 423–426.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., et al. (2010). Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLoS ONE, 5*(12).

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton: Chapman & Hall.

Hemmer, P., & Steyvers, M. (2009). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science, 1,* 189–202.

Huszár, F., Noppeney, U. & Lengyel, M. (2010). Mind reading by machine learning: A doubly Bayesian method for inferring mental representations. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* 2810-2815.

Huttenlocher, J., Hedges, L. V., & Vevea, J. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129,* 220–241.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition. *Behavioral and Brain Sciences, 34*(4), 169–88.

Kato, K., & Higashiyama, A. (1998). Estimation of height for persons in pictures. *Perception & Psychophysics, 60*, 1318–1328.

Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press: Elsevier.

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

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., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12*(4), 605–621.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology, 55,* 331–347.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of cognition. *Psychological Science, 24*(12), 2351–2360.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories. *Cognitive Science, 36,* 150–162.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

McDowell, M., Fryar, C. D., Ogden, C. L., & Flegal, K. M. (2008). Anthropometric Reference Data for Children and Adults: United States, 2003-2006. *National Health Statistics Reports, 10,* 1265–1272.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science, 32,* 1133–1147.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122.

Nelson, T. E., Biernat, M. R., & Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology, 59*(4), 664–675.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing.

Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds), *Advances in Neural Information Processing Systems, 20,* 369–376. Cambridge, MA: MIT Press.

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

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*(4), 583–639.

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*(7), 327–334.

Steyvers, M., & Griffiths, T. L. (2008). Rational Analysis as a Link between Human Memory and Information Retrieval. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects from Rational Models of Cognition* (pp. 327–347). Oxford: Oxford University Press.

Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53,* 168–179.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.

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

Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 33,* 1020–1034.

Vickers, D., Lee, M. D., Dry, M., Hughes, P., & McMahon, J. A. (2006). The aesthetic appeal of minimal structures: Judging the attractiveness of solutions to Traveling Salesperson problems. *Perception & Psychophysics, 68,* 32–42.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60,* 158–189.