Introduction

One goal of cognitive psychology is to understand the representations and processes that underly mental life, including those that underly vision, perception, memory, thought, decision-making, and language. This task is difficult because although the brain is vast in complexity, our methods for exploring it are limited. Many of us leverage the classical experimental method where stimuli are manipulated and a small number of output measures such as response choice and response time are observed. The obvious advantage of this method is on the input side—psychologists may creatively manipulate the environment and stimuli to derive causal insights unavailable from studying correlations alone. The difficulty of this approach is in interpretation. Because human cognition is complex and the brain is treated as a black box, the relationship between input and output is relatively indirect and impoverished. Consequently, it is often difficult to draw precise or well-localized inferences about underlying mental representation and processes. In our opinion, this difficulty in interpretation has been chronically understated throughout the field.

These difficulties in interpretation may be seen in the following classic example in learning theory. Estes (1956) was interested in inferring from learning curves whether learning happened gradually or, alternatively, all at once. These two accounts are shown in Figure 1A. Because these accounts are so different, adjudicating between them should be simple: 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 interpretation is tenuous, and that the gradual decrease in the pattern among the means does not necessarily imply that learning is gradual. Instead, learning may 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. Therefore, it is difficult to know whether the pattern of the means reflects a signature of cognitive processing or a signature of between-individual variation.

The main difficulty in Estes' example is the presence of two types of variation. One type is the form of the learning curve, whether it is graduation or all-or-none, which is the target of interest. The second type is variation in the parameters of the curve, but not the form, across individuals. For example, if learning is indeed all-or-none, there conceivably is variation in the transition time across participants. The result is that the analysis of any one curve is not informative for lack of data while the analysis of aggregated data confounds nuisance variation across participants with the cognitive process of interest. Unfortunately, this type of nuisance variation is common if not ubiquitous. Almost all experiments in psy-

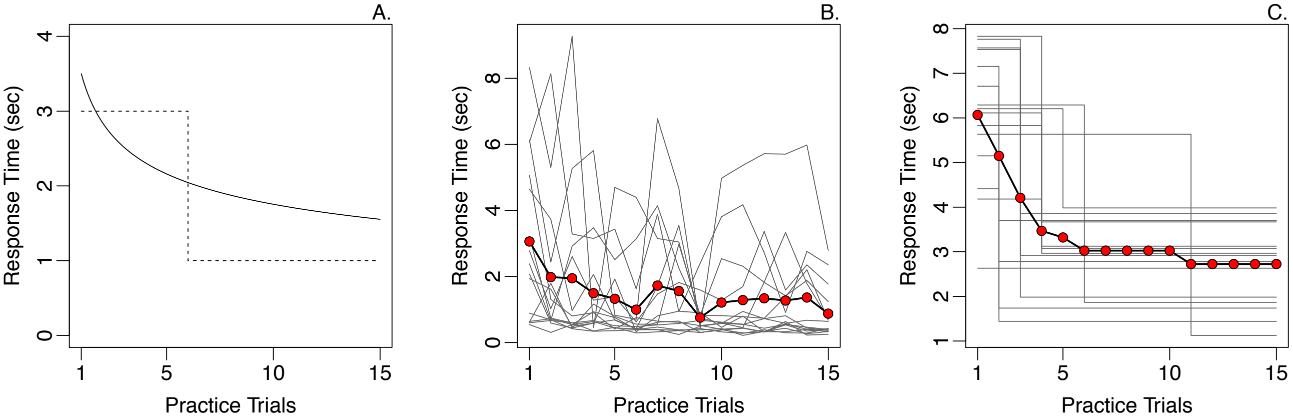

Figure 1. Estes' (1956) example of the difficulty of linking learning-curve data to learning theories. **A.** Predictions: The solid and dashed lines show predictions from the gradual-decrease and all-at-once models of learning, respectively . **B.** Data form Reder and Ritter (1992). The grey lines 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.

chological sciences contain variation across people or items or both that may be confounded with variation in the latent processes if not handled properly.

Many researchers know at a theoretical or abstract level of these problems of confounding variation at different levels. For example, the confounding in averaging learning curves has been expanded upon by Haider & Frensch (2002), Heathcote et al. (2000), Myung et al. (2000), and Rickard (2004). A similar worry is expressed in using multidimensional scaling to understand participants' mental representation of a domain (Ashby et al., 1994) where data are often averaged prior to scaling. Finally, the separation of cognitive variation from nuisance variation across participants and items has been expressed repeatedly in linguistics (Baayen et al., 2002; Clark, 1973) and recognition memory (Rouder et al., 2007; Pratte et al., 2010). Even though these problems are known and appreciated, researchers nonetheless routinely confound nuisance variation often by averaging data over items or participants. The reason that researchers confound levels is straightforward—they do not know what else to do! Our experience is that many researchers understand the problem of conflating levels but the lack of easily available alternatives offers little recourse.

The solution to the nuisance-variation problem is 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 inputs to these models are the raw, unaggregated data, and the outputs are individual-and-item specific estimates of psychological-process parameters. For example, if one fits a hierarchical all-or-none learning model, the outputs would be stable estimates of initial and final learning levels and transition points for each individual. More generally, hierarchical models provide both a clearer view of process and a method for exploring how these processes vary across populations of individuals or items. Hence, they turn a problem, how to account for nuisance variability, into a strength. Because they solve the nuisance-variation problem, hierarchal models are becoming increasingly popular in substantive domains in psychology.

The most popular and familiar application of hierarchical models are hierarchical linear models (HLM, e.g., Raudenbush and Bryk, 2002). These models are extensions of ANOVA and regression to contexts with multiple sources of variability. Unfortunately, theoretically interesting models of cognition are often nonlinear, and in these cases, HLM is not appropriate for capturing nuisance variation. Instead, cognitive psychologists need hierarchical nonlinear models, the hierarchal extension of nonlinear psychological process models. Hierarchical nonlinear models are, in general, difficult to analyze in conventional frameworks. This difficulty, however, is not present in the Bayesian framework. With modern advances in Bayesian analysis (e.g., Gelfand & Smith 1990), Bayesian analysis of hierarchical nonlinear models of psychological processes is possible and relatively straightforward. Consequently, 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 et al. (2007); Lee (2006); Farrell & Ludwig (2008); Merkle et al. (2011); Rouder et al. (2004, 2008); Vandekerckhove et al. (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; Rouder et al., in press; Shiffrin et al., 2008), and there is a special issue of the Journal of Mathematical Psychology (January 2011, Vol 55:1) devoted to the topic.

The remainder of this chapter provides a gentle introduction to Bayesian hierarchical modeling. Our goal is to make the material accessible to a wide audience including those who have a limited aptitude in mathematics and statistics. We use mathematics sparingly throughout and only where absolutely needed. We think this need is greatest for specifying models, and, consequently, devote the next session to model specification. Following this section is an introduction to Bayesian analysis, the core element in making hierarchical nonlinear models tractable. We provide a brief overview of Bayesian probability, Bayesian updating and estimation, and Bayesian model comparison. The fourth section on hierarchical models forms the core material for this chapter, and is followed by an in-depth application on characterizing the effect of word frequency on response time in lexical decision making.

Specifying Models

Many psychologists have not studied mathematics since high school, and many have not had a course in statistics that stresses the mathematical specification of models. Most are not inclined to slog through mathematically dense text, and reading mathematical equations does not evoke vivid mental images of psychological process. Furthermore, a bulk of psychologists are simply intimidated by math, and a sizable number will stop reading any article at the first equation. (As an aside, math intimidation is at some level a seeming constant. We have it on good authority that academic mathematicians are often intimidated by advanced mathematics, especially if the topic is outside their narrow specialty. It might

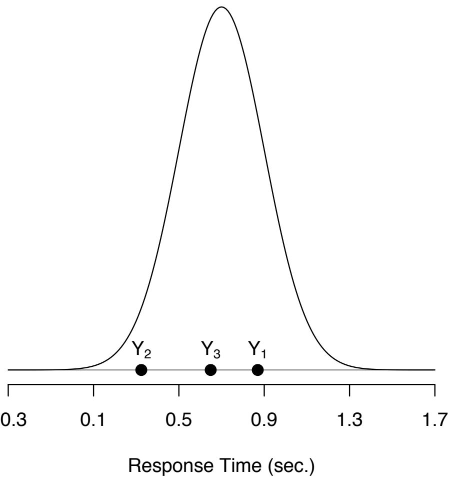

Figure 2. A normal curve representing the distribution of response times. The three dots labeled Y_1, Y_2, and Y_3 represent three random draws from this distribution.

be why mathematicians often look at their feet when talking.) Our goal in this chapter is to limit the role of mathematics. We will not be presenting detailed mathematical results, and will limit most of our discourse to verbal descriptions and figures.

The one area where we will use mathematics is in the specification of models. The term "model" has many meanings in psychology including a synonym for theory; an elaborate verbal description; a figure; and a set of mathematical equations. The more precisely specified a model is, the easier it is to assess the model's predictions and assess how well these predictions match real-world data. It is with dictum in mind that we use mathematical equations to describe models. Perhaps the most well-known model is the normal distribution. Figure 2 shows the common picture of the normal curve. This model is described by two parameters, the center, denoted μ, and the variability, denoted σ^2. The model of data is specified with mathematics as follows. Let i indicate the trial number for a set of observations. Let Y_i indicate the potential value or score of the ith observation. We can now place a model on each observation:

$$Y_i \sim \text{Normal}(\mu, \sigma^2). \tag{1}$$

The equation states that each observation is drawn from the same normal distribution with parameters μ and σ^2. The "$\sim$" denotes "is distributed as", and it means that the Y_i is variable and is drawn from a distribution, usually with parameters. Observations are independent in this notation unless otherwise noted.

Models are not limited to cases where each observation is a draw from the same distribution. Consider, for example, the case of regression. Let X_i be an independent variable. The dependence of Y_i on X_i may be modeled as

$$Y_i \sim \text{Normal}(\alpha + \beta X_i, \sigma^2), \tag{2}$$

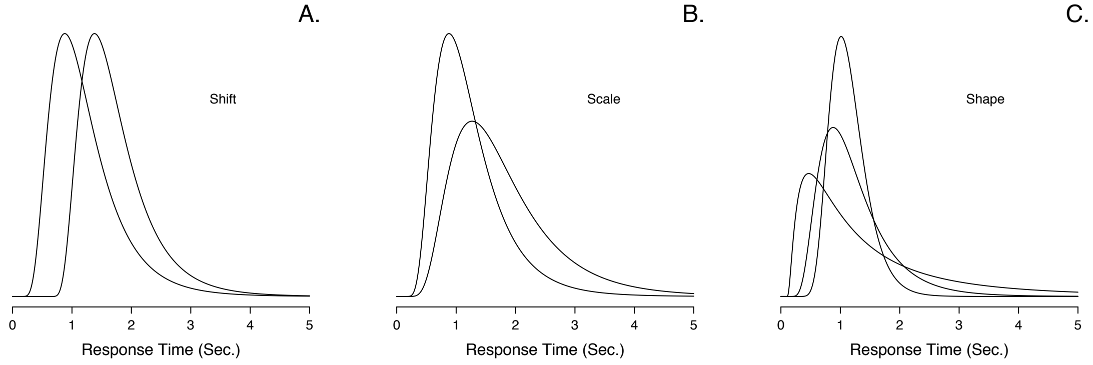

Figure 3. The Lognormal distribution. **A.-C.** Effects of varying shift, scale, and shape parameters in the lognormal model, respectively.

where parameters α and β are called the intercept and slope, respectively. Note here that the model specification is a bit different from the more common expression:

$$Y = \alpha + \beta X + \epsilon,$$

where ϵ is residual noise. We prefer the form in (2) for the following reasons: 1. The form in (2) makes the role of the normal more central, and, as will be shown, invites cognitive modelers to adapt it as needed; 2. The subscripts in (2) provide a convenient sense of what is varying in the experiment. In this case, it is clear there are a number of trials, and for each one there is a different value of the independent and dependent variable. Subscripts are the key to hierarchical model specification, and understanding how subscripts map onto different levels of variability is critical. We will stress both of these facets throughout.

Psychological process models are rarely as simple as models on the location of a normal distribution. Take, for example, response time, which is known to have a skewed right tail, and where effects that lengthen mean RT also make RT more variable. A more useful model of response time might be a three-parameter lognormal model. The lognormal is shown in Figure 3 and the parameters ψ, ν, and θ correspond in this case to shift, scale, and shape.[1] Figure 3A shows how changes in shift (ψ) correspond to literally a shift of the distribution; Figure 3B shows how changes in scale correspond to a stretching of the distribution; Figure 3C shows how changes in shape affect the skewness of the distribution. To specify a model of response time using a lognormal distribution, we may write

$$Y_i \sim \text{Lognormal}(\psi, \nu, \theta).$$

The specification of extensions to realistic cases is relatively simple. For example, let's consider a lexical decision task where participants are presented with a string of letters that

[1] The lognormal density is given by

$$f(t) = \frac{1}{(t-\psi)\sqrt{2\pi\theta}} \exp\left(-\frac{[\log(t-\psi) - \log\nu]^2}{2\theta}\right).$$

either forms a valid word, such as *CAFE*, or forms a pronounceable but invalid word, such as *MAFE*. Participants decide if the string is a valid word or not, and the time taken to do so serves as the dependent measure. Let's assume the words in the task are chosen to be high-frequency and a low-frequency condition, where frequency is the usual frequency of occurrence in standard corpora of written texts. For now, let's focus on variation across people and across frequency conditions. The response time to each word is denoted Y_{ijk}, where i indexes the participant, j indexes the condition (low frequency or high frequency), and k indexes the replicate. One model, a very general model, is that people can vary in all three parameters and that the effect of condition is in all three parameters. This most general model may be expressed as

$$Y_{ijk} \sim \text{Lognormal}(\psi_{ij}, \nu_{ij}, \theta_{ij}) \tag{3}$$

Here, the model is very general as there is a separate shift, scale, and shape for each person-by-condition combination, and the fact that there is a separate parameter for each person-by-condition combination is reflected in the subscripts on ψ, ν, and θ.

Psychologically-motivated constraint may be expressed as constraint on these parameters. For instance, in most psychological theories, the shift, ψ denotes an irreducible minimum Ratcliff (1978) that reflects the contribution of the time to encode the stimulus and the time for the motor system to execute the response. According to this interpretation, there should be variability in shift across people as different people will have different times motor and encoding times, but perhaps there should be no difference in shift across word frequency. This restriction is implemented by setting $\psi_{i1} = \psi_{i2} = \psi_i$, and this constraint is placed into the model statement:

$$Y_{ijk} \sim \text{Lognormal}(\psi_i, \nu_{ij}, \theta_{ij}) \tag{4}$$

Note that the lack of a second subscript on ψ conveniently carries the information that shifts do not vary across people.

A second constraint we may explore is whether the shape of the distributions varies across people or conditions. The shape-invariant constraint across conditions is given by $\theta_{i1} = \theta_{i2} = \theta_i$, and the corresponding model reflecting this restriction

$$Y_{ijk} \sim \text{Lognormal}(\psi_{ij}, \nu_{ij}, \theta_i). \tag{5}$$

The further restriction that shape is invariant across people is given by

$$Y_{ijk} \sim \text{Lognormal}(\psi_{ij}, \nu_{ij}, \theta). \tag{6}$$

The model we develop in this paper is the combination of shape invariance and that shifts reflect encoding and motor times. The resulting model is

$$Y_{ijk} \sim \text{Lognormal}(\psi_i, \nu_{ij}, \theta) \tag{7}$$

In this model, the effect of condition is only on the scale parameter, ν_{ij}, and the rationale is as follows. In the experimental record, the vast majority of stimulus variables affect both the mean and standard deviation of response time. For example, responses to low frequency words are not only slower but are more variable. This correlation between mean speed and variability has been noted repeatedly (Luce, 1986; Rouder et al., 2010), and has even been elevated to a "law" by Wagenmakers & Brown (2007). In the lognormal model, increases in scale correspond to both an increase in mean and a proportional increase in standard deviation, and, therefore, is an ideal location to model stimulus effects.

Even though it is relatively straightforward to specify the above models, analysis is surprisingly difficult through the usual packages. One difficulty in fact is the inclusion of the shift parameter ψ. Although this inclusion is both necessary to account for extant distributions (Rouder, 2005) and is motivated by psychological considerations (Dzhafarov, 1992; Hsu, 1999), it renders the model nonlinear and precludes analysis in the linear model framework. A second issue is that most of the models have parameters that are common across several individuals, and, therefore, cannot be fit to each individual one-at-a-time by maximizing likelihood. The models must be fit to all individuals and conditions simultaneously. These issues motivate our use of Bayesian analysis, discussed next.

The Basics of Bayesian Analysis

Bayesian Probability

Bayesian analysis starts with a different definition of probability than what most psychologists learn and teach in their statistics courses. In these courses, probabilities are conceived as proportions in the limit of much data. For example, if a coin ends up with 50% heads in the limit of all flips, the the probability of a head is defined as .5. Moreover, this probability has an objective basis, and may be thought of as a property of the coin independent of the observer, much like the weight or area of the coin. These two properties, that there is an objective probability that is a long-run proportion, form the basis of what is known as the *frequentist probability.* Although frequentist probability has the advantage of familiarity, it does lead to immediate difficulties as illustrated in the following example:

In October of 2012, many of us were watching Nate Silver's forecast for who would win the upcoming 2012 Presidential Election between President Barack Obama and challenger Mitt Romney (New York Times, five-xx blog). One statistic Silver reported was the probability that President Obama would capture more than 270 electoral votes and thus be reelected. For example, on October xx, Silver reported that the probability President Obama would be elected was .746. This prediction, however, was controversial as national polls at the time indicated an advantage for Mr. Romney. With Orwellian hypocrisy, conservative commentators argued that Silver's probability of President Obama's chances reflected a liberal ideological bias (e.g., "Morning Joe" Television Show on MSNBC, 10/29/2012). Perhaps the most interesting critique was from Dylan Byers of Politico, who argued that the real probability of Obama winning may be .501. The interesting part of Byers critique was not his justification for the .501 probability, which was an exercise in partisan propaganda. Instead, it was the implicit and perhaps unwitting criticism of the use of probability to describe one-off events such as a presidential election (Byers, 2012). Byers writes in the last line of his piece, "And even then [if Obama is reelected], you won't know if he actually had a 50.1% chance or a 74.6% of getting there." Here, it seems Byers believes that there really is a true probability of reelection that exists on October 28th as indicated by his use of the word "actually". But, he also believes that this probability, though it exists, is seemingly unattainable. Even after the fact, we will not know to any degree what this true probability is.

Byers is asking in what sense is Silver's probability estimate is applicable to the 2012 Presidential Election where there is no concept of long run or repeatability. Indeed, although probability has an intuitive appeal, and many of us find Silver's probabilities useful in the run up to the election, it is hard to understand how the traditional notion of probability may be applied in this case. Was there really a true probability of re-election for the President on October 28th, or was it just a construction in Silver's and his readers' minds?

Bayesian probability offers an elegant and intellectually pleasing alternative to traditional notions of probability. In the Bayesian framework, probabilities are statements of belief. When Silver states that the probability of Mr. Obama winning reelection is 74.6%, he is expressing a personal belief. We may even interpret Silver's beliefs in terms of wagers (de Finetti). Given these beliefs, Silver's probability of about .75 corresponds to odds of 3-to-1. If Silver holds these odds, then he might be inclined to take a bet where he loses $2.50 if Governor Romney is elected and gains $1 if President Obama is reelected as the expected value of the bet favors an overall monetary gain. Likewise, Silver might not be inclined to bet $3.50 to win a $1 as this bet is biased away from Silver beliefs. The key point here is that probabilities describe beliefs held by people rather than objective properties of objects or situations.

The key constraint in Bayesian inference is that beliefs are updated rationally or ideally in the light of new data. Before an experiment, a researcher holds a set of beliefs, and, indeed, two researchers may hold different beliefs. After the experiment, each researcher rationally updates these beliefs, and the influence of the data pushes these beliefs closer to each other and closer to the true state of affairs. Silver's beliefs are persuasive because they have been rationally updated from extant polling data.

Bayesian probability and subsequent analysis offers perhaps three main advantages to psychologists: **1.** Bayesian updating provides an appealing means of learning from data. The notion that beliefs are mutable may reflect data in ideal proportion and is attractive for scientific reasoning. **2.** Because probabilities are simply beliefs, they may be placed on data, parameters, models, hypotheses, and even theories. This ability to place probabilities on an expanded set of constructs compares favorable to frequentist probability where probabilities may only be placed on data and not on parameters, models, or hypotheses. This fact complicates model comparison for frequentist statisticians. As is discussed subsequently, model comparison in Bayesian probability follows directly from principles, and researchers may state how much they believe in or would bet on a particular hypothesis or model. **3.** There is a large class of models that are intractable in conventional settings but are straightforward to analyze in the Bayesian setting. This ease of analysis is especially true for hierarchical nonlinear models of psychological processing. We do not know how to analyze many of the models we present here outside the Bayesian framework.

Bayesian Updating

Bayesians update their beliefs optimally or rationally in light of data using Bayes' Rule. Figure 4 shows an example of Bayes Rule for accuracy where performance on a trial is coded as a success or a failure and is modeled as coin flip. The researcher's goal is to express beliefs on the underlying probability of success, and we suppose the researcher has observed 8 successes in 12 trials. Figure 4A shows a set of prior beliefs, beliefs expressed before

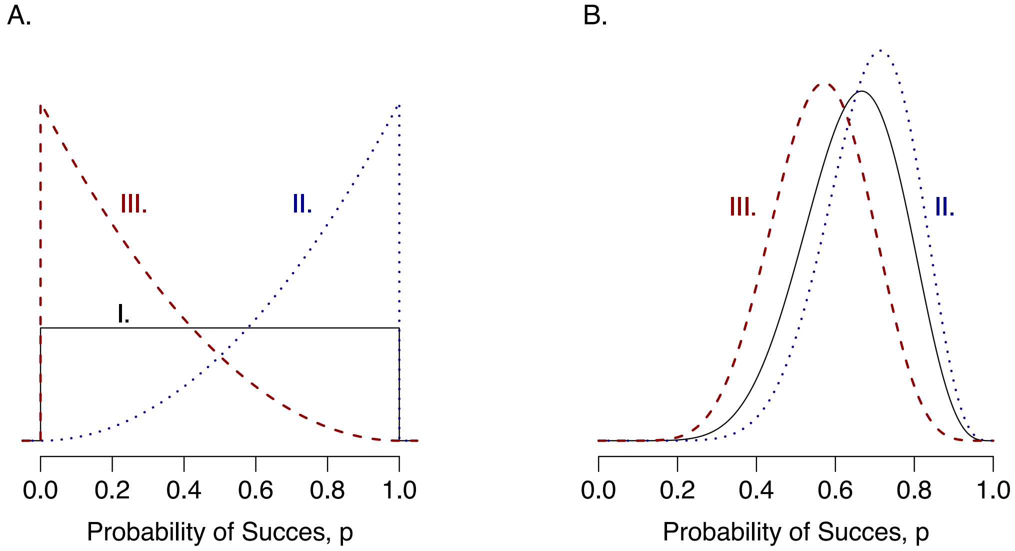

Figure 4. Prior and posterior beliefs from three analysts for the probability of success. **A.** Prior beliefs. Analyst I believes that all outcomes are equally plausible; Analyst II believes that successes are more likely than failures; and Analyst III believes that failures are more likely than successes. **B.** The updated posterior beliefs after observing 8 successes and 4 failures.

observing the data. We show the beliefs of three hypothetical researchers. Researcher I believes that all values of p are equally likely. This belief is shown by the solid flat line. Researcher II believes heads is more likely than tails, and this belief is shown by the dotted line. Researcher III believes that tails are more likely than heads, and this belief is shown by the dashed line. Bayes Rule, describes how these prior beliefs should be updated by data, which in this case are the 8 successes in 12 trials. These beliefs, called *posterior* beliefs, are shown in Figure 4B. There are three posterior distributions, one for each researcher. There are a few noteworthy points: First, the beliefs of all three analysts have been narrowed by the data; in particular, for Researcher 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 among the three researchers diverged markedly, the posterior beliefs are quite similar.

Figure 5 provides another example of updating by Bayes Rule which further elucidates the role of priors and data. 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, and we know from past research that the standard deviation of IQ across children is about 10.0. We consider two researchers that start with different prior beliefs. Researcher I is doubtful that Smarties have any effect at all, and has chosen a tightly constrained prior around 100.0, the mean value on IQ for the general population of children who presumably have not eaten Smarties. Researcher II on the other hand, is far less committal in her beliefs, and the prior beliefs have much greater variability. These choices are shown in Figure 5A.

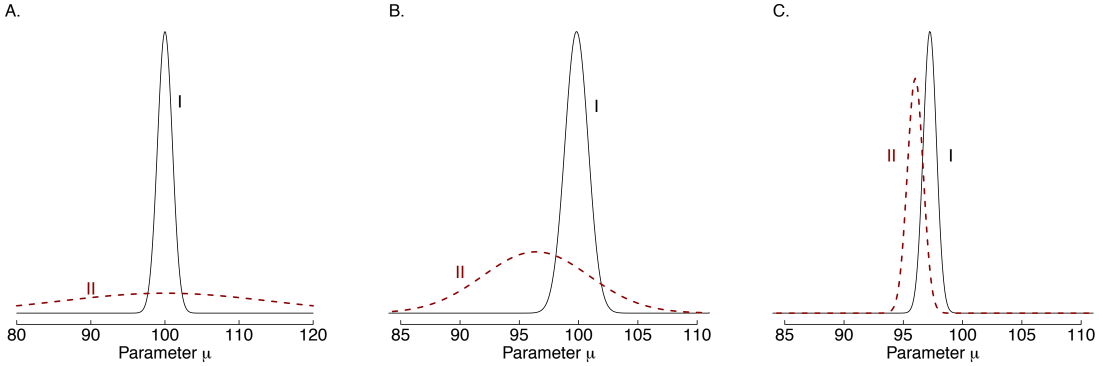

Figure 5. Prior and posterior beliefs on μ, the center of a normal distribution. **A.** Prior beliefs of two analysts. **B.** Posterior beliefs conditional on a sample mean of $\bar{Y} = 95$ and a small sample size of $N = 10$. **C.** Posterior beliefs conditional on a sample mean of $\bar{Y} = 95$ and a larger sample size of $N = 100$.

Figure 5B and 5C show the role of sample size in posterior beliefs. Figure 5B 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 5C 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 section, we demonstrated Bayesian updating from prior to posterior beliefs. We did so however for models with one parameter. In the previous IQ example, we updated μ, the center of IQ based on observations, but we did not update σ^2. A reasonable goal these is to state prior and posterior beliefs about both μ and σ^2 simultaneously. Figure 6 shows an example. Here, beliefs are expressed as a function on a plane: the input is an ordered pair (μ, σ^2) and the output is the relative belief at that point. The prior beliefs are fairly dispersed, and the posterior beliefs, in this case for 10 observations with sample mean and variance of 103.2, and 167.8, respectively, are constrained by rationale updating. Because the posterior and prior are functions of μ and σ^2 taken jointly, they are referred to as the *joint posterior* and the *joint prior*, respectively.

The joint expression of beliefs is fairly convenient for models with two parameter. But, as the dimensionality increases, it becomes difficult if not impossible to visualize the beliefs. 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 express *marginal posteriors*. The marginal posterior for one parameter, say μ, is obtained by averaging (integrating) the uncertainty in all other parameters. Marginal posteriors for the two

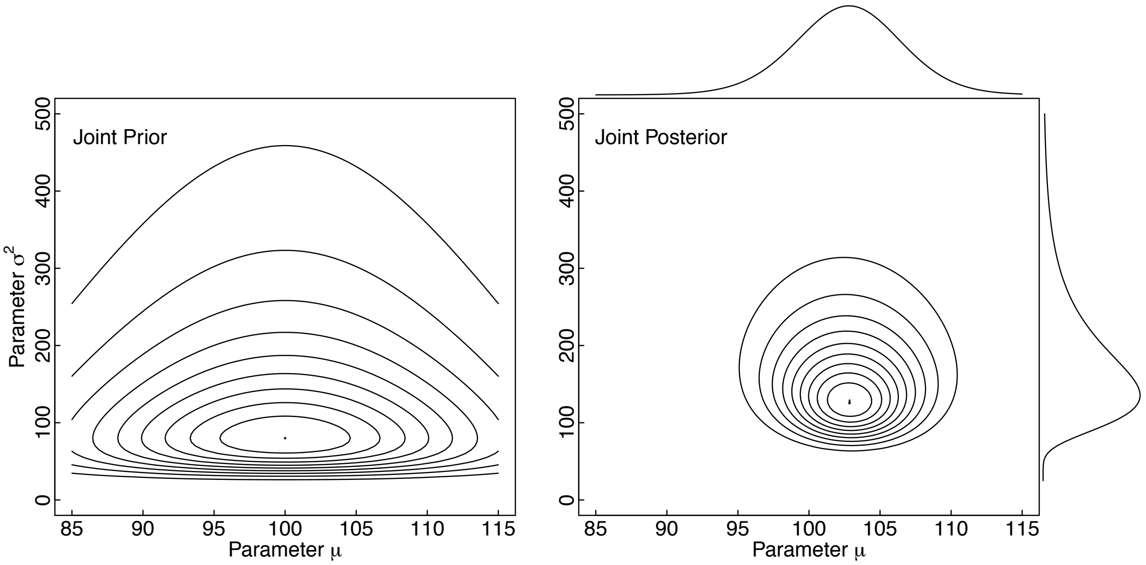

Figure 6. Joint prior (left) and joint posterior (right) distributions across normal-distribution parameters μ and σ^2. Also shown, in the margins are the marginal posterior distributions of μ (top) and σ^2 (right).

parameters are shown in Figure 6, right panel. As can be seen, these provide a convenient expression of posterior beliefs.

Although marginal posteriors are useful for expressing posterior beliefs, they are often 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 to not evaluate the joint posterior but instead to draw random samples from it. For instance, in Figure 6, one could at least in theory draw ordered pairs with probability determined by the joint posterior. We draw as many samples from the joint that is needed to characterize it to arbitrary precision. In general, each sample is a vector or parameter values. To characterized 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 $\mu^[1], \mu^{[2]}, \ldots$ characterized this distribution to arbitrary precision. So integrating the joint posterior may be reduced to sampling from it.

Directly sampling from a 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 such as Jackman (2009) and Gelman et al. (2004).

Researchers new to Bayesian analysis can use modern tools such as JAGS (Plummer, 2003) and WinBUGS (Lunn et al., 2000) to perform MCMC sampling without much special knowledge.[2] There are now a number of articles and books about using these tools including Kruschke (2011) and Ntzoufras (2009). Those with more experience can write their own code in high-level languages such as R or Matlab.

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. We develop here a series of models that demonstrate what hierarchical models are and how they may be used to improve analysis. Consider an experimenter who has asked 50 participants to make lexical decisions among high and low frequency words. For each person at each of the two levels, there are 25 replicates. In a general inquiry, we might wonder if word frequency effects are manifest on shift, scale, or shape parameters, and indeed Rouder et al. (2005) address whether covariates affect these parameter. For demonstration purposes here, we restrict the modeling of covariates like frequency to one locus, scale, and ask how scale varies across people and conditions. We start with the previously presented base model:

$$Y_{ijk} \sim \text{Lognormal}(\psi_i, \nu_{ij}, \theta).$$

Here there is a separate shift parameter for each individual, one shape parameter for all people and conditions, and a separate scale parameter for each participant-by-condition combination. We refer to this model as the *cell model* because there is a separate scale for each cell, where a cell refers to a specific participant-by-condition combination.

A diffuse prior on parameters

Prior beliefs are needed for all parameters, and in this first prior setup, we stress priors where wide ranges of parameter values are *a priori* plausible. Such priors are called *diffuse.* A few facts about RT help frame the discussion. RTs in quick, cognitive tasks like lexical decision range from about a quarter of a second to a few seconds in duration. For the shift parameter, it is possible to take a prior where all positive values are equally likely. The prior is shown in Figure 7A as the black line. This prior has an upper limit of 1.5 seconds, which is convenient to graph, but, in practice the prior on shift needs no upper limit. For the scale parameter, we chose priors that have plausibility across a wide range of scale values, say those from a millisecond to 1000s of seconds (Figure 7B, black near-horizontal line). Clearly, such a broad prior will encompass true underlying values of scale. The prior for shape is also informed somewhat by prior knowledge. Figure 3C shows distributions for three shapes, with shape-parameter values of $\theta = (.09, .25, 1.0)$. The value of .25 is about right for RT distributions; values less than .09 are too symmetric and values more than 1.0 are too skewed. We choose the distribution in Figure 7C, for which an extended range of

[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.

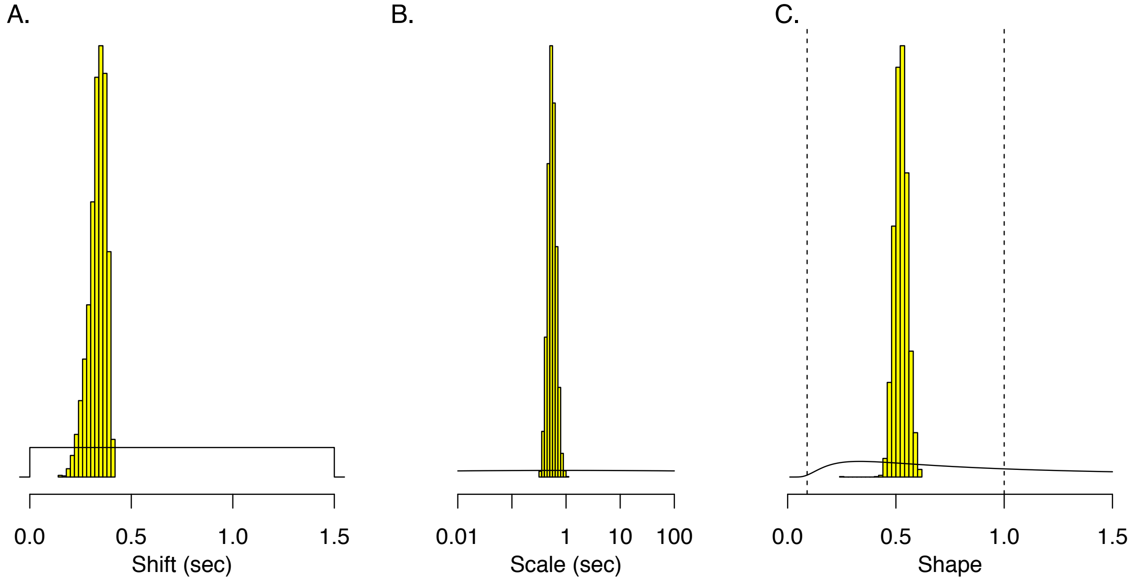

Figure 7. Prior and posterior distributions for shift, scale, and shape parameters.

values are plausible, but with the specific goal of having substantial mass over the bracketed range, which is indicated with dotted lines.

These prior beliefs may be updated using Bayes' Rule. In the above there are 50 participant-specific shifts, 100 participant-by-condition scales, and one shape, for a total of 151 parameters. The marginal posteriors for all 151 parameters are obtained through MCMC sampling. Figure 7 shows typical MCMC outputs. The one for shift is from the first participant, the one for scale is from the first participant in the first condition, and the one for shape applies to all participants and conditions as there is a single shape parameter in the model. The critical question is how have the data constrained beliefs about parameters. We may be predisposed to think that with only 25 observations per participant per condition, there is not much constraint. Yet, given the constraints in the model (a single shape, one shift per participant), and the rational updating from Baye' Rule, the resulting posteriors are fairly well localized. Even better localization through hierarchical models as discussed subsequently.

The model analysis provides insights not provided by simple statements about condition means. The model specifies that condition effects are in scale, and an overall group-level view of the condition effect on RT distributions is provided in Figure 8A. Here, we have constructed a group-level distribution by averaging parameters rather than combining data. These distributions are lognormals, and they share a common shift and shape much like for any participant. The shift value is the average of posterior means of participants' shifts; the shape is the posterior mean of the shape parameter. There are two scale values, one for each condition, and these are formed by averaging the relevant posterior mean of participant-by-condition scale parameters. The interpretation offered by the model is that the effect of condition is to slow the entire RT distribution above the shift point in a multiplicative manner. Fast and slow responses are slowed by about 14%, and this effect is greater for

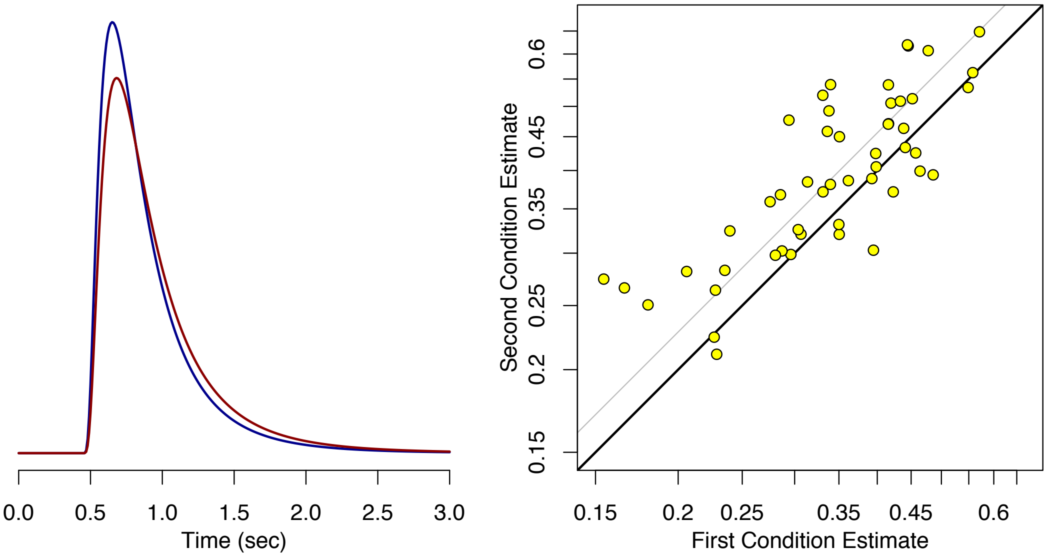

Figure 8. **A.** Group-level RT distributions show the appropriately averaged effect of condition. **B.** The condition effect on scale for each participant. The thick line shows no effect while the thinner shows a 14% increase in scale across conditions.

slow responses in absolute terms because 14% of a larger duration is a larger effect. The stability of the condition effect is shown in Figure 8B. Here, the posterior mean scale values are plotted for each participant as a scatter plot. The x-axis and y-axis values are for the two conditions, and the solid line is the diagonal indicating no condition effect. Most participants have a higher scale for Condition 2 than for Condition 1, although there is some variability. The thinner line corresponds to a 14% increase between Conditions 1 and 2, and if the scale of each participant increased by 14%, then the points would lie on this line.

A simple hierarchical prior

The lognormal cell model offers many advantages. Although the diffuse priors are useful, there are better choices to explore the structure in the data. We focus on the prior on scale: with the diffuse prior beliefs in Figure 7B, it is possible that one scale parameter could be very small, say 1 ms, while another could be quite large, say 1000 sec. In fact, we can use the same model notation to express prior beliefs; in fact, priors are models on parameters. The previous prior on scale, ν_{ij} was

$$\log \nu_{ij} \sim \text{Normal}(0, 132), \tag{8}$$

where the value of 0 is the mean and the value of 132 is the variance. We chose the value 132 because it encompassed a large range of scales but was still able to be seen in Figure 7B. In practice, there is no requirement that priors have enough mass to be seen, and far more diffuse values may be used as well.

Even though the prior is diffuse, there is an element that seems implausible. It seems difficult to believe that scales can really vary by several orders of magnitude. It seems unrealistic that one participant can have a characteristic scale of 1 ms and another of 1000 sec. Hierarchical priors may be used to limit a priori plausibility in smart ways. In a hierarchical prior, parameter values are modeled as coming from a common parent distribution. One such prior is

$$\log \nu_{ij} \sim \text{Normal}(\mu, \sigma^2), \tag{9}$$

where μ and σ^2 serve as parent parameters that describe the center and dispersion of the population of scales. These parent parameters need not be fixed *a priori*. Instead, they may be treated as parameters themselves upon which we may place priors and compute posteriors. Consider the following priors for parent parameters

$$\begin{aligned} \mu &\sim \text{Normal}(0, 132) \\ \sigma &\sim \text{Uniform}(0, 100) \end{aligned}$$

Here, we bring little if any *a priori* information about the population center and dispersion of scale effects as 132 and 100 are quite large numbers as exponents and encompass many orders of magnitude. In effect, all we commit to is that the effects themselves are similar in so much as they are samples from a common parent distribution without placing substantial *a priori* constraint on the parent distribution itself.

The hierarchical model allows information to be shared across participants and conditions. Here, all the participant-by-condition scales influence the localization of parent parameter, and these parent parameter serve in turn as constrained priors. Figure 9A shows the effects of this hierarchical structure. Scale estimates with the hierarchical structure as a function of that from the cell model. As can be seen, values that were extreme in the cell values for this hierarchical model are moderated; that is, they are modestly pulled toward the center. This effect is termed *hierarchical shrinkage*, and it is a good thing. The scale estimates in the cell model are variable because there are only 25 replicates per cell. The hierarchical structure is smoothing this variability, and this smoothing leads to posterior estimates that have lower error than nonhierarchical estimates (Efron & Morris, 1977; Jackman, 2009; Rouder & Lu, 2005).

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 model has 100 scale parameters and a variance parameter; the hierarchical model has these 100 parameters and additional parent mean and variance parameters. Yet, the cell model is more flexible as the 100 scale parameters are free to vary arbitrarily. In the hierarchical model, no one scale 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 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, μ and σ^2 in the above example, can be generalized to other participant-by-condition combinations.

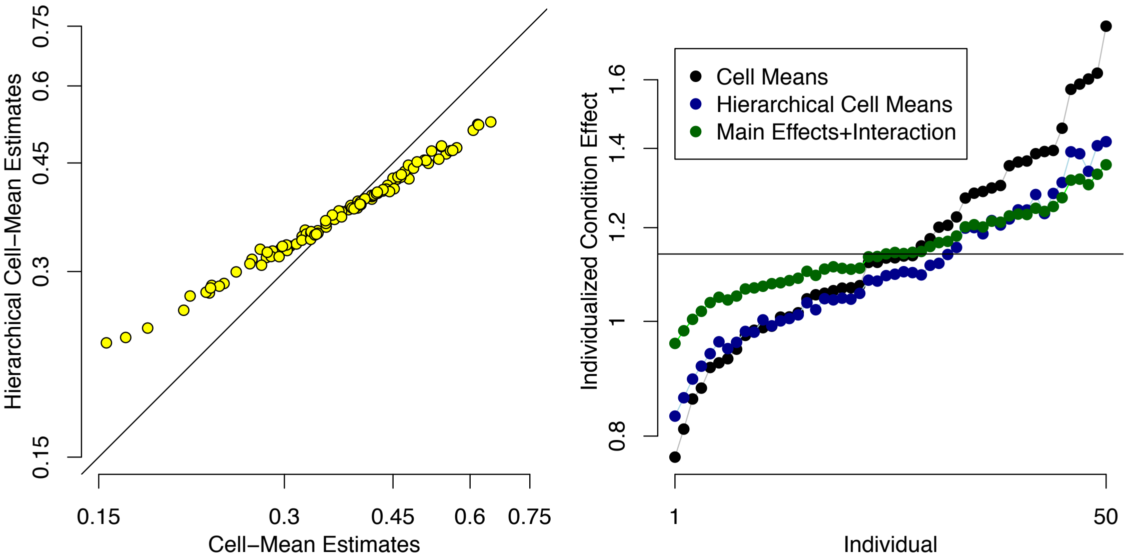

Figure 9. **A.** Parameter shrinkage resulting from hierarchical priors. Shown are posterior means of scale parameters with the hierarchical prior as a function of those in the cell model. **B.** Condition effects for individuals from three models. The effects are largest for the cell model because there is no smoothing. The unstructured and structured hierarchical models provide progressively more smoothing. The straight line at 1.14 shows a constant 14% difference in scale from the additive hierarchical model.

These parameters, therefore, provide a means of applying the results more broadly. Second, more advanced models may be placed on μ that includes subject and condition variables. In fact, given that this model does not place any models on participant or condition effects, it may be considered an unstructured hierarchical model. The following models add more structure.

A structured hierarchical prior with main effects and interactions

The above hierarchical prior is a conceptual step forward. Yet, it too may be improved. The shrinkage of each cell, that is, each participant-by-item combination, is to an overall grand mean. This shrinkage is crude if we think that people differ. We should be shrinking each cell to a person-specific mean, and the same holds true for condition effects as well. The following prior on log scale adds this specification:

$$\log \nu_{ij} \quad \sim \quad \text{Normal}(\alpha_i + \beta_j, \sigma^2). \tag{10}$$

This prior may be compared to that in (9), and the difference is in the center of the normal. Before there was a single center, μ, but here the center is the addition of participant and condition effects. The parameter α_i denotes the main effect of the ith participant on scale and parameter β_j denotes the main effect of the jth condition. There are even interactions implicit in the prior; they come about because $\log \nu_{ij}$ does not equal the sum of the main effects; instead it is just centered around them. The deviations from these centers are the interactions, and σ^2 is therefore a measure of the overall size of the interaction terms. In this model, shrinkage is toward the sum of main effects, that is outlying estimates of $\log \nu_{ij}$ are shrunk toward the sum of the participant effect and condition effect for that combination.

Priors are needed on α and β. We start with the condition effects:

$$\beta_j \sim \mathrm{Normal}(0, 132). \tag{11}$$

Because the variance is set large, at value 132, there is no shrinkage. We are letting condition effects estimate to what they may, but because there are only two of them, the estimates for each are affected by the totality of all participants data and are not likely to be effected by idiosyncratic noise in any one cell. The priors on individuals are quite a bit different:

$$\alpha_i \sim \mathrm{Normal}(\mu_\alpha, \sigma_\alpha^2). \tag{12}$$

Here, there is a parameterized parent distribution that will provide shrinkage. The rational is the same as before—participants cannot be arbitrarily different from one another, that is, it is unrealistic to think that one participant has a scale of 1 ms while another has a scale of 1000 sec. Instead, all participants are drawn from a parent distribution with parameters estimated by the data. The effect of this specification is two fold: First, participant effects will be shrunk toward the grand mean μ_α. Idiosyncratic noise in each participant's data will not be overweighted, and group level information is used in making estimates. Second, participants are modeled as random effects. The estimation of a parent distribution provides a means of generalization to a larger population of all potential participants. Potential participants have scales that enter on μ_α with variability described by σ_α^2. In contrast, the condition effects are modeled as fixed, and no information about other potential conditions may be inferred. Priors are needed for parent parameters μ_α and σ_α^2, and these may be made diffuse as before to convey little prior information about the parent distribution.

The effect of this prior structure is shown in Figure 9B. The plot is the condition effect per person. The condition effect for the ith person is given by ν_{i2}/ν_{i1}, which measures the scale larger for the ith person in Condition 2 relative to that in Condition 1. If the scale for all people is 14% larger in the second than first condition, then $\nu_{i2}/\nu_{i1} = 1.14$. The black points and line are from the cell model above where the priors were unstructured. Each point is from a different participant, and the participants are ordered from the smallest value of the ratio to the largest. For this series, the plot is a re-expression of the information in Figure 8, that for most individual's, the effects are larger scale in Condition 2, that there is moderate variability, and that the effects are centered on a 14% increase. The variation in this ratio is the interaction, that is, it is the variability of the condition effect across individuals. The blue and green points show the same ratio for the other models. As can be seen, the cell hierarchical model provides some shrinkage of the ratio. Because shrinkage is to the overall mean rather than to main effects, the shrinkage of the ratio is modest. The interaction model, however, has a large degree of shrinkage in the ratio—the hierarchical structure smooths the condition effect across participants. Here, the additive structure in the prior stabilizes the ratio estimates so it is not unduly influenced by noise in the cells.

One critical question is whether these shrinkage estimators shrink too much. It turns out not. As a matter of statistical theory, they improve estimation accuracy by borrowing information so long as the prior structure is approximately reasonable (Efron & Morris, 1977; Jackman, 2009). In our case, the hypothetical data were generated with no variation across participants and the structued model captures this truth best. Had we generated the data with variation, that is with a true interaction effect, the estimates would not shrink so thoroughly and would be closer to the cell model values.

Additive and Null Priors

If we were modeling experimental data, we would start with the prior structure in (10) because of its flexibility. The model with this prior not only provides posterior beliefs about shifts and shape, but it provides separate participant scale estimates, condition estimates, group parameters for generalization, and a measure of the stability of the condition effects across individuals. Nonetheless, simpler models that capture theoretically interesting positions are possible. The first is that there is no interaction, that is, the condition effect is constant for all people. This model may be implemented with a decomposition of cell scales into participant and condition effects:

$$\log \nu_{ij} = \alpha_i + \beta_j, \tag{13}$$

where priors on α and β parameters follow the specifications in (12) and (11), respectively. The difference between this model, without any interactions, and the previous one is seen by contrasting (10) and the above one. Figure 9B shows the comparable ratio, $\exp(\beta_2 - \beta_1)$. The solid line is the posterior mean of this ratio, and, as expected, it describes the average of log ratios. It is difficult to tell from looking at the parameter estimates whether there are interactions of not, and we approach the problem of comparing the evidence for the more general hierarchical model or the more constrained additive one in the next section. We also consider a prior that encodes no condition effect:

$$\log \nu_{ij} = \alpha_i. \tag{14}$$

The prior on participant effects are the same as previous. This prior, however, seems unlikely. Under this model, with no condition effects, the ratios in Figure 9B should be 1.0. We have added two dotted lines which show the 95% credible region on the ratio from the additive model. The bulk of posterior beliefs is quite far from the value of 1.0, which in this case serves as counter evidence to the model with no effects.

Comparing Hierarchical Models

The above analysis shows the usefulness of embedding constraints through prior structure. The most general model is the one in (10) because it models main effects and interactions. A restriction of it with consistent condition effects across participants is given in (13), and a restriction of both of these models with no condition effect is given in (14). In a real-world setting, it would be useful to formally compare these models. One could estimate the size of interaction or main effects, but that answers quite a different question than which model best captures the structure in the data.

Up to now, we have made a distinction between models on data, such as that in (7) and the specification of priors on the parameters, such as in (10). This distinction is useful for pedagogical purposes only, and going forward, we will dispense of it. The term *model* is used in an expanded connotation to refer to both the model on data and the specification of priors, or the model on parameters. The rationale is simple—because the prior encodes useful statements about the structure in the data, say whether effects are consistent or not across participants, it should be treated as part of the model. This expanded usage of the term *model* is common in statistics and is helpful for model comparison. We refer to the base model, the one with the base prior in (10) as $\mathcal{M}_1$; we refer to the model with the consistent conditions-effects in (13) as $\mathcal{M}_{add}$, and we refer to the model with no condition effect (14) as $\mathcal{M}_0$.

Model comparison is natural in the Bayesian framework. Bayesians may hold beliefs about models, and then may update these beliefs rationally in light of data. For example, if we want to express that we feel fairly certain that model $\mathcal{M}_{add}$ holds, we might express this belief as a probability statement, say $Pr(\mathcal{M}_{add}) = .9$ meaning that we might risk $9 to make $1 if Model $\mathcal{M}_{add}$ is indeed true. Yet, in most cases, placing probabilities on models is a bit strong because models might fail in uninteresting ways. One uninteresting misspecification might be that RT does not exactly follow a lognormal. Even with this misspecification, the comparisons between $\mathcal{M}_{add}$, $\mathcal{M}_1$ and $\mathcal{M}_0$ remains useful for assessing the existence and consistency of condition effects across participants. To avoid the problem of stating beliefs on models that may fail for unimportant or uninteresting reasons, it is common to express beliefs about one model relative to another. These relative beliefs may be expressed as odds ratios, e.g., $Pr(\mathcal{M}_{add})/Pr(\mathcal{M}_1)$, and they express how much more believable $\mathcal{M}_{add}$ is than $\mathcal{M}_1$ is.

These relative beliefs may be updated rationally in light of data using Bayes Rule. Let Model $\mathcal{M}_a$ and $\mathcal{M}_b$ refer to any two models. The equation to update relative beliefs is

$$\frac{Pr(\mathcal{M}_a \mid \text{Data})}{Pr(\mathcal{M}_b \mid \text{Data})} = \frac{Pr(\text{Data} \mid \mathcal{M}_a)}{Pr(\text{Data} \mid \mathcal{M}_b)} \times \frac{Pr(\mathcal{M}_a)}{Pr(\mathcal{M}_b)}. \tag{15}$$

The term $Pr(\mathcal{M}_a \mid \text{Data})/Pr(\mathcal{M}_b \mid \text{Data})$ is the *posterior odds*, the updated beliefs in light of data, and the term $Pr(\mathcal{M}_a)/Pr(\mathcal{M}_b)$ is the *prior odds*, the beliefs before observing data. The term $Pr(\text{Data} \mid \mathcal{M}_a)/Pr(\text{Data} \mid \mathcal{M}_b)$ is the updating factor—it describes how the data have changed the beliefs. This term is called the *Bayes factor*, and quantifies the relative evidence from data for competing models. Let B_{ab} denote the Bayes factor of Model $\mathcal{M}_a$ relative to Model $\mathcal{M}_b$. 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 relative support, respectively, for model $\mathcal{M}_a$ and $\mathcal{M}_b$. Bayes factors provide a principled method of inference, and advocacy in psychology is provided by Edwards et al. (1963); Gallistel (2009); Myung & Pitt (1997); R. D. Morey & Rouder (2011); Rouder et al. (2009); Wagenmakers (2007) among others.

The terms in the Bayes factor are $Pr(\text{Data} \mid \mathcal{M}_a)$ and $Pr(\text{Data} \mid \mathcal{M}_b)$. These terms may look innocuous enough, but evaluating them often presents substantial difficulties, especially in nonlinear hierarchical models. The key problem is that conditioning a probability on a model implies considering all possible values the parameters may take. This consideration is expressed as an integral over the parameter space, which is often of high dimensionality. For example, in some of the models we present in the next section, there are parameters for each participant and item, and this leads to a parameter space that spans thousands of dimensions. To make matters worse, the probabilities conditional on parameters are highly concentrated, and the integrand is highly peaked. The integration often becomes a matter of finding a needle in a large-dimensional haystack. Not surprising, the efficient computation of Bayes factors, especially in hierarchical models, remains challenging and topical in Bayesian analysis.

Though the integration problem is difficult, in some contexts, many of the parameters may be integrated out in analytical form. This sanguine case holds for linear models with normally distributed errors such as those behind t-tests, ANOVA, and regression. Seminal development was provided by Jeffreys (1961) and Zellner and Siow (1980). 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 Bayes factor development for linear models 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 computational approaches to the integration problem, and we mention them only 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 et al., 2001), and a new variation on importance sampling termed direct sampling (Walker et al., 2011). These techniques will undoubtably 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 et al. (2011); Wagenmakers et al. (2010); Wetzels et al. (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. (2012) uses it to discriminate between the power law and exponential law of learning in hierarchical settings. In the next section, we provide a similar example for understanding how lexical access occurs by studying how reading times are affected by word frequency covariates.

We used the Savage Dickey approach to compute the Bayes factors for the lognormal models presented above. The preferred model for the data in Figure 9 is the additive model $\mathcal{M}_{add}$. It is preferred to the more complicated interaction model, $\mathcal{M}_1$, by a factor of 548-to-1, indicating that the condition effect is the same across participants. Whether this consistent condition effect is zero or not is assessed by the Bayes factor between $\mathcal{M}_{add}$ and $\mathcal{M}_0$, and it evaluates to over 14,000, indicating overwhelming evidence for a condition effect. These effects while accounting for the fact that individuals have characteristic scale multipliers as well as characteristic shifts.

An Application: Word Frequency in Lexical Decision

To illustrate the advantages of Bayesian hierarchical modeling, we provide an application for the process of lexical access during reading. In the course of reading, printed strings on the page must be matched to words in the lexicon to build semantic meaning. Coarsely speaking, theories of access ascribe either a serial search through an ordered mental lexicon or a parallel search through all words in the lexicon. An example of a serial search theory is that from Murray and Forster (2004), who theorize that words are ordered in the lexicon by their frequency of occurrence in the natural language. Accordingly, the presented word might be matched against the word *chair* before it is matched against the word *conquer* because *chair* is more common in everyday usage than *conquer*. An example of parallel search comes form Morton's (1969) logogen model, where each word in the lexicon has an associated activation. These activations grow if the presented word matches the represented word, and the represented word is accessed if its activation exceeds a criterion. The logogen model also accounts for a word-frequency effect by positing that high-frequency words have a lower criterion for access than low-frequency words have. Hence, less activation is needed to access these high-frequency words.

One conventional approach to distinguish these types of models is to study the fine details of response-time distributions in lexical decision tasks. In these tasks, participants are asked whether strings are valid words or not (Meyer et al., 1975). Examples of invalid words, called *nonwords*, are *bunky* and *mafe*. The key target of inquiry are the response times to the words, and in particular, how they vary with word frequency. According to the serial access theory, the key covariate is the word frequency rank. Words are ranked from highest frequency to lowest, and this rank score serves as the relevant covariate because the first word is searched first, the second is searched second, and so on. Alternatively, parallel access models are assumed to be sensitive to the metric value of word frequency itself rather than rank score.

Assessing whether RT is more sensitive to rank scores or the frequency values is difficult because (i) these covariates are highly correlated, (ii) response times are neither symmetric nor homogeneous in variance, and (iii) there is much nuisance variation when data are collected across many people and items. Figure 10 shows these difficulties; plotted are mean-response times across the words (items) from Gomez et al.'s Experiment 1. The various lines are best fits from a few functional forms, and, as can be seen, there is much mimicry. The plots also show the lack of homogeneity, item means for a frequency of 1 are far more spread than those for a frequency of 20 (the bars show interquartile ranges of item means). Finally, it is unclear if misfits are influenced by systematic item or participant effects. For example, variability of the shift parameter across people will assuredly affect which functional form appears to fit better.

Model Specification

To better understand the word frequency effect in lexical decision, we use the hierarchical lognormal RT models to analyze the Gomez et al.'s data. Our approach is similar to that of Rouder et al. (2008), who used a Weibull model in a similar vein. The Weibull used

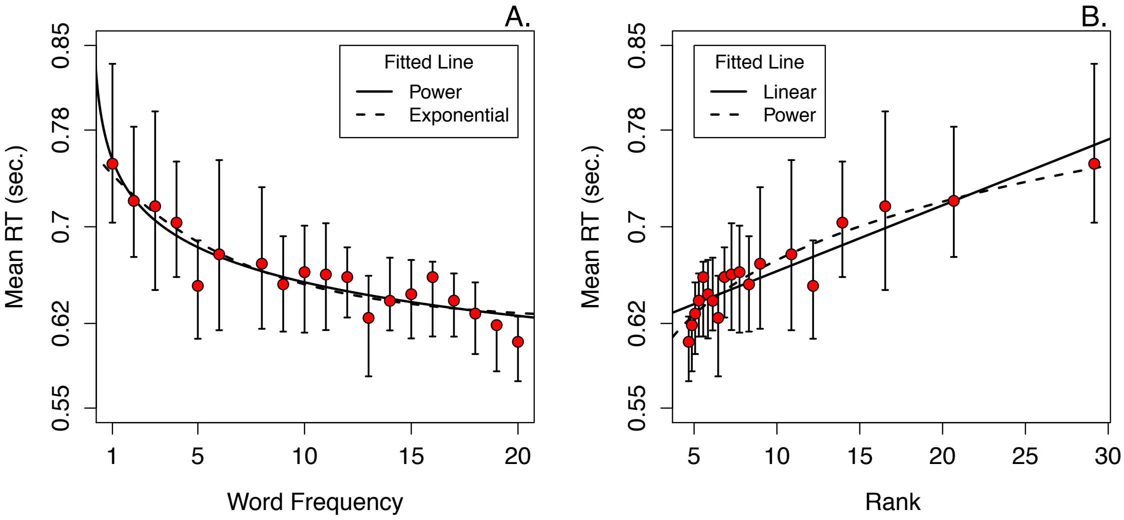

Figure 10. **A.** Mean RT (circles) plotted as a function of word-frequency. The vertical bars denote the interquartile range of item means, and these ranges clearly violate homoscadicity. The solid and dashed lines show the best power and exponential fits to these means, respectively. As can be seen, these fits are highly similar. **B.** Mean RT and interquartile range plotted as a function of the rank scores of word-frequency. The serial model of Murray and Forster (2004) predicts that mean RT is linear in rank and the solid line is the best fitting linear function for these data. The dashed line shows the best fitting power function alternative to this linear prediction.

by Rouder et al. has tails that fall off more rapidly than empirical RT distributions. The lognormal is a better choice because it has heavier tails that better match the tails of empirical distributions. Let RT_{ij} denote the response time for the ith participant and jth item. We start with a base model:

$$RT_{ij} \sim \text{Log-Normal}(\psi_i, \nu_{ij}, \theta).$$

The main target of inquiry is the scale parameter ν_{ij}, and the critical question is whether changes in scale are better accounted for word frequency or word-frequency ranks. We explore these scale effects by placing regression models on the logarithm of scale, much as we placed additive models on the logarithm of scale in the preceding sections. We construct separate regression models with frequency and rank serving alternatively as covariates, and then compare these models by Bayes factors. Our base regression model is

$$\log \nu_{ij} = \alpha_i + \beta_i x_j + \gamma_j \tag{16}$$

where x_j is a covariate for the jth item, say word frequency. The parameters α_i and β_i describe the ith person's performance. Parameter α_i is an intercept that describes the participant's overall speed; parameter β_i is a slope that describes the effect of the covariate on the participant's speed. Parameter γ_j serves as a residual for the jth item. It captures systematic item effects that are not captured by the covariate. For example, there is a known word-length effect in lexical decision. Long English words are classified as valid more slowly than short ones. The residual γ_j captures this and other systematic variation of items not associated with word frequency. In our subsequent analysis, word frequency accounts for about a quarter of all systematic variation across items.

The main target of inquiry is how slope varies with contrasting covariates. One covariate is Kucera-Francis word frequency. Let w_j denote the word frequency of the jth item. Then, x_j may be set to $(w_j - \bar{w})$, the zero-centered deviation in word frequency. Such a model where log-scale is linear in word frequency corresponds to an exponential law where mean RT follows the form $a + b\exp(-cw_j)$. Hence, we call the model where $x_j = (w_j - \bar{w})$ the *exponential-frequency model.* The word-frequency effect, captured by β in the exponential frequency model describes the relative change in scale for each additional occurrence in the language. For example a slope value of -.2 means there is 18% decrease in scale for each additional occurrence. An alternative is the *power-frequency model.* If $x_j = (\log w_j - \overline{\log w})$, then mean RT follows the power law given by the form $a + bw_j^{-c}$. Here, the slope describes the relative change in scale for each doubling of the rate of occurrence; for example, a slope value of -.2 means there is 13% decrease in scale for each doubling of word frequency.

The power-frequency and exponential-frequency models may be reduced in complexity. One possibility is that while people may vary in intercept (overall speed), they may not vary in slope (word-frequency effect on speed). Indeed, Rouder et al (2008) found that the word-frequency effect hardly varied across participants; they ascribed the same 11% decrease in scale per doubling of frequency across all participants. Such constancy denotes a theoretically important constraint, and it is accounted for by using a single slope rather than individual slopes. The corresponding model on log-scale is $\log \nu_{ij} = \alpha_i + \beta x_j + \gamma_j$. Therefore, we retain two power-frequency models, one with individual slopes and one with a common slope, and we retain two exponential-frequency models, one with individual slopes and one with a common slope.

The other covariate under consideration is the logarithm of the word-frequency rank. The serial model of Murray and Forster is implemented by setting $x_j = (\log r_j - \overline{\log r})$, and by setting β_i to 1.0 for all people. In this case, mean RT follows a lineaer form $a + br_j$ where b is the duration per item. We refer to this model as the *serial model.*

We also included a model without covariates given by $\log \nu_{ij} = \alpha_i + \gamma_j$. This model has fewer parameters than the above models, and all word frequency effects may be manifest in the item residual term γ_j. Word frequency effects if present inflate the posterior variance of these item residuals. We refer to this model as the *unstructured model* as effects are allowed but are not specifically modeled. This model is similar in spirit to the cell-means model in the previous section.

Priors are needed for regression parameters, and we use hierarchical priors to smooth participant and item effects:

$$\begin{aligned}
\alpha_i &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \\
\beta_i &\sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \\
\gamma_j &\sim \text{Normal}(0, \sigma^2_\gamma)
\end{aligned}$$

Results

The collection of 6 models described above were each analyzed using the MCMC methods previously outlined. The resulting Bayes factor model-comparison statistics were large in

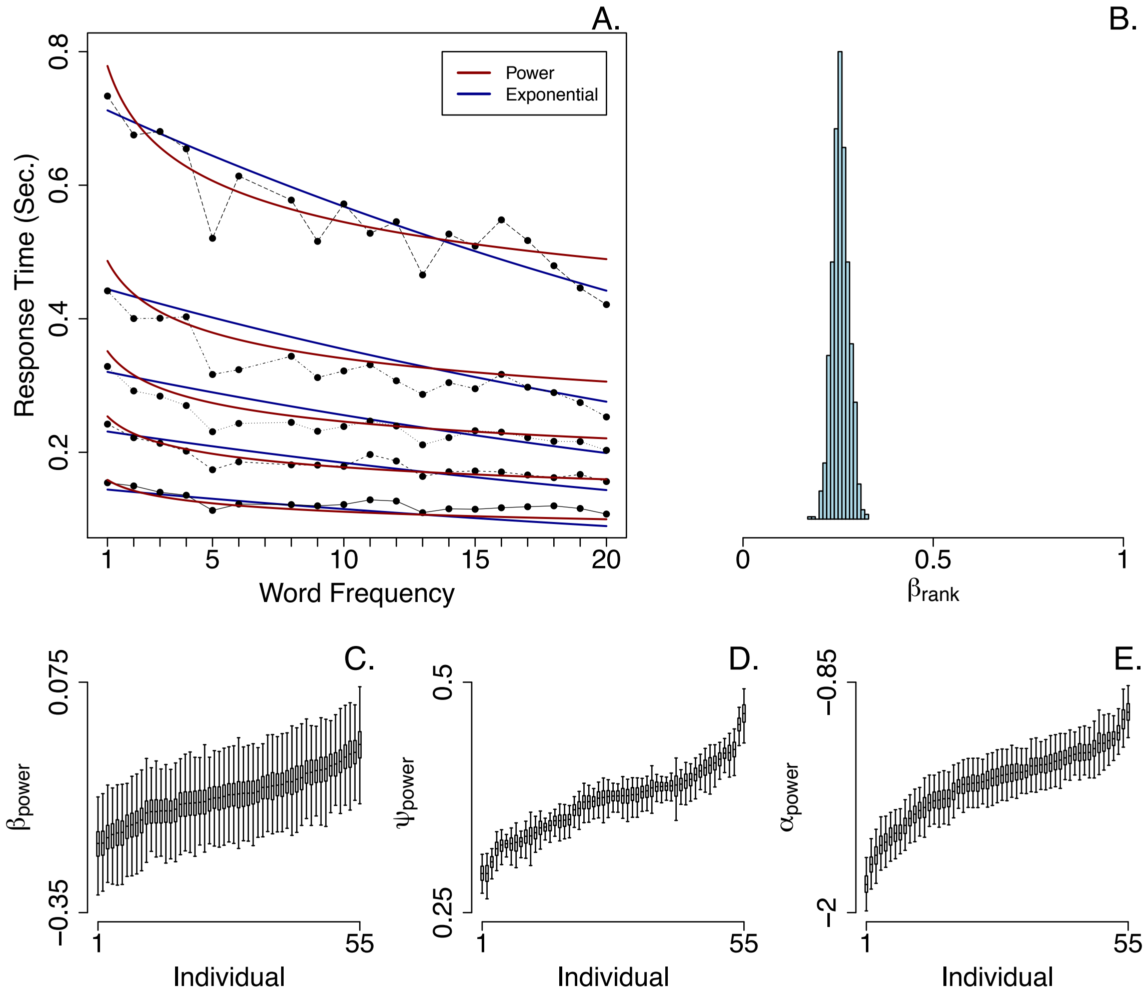

Figure 11. A graphical presentation of trends from the Bayes-factor analysis. **A.** Response time distributions after subtracting shift as a function of word frequency. Shown are the observed 10th through 90th percentile scores (black circles) as well as predicted values from the power-frequency (dark-red lines) and exponential-frequency (dark-blue lines) models. **B.** The poor performance of the serial model may be seen by estimating the slope with the rank covariate rather than setting it to 1.0 . The histogram displays the posterior distribution of slope, and there is virtually no mass near 1.0. **C.-E.** Posterior distibutions of individual slope, shift, and intercept parameters, respectively, for the power-frequency model. The whiskers, boxes, and dissecting lines indicate, respectively, interior 95% density regions, interquartile ranges, and medians of the posterior distribution for each parameter. Although individuals may vary in shift and intercept, the slope (word-frequency effect) appears nearly constant across individuals.

value reflecting that the analyzed experiment encompassed about 10,000 observations. The preferred model, that with the highest Bayes factor, was the power-frequency model with a common slope. This model had a Bayes factor that was about 11 orders-of-magnitude (100 billion) times larger than the second-most preferred model, the exponential model with a common slope. The power-frequency model with individual slopes fared about 13 orders-of-magnitude (10 trillion) worse than the winning power-freqeuncy law with common slope. The unstructured model was about 106 orders worse than the winning model, and the serial model fared worst of all, about 340 orders worse than the winning model.

With such large Bayes factors, the indicated structure should be obvious in the data, and Figure 11 shows the supporting trends. One finding is better performance for the power-frequency model than the exponential-frequency model. This performance difference may be seen in Figure 11A. Plotted are the 10th through 90th percentiles of the response time above shift as a function of word frequency. The degree of effect is greatest for the slowest responses and smallest for the fast ones, and the exponential can only account for this pattern with a fairly linear decline in word frequency. Hence, it tends to miss some of the curvature in the plots. The power law does better, and it is this improvement that drives the Bayes factor. A second finding is the abysmal performance of the serial. To confirm this level of performance, we fit an additional hierarchical RT model where the covariate was log rank ($x_j = \log r_j - \overline{\log r}$) and the slope was common but free to take on values other than 1.0. This model implies a power law on mean RT with rank though the serial interpretation no longer holds except when slope β is 1.0. The posterior distribution of the slope is plotted in Figure 11B. There is virtually no posterior mass near 1.0, indicating a gross violation of linearity, which in turn drives the exceedingly low Bayes factor. The final finding is that there is a single common slope across all participants and Figure 11D.-E. displays this result graphically. Posterior distributions of individual slope parameters are represented as box plots, and ordered from smallest slope to largest. Of course we expect some difference, but the range of differences is relatively small compared to the posterior spread. For comparison, we also made the same plots of individual variation in shift (ψ_i), and intercept (α_i). As can be seen, the range of variation across people is much greater than it is for slope. It is this diminished variation across people in slope that is driving the higher Bayes factors for common-slope models. This last result is consistent with the findings of Rouder et al. (2008) who used a Weibull rather than a lognormal model to assess the commonality of word-frequency effects across people.

Our final takeaway message is that lexical access follows a power-law function with common slope where RT scale decreases 13% for every doubling of word frequency. Descriptions like these are not possible without Bayesian hierarchical modeling.

Conclusions

Experimental psychologists have rich theoretical and experimental traditions. Nonetheless, testing psychological theory in real-world contexts is often difficult. 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 may distort the assessment of the underlying cognitive signatures and may lead to erroneous conclusions about theory. The solution is to jointly model the cognitive process of interest and and variation in this process across people, items, and conditions. These type of models are termed hierarchical, and hierarchical modeling allow researchers to uncover the rich cognitive structure and document how this structure varies across people, items, and conditions.

We have stressed in this chapter Bayesian hierarchical nonlinear models. The models are nonlinear because they capture psychological processes conceptualized as more than the sum of effects and noise. They are hierarchical to capture multiple sources of variability. And they are Bayesian because Bayesian probability is not only convenient in hierarchical settings, but form a coherent and intellectually rigorous system for understanding uncertainty. Our approach here has been to stress some of the foundational properties of Bayesian analysis, the meaning of hierarchical models, and the interface between the two. We have avoided the nuts and bolts of estimation and model comparison, and this avoidance leaves open the question of how interested researchers can develop and analyze their own models. There are now several excellent texts that cover Bayesian hierarchical models. Texts for advanced readers are Gelman et al. (2004) and Jackman (2009); one for beginners is Kruschke (2011). More recently there has been tutorials and texts specifically for psychology including Rouder & Lu (2005), Kruschke (2011), and the forthcoming book by Lee & Wagenmakers (2013). Bayesian model comparison remains thorny, but we believe that inference by Bayes factors will become more widely accepted. Readers wishing to understand more about Bayes factors will benefit from Gallistel (2009) and Wagenmakers (2007), and our work on the subject may be of interest as well (e.g., Morey & Rouder, 2011; Rouder et al., 2009; Rouder et al., 2012; Rouder & Morey, 2012).

Understanding hierarchical modeling and Bayesian analysis is no easy task. Substantive psychologists who are not methodologists will have to retrain to some degree. There will be an investment. Yet, what is at stake is a fundamental view of how we draw inferences about theory from experimental data, and this stake is well worth the investment.

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.

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

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.

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.

Dzhafarov, E. N. (1992). The structure of simple reaction time to step-function signals. *Journal of Mathematical Psychology*, *36*, 235-268.

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

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

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.

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.

Gomez, P., Ratcliff, R., & Perea, M. (2007). Diffusion model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*, 389-413.

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.

Hsu, Y. F. (1999). *Two studies on simple reaction times: I. On the psychophysics of the generalized Pieron's law. II. On estimating minimum detection times using the time estimation paradigm. Unpublished doctoral dissertation. University of California, Ir.*

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

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

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

Kruschke, J. K. (2011). *Doing Bayesian analysis: A tutorial with R and BUGS.* Academic Press.

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`

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

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

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

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

Meyer, D. R., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Local contextual effects on visual word recognition. In P. Rabitt & S. Dornic (Eds.), *Attention and performance V.* New York: Academic.

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

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Murray, W., & Forster, K. I. (2004). Serial mechanisms in lexical acces: The rank hypothesis. *Psychological Review*, *111*, 721-756.

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.

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., & 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.

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

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.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*, 377-381.

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., & Pratte, M. S. (in press). Bayesian hierarchical models. In W. H. Batchelder, H. Colonius, E. Dzhafarov, & J. I. Myung (Eds.), *The new handbook of mathematical psychology, volume 1: Measurement and methodology.* Cambridge.

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., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin and Review*, *17*, 427–435.

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.

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., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830-841.

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.

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.