Running head: BAYESIAN HIERARCHICAL MODELS

Bayesian Hierarchical Models

Jeffrey N. Rouder

University of Missouri

Richard D. Morey

University of Groningen

Michael S. Pratte

Vanderbilt University

Bayesian Hierarchical Models

Introduction: The need for hierarchical models

Those of us who study human cognition have no easy task. We try to understand how people functionally represent and processes information in performing cognitive activities such as vision, perception, memory, language, and decision making. Fortunately, experimental psychology has a rich theoretical tradition, and there is no shortage of insightful theoretical proposals. Also, it has a rich experimental tradition with a multitude of experimental techniques for isolating purported processes. What it lacks, however, is a rich statistical tradition to link theory to data. At the heart of the field is the difficult task of trying to use data from experiments to inform theory, that is, to understand accurately the relationships within the data and how they provide evidence for or against various theoretical positions.

The difficulty in linking data to theory can be seen in a classic example from Estes (1956). Estes considered two different theories of learning: one in which learning was gradual and another where learning happened all at once. These two accounts are shown in Figure 1A. Because these accounts are so different, adjudicating between them should be trivial: one simply examines the data for either a step function or a gradual change. Yet, in many cases, this task is surprisingly difficult. To see this difficulty, consider the data of Ritter and Reder (1992), who studied the speed up in response times from repeated practice of a mathematics tasks. The data are shown in Figure 1B, and the grey lines show the data from individuals. These individual data are highly variable making it impossible to spot trends. A first-order approach is to simply take the means across individuals at different levels of practice, and these means (red points) decrease gradually, seemingly providing support for the gradual theory of learning. Estes, however, noted that

this pattern does not necessarily imply that learning is gradual. Instead, learning might be all-at-once, but the time at which different individuals transition may be different. Figure 1C shows an example; for demonstration purposes, hypothetical data are shown without noise. If data are generated from the all-at-once model and there is variation in this transition time, then the mean will reflect the proportion of individuals in the unlearned state at a given level of practice. This proportion may decrease gradually, and consequently, the mean may decrease gradually even if every participant has a sharp transition from an unlearned state to a learned one. It is difficult to know whether the pattern of the means reflects a signature of cognitive processing or a signature of between-individual variation.

There are three critical elements in Estes' example: First, individuals' data are highly noisy, and this degree of noise necessitates combining data across people. Second, there is variability across individuals. For example, in the all-at-once model, people differ in their transition times. Finally, the theories themselves are nonlinear[1], and the all-at-once model in particular has a large degree of nonlinearity. It is the combination of these three factors—substantial variability within and across individuals that is analyzed with nonlinear models—that makes linking data to theory difficult. Unfortunately, the three elements that led to the difficulties in Estes' example are nearly ubiquitous in experimental psychology. Often data are too noisy to draw conclusions from consideration of single individuals; there is substantial variability across participants; and realistic models of cognition are nonlinear. Note that the problem of nuisance variation is not limited to individuals. In memory and language studies, for instance, there is nuisance variation across items. For instance, in the learning example, it is reasonable to expect that if the all-at-once model held, the time to transition across different problems (items) would vary as well.

Several psychologists have noted that drawing conclusions from aggregated data

may be tenuous. Estes' example in learning has been expanded upon by Haider & Frensch (2002), Heathcote, Brown, & Mewhort (2000), Myung, Kim, & Pitt (2000), and Rickard (2004). The worry about aggregation over individuals has also been expressed in the context of multidimensional scaling (Ashby, Maddox, & Lee, 1994), and the worry about aggregating over both individuals and items has been expressed in linguistics (Baayen, Tweedie, & Schreuder, 2002) and recognition memory (Rouder, Lu, et al., 2007; Pratte, Rouder, & Morey, 2010). Although the dangers of aggregation are widely known, researchers still routinely aggregate. For example, in studies of recognition memory, it is routine to aggregate data across individuals and items before fitting models. Even experienced researchers who fit sophisticated models to individuals routinely aggregate across some source of nuisance variation, typically items. The reason that researchers aggregate is simple—they do not know what else to do. Consider recognition memory tasks, where aggregation across items or individuals is seemingly necessary to form hit and false alarm rates. Without aggregation, the data for each item-by-individual combination is an unreplicated, dichotomous event. Our experience is that researchers would prefer to avoid aggregating data if alternatives are available.

In this chapter we present such an alternative: hierarchical modeling. In a hierarchical model, variability from the process of interest, as well as from nuisance sources such as from individuals and from items, are modeled simultaneously. The input to these models is the raw, unaggregated data, and the outputs are process-parameter estimates across individuals and items. In this regard, not only can the behavior of these process estimates be studied across conditions, but across individuals and items as well, and this later activity provides a process-model informed study of individual (and item) differences. Hence, hierarchical models turn a problem, how to account for nuisance variation that cloud our view of process, into a strength. Hierarchical models provide both a clearer view of process and a means of exploring how these processes vary across populations of

individuals or items. Not surprisingly, hierarchical linear models, models that extend ANOVA and regression to account for multiple sources of variance, are common in many areas of psychology as well as across the social sciences (Raudenbush & Bryk, 2002).

Although hierarchical linear models are suitable in several domains, they rarely make good models of psychological process. Instead, models that account for psychological processes are typically nonlinear. The appropriate extensions for these cases are *hierarchical nonlinear models.* It is difficult, however, to analyze hierarchical nonlinear models in conventional frameworks. As a result, the field has been moving toward Bayesian hierarchical models because hierarchical models, including hierarchical nonlinear models, are far more conveniently and straightforwardly analyzed in the Bayesian framework than in conventional ones. It is for this reason that there has been much recent development of Bayesian hierarchical models in the mathematical psychology community, the psychological community most concerned with developing models of psychological process. Recent examples of applications in psychologically substantive domains include Anders & Batchelder (2012); Averell & Heathcote (2011); Karabatsos & Batchelder (2003); Kemp, Perfors, & Tenenbaum (2007); Lee (2006); Farrell & Ludwig (2008); Merkle, Smithson, & Verkuilen (2011); Rouder, Morey, Cowan, & Pfaltz (2004); Rouder, Tuerlinckx, Speckman, Lu, & Gomez (2008); Vandekerckhove, Verheyen, & Tuerlinckx (2010) and Zeigenfuse & Lee (2010). Tutorial articles and chapters covering hierarchical cognitive process models are becoming numerous as well (e.g., Busemeyer & Diederich, 2009; Kruschke, 2011; Lee & Wagenmakers, 2013; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008), and there is a special issue of the Journal of Mathematical Psychology (January 2011, Vol 55:1) devoted to the topic.

In the next section, we cover the basics of Bayesian probability. Included is a comparison of the basic tenets of frequentist and Bayesian probability, examples of using data to update prior beliefs, and an introduction to Markov chain Monte Carlo sampling.

In Section 3, we show that the specification and analysis of hierarchical models is simple and natural with the Bayesian approach, and in Section 4 we provide a brief discussion of model comparison. Section 5 comprises our first example, and it is in the assessment of subliminal priming. Subliminal priming occurs when an undetectable stimulus nonetheless affects subsequent behavior. The methodological difficulty in establishing subliminal priming is proving that a set of participants cannot detect a stimulus at levels above chance. We show how previous approaches are woefully inadequate and demonstrate how a hierarchical approach provides a possible solution. We provide a second example of hierarchical modeling in Section 6. The example is from recognition memory, and shows how the estimation of parameters in Yonelinas' dual process model (Yonelinas, 1994) may be contaminated by aggregation bias. We develop a hierarchical model for uncontaminated assessment of the number of processes mediating recognition memory. Finally, in Section 7 we provide some advice on choosing computer packages and receiving training to perform Bayesian hierarchical modeling.

Bayesian Basics

In this paper we adopt a Bayesian rather than a conventional frequentist framework for analysis. One reason is pragmatic—the development of Bayesian hierarchical models is straightforward. Analysis of all Bayesian models, whether hierarchical or not, follows a common path. Bayesian techniques transfer seamlessly across different domains and models, providing a compact, unified approach to analysis. Because the Bayesian approach is unified, models that might be intractable in frequentist approaches become feasible with the Bayesian approach. The second reason we advocate Bayesian analysis is on philosophical grounds. The foundational tenets of Bayesian probability are clear, simple, appealing, and intellectually rigorous. In this section we review frequentist and Bayesian conceptualizations of probability. More detailed presentations may be found in

Bayesian textbooks such as Gelman, Shor, Bafumi, & Park (2007) and Jackman (2009).

Frequentist and Bayesian Conceptualizations of Probability

The frequentist conceptualization of probability is grounded in the Law of Large Numbers. Consider an event that may happen or not, and let Y be the number of occurrences out of N opportunities. The probability of an event is defined as the proportion when the number of opportunities is arbitrarily large; i.e.,

$$p = \lim_{N \to \infty} \frac{Y}{N}.$$

In this formulation, we may think of the probability as a physical property of the event. Consider, for example, the probability that a given coin results in a *heads* when flipped. This probability may be thought of as a physical property much like the coin's weight or chemical composition. And much like weight and chemical composition, the probability has an objective truth value even if we cannot measure it to arbitrary precision.

In both frequentist and Bayesian paradigms, useful models contain unknown parameters that must be estimated from data. For instance, if a participant performs N experimental trials on a task, we might model the resultant frequency of correct performance, Y, as a binomial random variable:

$$Y \sim \text{Binomial}(p, N),$$

where p serves as a parameter and denotes the probability of a correct response on a trial. Another simple, ubiquitous model is the normal. For example, Y might denote the mean response time of a participant in a task and be modeled as

$$Y \sim \text{Normal}(\mu, \sigma^2),$$

where μ and σ^2 serve as free parameters that denote the mean and variance of the distribution of people's mean response times. Although it is well known that response

times are not normals (Luce, 1986), the normal is a reasonable model of the distribution of mean RT across people. Consequently, the normal model is often useful for analyzing changes in mean RT as a function of experimental conditions or other covariates.

In the frequentist conceptualization, parameters are unknown fixed values which, in turn, are estimated from data. Because frequentist probability stresses the large-sample limit, the approach does not provide strict guidance on estimating parameters in finite samples sizes. Consequently, there are multiple approaches to estimation including finding estimates that maximize the likelihood (ML) or, alternatively, finding estimates that minimize squared error between predicted and observed data points (LS). These methods are not equivalent, and they may lead to different estimates in certain circumstances. For example, the ML estimator of σ^2 in the normal model is $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/N$ while the LS estimator is $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/(N-1)$. For frequentists, a minimal measure of acceptability of an estimator is its large-sample behavior. Principled estimators are *consistent*: they converge to true values in the large-sample limit. Both the ML and LS estimators of σ^2 are consistent because they converge to the true value as $N \to \infty$.

The Bayesian conceptualization of probability differs substantially from the frequentist one. Probabilities are statements of subjective belief held by observers about the occurrences of events. In the Bayesian formulation, probabilities describe the analyst's belief rather than a physical property of the system under study. Analysts may express their beliefs compactly as distributions. Figure 2A shows the beliefs of three analysts about a certain coin, or more specifically about p, the probability that a flip of a coin will result in heads rather than tails. Analyst I believes that all values of p are equally likely. This belief is shown by the solid flat line. Analyst II believes heads is more likely than tails, and this belief is shown by the dotted line. Analyst III believes not only that tails are more likely than heads, but that there is no chance whatsoever that the coin favors heads. This belief is shown by the dashed line. These beliefs are called *prior* beliefs,

because they are expressed before observing the data. After expressing these prior beliefs, the three analysts together flip the coin repeatedly and observe 8 heads in 12 flips. **The key tenet of Bayesian probability is that beliefs may be updated rationally in light of data.** To do so, one applies Bayes' Rule, which is discussed subsequently. The rationally updated belief distributions, called the *posterior* beliefs, are shown in Figure 2B. There are three posterior distributions, one for each analyst. There are a few noteworthy points: First, the beliefs of all three analysts have been narrowed by the data; in particular, for Analyst I, the beliefs have updated from a flat distribution to one that is centered near the proportion of heads and with narrowed variance. Second, even though the prior beliefs of Analysts I and Analyst II diverged markedly, the posterior beliefs are quite similar. Third, Analyst III had ruled out certain values, all those for $p > .50$ a priori. Indeed, because these have been ruled out, no result can make them probable, and the posterior has no density for $p > .50$.

In summary, Bayesian probability does not prescribe what beliefs analysts should hold. Instead, the emphasis is on how these beliefs should be updated in light of data. Posterior beliefs are still subjective even though they reflect data. For Bayesians, probabilities remain a construct of the observer rather than an objective property of the system, and this property holds regardless of how much data has been collected. However, because of the strong constraints imposed by Bayes' rule and their relationship to rational learning, Bayesian statistics offers a compelling, unified method for learning from data.

Bayes' Rule

The goal of Bayesian analysis is to update beliefs rationally with Bayes' Rule. Consider again the model of Y, the number of heads out of N coin flips, $Y \sim \text{Binomial}(p, N)$, where p is a free parameter. Bayes' Rule in this case is

$$\pi(p|Y) = \frac{Pr(Y|p)}{Pr(Y)}\pi(p). \tag{1}$$

The term $\pi(p|Y)$ is the posterior distribution of beliefs, that is, beliefs about the parameter conditional on the data. The term $\pi(p)$ is the prior distribution of beliefs. Three examples of prior and posterior beliefs are provided in Figure 2A and 2B, respectively. The term $Pr(Y|p)$ is the likelihood function and is derived from the model. For the binomial model, $Pr(Y|p) = \binom{N}{Y} p^Y (1-p)^{N-Y}$. The remaining term $Pr(Y)$, the probability of the data, may be re-expressed by the Law of Total Probability as

$$Pr(Y) = \int_0^1 Pr(Y|p)\pi(p)dp.$$

Fortunately, it is unnecessary to compute $Pr(Y)$ to express posterior beliefs. The distribution of posterior beliefs $\pi(p|Y)$ must be proper, that is, the area under the curve must be 1.0. The term $Pr(Y)$ is a normalizing constant on $\pi(p|Y)$ such that $\int_0^1 \pi(p|Y)dp = 1$. Often, the expression for this normalizing constant is obvious from the form of $Pr(Y|p)\pi(p)$ and need not be explicitly computed.

Let's use Bayes' Rule to express the posterior beliefs for the prior in Figure 2A for Analyst II. This prior is $\pi(p) = K_0 p(1-p)^3$, where K_0 is a constant that assures the prior integrates to 1.0. The data are 8 heads in 12 flips, and the likelihood $Pr(Y|p)$ is $\binom{12}{8} p^8 (1-p)^4$. Multiplying the prior and likelihood yields the following:

$$\pi(p|Y = 8) = K p^9 (1-p)^7,$$

where K is a constant of proportionality chosen such that $\int_0^1 \pi(p|Y = 8)dp = 1$. The dashed line in Figure 2B is the evaluation of $\pi(p|Y = 8)$ for all values of p.

Bayes' Rule is completely general, and may be extended to models with more than one parameter as follows: Let $\boldsymbol{Y}$ denote a vector of data which is assumed to be generated by some model $\mathcal{M}$ with a vector of parameters denoted by $\boldsymbol{\theta}$, i.e., $\boldsymbol{Y} \sim \mathcal{M}(\boldsymbol{\theta})$. Then

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) \propto Pr(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Once again, $Pr(\boldsymbol{Y}|\boldsymbol{\theta})$ is the likelihood function in this context, and Bayes' Rule may succinctly be stated as, "The posterior is proportional to the product of the likelihood and

prior." Bayesian updating, in contrast to frequentist parameter estimation, is highly constrained. There is only one Bayes' Rule, and it may be followed consistently without exception. One of the appeals of Bayesian updating is its conceptual simplicity and universal applicability.

The binomial model is useful for modeling dichotomous outcomes such as accuracy on a given trial. It is often useful to model continuous data as normally distributed. For example, suppose we wished to know the effects of "Smarties," a brand of candy, on IQ. Certain children have been known to implore their parents for Smarties with the claim that it assuredly makes them smarter. Let's assume for argument sake that we have fed Smarties to a randomly selected group of school children, and then measured their IQ. Let $\boldsymbol{Y} = (Y_i, \ldots, Y_n)$ be a vector that denotes the IQ of the children fed Smarties. We model these IQ scores as

$$Y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2),$$

where *iid* indicates that each observation is *independent* and *identically distributed*.

The goal is to derive posterior beliefs about μ and σ^2 given prior beliefs and the data themselves. For now, we focus on μ and, for simplicity, assume that σ^2 is known to equal the population variance of IQ scores, $\sigma^2 = 15^2 = 225$. In Section 2.3, we relax this assumption, and discuss how to update beliefs on multiple parameters simultaneously.

An application of Bayes' rule to update beliefs about μ yields

$$\pi(\mu|\boldsymbol{Y}) \propto L(\mu, \boldsymbol{Y})\pi(\mu),$$

where $\boldsymbol{Y}$ is the vector of observations and L is the likelihood function of μ. The likelihood for a sequence of independent and identically normally distributed observations is

$$L(\mu, \boldsymbol{Y}) = f(Y_1; \mu, \sigma^2) \times f(Y_2; \mu, \sigma^2) \times \ldots f(Y_n; \mu, \sigma^2) = \prod_i f(Y_i; \mu, \sigma^2)$$

where $f(x; \mu, \sigma^2)$ is the density function of a normal with mean μ and variance σ^2 evaluated at x.

A prior distribution on μ is needed. Consider prior beliefs to be distributed as normal:

$$\mu \sim \text{Normal}(a, b).$$

Constants a and b are the mean and variance of the prior, respectively, and must be chosen *a priori*. In this case, we consider two analysts with differing beliefs. Analyst I is doubtful that Smarties have any effect at all, and has chosen a tightly constrained prior with $a = 100$ and $b = 1$. Analyst II on the other hand, is far less committal in her beliefs, and chooses $\mu = 100$ and $b = 200$ to show this lack of commitment. These choices are shown in Figure 3A.

With this prior, the posterior beliefs may be expressed as

$$\pi(\mu|\boldsymbol{Y}) \propto \left(\prod_i f(Y_i; \mu, \sigma^2) \right) f(\mu; a, b).$$

The above equation may be expanded and simplified, and Rouder and Lu (2005) among many others show that

$$\pi(\mu|\boldsymbol{Y}) = f(cv, v),$$

where

$$c = \left(\frac{n\bar{Y}}{\sigma^2} + \frac{a}{b} \right), \tag{2}$$

$$v = \left(\frac{n}{\sigma^2} + \frac{1}{b} \right)^{-1}, \tag{3}$$

and n is the sample size and $\bar{Y}$ is the sample mean.

The posterior beliefs about μ follow a normal with mean cv and variance v, and this fact may equivalently be stated as

$$\mu|\boldsymbol{Y} \sim \text{Normal}(cv, v).$$

One property of the posterior is that it reflects information from both the prior and the data. Here, the posterior mean is a weighted average of the sample mean and the prior

mean, with the number of observations determining the relative weight. If there are few observations the prior has relatively higher weight than if there are many observations. A second property of the posterior is that it is the same functional form as the prior — both are normally distributed. When a prior and posterior have the same functional form, the prior is termed *conjugate*. Conjugate priors are desirable because they are computationally convenient. A third notable property of the posterior is that it may be well localized even if the prior is arbitrarily variable. The prior variance b reflects the certitude of prior information, with larger settings corresponding to less certitude. In fact, it is possible to set $b = \infty$, and the resulting prior may be called flat as all values of μ are equally plausible. This flat prior is *improper* – that is, it does not integrate to a finite value. Even though the prior is improper the posterior in this case is proper and is given by

$$\mu|\boldsymbol{Y} \sim \text{Normal}(\bar{Y}, \sigma^2/N).$$

For the flat prior, the posterior for μ corresponds to the frequentist sampling distribution of the mean.

Figures 3B and 3C show the role of sample size in posterior beliefs. Figure 3B shows the posterior beliefs of the two analysts for a very small set, $N = 10$, with a sample mean IQ score of $\bar{Y} = 95$. The data has slightly shifted and slightly widened the beliefs of Analyst I, the analyst who was *a priori* convinced there was little chance of an effect. It has more dramatically sharpened the beliefs of Analyst II, the less committed analyst. Figure 3C shows the case with a larger set, $N = 100$, and $\bar{Y} = 95$. Here the posterior beliefs are more similar because the data are sufficient in sample size to have a large effect relative to the prior. In the large-sample limit, these posterior distributions will converge to a point at the true value of μ.

Sampling: An Approach To Bayesian Analysis with more Than One Parameter

In the previous example, we modeled IQ scores as a normal under the assumption that σ^2 is known. Clearly such an assumption is too restrictive, and a more reasonable goal is to state posterior beliefs about both μ and σ^2, jointly. An application of Bayes' Rule yields

$$\pi(\mu, \sigma^2|\boldsymbol{Y}) \propto L(\mu, \sigma^2, \boldsymbol{Y})\pi(\mu, \sigma^2).$$

The prior density, posterior density, and likelihood functions in this case are evaluated on a plane and take as inputs ordered pairs. Examples of a prior, likelihood, and posterior are shown in Figure 4 as two-dimemsional surfaces. Because the posterior and prior in the above equation are functions of μ and σ^2 taken jointly, they are referred to as the *joint posterior* and the *joint prior*, respectively. Fortunately, deriving joint posteriors is straightforward as it is simply the result of Bayes' Rule: the posterior is the product of the likelihood and the prior.

Expressing joint posterior beliefs as surfaces may be reasonable for models with two dimensions, but becomes unwieldy as the dimensionality increases. For instance, in models with separate parameters for individuals and items, it is not uncommon to have thousands of parameters. The expression of joint posterior distributions over high dimensional parameter vectors is not helpful. Instead, it is helpful to plot *marginal posteriors*. The marginal posterior for one parameter, say μ, is denoted $\pi(\mu|\boldsymbol{Y})$, and is obtained by averaging (integrating) the uncertainty in all other parameters. The two marginal posteriors for this model are

$$\begin{aligned} f(\mu|\boldsymbol{Y}) &= \int_{\sigma^2} f(\mu, \sigma^2|\boldsymbol{Y})d\sigma^2 \\ f(\sigma^2|\boldsymbol{Y}) &= \int_{\mu} f(\mu, \sigma^2|\boldsymbol{Y})d\mu \end{aligned}$$

Marginal posteriors for the two parameters are shown in Figure 4, right panel. As can be seen, these provide a convenient expression of posterior beliefs.

Although marginal posteriors are useful for expressing posterior beliefs, they may be difficult to compute. In the two-parameter model, above, the computation was straightforward because the integration was over a single dimension and could be solved numerically. In typical models, however, there may be hundreds or thousands of parameters. To express each marginal, all other parameters must be integrated out, and the resulting integrals span hundreds or even thousands of dimensions. This problem of high dimensional integration was a major pragmatic barrier for the adoption of Bayesian methods until the 1980s, when new computational methods became feasible on low-cost computers.

A modern approach to the integration problem is sampling from the joint posterior distribution. We draw as many samples from the joint that is needed to characterize it to arbitrary precision. Each of these samples is a vector that has the dimensionality of the joint distribution. To characterize the marginal for any parameter, the corresponding element in the joint sample is retained. For example, if $(\mu, \sigma^2)^{[m]}$ is the mth sample from the joint, then the value of μ, which we denote as $\mu^{[m]}$, is a sample from the marginal posterior distribution of μ, and the collection $\mu^{[1]}, \mu^{[2]}, \ldots$ characterize this distribution to arbitrary precision. So integrating the joint posterior may be reduced to sampling from it.

Directly sampling from high-dimensional distributions is often difficult. To mitigate this difficulty, alternative indirect algorithms have been devised. The most popular class of these algorithms is called Markov chain Monte Carlo (MCMC) sampling. These techniques are covered in depth in many textbooks. (e.g., Jackman, 2009). Here, we cover the briefest outline. Those readers familiar with MCMC, or those who have no desire to learn about it may skip this outline without loss as the remainder of the chapter does not rely on understanding MCMC.

We focus here on the most common MCMC algorithm, the Gibbs sampler (Gelfand & Smith, 1990; Geman & Geman, 1984). When building a Gibbs sampler, researchers

focus on *conditional posterior distributions*. The conditional posterior distributions are the beliefs about one parameter if all others were known. For the normal model, there are two full conditional posteriors denoted $f(\mu|\sigma^2, \boldsymbol{Y})$ and $f(\sigma^2|\mu, \boldsymbol{Y})$. These are easily derived from an application of Bayes' rule:

$$\begin{aligned} \pi(\mu|\sigma^2, \boldsymbol{Y}) &= L(\mu, \sigma^2, \boldsymbol{Y})\pi(\mu|\sigma^2), \\ \pi(\sigma^2|\mu, \boldsymbol{Y}) &= L(\mu, \sigma^2, \boldsymbol{Y})\pi(\sigma^2|\mu). \end{aligned}$$

If the priors are independent of one another,

$$\begin{aligned} \pi(\mu|\sigma^2, \boldsymbol{Y}) &= L(\mu, \sigma^2, \boldsymbol{Y})\pi(\mu), \\ \pi(\sigma^2|\mu, \boldsymbol{Y}) &= L(\mu, \sigma^2, \boldsymbol{Y})\pi(\sigma^2). \end{aligned}$$

The reason researchers focus on the conditionals is that it is straightforward to analytically express these distributions. Moreover, and more importantly, it is often straightforward to sample from conditionals, which is the key to Gibbs sampling. For the normal-distribution case above, we denote samples of μ as $\mu^{[1]}|\sigma^2$, $\mu^{[2]}|\sigma^2$,$\ldots$,$\mu^{[M]}|\sigma^2$, where M is the total number of samples. Likewise, the samples of the conditional posterior distribution of σ^2 may be denoted $(\sigma^2)^{[1]}|\mu, \ldots, (\sigma^2)^{[M]}|\mu$. These samples, however, are conditional on particular values of μ and σ^2, and, consequently, are not so interesting.

The goal is to obtain marginal samples of μ and σ^2, rather than conditional ones. In our specific case, this goal may be achieved as follows: On the mth iteration, μ is sampled conditional on the previous value of σ^2, i.e., $\mu^{[m]}|(\sigma^2)^{[m-1]}$; then σ^2 is sampled conditional on the just-sampled value of μ, i.e., $(\sigma^2)^{[m]}|\mu^{[m-1]}$. In this manner, the samples are being conditioned on different values on every iteration, and if conditioning is done this way, the joint distribution of the samples approaches the true joint posterior as the number of samples grows infinitely large. If we have samples from the joint distribution, characterizing any marginal distribution is as easy as ignoring samples of all other

parameters. Researchers new to Bayesian analysis can use modern tools such as JAGS (Plummer, 2003) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) to perform MCMC sampling without much special knowledge.[2] Those with more experience can write their own code in high-level languages such as R or Matlab. We discuss these options further in the concluding remarks.

Bayesian Hierarchical Models Are Simple and Natural

There are several advantages of adopting a Bayesian perspective, and one of the most salient for cognitive modelers is the ease of building hierarchical models that may account for variation in real-world settings. Consider the following simple experiment where I individuals provide K replications in each of J conditions. To demonstrate the elegance and power of hierarchical modeling, we build a sequence of models, illustrating each with reference to an experiment where $I = 20$ individuals provided $K = 10$ replications in each of $J = 2$ conditions. Figure 5A shows the overall mean for each of these conditions (bars) as well as the participant-by-condition means (points and lines). As can be seen there is much participant variability as well as strong evidence for a condition effect.

Model $\mathcal{M}_1$: An Aggregation Model. One approach, which corresponds to aggregating, is to simply model condition means. Let Y_{ijk} denote the kth observation for the ith participant in the jth condition. The model is

$$Y_{ijk} \overset{iid}{\sim} \text{Normal}(\beta_j, \sigma^2). \tag{4}$$

where β_j is the condition effect. A prior is needed for each β_j and for σ^2, and we chose priors that makes no practical commitment to the location of these effects:

$$\begin{aligned} \beta_j &\overset{iid}{\sim} \text{Normal}(0, 10^6) \\ \sigma &\sim \text{Uniform}(0, 100) \end{aligned}$$

The model is designed to assess condition means, and the condition effect may be defined as the contrast $\beta_2 - \beta_1$.

We provide here the BUGS language code snippets for analysis of this and subsequent models, and these snippets may be used with WInBUGS, OpenBUGS, or JAGS. Those researchers who are not familiar with these packages will benefit from the well-written documentation (see Footnote 2) as well as the tutorials provided in Kruschke (2011) Ntzoufras (2009) and Lee & Wagenmakers (forthcoming book in press). The following `model` statement defines Model $\mathcal{M}_1$:

```
model {
#y is a vector of all observations
#cond is a vector that indicates the condition
#mu is a vector of J condition means

        # Model of Observations
        for (n in 1:N) {
                y[n] ~ dnorm(mu[cond[n], tau)
        }
        # note: BUGS uses precision to parameterize normal
        # note: tau is precision

        #Prior on mu
             for (j in 1:J){
             mu ~ dnorm(0, .0001)}

        #Prior on precision (std. dev.)
        tau <- pow(sigma, -2)
```

```
        sigma ~ dunif(0, 100)
}
```

Posterior beliefs about this contrast are shown as the dotted line in Figure 5B.[3] This model is not hierarchical as there is a single source of variability.

Model $\mathcal{M}_2$: A Cell Means Model. A more useful approach is to model the combination of participant and condition effects:

$$Y_{ijk} \overset{iid}{\sim} \text{Normal}(\mu_{ij}, \sigma^2). \tag{5}$$

The parameters μ_{ij} are the mean of the ith participant in the jth condition. In the example with 2 conditions and 20 participants, there are 40 of these effects, and each needs a prior. Again, we choose diffuse priors:

$$\begin{aligned} \mu_{ij} &\overset{iid}{\sim} \text{Normal}(0, 10^6) \\ \sigma &\sim \text{Uniform}(0, 100) \end{aligned}$$

The BUGS language snippet that defines this model is

```
model{
#y is a vector of all N observations
#sub is a vector that indicates the participant
#cond is a vector that indicates the condition
#mu is an I-by-J matrix

    #Model of observations
    for(n in 1:N){
            y[n] ~ dnorm(mu[sub[n],cond[n]], tau)
          }
```

```
    #Prior on mu
    for (i in 1:I){
    for (j in 1:J){
            mu[i,j] ~ dnorm(0, .01)
          }}

    #Prior on precision (std. dev.)
    tau <- pow(sigma, -2)
    sigma ~ dunif(0, 100)
}
```

The posterior means for the cell mean parameters are shown in Figure 5C as solid lines. As can be seen, participants often have higher mean scores in Condition 2 than in Condition 1, providing evidence for the condition effect. We can construct a contrast for this comparison: $\sum_i(\mu_{i2} - \mu_{i1})/I$, and the posterior for this contrast is shown as the solid line in Figure 5B.[4] Note that this posterior is better localized than the comparable contrast from Model $\mathcal{M}_1$. The reason is simple: individual variation is subtracted off leading to better parameter localization. It should be noted that these posterior beliefs, however, do not generalize to new participants. The reason is that people-by-condition effects are "fixed" in that they may vary arbitrarily and provide no information about a population of people, conditions, or their combination.

Model $\mathcal{M}_3$: A First Hierarchical Model. Although the interpretation of the cell means model is familiar and reasonable, we can make even more useful models. We start with the same data model:

$$Y_{ijk} \overset{iid}{\sim} \text{Normal}(\mu_{ij}, \sigma^2).$$

In the previous model the priors on μ_{ij} were very diffuse. Yet, it is unreasonable to think

that these cell mean parameters will arbitrarily differ from one another. For example, if we were studying IQ, it is hard to believe that participant-by-condition means vary by even a factor of two, much less orders of magnitude. One way of adding information without undue influence is through a hierarchical prior. Consider the prior

$$\mu_{ij} \overset{iid}{\sim} \text{Normal}(\nu, \delta^2) \tag{6}$$

where ν and δ describe the center and dispersion of the population of cell means. These values need not be fixed *a priori*. Instead, they may be treated as parameters upon which we may place priors and compute posteriors. Consider the following priors:

$$\begin{aligned} \nu &\sim \text{Normal}(0, 10^6) \\ \delta &\sim \text{Uniform}(0, 100) \end{aligned}$$

Here, we bring little if any *a priori* information about the population center and dispersion of effects. All we commit to is that the effects themselves are samples from this parent distribution. Hierarchical models are therefore implemented as hierarchical priors. Of course, a prior is still needed on σ^2, and we again use a diffuse prior:

$$\sigma \sim \text{Uniform}(0{,}100).$$

The hierarchical nature of model $\mathcal{M}_3$ is embedded in the relationships between parameters. The data Y_{ijk} are only explicitly dependent on the mean μ_{ij} and variance σ^2. If we know these two parameters, then the population from which Y_{ijk} is drawn is completely determined. Conversely, having observed Y_{ijk}, we constrain our beliefs about the parameters governing this population distribution. The hierarchy in $\mathcal{M}_3$ reflects the treatment of the collection μ parameters. These parameters are also treated as draws from a population. If we could observe the μ parameters directly, we could learn about ν, which is a parent parameter for this population of mean parameters. However, ν is one step removed from the data: we can only learn about ν through learning about the μ

parameters. Bayes' rule, the unifying rule for Bayesian inference, gives us a natural way of representing the way the information passes from level to level though the simple fact from probability theory that $p(a, b) = p(a|b)p(b)$. The posterior $p(\boldsymbol{\mu}, \nu|\boldsymbol{Y})$ is then proportional to

$$p(\boldsymbol{\mu}, \nu|\boldsymbol{Y}) \propto (\boldsymbol{Y}|\boldsymbol{\mu}, \nu)p(\boldsymbol{\mu}, \nu) = p(\boldsymbol{Y}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\nu)p(\nu)$$

(the parameters σ^2 and δ are assumed known for clarity). The right-hand side of the equation shows how knowledge about parameters is passed up through the hierarchy from the data to the higher-level parameters: the data $\boldsymbol{Y}$ and parameter $\boldsymbol{\mu}$ are connected through the likelihood, and $\boldsymbol{\mu}$ and ν are connected through the hierarchical prior on $\boldsymbol{\mu}$. Likewise, constraint from ν is passed down through the hierarchy from higher-level level parameters to the lower-level ones.

Figure 5D shows the effects of the constraint passed from the higher-level parameters. As can be seen, extreme cell mean values for this hierarchical model are somewhat moderated; that is, they are modestly pulled toward the population mean. This effect is often termed *hierarchical shrinkage*, and it leads to posterior estimates that have lower root-mean-squared error than nonhierarchical estimates. The effect here is modest because the data were generated with low noise for demonstration, but shrinkage can be especially pronounced in nonlinear models.

The use of hierarchical models has an element that is counterintuitive: one adds parameters to the prior to add constraint. In most models, adding parameters is adding flexibility, and more parameters implies a more flexible account of data. In hierarchical models, the opposite may hold when additional parameters are added to the prior. For instance, the cell means model has 40 cell mean parameters and a variance parameter; the hierarchical model has these 41 parameters and additional population mean and variance parameters. Yet, the cell means model is more flexible as the 40 cell mean parameters are free to vary arbitrarily. In the hierarchical model, no one cell mean can stray arbitrarily

from the others, and this behavior is a form of constraint even though it comes with more rather than less parameters. In Bayesian hierarchical modeling, flexibility is not a matter of the number of parameters, it is, instead, a matter of constraint or the lack there of in the priors. Principled Bayesian model comparison methods such as Bayes factors capitalize on this fact.

In addition to more accurate estimation of individual effects through shrinkage, hierarchical models offer two other benefits. First, posterior beliefs about group parameters, ν and δ^2 in the above example, can be generalized to other participant-by-condition combinations. These parameters, therefore, provide a means of applying the results more broadly. Second, more advanced models may be placed on ν that incorporate covariates.

Hierarchical models are straightforward to code in BUGS:

```
model{
#y is a vector of all N observations
#sub is a vector that indicates the participant
#cond is a vector that indicates the condition
#mu is an I-by-J matrix

    #Model of Observations
    for(n in 1:N){
            y[n] ~ dnorm(mu[sub[n],cond[n]], tau)
           }

    #Level 1: Prior on mu
    for (i in 1:I){
    for (j in 1:J){
```

```
            mu[i,j] ~ dnorm(nu,tauI)
          }}

  #Level 1: Prior on precision (std. dev.)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 100)

  #Level 2: Prior on nu, tauI
  nu~dnorm(0,.000001)
  tauI <- pow(delI, -2)
  delI ~ dunif(0, 100)

}
```

Model $\mathcal{M}_4$: A Hierarchical Model with Main Effects and Interactions. The shrinkage in Model $\mathcal{M}_3$ shrinks estimates toward the overall mean. Yet, there is clearly structure from participants and items. We add this structure into the prior as follows:

$$\mu_{ij} \overset{iid}{\sim} \text{Normal}(\alpha_i + \beta_j, \delta^2)$$

Priors are then needed for α_i, the effect of the ith participant, β_j, the effect of the jth condition, as well as δ^2 which now describes the variability of participant-by-condition

interactions. We use the following vaguely-informative priors:

$$\begin{aligned} \alpha_i &\sim \text{Normal}(0, \delta_\alpha^2) \\ \beta_j &= \text{Normal}(0, 10^6) \\ \delta &\sim \text{Uniform}(0, 100) \\ \delta_\alpha &\sim \text{Uniform}(0, 100) \end{aligned}$$

This new model treats participant effects as random effects drawn from a population distribution. Generalization to new people is possible through the inclusion of population variability parameter δ_α^2. This model also treats conditions as fixed effects, that is, conditions may differ from each other without constraint. Here, there is no concept of a population of conditions and, consequently, the results apply only to these two conditions. Finally, the interaction term reflects an asymptotic interaction between people and conditions; that is, it is the interaction that remains even in the limit that the number of replicates, K, increases without bound. We include this term hesitantly, because if it is too large, it is difficult to interpret the participant and condition effects. In these cases, we recommend that this interaction become more a target of inquiry, and models of patterned interactions be proposed and compared.

Even though this model is even more heavily parameterized than the previous hierarchical model, it is straightforward to estimate with the following BUGS snippet:

```
model{

    #Model of observations
    for(n in 1:N){
            y[n] ~ dnorm(mu[sub[n],cond[n]], tau)}

    #Level 1: Prior on mu
```

```
    for (i in 1:I){
    for (j in 1:J){
            mu[i,j] ~ dnorm(alpha[i]+beta[j],tauI)
          }}

    #Level 1: Prior on tau
    tau <- pow(sigma, -2)
    sigma ~ dunif(0, 100)

    #Level 2: Prior on alpha, beta
    for (i in 1:I){ alpha[i]~dnorm(0,tauA)}
    for (j in 1:J){ beta[j]~dnorm(100,.001)}

    #Level 2: Prior on tauI, scale of interactions.
    tauI <- pow(delI, -2)
    delI ~ dunif(0, 100)

    #Level 3: Prior on tauA, variability of individuals
    tauA <- pow(delA, -2)
    delA ~ dunif(0, 100)
}
```

The resulting values for the cell means, which are now treated hierarchically, are shown as dotted lines in Figure 5C. Notice that these are smoothed versions of the cell-means models. The shrinkage to main effects has smoothed away the interaction, making it easy to interpret the condition and participant effects. In fact, in this model,

the standard-deviation of these interactions ($\delta \approx 1.2$) is considerably less than the standard deviation of participant effects ($\delta_\alpha \approx 9.9$) or the difference between condition effects (≈ 4.0). The posterior beliefs about the condition effect is shown as the dashed line in Figure 5B.

Model $\mathcal{M}_5$: A Hierarchical Main-Effects Model. In many cases, it is desirable to remove the asymptotic interaction terms from the models. Not only do these make interpretation difficult, they may be unidentifiable, and this is certainly the case when there is a single replicate per cell ($K = 1$). Instead of modeling μ_{ij} as a random variable, we assume it is a deterministic sum:

$$\mu_{ij} = \alpha_i + \beta_j.$$

Missing is the parameter δ, which effectively is set to zero. The prior on the main effects is retained. This additive approach is taken in both applications in Sections 4 and 5. The following snippet is used for analysis:

```
model{

    #Model of observations
    for(n in 1:N){
            y[n] ~ dnorm(alpha[sub[n]]+beta[cond[n]], tau)}

    #Level 1: Prior on alpha and beta
    for (i in 1:I){ alpha[i]~dnorm(0,tauA)}
    for (j in 1:J){ beta[j]~dnorm(100,.001)}

    #Level 1: Prior on tau
    tau <- pow(sigma, -2)
```

```
    sigma ~ dunif(0, 100)


    #Level 2: Prior on tauA, variability across participants
    tauA <- pow(delA, -2)
    delA ~ dunif(0, 100)
}
```

Analysis of $\mathcal{M}_5$ results in essentially the same posterior beliefs as Model $\mathcal{M}_4$ for main effects of people (α_i), conditions (β_j) and their combinations (μ_{ij}). These results are omitted for clarity. Deciding whether asymptotic interactions are needed is not easy, and the topic of model comparison is discussed next.

Comparing Hierarchical Models

In most applications, it is useful to define a set of models and compare the relative evidence from the data for each. In the above case, for example, we might assess the evidence for interactions between people and condition by considering the relative evidence for Model $\mathcal{M}_4$, the model with interactions, and $\mathcal{M}_5$, the model without interactions. The condition main effect may be assessed if we compare Model $\mathcal{M}_5$ to a model without condition effects, specified by the constraint $\mu_{ij} = \alpha_i$. The critical question is how the relative evidence for such models may be stated.

Model comparison is a broad and expansive topic about which there is substantial diversity in the statistical and psychological communities. The chapter by Myung in this volume provides an overview of some of this diversity. Even though there is much diversity, we believe one model comparison method, comparison by Bayes factor (Jeffreys, 1961), is superior because it (a) directly provides a measure of evidence for models, and (b) is the unique, logical resultant of applying Bayes' Rule to model comparison. Although Bayes factors are ideal, they are associated with two issues. First, the Bayes

factor is sometimes difficult to compute, especially in hierarchical settings. Second, the Bayes factor is integrally dependent on the prior. We and others have argued that this dependency is necessary for valid model comparison (Gallistel, 2009; Jeffreys, 1961; Lindley, 1957; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Rouder, Morey, Verhagen, Province, & Wagenmakers, n.d.; Wagenmakers, 2007). Nonetheless, it often is not obvious how to structure priors to compare different nonlinear accounts of the same phenomena. In the following, (i) we define Bayes factors; (ii) discuss some of the difficulties in implementation including computational difficulties; (iii) mention some of the methods of circumventing these difficulties; and (iv) describe an alternative, Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002), a less desirable but more computationally feasible approach that may be used as a last resort.

The Bayesian interpretation of probability as subjective belief licenses the placing of probabilities (beliefs) on models themselves. To compare two models, denoted generically as $\mathcal{M}_A$ and $\mathcal{M}_B$, we may place the model probabilities in ratio. The ratios $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$ and $Pr(\mathcal{M}_A \mid \boldsymbol{Y})/Pr(\mathcal{M}_B \mid \boldsymbol{Y})$ are the prior and posterior odds of the models, respectively. Bayes' Rule for updating these prior odds is

$$\frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)} = \frac{f(\boldsymbol{Y} \mid \mathcal{M}_A)}{f(\boldsymbol{Y} \mid \mathcal{M}_B)} \times \frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)}. \tag{7}$$

The term $f(\boldsymbol{Y} \mid \mathcal{M}_A)/f(\boldsymbol{Y} \mid \mathcal{M}_B)$ is called the *Bayes factor*, and it describes the updating factor from the data (Kass & Raftery, 1995). We denote the Bayes factor by B_{AB}, where the subscripts indicate which two models are being compared. The term $f(\boldsymbol{Y} \mid \mathcal{M}_A)$ may be expressed as:

$$f(\boldsymbol{Y} \mid \mathcal{M}_A) = \int_{\boldsymbol{\theta} \in \Theta_A} L_A(\boldsymbol{\theta}, \boldsymbol{Y}) \pi_A(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{8}$$

where L_A is the likelihood of $\mathcal{M}_A$, and $\boldsymbol{\theta}$ and Θ_A are the parameters and parameter space, respectively, of the model. This term is called *the marginal likelihood*, and it is the weighted average of the likelihood over all possible parameter values. We use m_A and m_B

to denote the marginal likelihoods of models $\mathcal{M}_A$ and $\mathcal{M}_B$, respectively. The Bayes factor is

$$B_{AB} = \frac{m_A}{m_B}$$

A Bayes factor of $B_{AB} = 10$ means that prior odds should be updated by a factor of 10 in favor of model $\mathcal{M}_A$; likewise, a Bayes factor of $B_{AB} = .1$ means that prior odds should be updated by a factor of 10 in favor of model $\mathcal{M}_B$. Bayes factors of $B_{AB} = \infty$ and $B_{AB} = 0$ correspond to infinite support of one model over the other, with the former indicating infinite support for model $\mathcal{M}_A$ and the later indicating infinite support for model $\mathcal{M}_B$.

Bayes factors provide a principled method of inference, and advocacy in psychology is provided by Edwards, Lindman, & Savage (1963); Gallistel (2009); Myung & Pitt (1997); R. D. Morey & Rouder (2011); Rouder et al. (2009); Wagenmakers (2007) among others. The are two critical issues in use: (i) the choice of priors π_A and π_B, and (ii) the evaluation of the integrals in (8), and we discussed these issues in term. First, the choice of priors: The choice of priors will vary from situation to situation. In the case of linear models, such as those underlying t-test, linear regression, and ANOVA, researchers already know the range of plausible effect sizes. For instance, rarely do effect sizes exceed 5 or 10 in value, and we do not run experiments to search for effect sizes less than say .05 in value. These constraints may be capitalized upon to form reasonable and broadly acceptable priors for comparisons within the linear model (see Rouder et al., 2012; Rouder & Morey, 2012). The case for nonlinear models, however, is neither as straightforward nor as well explored. It is an area for future work.

The second issue is the evaluation of the integral in computing the marginal likelihoods in (8). Here, the parameter space is often of high dimension, especially in hierarchical models where there are several parameters for each participant and item. For example, in the subsequent recognition memory example in Section 6, there are over 2000 parameters. To compute the Bayes factor, the likelihood must be integrated across the

whole of the parameter space, and integration across high dimensional spaces is in general challenging. To make matters worse, the likelihood is often concentrated, and the integrand is highly peaked. The integration often becomes a matter of finding a needle in a multidimensional haystack. The topic of computing Bayes factors, especially in hierarchical models, remains topical in Bayesian analysis. Fortunately, Bayes factor computations for linear models underlying the t-test, ANOVA, and regression are well established. Seminal work was provided by Jeffreys (1961) and Zellner and Siow (1980). The key innovation from Zellner and Siow was specifying the problem in a manner so that the integration over most dimensions could be done analytically in closed form. The modern implementation of this work is provided among several others by Bayarri & Garcia Donato (2007) and Liang, Paulo, Molina, Clyde, and Berger (2008). Our group has translated and refined this approach, and we provide Bayes factor replacements for t-tests (Rouder et al., 2009), statistical-equivalence tests (Morey & Rouder, 2011), linear regression (Rouder & Morey, 2012), and ANOVA (Rouder, Morey, Speckman, & Province, 2012). We have also provided development of meta-analytic Bayes factors so researchers can assess the totality of evidence across several experiments (Rouder & Morey, 2011; Rouder, Morey, & Province, 2013).

Although this Bayes factor development covers a majority of statistical models used in psychology, current computational development does not cover a bulk of the psychological process model which tend to be nonlinear. There are a handful of advanced techniques that are potentially applicable, and we mention them in passing. Perhaps the most relevant is the *Laplace approximation*, where the likelihood is assumed to approach its asymptotic normal limits, and its center and spread are well approximated by classical statistical theory. Sarbanés Bové & Held (2011) use the Laplace approximation to provide a general Bayes factor solution for the class of generalized linear models. An alternative technique is to perform the integration by Monte Carlo sampling, and there has been

progress in a number of sampling based techniques including bridge sampling (Meng & Wong, 1996), importance sampling (Doucet, de Freitas, & Gordan, 2001), and a new variation on importance sampling termed direct sampling (Walker, Laud, Zanterdeschi, & Damien, 2011). These techniques assuredly will prove useful for future Bayes factor development in psychology. The final advanced technique in our survey is Bayes factor computation by means of Savage-Dickey density ratio (Dickey & Lientz, 1970; Verdinelli & Wasserman, 1995) which has been imported into psychology by C. C. Morey, Cowan, Morey, & Rouder (2011); Wagenmakers, Lodewyckx, Kuriyal, & Grasman (2010); Wetzels, Grasman, & Wagenmakers (2010). Under appropriate circumstances, this ratio is the Bayes factor and is convenient to calculate (see Morey, Rouder, Pratte, and Speckman, 2011). Wagenmakers et al. (2010) and Rouder et al. (2012) show how the Savage Dickey ratio can be used in the comparison of hierarchical models of psychological process, and Rouder et al. uses it to discriminate between the power law and exponential law of learning in hierarchical settings.

Even though there has been notable progress in developing Bayes factor solutions, there are several cases without such development, and, at present, Bayes factors are simply not available. For these cases, we have a backup, inference by *deviance information criterion* (DIC, Spiegelhalter et al., 2002). DIC is a Bayesian analog to AIC designed for hierarchical models. Unlike AIC (and BIC), DIC accounts for the flexibility of priors, and penalizes models with more flexible priors more heavily than those with more constrained priors. Such behavior is useful for hierarchical models where increased prior constraint is often accompanied by an increased number of parameters. The main advantage of DIC is computational ease; it is often computed in the same MCMC chain used to compute posterior beliefs about the parameters. The disadvantage is one of principle and calibration. DIC shares a calibration with AIC, and like AIC, tends to penalize flexibility too lightly (Rouder et al., 2009), especially for large sample sizes. The argument in favor

of BIC over AIC by Raftery can be applied to favor Bayes factor over DIC. Unlike Bayes factor, which is a principled direct result of Bayes' Rule, DIC is best viewed as a heuristic motivated by out-of-sample concerns. We use DIC only as a matter of last resort, and recommend Bayes factors be used without qualification when they are available.

Hierarchical Models For Assessing Subliminality

It is widely believed that a large portion of human cognition is unconscious (A. G. Greenwald, 1992). This unconscious cognition can manifest itself in many ways: for instance, we may have unconscious goals and motivations; we may be unaware of the effects of stimuli on these goals and motivations; we may even perceive and be affected by stimuli of which we are unaware. One example of this last category is the popular myth of subliminal advertisements in movie theaters: advertisement images were purportedly presented so quickly as to be consciously imperceptible, nonetheless these images supposedly changed the subsequent behavior of movie-goers by causing them to buy expensive snacks. This myth has been debunked (Rogers, 1992).

The fact that subliminal advertising was debunked does not mean that under controlled circumstances psychologists could not observe similar (if smaller) effects. In fact, many such claims have been made with demonstrations of subliminal priming (for examples, see Dehaene et al., 1998; Finkbeiner, 2011; A. Greenwald, Klinger, & Schuh, 1995; Merikle, Smilek, & Eastwood, 2001; Naccache & Dehaene, 2001). A subliminal prime is one that cannot be perceived, and yet has an effect on subsequent behavior. To answer the question of whether subliminal priming exists, one needs to show both that a prime cannot be identified at a rate greater than chance, and that this prime nonetheless affects behavior.

The priming task we model is a numerosity decision task. Participants are shown target numerals between 2 and 8 and judge whether the target is greater than or less than

5 in value. Preceding these targets are quickly-presented-and-subsequently masked prime numerals. When the prime has the same status as the target, that is, both are less than five or both are greater than five, responses are known to be speeded relative to the case where the prime and target do not have the same relation to five (e.g., Dehaene et al., 1998; Naccache & Dehaene, 2001; Koechlin, Naccache, Block, & Dehaene, 1999; Pratte & Rouder, 2009). The critical question is whether this priming persists even for presentations that are so fast that participants' ability to assess the prime's relation to five is at chance level.

We focus here on the difficult part of assessing subliminal priming: the assessment of whether a prime is identified at chance or above chance. Let p_i denote the true probability correct for the ith participant. Primes are subliminal for the ith participant if true performance is at chance, that is, if $p_i = .5$. One approach to assessing subliminality is to perform a null hypothesis significance test on the observed proportions against the null hypothesis that average performance across participants is at chance. If y_i and N_i are the number correct and the total sample size for participant i, and $q_i = y_i/N_i$, we might test the hypothesis that $\mu_q = .5$. If the sample sizes N_i are reasonably large and approximately the same, then $\bar{q}$ will be approximately normal, and we can apply a t test against $\mu_q = .5$. If the t test is not statistically significant, we conclude that performance is at chance. This logic has been used in several influential studies in the subliminal priming literature (Dehaene et al., 1998; Murphy & Zajonc, 1993).

There are at least two major flaws with this approach. First, there is the issue of acceptance of the null hypothesis. The t test essentially assumes that all participants are performing at chance unless there is sufficient evidence against that hypothesis. Thus, researchers who wish to show subliminality have an incentive to underpower their designs; after all, with sufficiently small sample sizes even very good average performance can be claimed to be subliminal simply because there is not enough evidence against it. For this

reason, a null result from a null hypothesis significance test can not be used to argue for the null hypothesis itself. We will use a Bayesian approach to overcome this fundamental limitation.

The second major flaw with this t test approach is a failure to properly separate between-participant and across-participant variability. Consider the sources of variability in estimated performances q_i: the statistic can vary due to natural sampling variability in the task, but also because people vary in their performance. We can reduce the first source of variability by increasing N_i, but not the second. Consider the extreme case where we have two participants, and they perform an arbitrarily large number of trials. Suppose that $q_1 = .6$ and $q_2 = .9$. We know that both participants are above chance with near-perfect certainty as $N_i \to \infty$, yet, we will always conclude that all participants are at chance because with two participants, the t test will not lead to a rejection of the null.

The failure to account for variability across participants leads not only to spurious acceptances of the null, but to spurious rejections as well. For example, suppose that to avoid the power problem outlined above, we obtain a large sample of participants. Suppose 99% truly perform at chance, and 1% of the population performs above chance at $p = .75$. Although 99% of our population is appropriate for assessing subliminal priming, we are guaranteed to reject all participants as we increase our sample size, because the true *average* performance is above 0.5.

To make these problems concrete, we consider data from a subliminal priming experiment reported in Rouder, Morey, Speckman, & Pratte (2007). In this experiment 27 participants performed 288 trials in a prime identification task. The primes were displayed briefly, only 22 ms, and were forward and backward masked. Performance was generally quite poor, with an average proportion correct of .53. A classical analysis of the accuracies reveals that average accuracy is significantly different from .5 ($t_{26} = 2.7, p = 0.011$) with a 95% CI of $(0.507, 0.551)$. Yet, a more complex story unfolds when participant variability is

examined (see Figure 6A). Although the majority of participants' observed accuracies are clustered around .5, there are two who score substantially higher than the rest. Under the logic outlined in the previous paragraphs, we would throw out the entire sample, even though the majority of participants' observed accuracies are concordant with chance performance. Instead, in the next section we present a hierarchical approach that overcomes this issue by modeling participant variability in assessing the subliminally of primes.

A Hierarchical Model

Our goal is to specify models of accuracy that include a psychological threshold. If activation from the stimulus is lower than this threshold, then performance is at chance. Conversely, if activation exceeds this threshold, then performance is above chance. The hierarchical model presented below is from Rouder, Morey, et al. (2007). At the first level of the model, we link the observed number correct y_i for each participant with an underlying true parameter, p_i:

$$y_i \overset{iid}{\sim} \text{Binomial}(p_i, N_i)$$

A hierarchical model is developed by specifying distributions on the individual performance parameters. In our case, we must carefully consider the parent population for the p_is. Since p is restricted to $[0, 1]$, it is inappropriate for a traditional normal population. Logit and probit models specify transformations of p into $(-\infty, \infty)$, making normal population distributions possible. For our purposes, however, these transformations are inappropriate because they allow true accuracy to be below $p = .5$. Instead, we use a half-probit transformation that restricts true accuracy to $p \geq .5$:

$$p_i = \begin{cases} \Phi(x_i) & x_i \geq 0 \\ .5 & x_i < 0 \end{cases}$$

where Φ is the CDF of the standard normal distribution. Rouder, Morey, et al. (2007) called this function the mass-at-chance (MAC) link, due to the fact that it allows participants to have true performance $p = .5$. We call x_i a "latent ability" because it indexes a person's ability even when $p_i = .5$. Figure 6B shows the relationship of latent ability to true accuracy. Consider two participants whose latent abilities are $x_1 = -.01$ and $x_2 = -2$. Participant 1 is very near the threshold of $x_i = 0$; perhaps a small increase in the duration of the prime stimuli would lead this participant to discriminate its less-than-five status more often than chance. Participant 2, however, is very far the threshold, and may need a larger increase in duration than Participant 1 to achieve above-chance performance.

The second level of the hierarchical model may be specified by placing a population distribution on the latent ability parameters:

$$x_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$$

The parameters μ and σ^2 together define the proportion of the participants whose performance is at chance. At the first level of the hierarchical model, we linked the observations with individuals' parameters; at the second level of the hierarchical model, we described how the individuals' parameters were distributed in a population. At the third and top level of the hierarchical model, we specify prior distributions for the parameters of the population of participants. There is some information for this specification from the context. In subliminal priming experiments, the goal is to make the primes difficult to see. It is therefore reasonable to place an informative prior on μ that is centered on the value of 0:

$$\mu \sim \text{Normal}(0, 1)$$

There is also natural constraint from the experimental context on parameter σ. If σ is too large, then bimodal distributions on p_i are likely with modes at chance and at ceiling. To

avoid bimodal distributions on performance, but still allow substantial variability across participants we choose a uniform prior on σ:

$$\sigma \sim \text{Uniform}(0, 1)$$

With all levels of the hierarchical model specified, joint and marginal posterior distributions may be computed. Of particular interest is the marginal posterior probability that the ith participant is performing at chance:

$$\omega_i = Pr(x_i \leq 0 \mid \boldsymbol{Y})$$

If this posterior probability is sufficiently high, then we should retain this participant to assess whether the primes truly influence judgments about the target. Also of interest are the marginal posteriors of the population level parameters μ and σ. A convenient statistic is the probability that any participant drawn from the population-level distribution is at chance. We denote this probability η, and it is

$$\eta = \Phi\left(-\frac{\mu}{\sigma}\right).$$

For example, if $\eta = .8$ for a given stimulus duration, then we expect that 80% of people will be at chance.

We can compute the marginal posterior distributions in a variety of ways: Rouder, Morey, et al. (2007) derived full conditional distributions and implemented a Gibbs sampler in R. Here, we present BUGS code model specification:

```
model {
  for( i in 1:M ){
    # Level 1: Binomial
    y[i] ~ dbin( p[i], N[i] )
```

```
    # Transformation between p and x, level 1
    p[i] <- phi( x[i] * step( x[i] ) )

    # Level 2: population on the latent abilities
    x[i] ~ dnorm( mu, precision )
  }

  # Level 3: Prior parameters
  mu ~ dnorm( 0, 1 )
  sig ~ dunif( 0, 1 )
  # BUGS uses precision, not std dev, to define normal
  precision <- 1 / ( sig * sig )
}
```

We fit the hierarchical model to the data of Rouder, Morey, et al. (2007) that is shown in Figure 6A. Figure 6C shows the resulting posterior distribution of the proportion of population judged to be performing at chance (η). Most of the mass is above .5, indicating that well over half of the population has performance at chance. Of particular interest are the posterior probabilities that the ith participant performs at chance (ω_i). Figure 6D shows the relationship between each participant's observed accuracy y_i/N_i and the corresponding posterior mean of ω_i.

One approach to selecting participants for subliminal priming analysis is to choose a criterion c such that if $\omega_i > c$, participant i is categorized as "at chance". The horizontal line in Figure 6D at .95 shows one possible criterion. The three points above the horizontal line represent participants whose priming effects we might examine; if we found evidence of priming for those participants, it could be used as evidence for subliminal priming.

The hierarchical model outlined above is quite simple, and allowed us to categorize by the plausibility that their ability correspond to at-chance levels given the assumptions of the model. Perhaps from a more broad perspective, it may viewed as a psychometric model of performance. The key innovation is the use of a half-probit link that accounts for a true psychological threshold. This threshold, unlike usual operationalizations in psychophysics, describes the point on latent ability where performance first rises above chance. One reasonable concern is the role of parametric assumptions, and the most salient is the half-probit mapping from latent ability to probability. To model the threshold, it seems necessary to have a link that maps many latent ability values to chance performance, but there are many alternatives to the half-probit link, such as the CDF of a Weibull which meet this requirement. We chose the half-probit for computational convenience, but there remains the question of whether this link is reasonable. Moreover, it is a somewhat open question of whether different links, such as that from the Weibull will lead to different assessments of which participants are at chance.

Unfortunately, it seems difficult to assess the fit of the half-probit and the dependence of conclusions of subliminality to parametric assumptions in typical priming studies. The reason for this difficulty is that in typical studies, many participants perform at near chance levels, and thus their performance offers little in the way of information to determine whether the link is reasonable. A better approach may be to change the paradigm to allow for a greater range of performance across individuals. In the current paradigm, stimulus difficulty reflects the duration of presentation, which was set to 22ms. In subsequent experiments (Morey et al., 2008), we asked participants to identify stimuli presented at durations from 17 ms to 167 ms. The model extension to this case is covered next.

Extending the hierarchical model

Extending the paradigm and model to multiple stimulus durations affords several advantages including the ability to state evidence that participants are at chance at specific durations. Participants who are particularly good at identifying the primes may require very short durations for chance performance, whereas participants who are not as good may be at chance to a wider range of prime durations. A second advantage is that the extended model covers the full range of performance of individuals across stimulus duration, and in this regard, may be treated as a psychophysical and psychometric model. The question of how well the probit-link is in accounting for performance may be assessed.

R. D. Morey, Rouder, & Speckman (2008) and R. D. Morey, Rouder, & Speckman (2009) developed several models that allow for multiple prime duration conditions. To demonstrate how hierarchical models can be naturally extended, we present the model of R. D. Morey, Rouder, & Speckman (2008) here, which is the simplest of the set. Consider an experiment in which J participants attempt to identify masked primes in I stimulus-duration conditions. In condition i, participant j performs N_{ij} prime identification trials, of which y_{ij} are correct. Figure 7A shows average accuracies in a prime identification task with $I = 6$ conditions (17ms, 25ms, 33ms, 58ms, 100ms, and 167ms). Identification for the shortest prime duration was extremely poor at 48%; the longest duration prime, however, was correctly identified an average of 85% of the time.

The first level of the extended hierarchical model is essentially the same as before, with the exception that now we index both participants and conditions. Observed accuracy for the jth participant in the ith condition, y_{ij}, is distributed as a binomial:

$$y_{ij} \overset{iid}{\sim} \text{Binomial}(p_{ij}, N_{ij}),$$

and, as before, we link true accuracy p with latent ability x_{ij} through the half-probit

transformation:

$$p_{ij} = \begin{cases} \Phi\left(x_{ij}\right) & x_{ij} \geq 0 \\ .5 & x_{ij} < 0 \end{cases}$$

In the previous model, all latent abilities x were drawn from a normal parent distribution. In this case, however, it is desirable to have this parent distribution depend systematically on the duration condition. We place an additive model on latent ability at the second level:

$$x_{ij} = \mu_i + \alpha_j.$$

where μ_i is the average ease with which primes in condition i are identified, and α_j is the identification ability of the jth participant. We have thus reduced the number of parameters underlying latent ability from ij to $i + j$. This type of reduction in complexity is one of the strengths of hierarchical modeling.

We assume that the participant ability parameters α_j are drawn from a normal population:

$$\alpha_j \mid \sigma_\alpha^2 \overset{iid}{\sim} \text{Normal}(0, \sigma_\alpha^2).$$

α_j can thus be interpreted as the random effect of participant j. We place an inverse gamma prior on the unknown variance σ_α^2:

$$\sigma_\alpha^2 \mid a, b \sim \text{Inverse Gamma}(a, b)$$

When parameters a and b are chosen to be small (e.g., 0.01), this prior is less constraining than the uniform prior on σ in the Rouder, Lu, et al. (2007) model. We can use a less constraining prior here because the data extend across multiple conditions including those where performance is definitively above chance.

The condition effect parameters μ can be interpreted as fixed effects; we thus place independent priors on each μ:

$$\mu_i \overset{iid}{\sim} \text{Normal}(0, 1),$$

where the prior parameters were selected to be similar to those for μ in the previously-presented model.

The full model can be described in the BUGS language:

```
model {
  for( j in 1:J ){
    for( i in 1:I ){

      # Binomial, level 1
          y[i,j] ~ dbin( p[i,j], N[i,j] )

          # Transformation between p and x, level 1
          p[i,j] <- phi( x[i,j] * step( x[i,j] ) )

      # latent ability is now a linear combination
      # of mu and alpha
          x[i,j] <- mu[i] + alpha[j]
    }
  }

  # Level 2 - define alpha
  for( j in 1:J ){
        alpha[j] ~ dnorm(0, precisionAlpha)
  }

  # Level 2 - define mu
  for(i in 1:I ){
```

```
        mu[i] ~ dnorm(0, 1)
    }

    # Prior parameters, level 3
    precisionAlpha ~ dgamma(aAlpha, bAlpha)
}
```

Figure 7, Panels B and C, show the results of fitting the model to the data shown in Panel A. Panel B shows the posterior probability on the proportion of participants in the six conditions who perform at chance. At the shortest duration, nearly all participants are predicted to be at chance; at the longest, the proportion at chance is surely less than 10%. Panel C shows the posterior probability that true performance is at chance for each participant by condition combination, that is, $Pr(x_{ij} < 0 \mid \boldsymbol{Y})$. Each line represents a participant, and each point on the line represents a condition.

As one would expect, the posterior probability of chance performance decreases for all participants as the prime duration increases. For most participants, the decrease occurs in a graded way. Interestingly, there are several participants whose curve is non-monotonic; that is, for some posterior probability by observed performance pairs, posterior probability increases as performance increases, which is the opposite of what one would expect. This is due to the fact that the additive nature of the model enforces the ordering of true performance to be the same for all participants across conditions. The model does not allow, for instance, one participant to improve their true performance as duration is increased from 25ms to 33ms, and another to get worse. However, because observed performance is subject to binomial noise, differences in performance across conditions may, for some participants, be the opposite of what one would expect. The hierarchical model allows us to use information from all participants to infer what the

ordering should be, and enforce it for all participants.

The extension of the hierarchical model – specifically, the addition of multiple conditions – allows for a more constrained, more easily tested model. In the simple hierarchical model, most participants were expected to be near chance performance, leaving us very little information by which to falsify the model. The extended model predicts a pattern of data for each participant across stimulus durations, and when this pattern is violated, it will be apparent.

Consider the participant represented by the right-most line in Figure 7C. In one condition, this participant is performing at an accuracy of .72, but the model says that this participant is almost surely at chance in that condition. This strange result led R. D. Morey et al. (2009) to further extend the model to allow for individual participant slopes:

$$x_{ij} = \theta_j(\mu_i + \alpha_j).$$

This model allows participants to improve at different rates as the stimulus intensity is changed, which improves model fit for some participants. Given the previous development, such an extension is conceptually straightforward. It requires an additional prior for θ; R. D. Morey et al. (2009) chose a normal distribution truncated below at 0, to require that θ be positive, and thus all participants must have the same ordering of true performance across conditions. The model can be easily defined in the BUGS language and fit with WinBUGS or JAGS.

The current set of hierarchical models are based on IRT type formulation with a novel link to account for thresholded behavior. Unlike IRT models, however, the effect of a person is modeled with two parameters while the effect of items (stimulus durations in this case) is modeled with just one. In this sense, these models may be considered perhaps the first set of *hierarchical psychophysical models*. We believe that such models may be of great use: they allow researchers to measure a truly at-chance threshold level of

performance across a large number of individuals with a limited numbers of experimental trials.

Subliminal priming remains a controversial topic. To assess whether it exists, R. D. Morey (2008) asked participants to both identify primes, and then identify primed targets. He first used the hierarchical model in R. D. Morey et al. (2009) to select participant-by-duration combinations for which it was more likely that latent ability was below rather than above chance. For these combinations, however, there was about 5-to-1 evidence by Bayes factor for a null priming effect on the time to identify primed targets. Hence, once one is somewhat sure that prime identification is at chance, the priming effect disappears! One notable study that contradicts this claim, however, comes from (Finkbeiner, 2011). Finkbeiner used two stimulus durations in a word priming experiment, and used the above extended hierarchical model to select participant-by-duration combinations as being at chance. With these combinations, Finkbeiner found about 10-to-1 evidence by Bayes factor for a priming effect. The approaches used by Morey and Finkbeiner provide for more rigorous assessment of subliminality and subliminal priming than previous methods, and further research with them will be a valuable part of unraveling the puzzle if and when subliminal priming occurs.

Hierarchical Models For Signal-Detection Experiments

In this section, we demonstrate how hierarchical modeling strengthens the inferential link between theory and data in understanding human memory. We focus on recognition memory, and a prevailing theoretical question is whether recognition memory is mediated by a single strength process or by the two processes of recollection and familiarity (Mandler, 1980). Aggregation, unfortunately, is the norm in recognition memory experiments. In these experiments, the basic unit of data is a dichotomous outcome. Either a participant indicates a test item is old or new, and to form hit and false

alarm rates, these outcomes seemingly must be aggregated across individuals or items. In the following section, we show how this aggregation may gravely distort conclusions about processing. We then introduce a hierarchical model that simultaneously accounts for participant and item variability, mitigating the need for aggregation. This hierarchical model provides for more valid assessment of processing, and we highlight our findings about the number and nature of processes underlying recognition memory.

Consequences of Aggregation In Memory Experiments

Recognition memory data have traditionally been modeled using the theory of signal detection (Green & Swets, 1966; Kintsch, 1967). Each tested item is assumed to generate some amount of mnemonic *strength*, which is graded and varies from trial to trial. This strength is compared to a criterion; an "old" response is produced if this strength is greater than the criterion, and a new response is produced otherwise. In the most conventional approach, called *equal variance signal detection*, the strength distribution for new items is a standard normal with a mean of 0 and variance of 1, and the strength of new items is shifted by an amount d', which serves as a sensitivity parameter. The corresponding hit and false alarm rates are given by

$$\begin{aligned} h &= \Phi\left(d' - C\right), \\ f &= \Phi\left(-C\right), \end{aligned}$$

where Φ denotes the cumulative distribution function (CDF) of the standard normal distribution, and C denotes the criterion. If hit rates are plotted as a function of false alarm rates for many values of the criterion, the resulting receiver operating characteristic (ROC) curve can be used to asses the veracity of the signal detection model of memory. In particular, this model predicts ROC curves that are curvilinear, as has now been observed in many recognition memory experiments. In addition, this model predicts that the ROC

curve will be symmetric about the negative diagonal. The solid black lines in Figure 8A correspond to equal-variance signal detection ROCs for $d' = 0.7, 1.6$, and 2.5.

The symmetric ROCs in Figure 8 are not characteristic of empirical ROC curves observed in recognition memory tasks. In almost all studies, observed ROCs are asymmetric with higher hit rates than expected for small values of false alarms (see Glanzer, Kim, Hilford, & Adams, 1999, for a review). This asymmetric pattern can be seen in the dashed line in Figure 8. There have been several models proposed to account for this asymmetry, including signal detection models with strength distributions of unequal variance across new and studied items (e.g., Ratcliff, Sheu, & Grondlund, 1992), and signal detection models that assume non-gaussian strength distributions (e.g., DeCarlo, 1998; Pratte & Rouder, 2009). Alternatively, Kellen, Klauer, & Broder (2013) argue that this asymmetry, indeed the curvature in general, is a result of aggregation and the true underlying curves are straight lines in accordance with a discrete-state model. Perhaps the most influential account, however, is a *dual-process* model proposed by Yonelinas (1994) and Yonelinas & Parks (2007). This model assumes that the recognition of a previously-studied item can come about by one of two separate processes: The item can be explicitly recollected in an all-or-none fashion, or failing recollection, it may be recognized based on its level of familiarity. Familiarity for both new and studied items follows the equal-variance signal detection model presented above. The hit and false alarm rates for this mixture model are given by:

$$\begin{aligned} h &= R + (1 - R) \times \Phi\left(d' - C\right), \\ f &= \Phi\left(-C\right), \end{aligned}$$

where d' and C are parameters of the signal detection process governing familiarity, and R is the probability of explicit recollection. The light, thick line in Figure 8A shows a typical ROC prediction for this model ($R = 0.29$, $d' = 1.0$). If $R = 0$, then the model reduces to

the equal-variance signal detection model, and the resulting ROC is symmetric. Recollection, R, has a one-to-one correspondence with the degree of asymmetry in the ROC curve. Accordingly, the ubiquitous finding of asymmetry in ROC curves is consistent with the presence of two processes mediating recognition memory.

We show here the potential distortions from aggregation in measuring the symmetry of ROC curves. Let's consider the role of item variability, as items are typically aggregated across to form hit and false-alarm rates. Suppose, for demonstration, that there is no recollection. That is, the data from each item follows an equal-variance signal detection model. Let's also suppose for demonstration that there are two items: an easy item with a true $d' = 2.5$, and a harder one, with true $d' = 0.7$. The ROCs for these items are shown as the solid lines labeled "Easy" and "Hard" in Figure 8A. Now, suppose the hit and false alarm events are averaged over these items. It is hoped that the resulting ROC would reflect the underlying structure, and perhaps be the middle solid line, which is the signal detection model with $d' = 1.6$, the average d' of the easy and hard items. Unfortunately, this ROC does not result from aggregating data. Instead, the dashed line occurs, and this line has a substantial degree of asymmetry. This asymmetry is distortion; an artifact of aggregation, and is not at all a signature of cognitive processing. Perhaps most unsettling is that this distortion is asymptotic — it will remain regardless of how much data are collected (Rouder & Lu, 2005). The dashed line is alarmingly close to the ROC prediction for the two-process model, and researchers who fit models to data aggregated across items run the risk of concluding that there are two processes with substantial recollection, when in fact there is only one process.

The question of whether the data are better described by the dual-process model or by simpler models is important and topical. It cannot be answered with data aggregated across items or individuals, as this aggregation may gravely distort the ROC patterns. To assess whether the asymmetry in ROC curves is a true signature of cognitive process or an

artifact of aggregation,we have constructed a series of hierarchical models (R. D. Morey, Pratte, & Rouder, 2008; Pratte et al., 2010; Pratte & Rouder, 2011). In this chapter, we use a hierarchical dual-process model (Pratte & Rouder, 2012) based on Yonelinas' model to assess ROC asymmetry. The degree of asymmetry in this model is indexed by the recollection parameter R, with $R = 0$ corresponding to the symmetric curves and greater values of R corresponding to greater degrees of asymmetry. The main feature of the model is that it accounts for variability across individuals and items, and there is no need to aggregate data for analysis. Consequently, estimates of recollection, which index asymmetry, are not distorted by these nuisance sources of variation.

A Hierarchical Dual-Process Model of Recognition Memory

Consider an experiment in which each of $i = 1, \ldots, I$ participants is tested on each of $j = 1, \ldots, J$ items. For each participant, some of these items were indeed studied, while the rest are novel. The participant responds by endorsing one of K confidence ratings options. In the signal detection approach, the multiple ratings options are modeled with multiple criteria: there are $K - 1$ criteria as shown in Figure 8B. In constructing the hierarchical model, it is useful to reparameterize the signal detection model such that one of the criteria is set to 0, and the center of the new-item distribution is free. We let $d^{(s)}$ and $d^{(n)}$ denote the centers of studied and novel-item distributions, respectively.

The hierarchical model is constructed by specifying parameters for each participant-by-item combination. Let R_{ij} be the participant-by-item recollection value, and let $d_{ij}^{(s)}$ and $d_{ij}^{(n)}$ be the participant-by-item values of the centers of the familiarity distribution for studied and novel items, respectively. The resulting hit and false alarm probabilities for each participant by item combination are

$$\begin{aligned} h_{ijk} &= R_{ij} + (1 - R_{ij}) \times \Phi\left(d_{ij}^{(s)} - C_{ik}\right), \\ f_{ijk} &= \Phi\left(d_{ij}^{(n)} - C_{ik}\right), \end{aligned}$$

where f_{ijk} and h_{ijk} are the false alarm and hit rates for the ith person responding to the jth item in the kth confidence rating. Individual criteria parameters C_{ik} are also free to vary across participants, reflecting individuals' response biases for particular confidence responses. The familiarity component of the model is depicted in Figure 8B.

In this model there are separate parameters for every participant by item combination for novel-item familiarity, studied-item familiarity and recollection. However, because each participant is tested on each item only once, there are no participant-by-item replicates in the data, and thus some constraint is needed. We assume that parameters are additive combinations of person and item effects in order to provide this constraint. The new-item familiarity follows:

$$d_{ij}^{(n)} \quad = \quad \mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)},$$

where $\mu^{(n)}$ denotes a grand mean, $\alpha_i^{(n)}$ denotes participant effects, and $\beta_j^{(n)}$ denotes item effects. Rather than place participant and item effects on the mean of studied-item familiarity $d_{ij}^{(s)}$, we place them on d'_{ij}, the difference between the studied-item and new-item distributions:

$$\log\left(d'_{ij}\right) = \mu^{(d)} + \alpha_i^{(d)} + \beta_j^{(d)}.$$

Placing an additive model on the log of d'_{ij} constrains the increase in sensitivity due to study to be positive for all participant by item combinations. Finally, the probability of recollection for each person and item is given by:

$$\Phi^{-1}\left(R_{ij}\right) \quad = \quad \mu^{(R)} + \alpha_i^{(R)} + \beta_j^{(R)},$$

where the inverse of the normal CDF (quantile) function is used to constrain the sum of participant and item effects to be between 0.0 and 1.0, as recollection is a probability.

Although these additive structures greatly simplify the model, there are still a large number of parameters to be estimated. Further constraint is achieved by placing

hierarchical structures on participant and item effects. For example, new-item familiarity values follow:

$$\begin{aligned} \alpha_i^{(n)} &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \beta_j^{(n)} &\sim \text{Normal}(0, \sigma_\beta^2). \end{aligned}$$

where the variance parameters are estimated from the data, and provide measures of participant and item variability in new-item familiarity. Similar hierarchical structures are placed on participant and item effects in studied-item familiarity and recollection, providing for efficient parameter estimation even with small numbers of participants and items.

Applications of the Hierarchical Memory Model

The hierarchical model allows for the estimation of underlying mnemonic processes from recognition memory data without recourse to aggregation and the accompanying distortions. The presented model is discussed in detail in Pratte & Rouder (2011, 2012), and estimation may be performed with the R package *HBMEM*, available on CRAN.

One of the main questions is whether the asymmetry in ROC curves is truly the result of cognitive processing, such as all-or-none recollection, or reflects distortion that results from averaging data over participants or items, as is demonstrated in Figure 8. This question can be answered by consideration of the mean recollection parameter ($\mu^{(R)}$), a measure of ROC asymmetry that in the hierarchical model is uncontaminated by participant and item variability. If posterior beliefs are centered far from zero, then the ubiquitous ROC asymmetry is indeed a cognitive signature rather than an artifact. We applied the model to Experiment 1 in Pratte et al. (2010), a large recognition memory experiment in which 94 participants were tested on 480 items. The resulting posterior distribution for the mean recollection parameter ($\mu^{(R)}$) is shown in Figure 9A. All of the posterior mass is substantially above zero, implying that ROC asymmetry is present even

when participant and item variability are modeled. This asymmetry is seemingly a cognitive signature rather than an artifact of aggregation (see R. D. Morey, Pratte, & Rouder, 2008; Pratte et al., 2010), and should be treated as an important benchmark in theory construction.

Although both the aggregated and hierarchical analysis of these data provide the same qualitative conclusion of ROC asymmetry, aggregation nonetheless leads to distorted parameter estimates and a dramatic overestimation of precision of these estimates. Familiarity, for example, is overestimated by 15% when data are aggregated over both participants and items, compared to mean familiarity in the hierarchical model. More alarming is the dramatic overestimation of precision from aggregation. For example, the 95% credible interval on mean recollection in Figure 9A is 2.7 times larger than the 95% confidence interval resulting from the aggregated analysis. This overstatement of precision from aggregation is a direct result of mismodeling multiple sources of variation and is well known (Clark, 1973; Raudenbush & Bryk, 2002; Rouder & Lu, 2005). Conversely the wider credible intervals from the hierarchical estimates directly represent the uncertainty from accounting for multiple sources of variation. The differing degrees of precision has dramatic effects on assessing whether mean recollection or familiarity changes with condition variables, and it is possible that previous demonstrations of effects with aggregated data are overstatements of the true significance of condition effects (Pratte & Rouder, 2012).

The above assessment shows that ROC asymmetry is a signature of the cognitive processes subserving recognition memory, but does not necessarily imply that recognition memory is mediated by recollection and familiarity. The hierarchical model provides additional insights because it provides for separate assessment of recollection and familiarity for each individual and for each item. If recollection and familiarity are statistically independent processes, then the recollection and familiarity across individuals

should be uncorrelated; likewise, recollection and familiarity across items should be uncorrelated. The dark points in Figure 9B show the relationship between recollection and familiarity for people ($\alpha_i^{(r)}$ vs. $\alpha_i^{(d)}$), the light points show the relationship for items ($\beta_j^{(r)}$ vs. $\beta_i^{(d)}$). Two trends are evident. First, there is substantial variability in both individuals' mnemonic abilities and how easily items are remembered. Second, there is substantial correlation: people with high recollection also have high familiarity ($r = .48$), and items with high recollection also have high familiarity ($r = .49$). Pratte & Rouder (2011) found that the degree of correlation is statistically significant, but, nonetheless, a model with only shared variability does not do as well as a model with both shared and unique variability for recollection and familiarity.

One of the main sources of evidence for two separable processes has been the demonstrations of dissociations across experimental conditions. One classic dissociation is between a levels-of-processing manipulation and a perceptual-feature manipulation. Deep levels of study, such as producing a related word to an item at study, should lead to an increase in recollection over shallow levels of study, such as counting vowels in study items. Conversely, changing perceptual features between study and test, such as font or color, should attenuate familiarity rather than recollection. Some researchers have had success in generating these dissociations, but they seemingly occur only under special circumstances. In particular, perceptual effects are difficult to obtain (Hockley, 2008; Mulligan, Besken, & Peterson, 2010; Murnane & Phelps, 1995), and tend to occur only in experiments with poor overall performance (e.g. Boldini, Russo, & Avons, 2004).

In Pratte & Rouder (2012), we used the hierarchical model to assess recollection and familiarity across 13 conditions in 4 experiments. Our manipulations produced effects that were as large or larger than previous ones in the literature. Figure 9C shows joint posterior distributions of mean recollection as a function of mean familiarity across the conditions. Each ellipse is a 95% credible region. If there was evidence for two distinct

processes, then these conditions should lie in a plane rather than on a monotonic curve (see Bamber (1979), and Newell & Dunn (2008) for an overview of the logic in interpreting such *state-trace* plots). Note that the curve is not incompatible with a double dissociation: some pairs of points differ more in familiarity than recollection (see the poorest performing points), whereas others differ more in recollection than familiarity (see the best performing points). Yet, all of the condition effects can be connected by an increasing curve, suggesting that a single factor, the location on the curve, is needed to account for these data. We think the relative attenuation of recollection in conditions with poor performance reflects the nature of ROC space. When performance is poor, the ROCs are near the diagonal and it is easier to detect small overall sensitivity effects (familiarity) than to detect small changes in asymmetry (recollection). Hence, even though our data has degrees of dissociation as large as any in the comparable literature, they are more compatible with a single-process approach than a dual-process approach.

Concluding Remarks

In this chapter, we have shown that while experimental psychologists have a rich theoretical and experimental tradition, the link between theory and data often presents difficulties in real-world contexts. These difficulties arise because theories are nonlinear, and there is often substantial nuisance variation across individuals and items. If these sources of nuisance variation are not appropriately modeled, they will distort the assessment of the underlying cognitive signatures, and lead to erroneous conclusions about theory. These potential problems occur across psychology, and here we have presented examples in assessing learning, subliminal priming, and recognition memory.

We advocate a Bayesian hierarchical approach for linking theory and data. These models provide for the simultaneous assessment of nuisance variation and variation from the target cognitive process of interest. They not only allow researchers to uncover the

rich cognitive structure in their data without aggregation artifacts, but allow for an understanding of how this structure varies across individuals and items.

In this chapter, we have tried to focus on the types of problems hierarchical modeling can solve, as well as an introduction to Bayesian probability. We have avoided the nuts and bolts of estimation, and this avoidance leaves open the question of how interested researchers can develop and analyze their own models. There are now several excellent texts on Bayesian modeling that include development of Bayesian hierarchical models, and advanced texts include Gelman, Carlin, Stern, & Rubin (2004) and Jackman (2009). More recently there have been tutorials and texts specific for psychology including Rouder & Lu (2005), Kruschke (2011), and the forthcoming book by Lee & Wagenmakers (2013). Here, we tackle more global questions about how researchers should learn Bayesian hierarchical modeling.

One question that arises is about software: which language and packages should researchers use? We think researchers should invest in three classes of languages. At the highest level, there are specialty languages developed especially for Bayesian hierarchical modeling, of which JAGS (Plummer, 2003) and WinBUGS (Lunn et al., 2000) are the most popular. These languages allow researchers to specify models and priors as input in a natural random-variable notation, and provide samples from posterior distributions as output. When they work, they often work well and save much development time. Therefore, these specialty languages serve as an excellent first option, and, importantly, require little special knowledge above and beyond the skills needed to specify models. Unfortunately, as general-purpose sampling solutions, they sometimes do not work well in specific situations: they may lack a feature necessary to define a model, or take an exceedingly long time to sample[5]. Determining whether a specialty language such as JAGS or WinBUGS will work is often fast and should be a first step for most researchers.

In cases where the general-purpose solutions fail, researchers may need to derive

conditional posterior distributions, develop sampling routines, and implement them. Data-analytic languages such as R (R Development Core Team, 2009) and MATLAB (MATLAB, 2010) are ideal for implementation, and often contain useful routines for MCMC sampling. Sometimes, however, the speed of R and MATLAB can be improved by implementing the sampling in a fast, low-level language such as C or Fortran. We use JAGS as our high-level specialty language, R as our mid-level data-analytic language, and C as our fast, low-level language, and we routinely move between these three as dictated by the model we wish to analyze. The hierarchical normal model and the mass-at-chance model in this chapter are both implementable in JAGS; analysis of the hierarchical dual-process model, however, was more convenient using a combination of R and C routines for efficiency. Our hope is that as more researchers use hierarchical models, they will develop the skills to go beyond WinBUGS or JAGS implementations as needed.

Perhaps the most important question is how should young scholars be trained so that they may use Bayesian hierarchical models. In our view, it is hard to overstate the usefulness of solid training in statistics including courses in calculus-based mathematical statistics, linear algebra, and Bayesian analysis. We realize, that many talented students will not have the aptitude or time for such study, and so it is worthwhile to consider alternatives. A good course would be one that stresses the logic of modeling. This course would focus on the basics of probability and statistics, and promote a deep understanding of conditional probability. Course objectives would include the ability to specify models, and write down and visualize likelihoods, and would provide an overview of the issues in model comparison. We hope the appeal of Bayesian hierarchical models will motivate more rigorous general statistical training in psychology.

References

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452-469.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144-151.

Averell, L., & Heathcote, A. (2011). The form of forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*, 25-35.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, *81*, 55-65.

Bamber, D. (1979). State trace analysis A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137-181.

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135-152.

Boldini, A., Russo, R., & Avons, S. E. (2004). One-process is not enough! A speed-accuray tradeoff study of recognition memory. *Psychonomic Bulletin and Review*, *11*, 353-361.

Busemeyer, J. R., & Diederich, A. (2009). *Cognitive modeling*. Sage.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

DeCarlo, L. M. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186-205.

Dehaene, S., Naccache, L., Le Clech, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597-600.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), pp. 214-226. Retrieved from `http://www.jstor.org/stable/2239734`

Doucet, A., de Freitas, N., & Gordan, N. (2001). *Sequential monte carlo methods in practice*. Springer.

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

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

Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin and Review*, *15*, 1209–1217.

Finkbeiner, M. (2011). Subliminal priming with nearly perfect performance in the prime-classification task. *Attention, Perception, & Psychophysics*, *73*, 1255-1265.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439-453. Retrieved from `http://psycnet.apa.org/doi/10.1037/a0015251`

Gelfand, A., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman and Hall.

Gelman, A., Shor, B., Bafumi, J., & Park, D. (2007). Rich state, poor state, red state, blue state: Whats the matter with Connecticut? *Quarterly Journal of Political Science*, *2*, 345-367.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 500-513.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greenwald, A., Klinger, M., & Schuh, E. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General*, *124*, 22-42.

Greenwald, A. G. (1992). New look 3: Unconscious cognition reclaimed. *American Psychologist*, *47*, 766-779.

Haider, H., & Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 392-406.

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

Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1412–1429.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Chichester, United Kingdom: John Wiley & Sons.

Jeffreys, H. (1961). *Theory of probability (3rd edition).* New York: Oxford University Press.

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation methods for test theory without an answer key. *Psychometrika*, *68*, 373-389.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kellen, D., Klauer, K., & Broder, A. (2013). Recognition memory models and binary-response rocs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, 1-27.

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

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, *74*, 496-504.

Koechlin, E., Naccache, L., Block, E., & Dehaene, S. (1999). Primed numbers: Exploring the modularity of numerical representations with masked and unmasked semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1882-1905.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models*. Chicago: McGraw-Hill/Irwin.

Lee, M. D. (2006). A hierarchical bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 126.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410-423. Retrieved from `http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337`

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187-192.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Lunn, D. (2003). WinBUGS development interface (WBDev). *IBSA Bulletin*, *10*.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mandler, G. (1980). Recognizing: The judgment of previous occurence. *Psychological Review*, *87*, 252-271.

MATLAB. (2010). *version 7.10.0 (r2010a)*. Natick, Massachusetts: The MathWorks Inc.

Meng, X., & Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*, 831-860.

Merikle, P., Smilek, D., & Eastwood, J. (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition*, *79*, 115-134. Retrieved from `http://dx.doi.org/10.1016/S0010-0277(00)00126-8`

Merkle, E., Smithson, M., & Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, *55*, 57-67.

Morey, C. C., Cowan, N., Morey, R. D., & Rouder, J. N. (2011). Flexible attention allocation to visual and auditory working memory tasks: Manipulating reward induces a trade-off. *Attention, Perception & Psychophysics*, *73*, 458-472.

Morey, R. D. (2008). *Item response models for the measurement of thresholds.* Unpublished doctoral dissertation, University of Missouri.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376-388.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21-36.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2009). A truncated-probit item response model for estimating psychophysical thresholds. *Psychometrika*, *74*, 603-618.

Mulligan, N. W., Besken, M., & Peterson, D. (2010). Remember-know and source memory instructions can qualitatively change old-new recognition accuracy: The

modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 558-566.

Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 158–172.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723-739.

Myung, I.-J., Kim, K., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*, 832-840.

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79-95.

Naccache, L., & Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*, 215-229.

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*, 285–290.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Hoboken, New Jersey: Wiley.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing.*

Pratte, M. S., & Rouder, J. N. (2009). A task-difficulty artifact in subliminal priming. *Attention, Perception, & Psychophysics*, *71*, 276-283.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224-232.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Ratcliff, R., Sheu, C. F., & Grondlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518-535.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods. second edition.* Thousand Oaks, CA: Sage.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 435-451.

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 65-82.

Rogers, S. (1992). How a publicity blitz created the myth of subliminal advertising. *Public Relations Quarterly*, *37*, 12-17.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573-604.

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

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689. Retrieved from `http://dx.doi.org/10.3758/s13423-011-0088-7`

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877-903.

Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin and Review*, *11*, 938-944.

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: A rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*, 241-247.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597-605.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (n.d.). *The $p < .05$ rule and the hidden costs of the free lunch in inference.* (Paper under review)

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237. Retrieved from `http://dx.doi.org/10.3758/PBR.16.2.225`

Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*(1201-1208).

Sarbanés Bové, D., & Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, *6*, 1–24.

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

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

Stan Development Team. (2013). *Stan: A c++ library for probability and sampling, version 1.1.* Retrieved from `http://mc-stan.org/`

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

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618. Retrieved from `http://www.jstor.org/stable/2291073`

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

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.

Walker, S. G., Laud, P. W., Zanterdeschi, D., & Damien, P. (2011). Direct samping. *Journal of Computational and Graphical Statistics*, *20*, 692-713.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio. *Computational Statistics and Data Analysis*, *54*, 2094-2102.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1341-1354.

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

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

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian*

statistics: Proceedings of the First International Meeting held in Valencia (Spain) (pp. 585–603). University of Valencia.

Footnotes

[1]Linear models are those where the expected value of the data is a linear function of the parameters (Kutner, Nachtsheim, Neter, & Li, 2004). Examples include ANOVA and regression. Nonlinear models violate this basic tenet: the expected value of the data cannot be expressed as a linear function of parameters.

[2]JAGS may be obtained at `http://mcmc-jags.sourceforge.net`. WinBUGS and OpenBUGS (for non-Windows operating systems) may be obtained at `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml` and `http://www.openbugs.info/w/`, respectively.

[3]Posterior beliefs may be computed by subtracting MCMC samples. For the mth iteration, let $c^{(m)} = \beta_2^{(m)} - \beta_1^{(m)}$. The dotted line in Figure 5B is the smoothed histogram of $c^{(m)}$.

[4]The posterior for this contrast is computed in MCMC as $c^{(m)} = (\sum_i (\mu_{i2}^{(m)} - \mu_{i1}^{(m)})/I$, and the solid line in Figure 5B is the smoothed histogram.

[5]Fortunately, these general-purpose samplers are extensible (Lunn, 2003) and have improved greatly in recent years. In addition, newcomers such as Stan (Stan Development Team, 2013) show promise.

Figure Captions

Figure 1. Estes' (1956) example of the difficulty of linking learning-curve data to learning theories. **A.** Predictions: The solid and dashed lines show predictions from the gradual-decrease and all-at-once models of learning, respectively . **B.** Data form Reder and Ritter (1992). The grey lines show the times for 15 individuals as a function of practice; the red circles are means across individuals, and these means decrease gradually with practice. **C.** Hypothetical noise-free data from the all-at-once learning model. Individuals' data are shown as thin grey lines. The mean, shown with red points, nonetheless decreases gradually. This panel shows that the mean over individuals does not reflect the structure of any of the individuals.

Figure 2. Prior and posterior beliefs from three analysts for the probability of heads. **A.** Prior beliefs. Analyst I believes that all outcomes are equally plausible; Analyst II believes that heads are more likely than tails; and Analyst III not only believes that tails are more likely than heads, but that the coin has no chance of favoring heads. **B.** The updated posterior beliefs after observing 8 heads and 4 tails.

Figure 3. Prior and posterior beliefs on μ, the center of a normal distribution. **A.** Prior beliefs of two analysts. **B.** Posterior beliefs conditional on a sample mean of $\bar{Y} = 95$ and a small sample size of $N = 10$. **C.** Posterior beliefs conditional on a sample mean of $\bar{Y} = 95$ and a larger sample size of $N = 100$.

Figure 4. Joint prior (left), likelihood (center), and joint posterior (right) distributions across normal-distribution parameters μ and σ^2. Also shown, in the margins are the marginal posterior distributions of μ (top) and σ^2 (right).

Figure 5. The advantages of hierarchical modeling. **A**. Hypothetical data from 20 individuals each providing observations in 2 conditions. The bars show overall condition

means; the points and lines show individual's condition means. **B.** Posterior distributions of the condition effect from Model $\mathcal{M}_1$, the aggregation model (dotted line), Model $\mathcal{M}_2$, the cell means model (solid line), and Model $\mathcal{M}_4$, the hierarchical model with main effects and interactions (dashed line). Localization is worse for $\mathcal{M}_1$ because participant variability is not modeled. **C.** Solid lines show participant-by-condition point estimates from $\mathcal{M}_2$; the dotted lines show the same from $\mathcal{M}_4$. The shrinkage in $\mathcal{M}_4$ to main effects imposed by the hierarchal prior smooths these estimates. **D.** A comparison of individual-by-condition estimates from the $\mathcal{M}_2$, the cell-means model, and $\mathcal{M}_4$, the hierarchical model with main effects and interactions. There is modest shrinkage for extreme estimates.

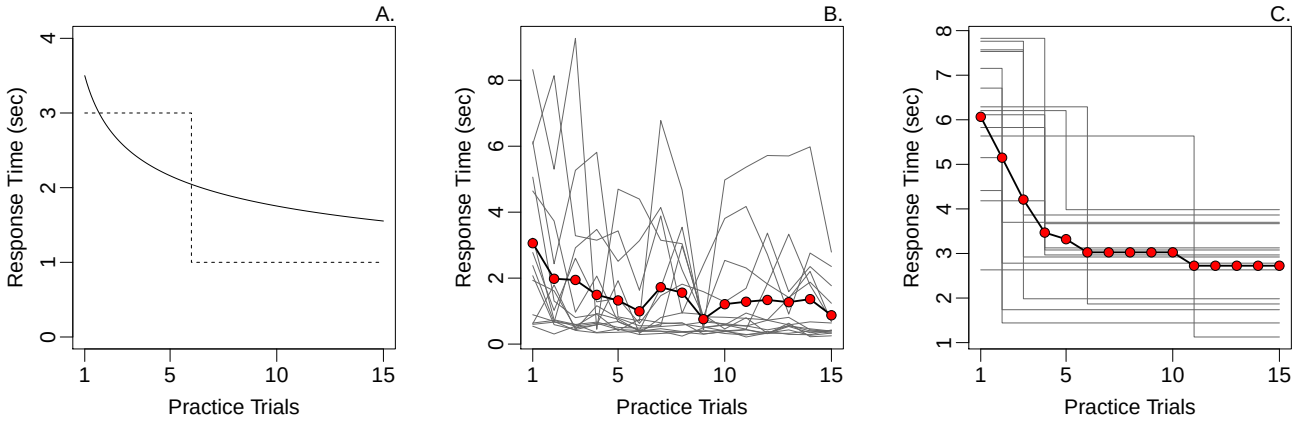
Figure 6. A: Violin plot of 27 participants' performance in a prime identification task. The confidence interval within the violin plot is the 95% CI on the mean accuracy; the horizontal line at 0.5 represents chance performance, and the horizontal dashed lines bound the interval within which we would expect 95% of participants to perform if they were truly at chance. B. The mass-at-chance link function. C: Posterior distribution of the population proportion at chance. D: Posterior probability that individuals are at chance as a function of their observed performance.

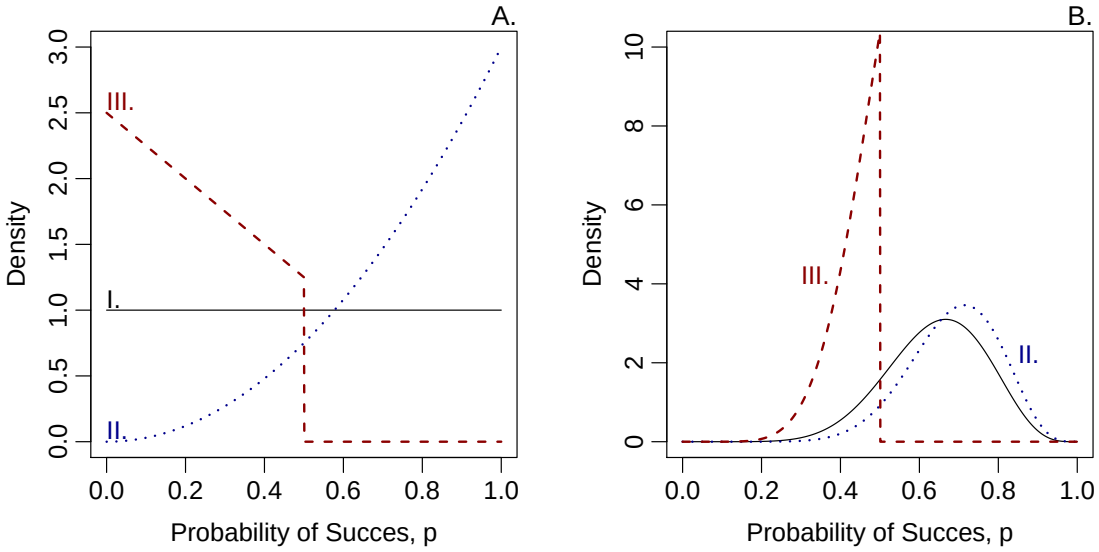
Figure 7. A: Mean performance by duration condition in a prime identification task. Error bars are standard errors of the mean. B: The posterior distribution, for each duration condition, of the proportion of the population that would perform at chance in that condition. C: Posterior probability that individuals are at chance as a function of their observed performance. Lines represent participants, and each point a condition. The top/left-most point for each participant is the briefest duration condition, and subsequent points along the lines are increasingly higher-duration conditions.

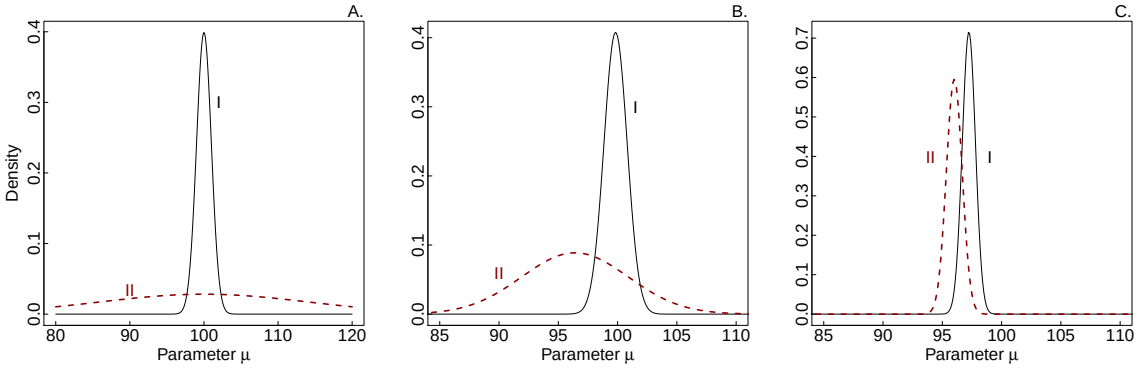
Figure 8. **A.** ROC curves from the equal variance signal detection model (solid black

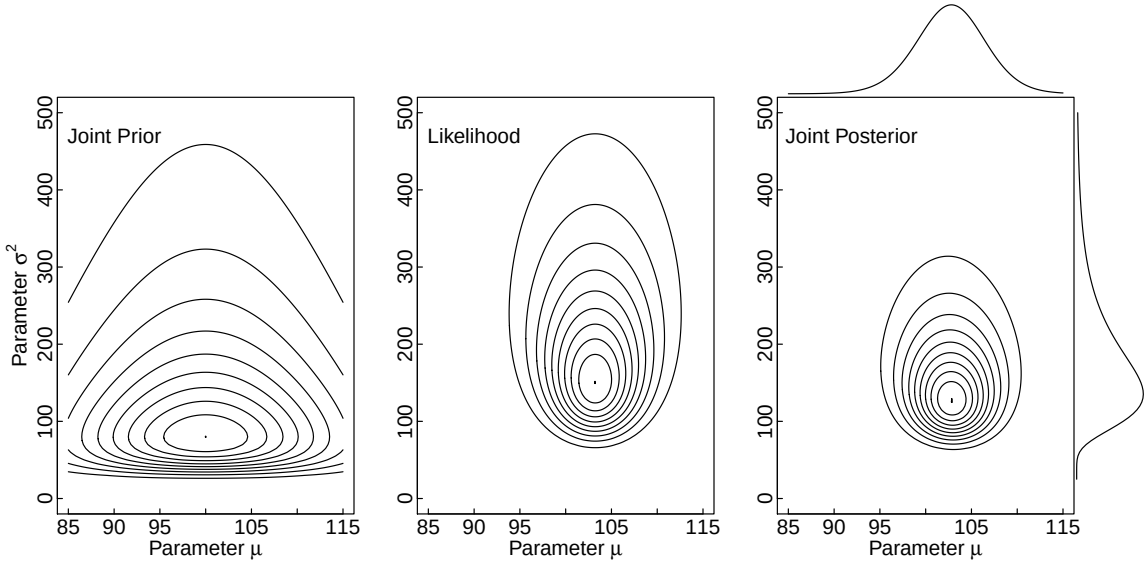
lines), the distorted data from this model averaged over participants (dashed line), and the dual-process model fit to these distorted data (thick grey line). **B.** The signal detection component of the hierarchical dual-process model.

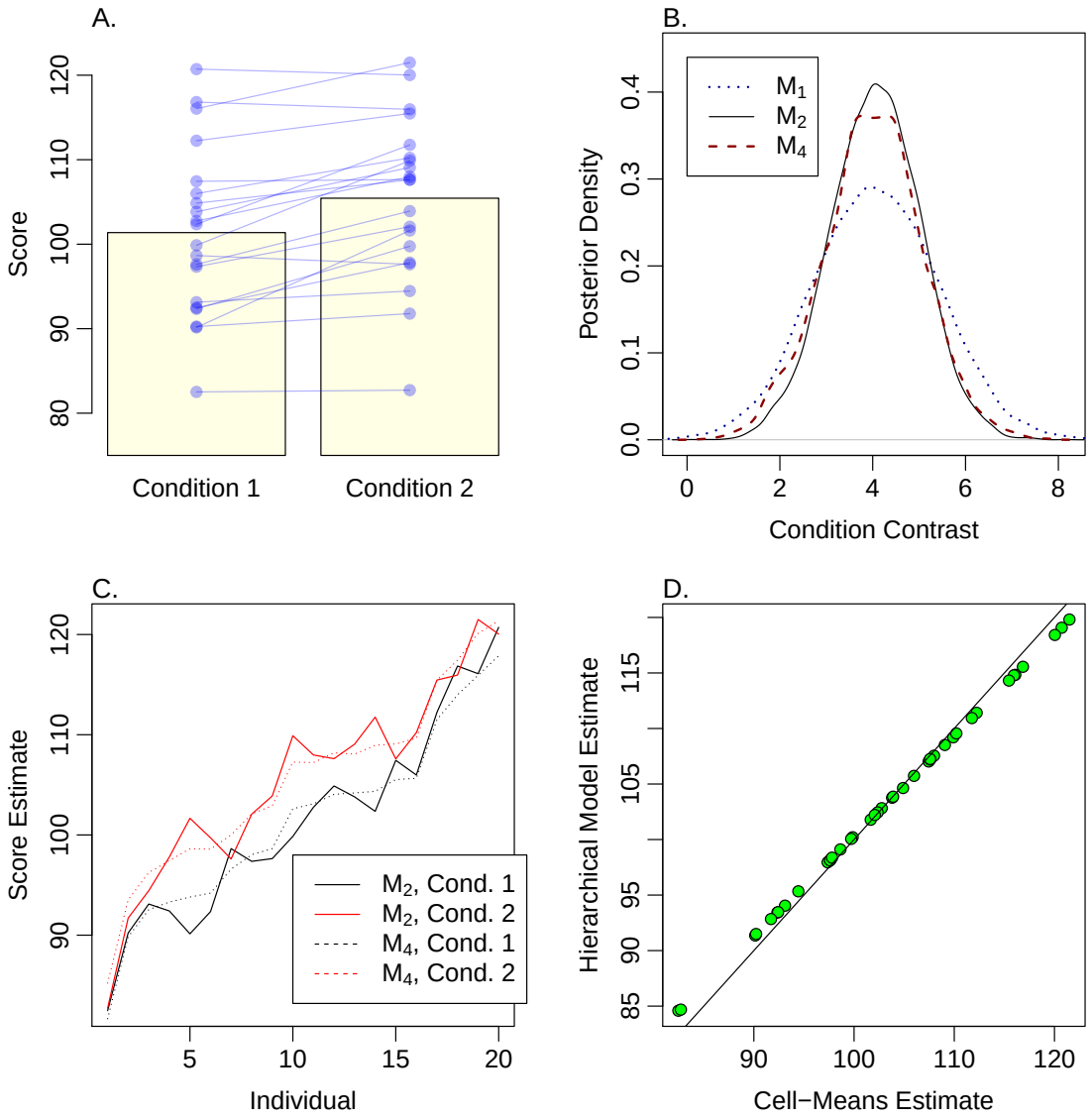
Figure 9. **A.** Posterior distribution of mean recollection, estimated with the hierarchical dual-process model. **B.** Participant and item effects in recollection plotted as a function of effects in familiarity. **C.** Joint posterior distributions of recollection and familiarity for 13 experimental conditions. The line is a non-parametric fit, highlighting the monotonic relationship between recollection and familiarity estimates.

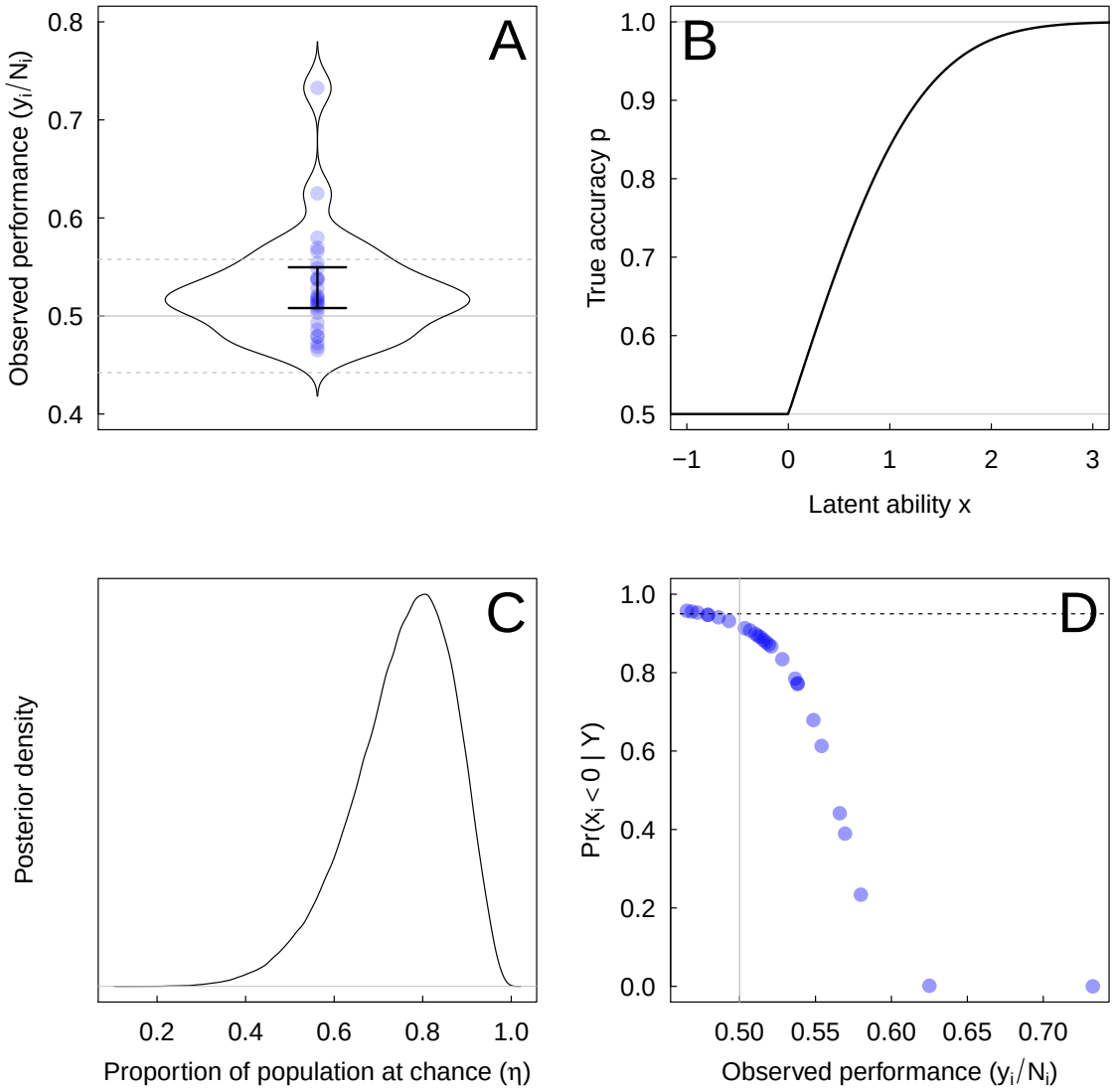

A.
B.
C.
Response Time (sec)
Practice Trials
0
1
2
3
4
5
6
7
8
1
5
10
15

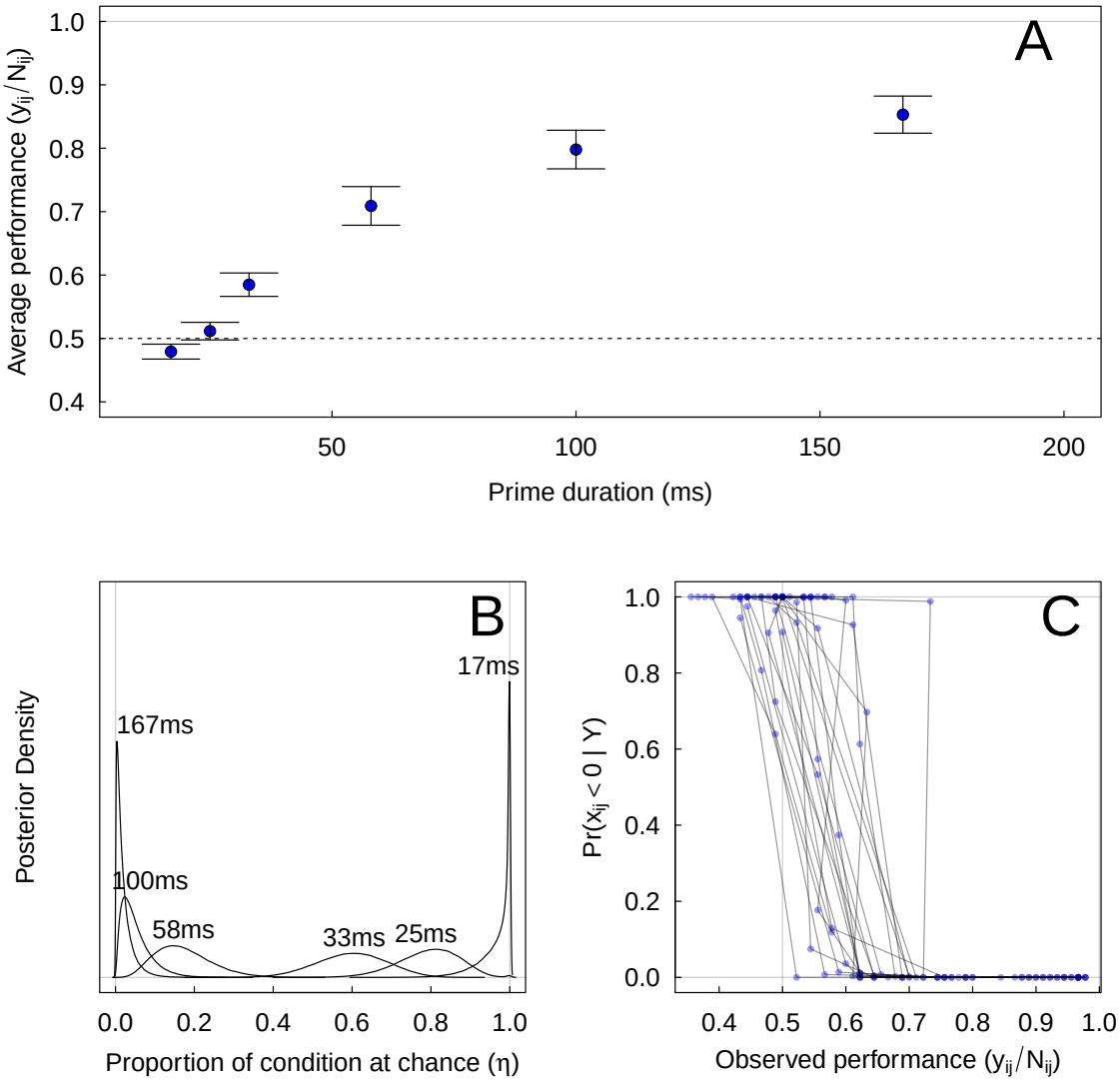

A.
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
III.
I.
II.
0.0
0.2
0.4
0.6
0.8
1.0
Probability of Succes, p
B.
Density
0
2
4
6
8
10
III.
II.
0.0
0.2
0.4
0.6
0.8
1.0
Probability of Succes, p

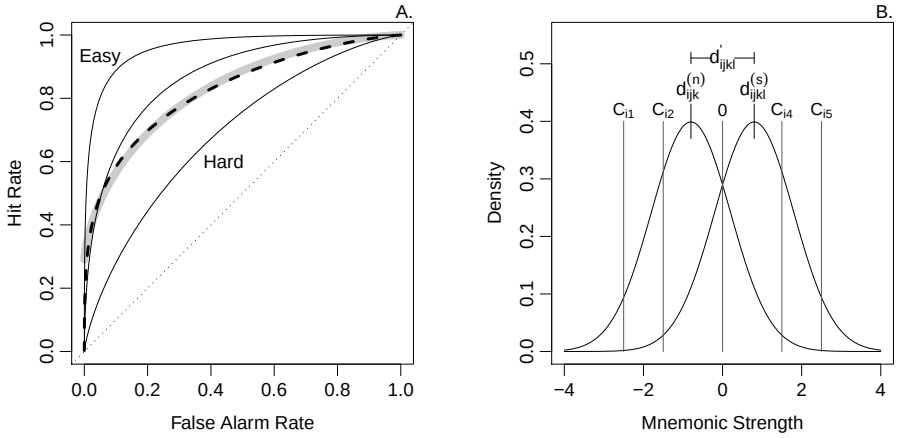

A.
B.
C.
Density
Parameter μ
I
II

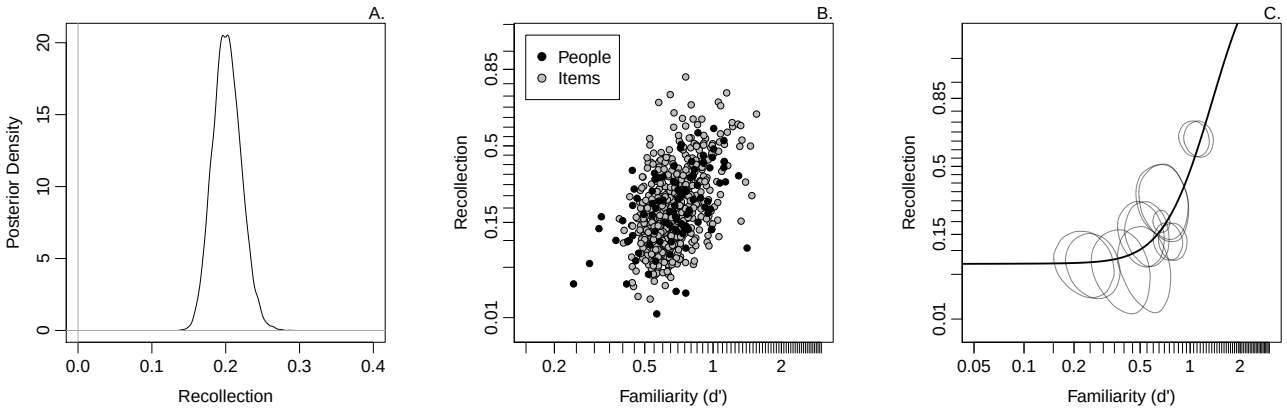

Joint Prior
Likelihood
Joint Posterior
Parameter σ2
Parameter μ
0
100
200
300
400
500
85
90
95
105
115


A.
Score
120
110
100
90
80
Condition 1
Condition 2
B.
Posterior Density
0.4
0.3
0.2
0.1
0.0
M1
M2
M4
0
2
4
6
8
Condition Contrast
C.
Score Estimate
120
110
100
90
M2, Cond. 1
M2, Cond. 2
M4, Cond. 1
M4, Cond. 2
5
10
15
20
Individual
D.
Hierarchical Model Estimate
115
105
95
90
85
90
100
110
120
Cell−Means Estimate


A
Observed performance (y_i/N_i)
0.8
0.7
0.6
0.5
0.4
B
True accuracy p
1.0
0.9
0.8
0.7
0.6
0.5
−1
0
1
2
3
Latent ability x
C
Posterior density
0.2
0.4
0.6
0.8
1.0
Proportion of population at chance (η)
D
$Pr(x_i < 0 \mid Y)$
1.0
0.8
0.6
0.4
0.2
0.0
0.50
0.55
0.60
0.65
0.70
Observed performance (y_i/N_i)


A
Average performance (y_{ij}/N_{ij})
1.0
0.9
0.8
0.7
0.6
0.5
0.4
50
100
150
200
Prime duration (ms)
B
17ms
167ms
100ms
58ms
33ms
25ms
Posterior Density
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of condition at chance (η)
C
$Pr(x_{ij} < 0 \mid Y)$
1.0
0.8
0.6
0.4
0.2
0.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Observed performance (y_{ij}/N_{ij})


A.
Easy
Hard
Hit Rate
False Alarm Rate
0.0
0.2
0.4
0.6
0.8
1.0
B.
Density
Mnemonic Strength
d'_{ijkl}
$d^{(n)}_{ijk}$
$d^{(s)}_{ijkl}$
C_{i1}
C_{i2}
0
C_{i4}
C_{i5}
−4
−2
2
4
0.1
0.3
0.5


A.
Posterior Density
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
Recollection
B.
People
Items
Recollection
0.01
0.15
0.5
0.85
0.2
0.5
1
2
Familiarity (d')
C.
Recollection
0.01
0.15
0.5
0.85
0.05
0.1
0.2
0.5
1
2
Familiarity (d')